



# **ASSESSING INFORMATION PROCESSING AND ONLINE REASONING AS A PREREQUISITE FOR LEARNING IN HIGHER EDUCATION**

EDITED BY: Olga Zlatkin-Troitschanskaia, Patricia A. Alexander and  
James W. Pellegrino

PUBLISHED IN: Frontiers in Education and Frontiers in Psychology



# frontiers

## Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-83250-164-1

DOI 10.3389/978-2-83250-164-1

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: [frontiersin.org/about/contact](https://frontiersin.org/about/contact)



# ASSESSING INFORMATION PROCESSING AND ONLINE REASONING AS A PREREQUISITE FOR LEARNING IN HIGHER EDUCATION

Topic Editors:

**Olga Zlatkin-Troitschanskaia**, Johannes Gutenberg University Mainz, Germany

**Patricia A. Alexander**, University of Maryland, College Park, United States

**James W. Pellegrino**, University of Illinois at Chicago, United States

**Citation:** Zlatkin-Troitschanskaia, O., Alexander, P. A., Pellegrino, J. W., eds. (2022).  
Assessing Information Processing and Online Reasoning as a Prerequisite for  
Learning in Higher Education. Lausanne: Frontiers Media SA.  
doi: 10.3389/978-2-83250-164-1

# Table of Contents

- 05 Editorial: Assessing Information Processing and Online Reasoning as a Prerequisite for Learning in Higher Education**  
Olga Zlatkin-Troitschanskaia, Patricia A. Alexander and James W. Pellegrino
- 09 Do Individual Differences in Cognition and Personality Predict Retrieval Practice Activities on MOOCs?**  
Daniel Fellman, Alisa Lincke and Bert Jonsson
- 19 Changes in Students' Understanding of and Visual Attention on Digitally Represented Graphs Across Two Domains in Higher Education: A Postreplication Study**  
Sebastian Brückner, Olga Zlatkin-Troitschanskaia, Stefan Küchemann, Pascal Klein and Jochen Kuhn
- 39 Calibrating the Test of Relational Reasoning: New Information From Oblique Bifactor Models**  
Denis Federiakin
- 49 Performance Assessment of Critical Thinking: Conceptualization, Design, and Implementation**  
Henry I. Braun, Richard J. Shavelson, Olga Zlatkin-Troitschanskaia and Katrina Borowiec
- 59 The Role of Students' Beliefs When Critically Reasoning From Multiple Contradictory Sources of Information in Performance Assessments**  
Olga Zlatkin-Troitschanskaia, Klaus Beck, Jennifer Fischer, Dominik Braunheim, Susanne Schmidt and Richard J. Shavelson
- 78 Learning to Fly Through Informational Turbulence: Critical Thinking and the Case of the Minimum Wage**  
Gerhard Minnameier and Rico Hermkes
- 89 Strategy Use in Learning From Multiple Texts: An Investigation of the Integrative Framework of Learning From Multiple Texts**  
Alexandra List and Patricia A. Alexander
- 105 More Than (Single) Text Comprehension? – On University Students' Understanding of Multiple Documents**  
Nina Mahlow, Carolin Hahnel, Ulf Kroehne, Cordula Artelt, Frank Goldhammer and Cornelia Schoor
- 122 Multiple Texts as a Limiting Factor in Online Learning: Quantifying (Dis-)similarities of Knowledge Networks**  
Alexander Mehler, Wahed Hemati, Pascal Welke, Maxim Konca and Tolga Uslu
- 153 How Do University Students' Web Search Behavior, Website Characteristics, and the Interaction of Both Influence Students' Critical Online Reasoning?**  
Marie-Theres Nagel, Svenja Schäfer, Olga Zlatkin-Troitschanskaia, Christian Schemer, Marcus Maurer, Dimitri Molerov, Susanne Schmidt and Sebastian Brückner

- 168** *Narratives and Their Impact on Students' Information Seeking and Critical Online Reasoning in Higher Education Economics and Medicine*  
Mita Banerjee, Olga Zlatkin-Troitschanskaia and Jochen Roeper
- 185** *Test-Taking Motivation in Education Students: Task Battery Order Affected Within-Test-Taker Effort and Importance*  
Anett Wolgast, Nico Schmidt and Jochen Ranger
- 201** *Undergraduate Students' Critical Online Reasoning—Process Mining Analysis*  
Susanne Schmidt, Olga Zlatkin-Troitschanskaia, Jochen Roeper, Verena Klose, Maruschka Weber, Ann-Kathrin Bültmann and Sebastian Brückner
- 222** *Constraints and Affordances of Online Engagement With Scientific Information—A Literature Review*  
Friederike Hendriks, Elisabeth Mayweg-Paus, Mark Felton, Kalypso Iordanou, Regina Jucks and Maria Zimmermann
- 243** *Assessing University Students' Critical Online Reasoning Ability: A Conceptual and Assessment Framework With Preliminary Evidence*  
Dimitri Molerov, Olga Zlatkin-Troitschanskaia, Marie-Theres Nagel, Sebastian Brückner, Susanne Schmidt and Richard J. Shavelson
- 272** *Evaluation of Online Information in University Students: Development and Scaling of the Screening Instrument EVON*  
Carolin Hahnel, Beate Eichmann and Frank Goldhammer
- 288** *Computational Linguistic Assessment of Textbooks and Online Texts by Means of Threshold Concepts in Economics*  
Andy Lücking, Sebastian Brückner, Giuseppe Abrami, Tolga Uslu and Alexander Mehler
- 317** *Keep Calm in Heated Debates: How People Perceive Different Styles of Discourse in a Scientific Debate*  
Juliane Tkotz, Dorothe Kienhues, Regina Jucks and Rainer Bromme
- 328** *Key Information Processes for Thinking Critically in Data-Rich Environments*  
Jacqueline P. Leighton, Ying Cui and Maria Cutumisu
- 343** *The Semiotics of Test Design: Conceptual Framework on Optimal Item Features in Educational Assessment Across Cultural Groups, Countries, and Languages*  
Guillermo Solano-Flores
- 359** *Patterns of Domain-Specific Learning Among Medical Undergraduate Students in Relation to Confidence in Their Physiology Knowledge: Insights From a Pre–post Study*  
Jochen Roeper, Jasmin Reichert-Schlax, Olga Zlatkin-Troitschanskaia, Verena Klose, Maruschka Weber and Marie-Theres Nagel



## OPEN ACCESS

EDITED AND REVIEWED BY  
Chung Kwan Lo,  
The Education University of Hong  
Kong, China

\*CORRESPONDENCE  
Olga Zlatkin-Troitschanskaia  
troitschanskaia@uni-mainz.de

SPECIALTY SECTION  
This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Education

RECEIVED 25 August 2022  
ACCEPTED 09 August 2022  
PUBLISHED 25 August 2022

CITATION  
Zlatkin-Troitschanskaia O,  
Alexander PA and Pellegrino JW (2022)  
Editorial: Assessing information  
processing and online reasoning as a  
prerequisite for learning in higher  
education. *Front. Educ.* 7:1014654.  
doi: 10.3389/feduc.2022.1014654

COPYRIGHT  
© 2022 Zlatkin-Troitschanskaia,  
Alexander and Pellegrino. This is an  
open-access article distributed under  
the terms of the [Creative Commons  
Attribution License \(CC BY\)](#). The use,  
distribution or reproduction in other  
forums is permitted, provided the  
original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which  
does not comply with these terms.

# Editorial: Assessing information processing and online reasoning as a prerequisite for learning in higher education

Olga Zlatkin-Troitschanskaia<sup>1\*</sup>, Patricia A. Alexander<sup>2</sup> and James W. Pellegrino<sup>3</sup>

<sup>1</sup>Department of Law and Economics, Johannes Gutenberg University Mainz, Mainz, Germany, <sup>2</sup>Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, United States, <sup>3</sup>Department of Psychology and Learning Sciences Research Institute, University of Illinois at Chicago, Chicago, IL, United States

## KEYWORDS

online reasoning, student learning, online information use, higher education, information landscape, internet-based learning

## Editorial on the Research Topic

[Assessing information processing and online reasoning as a prerequisite for learning in higher education](#)

## A critical need

Over the last decades, the World Wide Web (WWW) has created new opportunities but also many challenges for teaching and learning in higher education. To build a coherent, well-informed knowledge base, university students must know how to effectively search for, select, and critically evaluate online information that is of extremely varied quality and credibility (Rouet and Britt, 2011). Students must also be able to analyze, synthesize, and integrate the information from multiple sources into some external product such as a written summary or argumentative essay, even sources espousing contradictory data and views (List and Alexander). However, in a review of over 500 studies of online information processing, Zlatkin-Troitschanskaia et al. (2021) found that university students habitually rely on the first few search results, evaluate information using inappropriate criteria, and systematically avoid information that contradicts their beliefs. Perhaps for these reasons, they can easily overlook important, factually correct content, and fall prey to biased information.

Paradoxically, recent studies indicate a decrease in students' acquisition of domain-specific knowledge over the course of their university studies, juxtaposed with an increase in the development of (counterfactual) misconceptions and false (inter-)disciplinary concepts (Schmidt et al., 2016). The acquisition of erroneous knowledge seems to be specifically pronounced among students who report that they predominantly use online sources for learning (Maurer et al., 2020), while also claiming to be confident in their knowledge and skills despite its inadequacies and errors (Brückner and Pellegrino, 2016).

Simply "googling" without critical reflection on the quality of sources or their contents is likely to result in the acceptance of unwarranted claims and inaccurate or

misleading information. Zlatkin-Troitschanskaia et al. (2020) termed this phenomenon *negative learning*, which can occur without students' awareness. In contrast, *positive learning* can be defined as *the acquisition of academically or scientifically substantiated conceptual, procedural, and transferable knowledge and understanding that has a long half-life, e.g., is flexible in adapting to new information, meets epistemic standards, and can be reconciled with ethical norms and moral values* (Zlatkin-Troitschanskaia et al., 2020, p. 2).

Although negative learning is a general problem, university students are confronted with an internet-based information and learning environment that can increase negative learning's occurrence and/or amplify its effects. For example, the radius and speed of the distribution of distorted and false information is substantial and continuously increasing on the WWW. Also, the dissemination mechanisms are not transparent on various levels, including algorithmic sorting and personalization, social recommendation and sharing of anonymous sources, commercial amplification, shifting of gatekeeping functions, decontextualization and cross-mediatization of content, and, in some areas, orchestrated censorship and propaganda. University students' skills and strategies for selecting, processing, and learning with online information have proven insufficient for what is required for knowledge development in a complex and ever-changing online environment (Zlatkin-Troitschanskaia et al., 2021). Consequently, when students do not recognize biased or false information and incorporate it into their knowledge base, negative learning occurs. This negative learning can then inhibit or distort subsequent information processing and knowledge acquisition over the course of their university studies (List and Alexander, 2019).

Current online learning environments also contribute substantially to cognitive overload and cognitive dissonance, increasing the danger that learners will commit reasoning errors or operate from biased perspectives. It has been shown that university students often neglect complex, more abstractly presented content in favor of less credible but quicker to access, and easier to comprehend information that tends to be consistent with their beliefs and biases (Goldman et al., 2016). No matter what field they decide to pursue, university students begin their studies after years of prior (in-)formal learning and knowledge gained from the Internet and after having been exposed to the information structures and engagement mechanisms of online media that by their very nature do not observe disciplinary boundaries. Domain-specific misconceptions and erroneous beliefs about the nature of knowledge and knowing (i.e., epistemic) are nothing new. Yet, such distorted notions seem far more entrenched and thus harder to eliminate these days.

Further, established theories and models aiming to explain, predict, or even influence learning in higher education stem mainly from an era in which learning was primarily institutionalized and moderated, technologically limited, highly

disciplinary, and characterized by minor variations in teaching methodology. It is therefore evident that a thorough overview of theoretical and empirical research that serves to describe, assess, and predict online information processing and reasoning for students in higher education contexts is urgently required (Zlatkin-Troitschanskaia et al., 2021). The purpose of this issue is to provide such an overview.

## The goals of this Research Topic

Our goals for this issue were to share cutting-edge research that examines important factors and forces that not only illuminate the general challenges that university students face when engaged in online searches for relevant and credible information, but also detail the effects that their pre-existing knowledge, beliefs, language background, and computer use can have on that search process. Most of the research presented in this issue focuses on the preconditions and processes of *self-directed and independent learning* of university students in Internet-based environments, both as part of university courses and outside regular courses. Contributors to this issue also explore the tools and techniques for gathering rich data on what is transpiring at each phase of information processing—from the search for documents to the way students' read and reflect on those documents. Finally, this overview of information processing and online reasoning considers students in higher education generally as well as special populations, e.g., students pursuing medical education.

## Emerging themes

There are also themes that emerge across the 21 studies that form this issue. One such theme includes contributions that provide a profile of the information landscape that today's university students encounter. *Information landscape* refers to the online learning space freely available to students for their learning, which comprises all locatable online information resources for a given domain or topic (List and Alexander). The information landscape is analyzed using (computer-based) data and text mining technologies from linguistics (Mehler et al.) as well as qualitative content analyses, e.g., using established methods from media and communication sciences (Nagel et al.) and narrative analyses (Banerjee et al.). These studies, in particular, describe and analyze the sources and types of information university students select and use for learning and identify mis-information that may introduce misconceptions related to given concepts be they domain specific or otherwise.

Two other themes within the issue pertain to the learning processes and learner characteristics that are relevant to the execution and outcomes of online learning. *Learning processes* represent cognitive, metacognitive, motivational, and affective procedures enacted during online information processing. The

observable activities to which contributors to this issue refer include search for information, navigation on and between websites, and evaluation of information. In several articles, those activities are recorded by logging online activities and by means of observational techniques like eye tracking (Hahnel et al.; Leighton et al.; Mahlow et al.). The authors also share innovative quantitative and qualitative methods for analyzing the small and large data sets that result (e.g., process mining, Schmidt et al.). In addition, a range of data sources were used to craft a rich picture of these university students' learning processes, including data from cognitive labs on students' use of verified knowledge versus specific misconceptions, and their related attitudes such as overconfidence in incorrect answers (Zlatkin-Troitschanskaia et al.).

*Learner characteristics* or individual differences form the third theme within this issue. The characteristics that contributors investigate include students' figural-spatial abilities, linguistic facility, argumentation skills, socioeconomic background, domain-specific knowledge, and knowledge of computers (CITES). The range of learner characteristics was measured in a variety of ways, including tests, questionnaires, or behavioral patterns documented in log files (Fellman et al.; Wolgast et al.). For the big data sets used in several studies, machine learning techniques were employed to extract key learner characteristics (Lücking et al.).

Finally, *learning outcomes* (e.g., acquired domain-specific concepts), constitutes a fourth thematic element, primarily serving as dependent variables in various studies. Such outcomes are assessed using various types of achievement tests, including rubric-based and automated analyses of texts written by students (Brückner et al.; Roeper et al.).

## Contributions to understanding information processing and online reasoning

Overall, the 21 studies in this issue present interesting and important results from contemporary international research that identifies and systematically describes properties of the various learning sources and information university students use for learning. For instance, researchers systematically examine key instructional texts, assessments, and thematically related online information used by students as well as their cognitive and non-cognitive effects that hinder or promote learning in higher education. Some studies in this issue also examine the interplay between text structures and features and test-takers' responses as well as variance depending on presented information in sources for learning. By systematically and comprehensively investigating student learning, the results from these studies have identified online information processing and online reasoning as a crucial prerequisite for successful learning in higher education in the age of mis-information.

Structurally, this issue is composed of empirical studies, combined with conceptual and literature reviews, grounded not only in higher education research but also in various intersectional disciplines (e.g., communication sciences). The empirical studies present innovative conceptual and measurement approaches, linking educational results with analysis methods from linguistics, computational linguistics, media science etc., which have not previously been applied and combined in research in higher education. Remarkably, the explanatory power of the new integrative, multi- and interdisciplinary approaches applied in these studies has exceeded that of typical explanatory variables and approaches from the educational and learning sciences alone. Overall, these studies illustrate how the new methods presented tie in with current challenges as well as current developments in higher education research and practice.

Overall, this issue illuminates a controversial and very timely topic in higher education of international importance, and addresses and investigates it from different cross-disciplinary perspectives. Original theoretical, conceptual, and empirical studies are presented that offer examinations and explanations of *Information Processing and Online Reasoning and their Effect on Learning in Higher Education* in the age of mis-information. This issue contains studies related to teaching and learning across different environments in the digital age, the generation and dissemination of knowledge, and modes of inquiry. Moreover, the work described in this issue comes from different countries and encompasses analyses in several disciplines related to higher education learning and its assessment. All contributors to this issue, which provides complementary and diverse perspectives and methodologies, are international scholars whose empirical and theoretical work is centered around the processing of digital content and online reasoning within higher education and their assessment. In this way, this issue serves as a benchmark contribution in this emerging, crucial new field of learning research. The work is foundational for addressing extremely controversial developments regarding students' use of online media for learning and helps to close the gap in corresponding learning research to date.

## Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Brückner, S., and Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multilevel models to validate an assessment of higher education students' competency in business and economics. *J. Educ. Measur.* 53, 293–312. doi: 10.1111/jedm.12113
- Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., and Greenleaf, C. (2016). Disciplinary literacies and learning to read for understanding: a conceptual framework for disciplinary literacy. *Educ. Psychol.* 51, 219–246. doi: 10.1080/00461520.2016.1168741
- List, A., and Alexander, P. A. (2019). Toward an integrated framework of multiple text use. *Educ. Psychol.* 54, 20–39. doi: 10.1080/00461520.2018.1505514
- Maurer, M., Schemer, C., Zlatkin-Troitschanskaia, O., and Jitomirski, J. (2020). "Positive and negative media effects on university students' learning: preliminary findings and a research program," in *Frontiers and Advances in Positive Learning in the Age of Information*, ed. O. Zlatkin-Troitschanskaia (Berlin: Springer), 109–119.
- Rouet, J., and Britt, M. A. (2011). "Relevance processes in multiple document comprehension," in *Text Relevance and Learning From Text*, eds. M. T. McCrudden, J. P. Magliano, and G. Schraw (Charlotte, NC: Information Age), 19–52.
- Schmidt, S., Zlatkin-Troitschanskaia, O., and Fox, J.-P. (2016). Pretest-posttest-posttest multilevel IRT modeling of competence growth of students in higher education in Germany. *J. Educ. Meas.* 53, 332–351. doi: 10.1111/jedm.12115
- Zlatkin-Troitschanskaia, O., Bisang, W., Mehler, A., Banerjee, M., and Roeper, J. (2020). "Positive learning in the Internet age: Developments and perspectives in the PLATO program," in *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)*, ed. O. Zlatkin-Troitschanskaia (Cham: Springer), 1–5.
- Zlatkin-Troitschanskaia, O., Hartig, J., Goldhammer, F., and Krstev, J. (2021). Students' online information use and learning progress in higher education—a critical literature review. *Stud. High. Educ.* 46, 1996–2021. doi: 10.1080/03075079.2021.1953336





# Do Individual Differences in Cognition and Personality Predict Retrieval Practice Activities on MOOCs?

Daniel Fellman<sup>1\*</sup>, Alisa Lincke<sup>2</sup> and Bert Jonsson<sup>1</sup>

<sup>1</sup> Department of Applied Educational Science, Umeå University, Umeå, Sweden, <sup>2</sup> Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden

## OPEN ACCESS

### Edited by:

Olga Zlatkin-Troitschanskaia,  
Johannes Gutenberg University  
Mainz, Germany

### Reviewed by:

Sergio Luján-Mora,  
University of Alicante, Spain  
Li Neng Lee,  
National University of Singapore,  
Singapore

### \*Correspondence:

Daniel Fellman  
daniel.fellman@umu.se

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Psychology

**Received:** 03 June 2020

**Accepted:** 27 July 2020

**Published:** 18 August 2020

### Citation:

Fellman D, Lincke A and  
Jonsson B (2020) Do Individual  
Differences in Cognition  
and Personality Predict Retrieval  
Practice Activities on MOOCs?  
Front. Psychol. 11:2076.  
doi: 10.3389/fpsyg.2020.02076

Online quizzes building upon the principles of retrieval practice can have beneficial effects on learning, especially long-term retention. However, it is unexplored how interindividual differences in relevant background characteristics relate to retrieval practice activities in e-learning. Thus, this study sought to probe for this research question on a massive open online course (MOOC) platform where students have the optional possibility to quiz themselves on the to-be-learned materials. Altogether 105 students were assessed with a cognitive task tapping on reasoning, and two self-assessed personality measures capturing need for cognition (NFC), and grittiness (GRIT-S). Between-group analyses revealed that cognitively high performing individuals were more likely to use the optional quizzes on the platform. Moreover, within-group analyses ( $n = 56$ ) including those students using the optional quizzes on the platform showed that reasoning significantly predicted quiz performance, and quiz processing speed. NFC and GRIT-S were unrelated to each of the aforementioned retrieval practice activities.

**Keywords:** retrieval practice, test-enhanced learning, e-learning, MOOC, personality, cognition

## INTRODUCTION

Learning via Internet has increased its popularity during the past decades due to its advantages it offers with respect to flexibility of time (i.e., studying can be carried out at any time) and space (i.e., studying can be carried out anywhere). In this vein, a new concept, denoted as e-learning has arisen, which is an umbrella term that covers all aspects related to individualized instructions distributed over public or private computer networks intended to promote learning (Manochehr, 2006; Clark and Mayer, 2016). One particularly fast-growing learning format pertains to massive open online courses (MOOCs). MOOC refers to learning platforms to which an unlimited amount of students can enrol (either paid or unpaid), and access a wide range of courses materials, including additional learning resources such as interactive courses, problems sets (e.g., quizzing), and filmed lectures (Kaplan and Haenlein, 2016). The advantages



with MOOCs lies in its flexibility, allowing students to take courses independently at their own pace, without being bound by time and place.

## Retrieval Practice

Along with the increased popularity in e-learning, several MOOC platforms have also started to apply features on their platforms with the purpose to boost learning outcomes. One such feature pertains to the opportunity to quiz oneself on the learned materials (van der Zee et al., 2018). A large body of evidence in experimental settings shows that self-testing of the to-be-learned material, typically denoted as *retrieval practice*, increase students' long-term retention and transfer of knowledge to new situations (Roediger and Karpicke, 2006; Butler, 2010; Weinstein et al., 2010; Agarwal et al., 2012). Moreover, the benefit of retrieval practice over other study strategies (e.g., summaries, note-taking), often referred to *testing effects*, are typically not visible when knowledge is tested immediately after learning (e.g., Roediger and Karpicke, 2006). Rather, the effects are prominent over lengthier retention intervals, for instance, when students are tested a week after the learning phase (Karpicke and Roediger, 2007). Although the effects of retrieval practice on memory retention occur independently on feedback (Roediger and Butler, 2011), the inclusion of feedback strengthens learning and provide a formative component through which students can monitor their accuracy and thus prevent that erroneous learning (Roediger and Marsh, 2005). The mechanisms underlying retrieval practice remains unclear (see Rowland, 2014 for an overview), but the effectiveness has been studied and confirmed in different experimental settings, educational contexts, across a range of materials and by brain imaging studies (Dunlosky et al., 2013; van den Broek et al., 2016; Adesope et al., 2017). Thus, retrieval practice is relatively well-established and that the act of retrieving memory from information seems to strengthen memory and to reduce forgetting (Kornell et al., 2011; Rowland, 2014). The features of self-testing with feedback appears to be a very promising study technique, especially for MOOCs, as the content, typically is directed to lifelong learning (van der Zee et al., 2018). In this vein, the present study set out to test how individual differences in cognition and personality is associated with retrieval practice activities on a MOOC platform targeted for university students, providing new insights to the body of research within teaching and learning across different environments in the digital age (explicit link to the special issue: <https://www.frontiersin.org/research-topics/12111/assessing-information-processing-and-online-reasoning-as-a-prerequisite-for-learning-in-higher-educ>).

Retrieval practice is typically implemented in various formats on MOOCs using quizzing with and without the support of images and video clips, utilizing different response formats such as multiple-choice and short and open-answer responses for boosting learning outcomes (van der Zee et al., 2018). Such quizzing features are often optionally implemented, that is, students can complete quizzes as an additional support for their learning. However, several studies indicate that quizzes remain highly unutilized when they are optional in online course (Olson and McDonald, 2004; Kibble, 2007; Carpenter et al., 2017;

Corral et al., 2020). Corral et al. (2020), for instance, showed that nearly 45% of students did not complete a single quiz during an online course covering introductory psychology and that 88% of these quizzes were not completed. Furthermore, Carpenter et al. (2017) examined the quizzing frequency among the students that took an online biology course, showing that about 50% of the students completed the practice quizzes that were made available. However, the findings reported above have been observed at the group level, and it is still unclear whether individuals with certain traits are more likely to engage in quizzing than others are, thus prompting further research.

Besides that quizzing remains largely unexploited by students, it is also unclear in what way, and in what volume the quizzes are used by those individuals that actually engage in retrieval practice on MOOC platforms. As previously mentioned, the majority of previous retrieval practice studies on this topic are experimental, typically applied in the context of laboratory or classroom settings (for a meta-analysis, see Adesope et al., 2017), whereas only a few ones have focused on examining how retrieval practice on MOOC platforms are related to learning outcomes (Davis et al., 2016, 2018). Davis et al. (2016) examined whether students' learning outcomes on a MOOC functional programming course would be altered if the participants ( $n = 2166$ ) were prompted with retrieval practice cues following each lecture. Compared to a control group receiving no quizzes, the results showed no beneficial effects of retrieval practice neither in test performance or actual course grades. In another study, Davis et al. (2018) prompted participants to write summaries of the content following each video clip on a MOOC course on Coursera. The results showed that the amount of written summaries were associated with a better performance in the weekly quiz assessments, but not in a better performance in the final course exam.

Albeit retrieval practice has been extensively examined, few studies have focused on for which individuals this learning technique is beneficial. Indeed, there are large inter-individual differences in most of human-related behavior, but one background factor that consistently has shown to influence learning outcomes is cognitive ability. Study results show that fluid intelligence (i.e., the ability to solve problems in novel situations) and working memory (i.e., the ability to maintain and manipulate information over a short period before it decays) are both reliable predictors of academic attainment (Turner and Engle, 1989; Cowan et al., 2005; Krumm et al., 2008; Furnham and Monsen, 2009; Ren et al., 2015). The few studies examining cognition in relation to retrieval practice have been somewhat mixed, with some studies showing that cognitively strong individuals show greater testing effects (Tse and Pu, 2012; Agarwal et al., 2017), especially when prompted with more difficult items (Minear et al., 2018), or results that support effects in neither direction (Brewer and Unsworth, 2012; Wiklund-Hörnqvist et al., 2014; Bertilsson et al., 2017). With respect to cognitive ability and retrieval practice on MOOCs, the evidence is scarce, albeit one study investigating this relationship shows that better cognitive ability is associated with higher accuracies in

quizzes, and tend to spend more time on quizzing themselves (Fellman et al., 2020).

Besides cognitive abilities, personality characteristics are important for learning outcomes as well. Especially on MOOCs where quizzing is optional, it is plausible to assume that individual characteristics tapping on motivation, openness and curiosity for learning new things are important traits for maximizing the utility of the platform. One personality trait shown to be important for learning is “the tendency to engage in and enjoy thinking”. This ability is typically referred to as *need for cognition* (NFC; Cacioppo and Petty, 1982). Individuals with high NFC typically analyze and seek to understand information and events in their surroundings, whereas low NFC individuals are more likely to rely on experts or cognitive heuristics. Hence, high NFC individuals typically approach problem-solving tasks more positively than those with low NFC (Cacioppo et al., 1996). In traditional classroom settings, high NFC has been found to result in better performance when solving math problems (Dornic et al., 1991), and to predict academic performance (Sadowski and Gülgös, 1996). With respect to retrieval practice in experimental settings, NFC appears to be weakly related to recall performance in quizzes (Bertilsson et al., 2017; Stenlund et al., 2017), but to our knowledge, no previous study has specifically examined the relationship between NFC and retrieval practice activities on MOOCs.

Another personality trait that has been shown to be critical for learning outcomes is the perseverance and passion for long-term goals. This ability, denoted as *GRIT* (Duckworth et al., 2007), has shown to be a reliable predictor for several important outcomes such as academic achievement and life success (Duckworth et al., 2007; Eskreis-Winkler et al., 2014). It has been suggested that GRIT contribute uniquely to learning outcomes as it works independently of intelligence, and that both talent and GRIT is necessary to become highly competent in a specific skill (Duckworth et al., 2007). To our knowledge, only one study has examined the relationship between GRIT and retrieval practice in educational classroom settings (Bertilsson et al., 2017). In that study, the authors conducted two between-subjects design experiments where Swedish participants were to learn novel Swahili words either in a re-study condition or in a retrieval practice condition. While both experiments showed that those receiving retrieval practice outperformed those receiving re-study in recall following 4 weeks, the results showed no evidence that NFC would have any moderating role in these gains. However, to our knowledge, no study has investigated GRIT in relation to retrieval practice activities on MOOC platforms, deserving further scrutiny.

Lastly, it is worth pointing out that both NFC and GRIT are personality characteristics suggested to be stable over time and thus influence learning (Duckworth et al., 2007; Stenlund and Jonsson, 2017). Within the context of students using MOOCs platforms where the student has a greater autonomous responsibility for his/her studies, personality factors are potentially even more critical. Hence, the use of these platforms are often (as in the present study) not mandatory for the students, and as shown by Corral et al. (2020), the majority of students do not complete their online quizzing. Potentially,

the likelihood of using retrieval practice in MOOCs platforms is associated with personality characteristics.

## MATERIALS AND METHODS

### Study Design

The research question probed for in the present study was: Are individual differences in cognitive ability, and personality characteristics are related to retrieval practice activities on a MOOC platform? The data for this study stems from an interactive MOOC platform in Sweden titled Hippocampus (see <https://www.hypocampus.se>). Approximately 15 000 students use Hippocampus as a fee-based platform (99 Swedish SEK/month; approximately \$10.49/month) for carrying out university courses, with most of the users consisting of medical students. The MOOC platform provides the students with compressed course materials that are highly relevant for the to-be-completed courses at their universities. Specifically, instead of completing the course by reading from the course books, the content of the course is transferred to the interactive MOOC platform. Hippocampus also provides a high degree of learner control, offering more than 50 interactive courses covering different topics in medicine that students can complete non-linearly at their own pace (i.e., they can choose to jump back and forth from a course to another). With most relevance for the present study, the students also have the optional possibility to quiz themselves on the materials they just read, building upon the principles of *retrieval practice* (Dunlosky et al., 2013). These optional quizzes are implemented at the end of each learning section. Altogether 105 university-dwelling participants that were carrying out studies on the MOOC took part in this study. Cognitive ability among the participants was measured with the Raven's Advanced Progressive Matrices (RAPM; Raven et al., 1991). For measuring the tendency to engage in and enjoy thinking, and the perseverance and passion for long-term goals, participants were assessed with the questionnaires Need For Cognition (NFC; Dornic et al., 1991), and the Short Grit Scale (GRIT-S; Duckworth and Quinn, 2009), respectively.

The relationship between individual characteristics (i.e., cognitive ability, personality) and retrieval practice activities on the MOOC was examined in a two-fold way. First, using between-group analyses, we examined whether individuals with high usage of the optional quizzes (henceforth *high retrieval practice*; high-RP) differed from the individuals with low usage of the optional quizzes (henceforth *low retrieval practice*; low-RP) with respect to our three predictors RAPM, GRIT-S, and NFC. Second, using within-group analyses including only the high-RP group, we extracted three measures of relevant retrieval practice activities, which we presupposed that could be related to cognitive- and personality measures, and those variables were regressed on our three predictors. The three target outcomes of retrieval practice were: (1) number of quizzes taken per study session, (2) accuracy in taken quizzes and (3) quiz processing speed per study session. Note that the reason for excluding the low-RP group in the within-group analyses were justified, as this group had barely engaged in retrieval

practice activities on the MOOC platform (see “Between-Group Analyses” in the Results section for more details). For the between-group analyses, we hypothesized that higher cognitive ability, as well as higher grittiness and need for cognition, would increase the likelihood for belonging to the high-RP group. For the within-group analyses, we surmised that the cognitively high-performing individuals, individuals with high GRIT-S, and individuals with high NFC would use the optional quizzes more persistently, show higher quiz accuracies, and exhibit faster reaction times in the quizzes. Our attempt to unravel individual characteristics that bear importance for retrieval practice activities on MOOCs will hopefully yield new insights to the body of research within teaching and learning across different environments in the digital age (explicit link to the special issue: <https://www.frontiersin.org/research-topics/12111/assessing-information-processing-and-online-reasoning-as-a-prerequisite-for-learning-in-higher-educ>).

## Data Description, Participants and Methods

Regarding the technical aspects, each day that a student login and use the MOOC (i.e., Hippocampus), a large amount of interactional data is generated. The data is collected using JavaScript methods available in the user's browser and stored in the backend system in a database. The log-files retrieved from the database are organized into two tables: *reading\_material* and *quiz\_material*. The *reading\_material* table contains data related to student interaction with learning materials in a course and can be used to identify reading time information (e.g., the amount of time the student was active on a particular page). The *quiz\_material* table contains information regarding quiz activity such as the number of quizzes taken, and total time spent on quizzes. As this study focus solely on retrieval practice activities, only data stemming from the *quiz\_material* table was analyzed. All available data from the *quiz\_material* table within the date range 01.01.2019 – 02.02.2020 was extracted. Feature extraction was computed by aggregating scores as a function of a particular student (labeled as ‘user Id’ in the dataset).

The participants in the present study consisted of medical students who were studying at Hippocampus platform to prepare themselves for the actual exam at their university. The study was approved by the Regional Vetting Committee (2017/517-31), Sweden, and informed consent was obtained from all participants. All students on Hippocampus were invited to complete the test session consisting of a background questionnaire, personality questionnaires, and a reasoning task capturing cognitive ability<sup>1</sup>. The test session was administered online using an in-house developed web-based test platform by sending a link to the students via email (i.e., the participants could complete the experiment on a computer of their choosing) (Röhlcke et al., 2018; Fellman et al., 2020). Those who completed the test session were allowed to participate in a

lottery of two premium accounts, consisting of 6 months of free use on Hippocampus.

Altogether 185 students completed the test session to the end. However, as is common on MOOC platforms, the test takers were highly varying in terms of how much time they had been spent studying at Hippocampus. For leveling out those who only was visiting the platform from those that actually used the platform for studying, we followed the threshold criteria used in Fellman et al. (2020). First, we excluded participants that had been active less than 10 times during the first 100 days since registering themselves on the system (i.e., only one login session), resulting in the exclusion of 80 students. For the remaining participants ( $N = 105$ ), we split the data into two groups with respect to retrieval practice activity as follows: those students that had completed  $\geq 50$  quizzes formed a group coined as *high retrieval practice group* (high-RP) whereas those that had completed  $< 50$  quizzes formed a separate group coined as *low retrieval practice group* (low-RP)<sup>2</sup>.

Together, these criteria resulted in a total sample size of 105 participants, with 56 of the participants belonging to the high-RP group and 49 of the participants to the low-RP group. As such, the participation rate was very low, considering that as many as 15,000 students are registered users. The mean age of the participants included in the present study was 30.29 years ( $SD = 7.06$ ) out of which 49.52% were females. An independent samples *t*-test verified that the groups did not differ significantly in terms of age [ $t(104) = 0.682$ ,  $p = 0.50$ ], and there were no statistically significant differences between the two groups with respect to gender ( $\chi^2 = 9.153$ ,  $df = 2.380$ ,  $p = 0.67$ ), and education ( $\chi^2 = 2.429$ ,  $df = 4$ ,  $p = 0.66$ ). See also Table 1 that summarizes the demographical data of the participants.

## Target Outcomes of Retrieval Practice Activities

As previously mentioned, participants were prompted with optional quizzes following each study session at Hippocampus. These quizzes could be either in multiple-choice format or open-ended format. In the multiple choice quizzes, the participants

<sup>2</sup>After a careful exploratory data analysis, this threshold proved to be optimal based on two important criteria: (1) the median value of this variable was 58, thus very close to 50 completed quizzes used as cut-offs, and (2) this threshold proved to spread the participants to the respective groups fairly evenly.

**TABLE 1 |** Background characteristics of the study sample.

	High-RP	Low-RP
Sample size ( $n$ )	56	49
Gender (F/M)	27/29	25/24
Age (M, SD)	29.80 (6.48)	30.80 (7.06)
Education		
	Basic vocational 12.5%	Basic vocational 8.16%
	Bachelor's degree 25.0%	Bachelor's degree 28.6%
	Master's degree 55.4%	Master's degree 59.2%
	Doctoral degree 3.6%	Doctoral degree 4.1%
	Other 3.6%	Other 0.0%

*High-RP = High retrieval practice group; Low-RP = Low retrieval practice group.*

<sup>1</sup>Besides the reasoning task and the questionnaires, participants also completed several other tasks tapping on working memory and episodic memory. However, results attributed to these tasks will be reported elsewhere.



were asked about specific information concerning the learning section followed by four alternatives out of which one was correct. Correctly recalled quiz responses were logged as ‘True’ whereas incorrectly recalled quiz responses were logged as ‘False.’ In the self-assessed quiz format, participants were prompted with a quiz in a similar fashion as in the multiple-choice quizzes. However, instead of being prompted with four alternatives, they were now asked to respond to the quiz in a written format by typing down their response in an empty box. Following the response, the system showed the correct answer. Thus, the scorings of the responses were self-corrected, meaning that the participants were to tick either on a red box with a text stating *Read more* (corresponding to an incorrectly recalled quiz and marked as *False* in the log file) or a green box with a text stating *I knew this* (corresponding to a correctly recalled quiz and marked as *True* in the log file).

We extracted three outcome variables from the Hippocampus platform that captured different aspects of retrieval practice activities: (1) *Number of taken quizzes per study session* (Quizzes per session), (2) *accuracy in taken quizzes* (Quiz performance), and (3) *processing speed in quizzes* (Quiz processing speed). Quizzes per session were calculated by averaging the number of taken quizzes (including both multiple-choice and self-assessed items) across all login sessions (formula: *Quizzes per session = total number of quizzes/total number of login sessions*). Quiz performance encompassed only the multiple-choice items (the self-assessed items were excluded as students could self-correct the responses *a posteriori*) and was calculated as a proportion score of correct responses (formula: *Quiz performance = number of correctly recalled quizzes/total number of completed quizzes*). Quiz processing speed comprised of the average time spent on a given quiz (formula: *Quiz processing speed = total quiz time/number of completed quizzes*).

## Predictors of Individual Differences in Cognition and Personality

### *Raven’s advanced progressive matrices (RAPM)*

For capturing cognitive ability, the participants were measured with Raven’s Advanced Progressive Matrices (RAPM) (Raven et al., 1991). In this task, 24 items were presented in ascending order (i.e., item difficulty increased progressively), each of which consisted of a  $3 \times 3$  matrix of geometric patterns with the bottom-right area missing a pattern. The participants were asked to complete the pattern by picking one option among eight alternatives. The participants had 20 min to complete the task. As the dependent variable, we used the total number of correctly recalled items (score range 0–24), with higher scores indicating better reasoning ability. Internal consistency was good for RAPM in the present study (Cronbach’s  $\alpha = 0.83$ ).

### *Short grit scale-s (GRIT-S)*

A Swedish version of the short version of GRIT (GRIT-S; Bertilsson et al., 2017) was used in the present study. GRIT-S includes eight items. Four of the items reflect participants’ ability to maintain interest (e.g., “I often set a goal but later choose to pursue a different one”) whereas the four other items capture participants’ ability to maintain effort (e.g., “I have achieved a goal

that took years of work”). Each item is rated on a 5-point Likert-like scale (1 = strongly disagree, 3 = neutral, and 5 = strongly agree). The scores from each individual item were averaged together and served as our dependent variable, with higher scores indicating more GRIT-S. Cronbach’s  $\alpha$  for GRIT-S in the present study was 0.76, indicating acceptable internal consistency.

### *Need for cognition (NFC)*

Need for cognition (NFC) was measured with the Mental Effort Tolerance Questionnaire (METQ; Stenlund and Jonsson, 2017), which is a Swedish adaptation of the original Need for Cognition Scale (Cacioppo and Petty, 1982). The NFC questionnaire encompasses 30 items, each of which is rated on a five-point Likert scale (1 = strongly disagree; 3 = neutral; 5 = strongly agree), yielding a possible score range from between 30 and 150. Twelve of the items represent positive attitudes toward engaging and enjoying thinking, whereas the remaining items indicate negative attitudes. Thus, the items capturing negative attitudes were reversed before calculating our dependent variable (i.e., the sum score of after all items were summed together), with higher scores indicating more NFC. Internal consistency was acceptable for NFC in the present study (Cronbach’s  $\alpha = 0.75$ ).

## RESULTS

### Between-Group Analyses

First, we examined whether the low-RP individuals ( $n = 49$ ) differed from the high-RP individuals ( $n = 56$ ) with respect to our three predictors. We employed logistic regression analyses where the group served as the dependent variable and the predictor of interest as the independent variable. Moreover, number of login sessions served as the covariate in the models to control for activity effects (i.e., it is likely that those having more login sessions also have a higher probability of belonging to the high RP group). The results showed that, after controlling for number of study sessions ( $p < 0.001$ ), RAPM had a statistically significant effect on group ( $\beta = 0.67, p = 0.011$ ). Specifically, one unit increase in RAPM increased the odds ratio for being a high-RP individual with 1.18 (95% CI: 1.04–1.35). The personality predictors GRIT-S ( $\beta = -0.48, p = 0.592$ ), and METQ ( $\beta = -0.11, p = 0.636$ ) did not significantly predict group affiliation after controlling for number of login sessions.

### Within-Group Analyses of the High-RP Group

We further investigated how the three retrieval practice activities (i.e., quizzes per session, quiz performance, quiz processing speed) in the high RP group. Of note, we decided not to include the low-RP group in the within-group analyses, as the distribution in the three dependent variables of retrieval practice activities were highly non-normal. Specifically, most participants in the low-RP group had taken  $\leq 1$  quizzes, yielding unreliable results in the two other retrieval practice outcomes quiz performance (e.g., an individual with 1/1 correct quizzes obtains 100% accuracy) and quiz processing speed

(e.g., quiz response time is calculated based on only one or a few items) as well.

We employed multiple regression analyses for investigating the relationship between the predictors and the retrieval practice variables. Specifically, this yielded three different models where a given retrieval practice variable was regressed on all predictors. Prior to analyses, the three retrieval practice measures were screened for multivariate outliers using the Mahalanobis distance value  $\chi^2$  table ( $p < 0.001$ ; Tabachnick and Fidell, 2007). We also screened each predictor variable (i.e., NFC, GRIT-S, RAPM) and dependent variable (i.e., retrieval practice activities) for univariate outliers (scores on any online activity feature that deviated more than 3.5 *SD* from the z-standardized group mean were defined as univariate outliers). All identified outliers from the aforementioned screening analyses were imputed using multivariate imputations by chained equations (MICE) (van Buuren and Groothuis-Oudshoorn, 2011). Following data cleaning, the assumptions for multiple regression (multicollinearity, homoscedasticity, multivariate normality, lack of outliers in standardized residuals) were met in all three models. **Table 2** depicts descriptive statistics for the extracted retrieval practice activity variables and the three predictors, whereas zero-order correlations between the predictors and the retrieval practice variables can be found in **Table 3**. With respect to correlational relationships, we observed a statistically significant association between quizzes per session and quiz processing speed ( $r = -0.337$ ,  $p = 0.012$ ), between quiz performance and RAPM ( $r = 0.512$ ,  $p < 0.001$ ), between quiz processing speed and RAPM ( $r = 0.356$ ,  $p = 0.007$ ), and between RAPM and GRIT-S ( $r = -0.265$ ,  $p = 0.048$ ).

**TABLE 2 |** Descriptive statistics for the extracted retrieval practice activity variables and the predictors.

Variable	M	SD	Skew	Kurtosis
Number of quizzes per session	27.13	16.79	1.12	0.54
Quiz performance	0.76 <sup>a</sup>	0.11	-0.57	-0.1
Quiz processing speed	0.74 <sup>b</sup>	0.24	0.34	-0.28
RAPM	18.79	4.33	-1.52	2.66
GRIT-S	3.30	0.62	-0.12	0.08
NFC	114.71	10.13	-0.38	-0.63

<sup>a</sup>Proportion of correctly recalled quizzes; <sup>b</sup>Depicted in minutes,  $N = 56$ .

**TABLE 3 |** Intercorrelations between the retrieval practice variables and the predictors.

Variable	1	2	3	4	5
1. Quizzes per session	—				
2. Quiz performance	0.16	—			
3. Quiz processing speed	-0.34*	-0.25	—		
4. RAPM	0.10	0.36**	-0.51**	—	
5. GRIT-S	0.14	0.06	-0.06	-0.27*	—
6. NFC	0.02	0.22	-0.17	0.21	0.10

\*indicates  $p < 0.05$ , \*\* indicates  $p < 0.01$ .

## Quizzes per Study Session

The regression model with quizzes per session as the dependent variable, and RAPM, GRIT-S and NFC as predictors was statistically non-significant [ $F(4, 52) = 1.472$ ,  $p = 0.536$ ,  $R^2_{\text{Adjusted}} = -0.015$ ]. A closer examination of the coefficients (see **Table 4**) showed that none of the predictors were significantly related to quizzes per session (all  $p$ -values  $\geq 0.198$ ).

## Quiz Performance

When quiz performance served as the dependent variable, the predictors together explained 11.9% of the variance and the regression equation was statistically significant [ $F(4, 52) = 3.478$ ,  $p = 0.022$ ]. A closer inspection of the coefficients (see **Table 5**) showed that RAPM was significantly related to quiz performance ( $\beta = 0.368$ ,  $p = 0.009$ ) such that those with better reasoning performance having higher quiz performance scores. Neither GRIT-S nor NFC were significantly related to quiz performance ( $p$ 's  $\geq 0.29$ ).

## Quiz Processing Speed

In the regression model with quiz processing speed as the dependent variable, the results showed a statistically significant regression equation [ $F(4, 52) = 7.494$ ,  $p < 0.001$ ]. Together, the three predictors explained 26.2% of the variance in quiz processing speed. As depicted in **Table 6**, RAPM significantly predicted quiz processing speed ( $\beta = -0.559$ ,  $p < 0.001$ ), with those performing better in the reasoning task had faster quiz processing speed. Neither GRIT-S nor NFC were significantly related to quiz processing speed.

**TABLE 4 |** Regression coefficients with quizzes per study session as the outcome variable.

	B	SE B	$\beta$	t-value	Sig.
RAPM	0.593	0.565	0.153	1.05	0.298
GRIT-S	5.067	3.89	0.186	1.303	0.198
NFC	-0.051	0.234	-0.031	-0.218	0.828

$R^2_{\text{Adjusted}} = -0.015$

RAPM = Raven's Advanced Progressive Matrices, GRIT-S = Short Grit Scale-S, NFC = Need for Cognition.

**TABLE 5 |** Regression coefficients with quiz performance as the outcome variable.

	B	SE B	$\beta$	t-value	Sig.
RAPM	0.009	0.003	0.368	2.711	0.009
GRIT-S	0.026	0.024	0.143	1.07	0.290
NFC	0.001	0.001	0.126	0.961	0.341

$R^2_{\text{Adjusted}} = 0.119^*$

RAPM = Raven's Advanced Progressive Matrices, GRIT-S = Short Grit Scale-S, NFC = Need for Cognition.

**TABLE 6 |** Regression coefficients with quiz processing speed as the outcome variable.

	<i>B</i>	<i>SE B</i>	$\beta$	<i>t</i> -value	Sig.
RAPM	−0.032	0.007	−0.559	−4.506	< 0.001
GRIT-S	−0.08	0.048	−0.202	−1.653	0.104
NFC	−0.001	0.003	−0.027	−0.22	0.827
$R^2$ Adjusted	0.262***				

RAPM = Raven's Advanced Progressive Matrices, GRIT-S = Short Grit Scale-S, NFC = Need for Cognition.

## Follow-Up Analysis: Moderation Analyses

For examining whether the personality measures GRIT-S and NFC moderated the relationship between RAPM and retrieval practice activities, we followed up the previous analyses with moderation analyses. GRIT-S and NFC, which were fed into separate models with RAPM in these analyses, were transformed into binary variables using median splits prior to model computation (i.e., those with scores above median were defined as high GRIT-S/high NFC, whereas those having scores below the median were defined as low GRIT-S/low NFC). As we were interested in examining whether GRIT-S and NFC moderated the relationship between RAPM and each of the three retrieval practice activity variables, altogether six separate models were computed (for more information, see **Supplementary Material**). The results of the moderation analyses are summarised in **Supplementary Material (Appendix A, Table A1)**, showing that neither personality variable moderated the relationship between RAPM and retrieval practice activities.

## DISCUSSION

Retrieval practice is a well-established evidence-based study technique shown to have facilitating effects on long-term memory retention of information (Roediger and Karpicke, 2006; Butler, 2010; Weinstein et al., 2010; Agarwal et al., 2012), which have led several MOOC administrators to implement features tapping on retrieval practice on their platform. However, optionally based quizzes on MOOCs tend to be highly unutilized, and it is scarcely unknown which individuals, and in what way retrieval practice on in e-learning is used. This study set out to test how individual differences in cognition and personality is associated with retrieval practice activities on a MOOC platform targeted for university students. As a first step, we employed logistic regression analyses to examine whether low retrieval practice individuals (low-RP) differed from high retrieval practice individuals (high-RP) with respect to our three predictors tapping on reasoning (RAPM), and two personality measures capturing students ability to maintain interest over time (GRIT-S), and the tendency to engage in and enjoy thinking (NFC). As a second step, we conducted multiple regression analyses within the high-RP group where three relevant target outcomes of retrieval practice activities (number of taken quizzes per session, accuracy

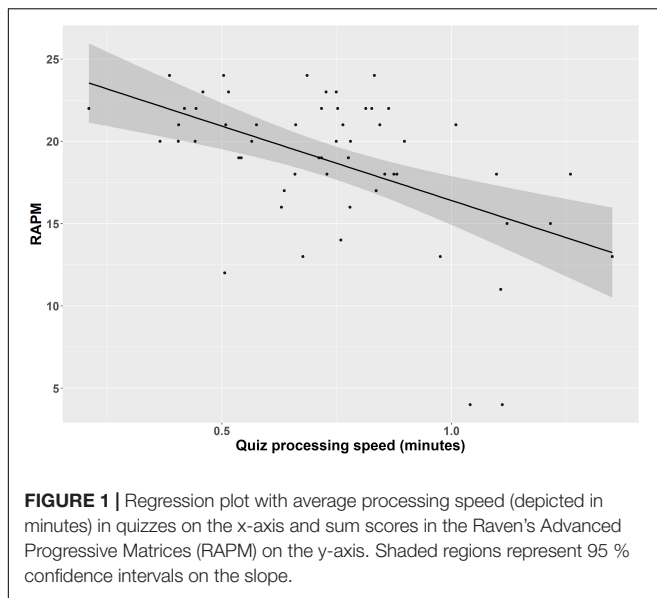
in quizzes, quiz processing speed) were regressed on our three predictors.

## Cognitive Ability and MOOC Retrieval Practice Activities

The results from the between-group analyses showed that fluid reasoning was a significant predictor for what group a given student belonged to after controlling for user activity. More specifically, a one point increase in the reasoning task increased the odds ratio of being a high-RP individual with 1.18. This finding aligns well with previous MOOC evidence, showing that cognitively high performing individuals typically tend to use optionally based quizzes more extensively than low-performing individuals on e-learning platforms (Fellman et al., 2020). Also in experimental settings, high-performing individuals typically use efficient study techniques (Barnett and Seefeldt, 1989), and strategies (Bailey et al., 2009) to a greater extent as compared with cognitively low-performing individuals.

The results from the within-group analyses showed that reasoning had a weak impact on the number of quizzes students took per reading session. Hence, this result is in discrepancy to the ones obtained from the between-group analysis. However, there might be other extraneous factors which potentially masks the true relationship between reasoning and quiz volumes in the latter analysis. First, our sample size was small in the within-group analysis, which increases the risk of making type II errors. Ideally, an inclusion of the low-RP group would both have increased the statistical power of the within-group analysis, and mimicked the between-group analysis to a greater extent, but as the participants in the low-RP group had barely engaged in retrieval practice activities on the MOOC platform, it prohibited us to include them in the analysis. Second, quizzing typically remains highly unutilized on MOOCs when the items remain only optionally available (Corral et al., 2020). Thus, it is also possible that quiz volume effects are difficult to observe when students merely use retrieval practice on MOOCs even if they have the possibility to do so.

As regards quiz performance, the results from the within-group analysis showed that those with better reasoning abilities had quiz higher accuracy scores on the MOOC. This result aligns well with a body of experimental evidence, showing that high performing individuals typically have better recall performance in retrieval practice items (Tse and Pu, 2012; Agarwal et al., 2017; Minear et al., 2018). Moreover, those performing better in the reasoning task processed the quizzes on the MOOC platform more rapidly as compared to those with lower reasoning scores (see **Figure 1**). This result is supported with factor-analytical evidence, showing that cognitive abilities and processing speed are correlated, yet separable constructs (Conway et al., 2002; Martínez and Colom, 2009). Thus, the relationship observed here does not deviate from findings typically obtained in laboratory settings.



## Personality and MOOC Retrieval Practice Activities

Regarding the relationship between retrieval practice activities and our personality measures were generally weak. More specifically, between-group analyses showed that both NFC and GRIT-S did not significantly predict to which group individuals belonged to. Accordingly, the within-group analyses revealed that the personality measures were not significantly related to neither quizzes taken per session, quiz performance, nor quiz processing speed. Our follow-up analyses also showed that our personality measures had no moderating effects on the relationship between RAPM and retrieval practice activity, which further underscores their insignificant influence on how individuals engage in retrieval practice activities online. To our knowledge, it is rather unstudied how personality characteristics relate to quiz performance in retrieval practice items on MOOCs. The only comparable data stems from studies in experimental settings (Bertilsson et al., 2017; Stenlund et al., 2017), indicating retrieval practice accuracies are weakly related to individual differences in personality. Thus, although the relationship between personality and retrieval practice activities was weak in the present study, it both supports and extends existing research by showing that personality measures are weakly related to retrieval practice activities on e-learning platforms as well.

## Limitations

There are several limitations that should be regarded as shortcomings in the present study, and that could be addressed in a better way in future studies. First, the present study encompassed a relatively homogenous sample, mainly consisting of medical students, with the majority of them probably belonging to the most gifted individuals in the normal population. Thus, the generalization of the findings of the present study to other online groups should be interpreted with caution. Second, and partly related to the issue above, 185 participants

out of a total of 15,000 students at Hippocampus took part in the study, with 105 participants included in the analyses. This is clearly a shortcoming, and which further underscores a tentative generalization of our results. Third, the study exhibited low statistical power, and thus the lack of effects, especially in the within-group analyses, can be questioned with respect to potential type II errors (Faber and Fonseca, 2014). The inclusion of the low-RP group would indeed have increased the sample size, but as mentioned earlier, their low engagement in retrieval practice activities on the MOOC platform prohibited us to include them in the analysis. Fourth, due to the lack of reliability and validity values of the retrieval practice activity variables that we extracted in this study, one can question what these outcome variables in fact capture. Although we cannot be entirely sure that each of them is tapping on relevant retrieval practice activities, we can, however, be confident that they at least measure different aspects of activity due to their relatively low intercorrelations with each other. Fifth, a shortcoming that pertains to all MOOCs is the lack of experimental control. The participants exhibited high independence when using the online platform, having the possibility to jump back and forth from a course to another, and complete courses and quizzes at their own pace. Future studies could assess the same research question as we did in the present study, yet with a more controlled user navigation and where participants receive identical stimuli during course completion.

## Conclusion and Future Directions

This study examined whether interindividual differences in cognitive ability, and personality characteristics were related to retrieval practice activities on a MOOC platform where students have the optional possibility to quiz themselves following each study session. Between-group analyses revealed that cognitively high performing individuals were more likely to use the optional quizzes on the platform. Moreover, within-group analyses including those students using the optional quizzes on the platform, showed that reasoning significantly predicted quiz performance, and quiz processing speed, but not number of quizzes. However, NFC and GRIT-S were unrelated to each of the aforementioned retrieval practice activities. From a more broad perspective, it appears that reasoning is a stronger predictor for retrieval practice usage on MOOCs as compared to self-assessed personality measures. Moreover, our results contribute to the research within teaching and learning across different environments in the digital age, by implicating that retrieval practice tend to be more used by cognitively high-performing individuals, bearing importance for MOOC administrators, especially from a personalization perspective (i.e., tailor-made learning in relation to students' personal profiles).

Furthermore, we hope that our obtained results could serve as a framework for forthcoming studies that examines individual differences in cognition, personality together with other potentially relevant background factors, and how these relates to retrieval practice activities on MOOCs. One interesting topic for further studies could be to specifically elucidate how other personality measures, such as Openness and Conscientiousness from the "Big five" personality inventory



(Goldberg et al., 2006), are related to retrieval practice activities on MOOCs.

## DATA AVAILABILITY STATEMENT

The datasets and scripts for feature extraction and analyses will be made available by the authors upon request.

## ETHICS STATEMENT

This study involving human participants was reviewed and approved by the Regional Vetting Committee (2017/517-31), Sweden. The participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

DF and BJ developed the study concept, conducted the feature extraction- and data preprocessing, performed the data analysis and interpretation, and drafted the manuscript. All

coauthors provided critical revisions and approved the final version of the manuscript for submission and contributed to the study design.

## FUNDING

This work was funded by the Swedish Research Council (2014–2099) to the third author.

## ACKNOWLEDGMENTS

We wish to thank Hippocampus AB for providing the researchers with data from the e-learning platform.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.02076/full#supplementary-material>

## REFERENCES

- Adesope, O. O., Trevisan, D. A., and Sundararajan, N. (2017). Rethinking the use of tests: a meta-analysis of practice testing. *Rev. Educ. Res.* 87, 659–701. doi: 10.3102/0034654316689306
- Agarwal, P. K., Bain, P. M., and Chamberlain, R. W. (2012). The value of applied research: retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educ. Psychol. Rev.* 24, 437–448. doi: 10.1007/s10648-012-9210-2
- Agarwal, P. K., Finley, J. R., Rose, N. S., and Roediger, H. L. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory* 25, 764–771. doi: 10.1080/09658211.2016.1220579
- Bailey, H., Dunlosky, J., and Hertzog, C. (2009). Does differential strategy use account for age-related deficits in working-memory performance? *Psychol. Aging* 24, 82–92. doi: 10.1037/a0014078
- Barnett, J. E., and Seefeldt, R. W. (1989). Read something once, why read it again?: repetitive reading and recall. *J. Liter. Res.* 21, 351–361. doi: 10.1080/10862968909547684
- Bertilsson, F., Wiklund-Hörnqvist, C., Stenlund, T., and Jonsson, B. (2017). The testing effect and its relation to working memory capacity and personality characteristics. *J. Cogn. Educ. Psychol.* 16, 241–259. doi: 10.1891/1945-8959.16.3.241
- Brewer, G. A., and Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *J. Mem. Lang.* 66, 407–415. doi: 10.1016/j.jml.2011.12.009
- Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 1118–1133. doi: 10.1037/a0019902
- Cacioppo, J. T., and Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology* 42, 116–131. doi: 10.1037/0022-3514.42.1.116
- Cacioppo, J. T., Petty, R. E., Feinstein, J. A., and Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: the life and times of individuals varying in need for cognition. *Psychol. Bull.* 119, 197–253. doi: 10.1037//0033-2909.119.2.197
- Carpenter, S. K., Rahman, S., Lund, T. J. S., Armstrong, P. I., Lamm, M. H., Reason, R. D., et al. (2017). Students' use of optional online reviews and its relationship to summative assessment outcomes in introductory biology. *CBE Life Sci. Educ.* 16, 1–9. doi: 10.1187/cbe.16-06-0205
- Clark, R., and Mayer, R. (2016). *e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning*, 4th Edn. Hoboken, NJ: Wiley.
- Conway, A. R. A., Cowan, N., Bunting, M. F., Theriault, D. J., and Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence* 30, 163–183. doi: 10.1016/S0160-2896(01)00096-4
- Corral, D., Carpenter, S. K., Perkins, K., and Gentile, D. A. (2020). Assessing students' use of optional online lecture reviews. *Appl. Cogn. Psychol.* 34, 318–329. doi: 10.1002/acp.3618
- Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., et al. (2005). On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. *Cogn. Psychol.* 51, 42–100. doi: 10.1016/j.cogpsych.2004.12.001
- Davis, D., Chen, G., van der Zee, T., Hauff, C., and Houben, G. J. (2016). "Retrieval practice and study planning in MOOCs: exploring classroom-based self-regulated learning strategies at scale," in *Adaptive and Adaptable Learning. EC-TEL 2016. Lecture Notes in Computer Science*, Vol. 9891, eds K. Verbert, M. Sharples, and T. Klobučar (Cham: Springer).
- Davis, D., Kizilcec, R. F., Hauff, C., and Houben, G. J. (2018). "The half-Life of MOOC knowledge a randomized trial evaluating knowledge retention and retrieval practice in MOOCs," in *Proceedings of the ACM International Conference Sydney*, NSW, 216–225.
- Dornic, S., Ekehammar, B., and Laaksonen, T. (1991). Tolerance for mental effort: self-ratings related to perception, performance and personality. *Pers. Individ. Differ.* 12, 313–319. doi: 10.1016/0191-8869(91)90118-U
- Duckworth, A. L., Peterson, C., Matthews, M. D., and Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *J. Pers. Soc. Psychol.* 92, 1087–1101. doi: 10.1037/0022-3514.92.6.1087
- Duckworth, A. L., and Quinn, P. D. (2009). Development and validation of the short Grit Scale (Grit-S). *J. Pers. Assess.* 91, 166–174. doi: 10.1080/00223890802634290
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., and Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol. Sci. Public Inter.* 14, 4–58. doi: 10.1177/1529100612453266



- Eskreis-Winkler, L., Shulman, E. P., Beal, S. A., and Duckworth, A. L. (2014). The grit effect: predicting retention in the military, the workplace, school and marriage. *Front. Psychol.* 5:36. doi: 10.3389/fpsyg.2014.00036
- Faber, J., and Fonseca, L. M. (2014). How sample size influences research outcomes. *Dent. Press J. Orthodont.* 19, 27–29. doi: 10.1590/2176-9451.19.4.027-029.ebo
- Fellman, D., Lincke, A., Berge, E., and Jonsson, B. (2020). Predicting Visuospatial and Verbal Working Memory by Individual Differences in E-Learning Activities. *Front. Educ.* 5:5–11. doi: 10.3389/educ.2020.00022
- Furnham, A., and Monsen, J. (2009). Personality traits and intelligence predict academic school grades. *Learn. Individ. Differ.* 19, 28–33. doi: 10.1016/j.lindif.2008.02.001
- Goldberg, L., Johnson, J., Eber, H., Hogan, R., Ashton, M., and Cloninger, R. H. G. (2006). The international personality item pool and the future of public-domain personality measures. *J. Res. Pers.* 40, 84–96. doi: 10.1016/j.jrjp.2005.08.007
- Kaplan, A. M., and Haenlein, M. (2016). Higher education and the digital revolution: about MOOCs, SPOCs, social media, and the Cookie Monster. *Bus. Horiz.* 59, 441–450. doi: 10.1016/j.bushor.2016.03.008
- Karpicke, J. D., and Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *J. Mem. Lang.* 57, 151–162. doi: 10.1016/j.jml.2006.09.004
- Kibble, J. (2007). Use of unsupervised online quizzes as formative assessment in a medical physiology course: effects of incentives on student participation and performance. *American J. Physiol. Adv. Physiol. Educ.* 31, 253–260. doi: 10.1152/advan.00027.2007
- Kornell, N., Bjork, R. A., and Garcia, M. A. (2011). Why tests appear to prevent forgetting: a distribution-based bifurcation model. *J. Mem. Lang.* 65, 85–97. doi: 10.1016/j.jml.2011.04.002
- Krumm, S., Ziegler, M., and Buehner, M. (2008). Reasoning and working memory as predictors of school grades. *Learn. Individ. Differ.* 18, 248–257. doi: 10.1016/j.lindif.2007.08.002
- Manochehr, N.-N. (2006). The influence of learning styles on learners in e-learning environments: an empirical study. *Comp. High. Educ. Econ. Rev.* 18, 10–14.
- Martínez, K., and Colom, R. (2009). Working memory capacity and processing efficiency predict fluid but not crystallized and spatial intelligence: evidence supporting the neural noise hypothesis. *Pers. Individ. Differ.* 46, 281–286. doi: 10.1016/j.paid.2008.10.012
- Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., and Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *J. Exp. Psychol. Learn. Mem. Cogn.* 44, 1474–1486. doi: 10.1037/xlm0000486
- Olson, B. L., and McDonald, J. L. (2004). Influence of online formative assessment upon student learning in biomedical science courses. *J. Dent. Educ.* 68, 656–659. doi: 10.1002/j.0022-0337.2004.68.6.tb03783.x
- Raven, J., Raven, J., and Court, J. (1991). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 1*. Oxford: Oxford University Press.
- Ren, X., Schweizer, K., Wang, T., and Xu, F. (2015). The prediction of students' academic performance with fluid intelligence in giving special consideration to the contribution of learning. *Adv. Cogn. Psychol.* 11, 97–105. doi: 10.5709/acp-0175-z
- Roediger, H. L., and Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends Cogn. Sci.* 15, 21–27. doi: 10.1016/j.tics.2010.09.003
- Roediger, H. L. and Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 1155–1159. doi: 10.1037/0278-7393.31.5.1155
- Roediger, H. L., and Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychol. Sci.* 2, 181–210. doi: 10.1111/j.1467-9280.2006.01693.x
- Röhlcke, S., Bäcklund, C., Sörman, D. E., and Jonsson, B. (2018). Time on task matters most in video game expertise. *PLoS One* 13:e0206555. doi: 10.1371/journal.pone.0206555
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychol. Bull.* 140, 1432–1463. doi: 10.1037/a0037559
- Sadowski, C. J., and Gülgös, S. (1996). Elaborative processing mediates the relationship between need for cognition and academic performance. *J. Psychol. Interdiscipl. Appl.* 130, 303–307. doi: 10.1080/00223980.1996.9915011
- Stenlund, T., and Jonsson, B. (2017). Assessing the willingness to elaborate among young students: psychometric evaluation of a swedish need for cognition scale. *Front. Educ.* 2:2. doi: 10.3389/educ.2017.00002
- Stenlund, T., Jönsson, F. U., and Jonsson, B. (2017). Group discussions and test-enhanced learning: individual learning outcomes and personality characteristics. *Educ. Psychol.* 37, 145–156. doi: 10.1080/01443410.2016.1143087
- Tabachnick, B. G., and Fidell, L. S. (2007). *Using Multivariate Statistics*, 5th Edn. Boston, MA: Pearson/Allyn & Bacon.
- Tse, C. S., and Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *J. Exp. Psychol. Appl.* 18, 253–264. doi: 10.1037/a0029190
- Turner, M. L., and Engle, R. W. (1989). Is working memory capacity task dependent? *J. Mem. Lang.* 28, 127–154. doi: 10.1016/0749-596X(89)90040-5
- van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* 45, 1–67. doi: 10.18637/jss.v045.i03
- van den Broek, G., Takashima, A., Wiklund-Hörnqvist, C., Wirebring, L. K., and Nyberg, L., et al. (2016). Neurocognitive mechanisms of the “testing effect”: A review. *Trends Neurosci. Educ.* 5, 52–66. doi: 10.1016/j.tine.2016.05.001
- van der Zee, T., Davis, D., Saab, N., Giesbers, B., Ginn, J., van der Sluis, F., et al. (2018). “Evaluating retrieval practice in a MOOC,” in *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, Sydney, NSW, 216–225.
- Weinstein, Y., McDermott, K. B., and Roediger, H. L. (2010). A comparison of study strategies for passages: rereading, answering questions, and generating questions. *J. Exp. Psychol. Appl.* 16, 308–316. doi: 10.1037/a0020992
- Wiklund-Hörnqvist, C., Jonsson, B., and Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scand. J. Psychol.* 55, 10–16. doi: 10.1111/sjop.12093

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Fellman, Lincke and Jonsson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Changes in Students' Understanding of and Visual Attention on Digitally Represented Graphs Across Two Domains in Higher Education: A Postreplication Study

Sebastian Brückner<sup>1\*</sup>, Olga Zlatkin-Troitschanskaia<sup>1</sup>, Stefan Küchemann<sup>2</sup>, Pascal Klein<sup>3</sup> and Jochen Kuhn<sup>2</sup>

<sup>1</sup> Chair of Business and Economics Education, Johannes Gutenberg-University Mainz, Mainz, Germany, <sup>2</sup> Physics Education Research Group, Technische Universität Kaiserslautern, Kaiserslautern, Germany, <sup>3</sup> Faculty of Physics, Georg August University Göttingen, Göttingen, Germany

## OPEN ACCESS

### Edited by:

Raquel Cerdan,  
University of Valencia, Spain

### Reviewed by:

Gaston Saux,  
Consejo Nacional de Investigaciones  
Científicas y Técnicas (CONICET),  
Argentina  
Dan Zhang,  
Tsinghua University, China

### \*Correspondence:

Sebastian Brückner  
brueckner@uni-mainz.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Psychology

Received: 13 May 2020

Accepted: 28 July 2020

Published: 27 August 2020

### Citation:

Brückner S,  
Zlatkin-Troitschanskaia O,  
Küchemann S, Klein P and Kuhn J  
(2020) Changes in Students'  
Understanding of and Visual Attention  
on Digitally Represented Graphs  
Across Two Domains in Higher  
Education: A Postreplication Study.  
Front. Psychol. 11:2090.  
doi: 10.3389/fpsyg.2020.02090

Domain-specific understanding of digitally represented graphs is necessary for successful learning within and across domains in higher education. Two recent studies conducted a cross-sectional analysis of graph understanding in different contexts (physics and finance), task concepts, and question types among students of physics, psychology, and economics. However, neither changes in graph processing nor changes in test scores over the course of one semester have been sufficiently researched so far. This eye-tracking replication study with a pretest–posttest design examines and contrasts changes in physics and economics students' understanding of linear physics and finance graphs. It analyzes the relations between changes in students' gaze behavior regarding relevant graph areas, scores, and self-reported task-related confidence. The results indicate domain-specific, context- and concept-related differences in the development of graph understanding over the first semester, as well as its successful transferability across the different contexts and concepts. Specifically, we discovered a tendency of physics students to develop a task-independent overconfidence in the graph understanding during the first semester.

**Keywords:** graph understanding, pretest–posttest, eye-tracking, dwell times, confidence rating, university students

## RESEARCH FOCUS AND OBJECTIVE

The ability to understand digitally represented graphs is a necessary prerequisite for (online) learning in most disciplines in higher education<sup>1</sup> (Bowen and Roth, 1998). In general, graphs are used to simplify the presentation of (complex) concepts and to facilitate the exchange of information between individuals (Curcio, 1987; Pinker, 1990; Freedman and Shah, 2002). Because

<sup>1</sup> In biology, for example, it is important to make developments (e.g., cell division) visible (Bergey et al., 2015; Kragten et al., 2015); in mathematics and statistics, relationships between variables, their distributions and progressions can be graphically visualized (Lichti and Roth, 2019).

the graphical representation of information is increasingly becoming more important than texts in the online information landscape (Moghavvemi et al., 2018), the ability to interpret graphs is considered a central facet of cross-domain generic skills such as online reasoning (Wineburg et al., 2018), media literacy (Shah and Hoeffner, 2002), data literacy (Cowie and Cooper, 2017), and information problem-solving (Brand-Gruwel et al., 2009).

Because graphs and other types of diagrams are an instructional method for representing both domain-specific and generic knowledge, they are the main focus in teaching, especially at the beginning of university studies (Heublein, 2014; Laging and Voßkamp, 2017). Usually, the graphs are embedded in text-based instructions to aid the comprehension of textual descriptions and to supplement these descriptions by providing the learner with further visually structured information (Stern et al., 2003).

Line graphs, in particular, are frequently used in higher education. For example, the relationships between distance and speed in physics or between time and stock prices in finance can both be illustrated with a line graph (Bowen and Roth, 1998; Benedict and Hoag, 2012; Susac et al., 2018; Becker et al., 2020a,b; Hochberg et al., 2020; Klein et al., 2020). More recently, a number of studies investigated and compared university students' understanding of graphs in mathematics, physics, and other contexts using parallel (isomorphic) tasks (Christensen and Thompson, 2012; Planinic et al., 2012, 2013; Wemyss and van Kampen, 2013; Bollen et al., 2016; Ivanjek et al., 2016, 2017). These studies have shown that parallel tasks with an added context (physics or other context) were more difficult to solve than the corresponding mathematics problems and that students who successfully solve problems in (purely) mathematical contexts often fail to solve corresponding problems in physics or other contexts. Other studies have discovered that students often struggle to interpret line graphs or solve problems based on line graphs (Canham and Hegarty, 2010; Kragten et al., 2015; Miller et al., 2016). Students do not succeed in transforming data into line graphs (Bowen and Roth, 1998); they do not spend sufficient time trying to understand the depicted concepts (Miller et al., 2016) or have difficulties comprehending the underlying concept.

Although the major importance of being able to correctly interpret visual representations and graphs within and across domains (Beichner, 1994; Stern et al., 2003; Planinic et al., 2013), which must be distinguished from the ability to understand textual representations (Mayer, 2009), is widely known and recognized, research on the ability of students in higher education to solve problems with digitally represented graphs combined with results on how students extract information from graphs within and across domains is still scarce. In particular, there are only very few studies on the development of students' graph understanding over a degree course.

In this paper, we address this research deficit in a post-replication study by following up on two existing studies by Susac et al. (2018) and of Klein et al. (2019). Both studies investigated students' allocation of visual attention, i.e., how students extract information from graphs, during problem-solving in relation to their scores. In our study, we extend this approach by including

a comparison of pre- and post-test results. For this purpose, we use the same graph tasks from the two domains (physics and economics) that were chosen in the two reference studies. To gain initial insights about a change in students' graph comprehension within and across domains, we also retest a subset of the same students who previously participated in Klein et al.'s study (2019) at the end of their first semester.

To achieve a higher degree of (external) validity and generalizability, the replication of a study requires a comprehensive presentation of the control variables and can expand the original study in some aspects (Schmidt, 2009). The study presented here, in addition to a replicating previous research, was expanded through the addition of the second measurement point. As learning with graph tasks, especially in first semester lectures, is an integral part of the curriculum and instruction in both domains examined here (e.g., Jensen, 2011), more in-depth knowledge and skills can be acquired by attending such lectures, and a change in graph understanding in these two domains can be expected. Thus, in this post-replication study, changes in students' understanding of graphs are investigated within and across the two domains physics and economics. Moreover, previous research indicates that while students' understanding of graphs can improve after a targeted intervention, students did not improve in transferring this ability to different task contexts (e.g., Klein et al., 2015). Therefore, in this study, we investigate whether eye movements are indicative of increases in graph understanding and potential weaknesses in transferring graph understanding across different domains and contexts.

Based on these studies, we developed the following research questions (RQ) for this article, which focus on the theoretically expected (i) time effects (measurements t1 and t2), (ii) domain effects (physics and economics), (iii) (task) context effects, and possible (vi) interaction effects:

- RQ1: To what extent does the ability of students from both domains to solve line-graph problems in physics and finance contexts change over the course of the first semester?
- RQ2: Are the confidence ratings of graph task solutions in physics and finance contexts of students from both domains higher at the end of the semester, and how do they change with respect to correct and incorrect responses?
- RQ3: How do the dwell times on specific parts of graph tasks in physics and finance contexts of students from both domains change between the beginning and the end of the semester?

In the following, the two studies by Susac et al. (2018) and Klein et al. (2019) that this replication study is based on are described in detail. Next, we expand the focus on the two domains examined and theoretically ground the additional research focus on the development of the students' graph understanding. The hypotheses for this study are formulated based on the defined conceptual and methodological frameworks. These, in turn, are based on the method of eye-tracking (ET).

## BACKGROUND OF THE POST-REPLICATION STUDY

### Cross-Sectional Studies by Susac et al. (2018) and Klein et al. (2019)

In a recent study in Croatia (Susac et al., 2018) and a German replication study (Klein et al., 2019), students' graph understanding in physical and economic tasks was experimentally investigated for the first time. In a  $2 \times 2 \times 2$  study design, permuted systematically according to three characteristics, the graphical concept (graph "slope" vs. "area" under the curve), the type of question (quantitative vs. qualitative), and the domain-specific context (physics vs. finance) were differentiated into four isomorphic task pairs (eight tasks in total; Klein et al., 2019; Susac et al., 2018). Comparing the students from two different domains (in Susac et al., students of physics and psychology responding to graph tasks from physics and finance; in Klein et al., students of physics and economics responding to graph tasks from physics and finance), both studies confirm the differences between the two domains; for instance, physics students spend a longer period of time on unknown axis tick labels and analyzing the curve, whereas within the domain of economics, students there were no significant differences (Susac et al., 2018; Klein et al., 2019).

In the cross-sectional study by Susac et al. (2018), students often found it more difficult to make calculations based on the graph concept "area" (e.g., using an integral) than to determine the "slope" of a graph. This result confirms existing findings and theoretical assumptions (Cohn et al., 2001; Benedict and Hoag, 2012). Klein et al. (2019) found that "area" tasks required more time and were therefore cognitively more demanding than "slope" tasks for both domains (physics and economics).

With regard to the transfer of task solutions across domains, both studies found that physics students, who are the better task solvers in one task context (physics), also performed better in another context (finance). For instance, physics students achieved similarly good results in graph understanding in both examined domains, however, they solved the tasks from the domain of physics more quickly than the tasks from the domain of economics. Psychology students generally scored comparatively lower in graph understanding (Susac et al., 2018). Klein et al. (2019) found similar differences.

Comparing tasks that require calculations (quantitative) and those that require only verbal interpretation (qualitative), both studies demonstrated that quantitative tasks are generally more challenging as the students achieved lower scores and at the same time took longer to complete these tasks (Susac et al., 2018; Klein et al., 2019). This finding is in line with existing research, indicating that students have specific difficulties when solving tasks with numerical or mathematical requirements (Planinic et al., 2012; Shavelson et al., 2019).

In addition to an analysis of task scores and retention times, Klein et al. (2019) also collected the students' self-assessments of their task solutions and compared them with the actual scores. The metacognitive assessment provided further significant insights into the students' expertise, in

particular between high- and low-performing students. In line with this existing research, Klein et al. (2019) found better self-assessments among high-performing students (Brückner and Zlatkin-Troitschanskaia, 2018) and a systematic overestimation of their own abilities among low-performing students (Kruger and Dunning, 1999). The physics students provided correct answers with higher confidence ratings in comparison to instances when they gave incorrect answers, whereas economics students who achieved lower scores also gave lower confidence ratings with regard to their own performance.

For the postreplication study, the following assumptions can be summarized:

- There are significant differences between students from the two domains when it comes to solving graph problems from familiar versus unfamiliar contexts.
- Students with high test scores assess the correctness of their solutions more precisely than students with lower test scores.
- Graph tasks with a focus on the "area under the curve" or with quantitative requirements are more difficult for students from both domains than tasks on the concept of "slope" or without mathematical requirements. This applies to both task contexts (physics and finance).

Because the data of the study by Susac et al. (2018) and the replication study (Klein et al., 2019) only originated from assessment at one point in time, changes that must be expected over the course of a semester cannot be described. As longitudinal studies indicate a significant change in knowledge during the first semester (Happ et al., 2016; Chen et al., 2020; Schlax et al., 2020), our postreplication study was expanded to include a so far underresearched developmental focus.

### Development of Graph Understanding

Through the systematic use of learning tasks comprising graph representations in different domains, especially at the beginning of studies, a positive development of graph understanding can be assumed because the acquisition of domain-specific knowledge is expected to support students in solving typical domain-specific problems related to graphs (e.g., McDermott et al., 1987). However, there are currently only few studies with a pretest–posttest assessment design focusing on the changes in graph understanding and how to foster this understanding. Digital learning environments, learning from examples, and using instructional material showed an impact on students' graph comprehension (Bell and Janvier, 1981; Bergery et al., 2015; Becker et al., 2020a,b; Hochberg et al., 2020). For example, the impact of instruction on graph construction conventions (e.g., on legends and labels) on students' graph understanding was confirmed in a control group design (Miller et al., 2016). By systematically training the (prospective) teachers as well as the students over several weeks, graph understanding in biology was promoted (Cromley et al., 2013).

In supplementing the instruction of graph use with material for understanding multiple representations (e.g., how data can be visualized in a graph or how information for graph use



can be meaningfully extracted from a text), multiple ways to promote graph understanding over a period of 2 months were identified (Bergey et al., 2015). In a study with an augmented reality intervention over a period of 3 weeks, students of the intervention group showed an improvement in understanding that exceeded the increase in understanding of a control group with no such intervention (Jerry and Aaron, 2010). In some studies, a positive relationship between textual perception and the understanding of visual representations of domain-related concepts was found only among students with poor spatial abilities (e.g., Bartholomé and Bromme, 2009). Overall, a positive, instructionally initiated development of graph comprehension was found for several educational levels, types of instruction, and domains (Miller et al., 2016).

The development of graph understanding is often considered a generic skill that is also transferable to graphs in other contexts and domains (Miller et al., 2016). So far, however, recent research indicates learner difficulties in transferring graph understanding across problems and domains (Bergey et al., 2015; Cromley et al., 2013). For instance, one pretest–posttest study investigated to what extent university students succeed in applying mathematical functions for the “slope” of a curve to the context of physics (Woolnough, 2000). The posttest after 1 year showed an improved calculation and interpretation of the gradient, as well as a frequent use of the concept of proportionality, but the students had difficulties with the transfer from model to real world. Bergey et al. (2015) and Cromley et al. (2013) showed that training based on graph uses and learning with multiple representations can improve the understanding of graphs in biology, but the students did not succeed in transferring their skills to graphs in geoscience.

Although the ability to understand graphs is necessary for the development of domain-specific knowledge and conceptual change, especially in the domains of physics and economics<sup>2</sup>, so far little research has been conducted on the development of graph understanding in these two domains. Whereas in physics, and especially in physics education, there are many studies on graph understanding in kinematics (McDermott et al., 1987; Beichner, 1994; Planinic et al., 2013; Wemyss and van Kampen, 2013; Klein et al., 2020), in economics, no research field has yet been established that explicitly analyzes graph understanding (Cohn et al., 2001; Stern et al., 2003; Benedict and Hoag, 2012). In particular, it is yet underresearched to what extent a transfer between more distant disciplines, such as between natural and social science disciplines, can succeed.

While Klein et al. (2019) showed the connection between the self-assessment of solutions and the correct answers to these graph tasks, there are hardly any studies that investigate this relationship over time. However, prior, longitudinal research has identified correlations of this kind in studies using general knowledge tests (without graphs; Cordova et al., 2014; Brückner

and Zlatkin-Troitschanskaia, 2018). Moreover, the Dunning–Kruger effect (Kruger and Dunning, 1999) suggests that learners with a low level of knowledge struggle to rate their own performance accurately in self-assessments. Based on these results, it can be assumed that an increase in (conceptual) knowledge and (graph) understanding is accompanied by a more precise self-assessment of knowledge.

In summary, based on research questions 1 and 2, the following hypotheses (H) can be formulated and will be examined in this study with regard to the changes in graph understanding:

- H1: Physics and economics students solve graph tasks related to the subject they are enrolled in more successfully at the second measurement point than at the first measurement point.
- H2: Physics and economics students rate their confidence in their solution of tasks related to the subject they are enrolled in more accurately at the second measurement point than at the first measurement point.

## Eye-Tracking and Graph Understanding

In recent years, ET is increasingly used to study visual representations in general (e.g., Küchemann et al., 2020b) and graph understanding in particular (Madsen et al., 2012; Planinic et al., 2013; Klein et al., 2017, 2020) as it offers many advantages, especially for uncovering the systematics underlying the perception of different graphical representations (Küchemann et al., 2020a) and can also supplement the findings on changes in test scores and self-assessments with evidence obtained from changes in eye movements. This method is also used in the two studies by Susac et al. (2018) and Klein et al. (2019) referenced here.

According to the Eye-Mind-Hypothesis (Just and Carpenter, 1980), there is a strong spatiotemporal and causal connection between visual attention and the associated cognitive processes. The visual representation of graphs includes, for instance, axes and labels, which can be arranged in different ways and, depending on the intensity and duration of the observation, can also impact understanding of the graph. For example, a comparatively longer fixation time on relevant areas of a graph was mainly observed in students who solved a task correctly (Madsen et al., 2012; Susac et al., 2018; Klein et al., 2019). Regarding the dwell time for processing one task, students' previous experience and familiarity with tasks of this kind ease their comprehension; thus, it can be expected that such effects also develop over time and that students need less time overall for solving a task. The transfer between contexts can also be made visible by analyzing the corresponding eye movements on components of the graph (Susac et al., 2018; Klein et al., 2019). However, to date, there is no ET study that analyzes changes in students' problem-solving of digitally presented graphs across two domains using pretest–posttest measurements at the beginning and end of a semester.

With regard to the additional ET data from the second measurement point, the factor “time” will be integrated into the previous models by Susac et al. (2018) and Klein et al. (2019) to analyze the following hypothesis with regard to the expected

<sup>2</sup>In economics, graphs are systematically used in learning situations to explain complex phenomena (e.g., economic developments, inflation, gradients) and also in (non)standardized examinations. Graphs are also an integral part of methodological lessons focusing on modeling economic content and its graphical representation, especially in the first semesters of an economics degree course (e.g., Jensen, 2011).

developmental effects within and across the two domains and contexts on the relationship between dwell times and test scores:

- H3: The dwell time on the tasks and the individual graph components [areas of interest (AOIs)] is lower at the second measurement point for students from both domains and in both contexts.

MATERIALS AND METHODS

Sample

As a postreplication study, we based the present article on the sample of the replication study by Klein et al. (2019) and carried out a second measurement at the end of the winter semester 2018/19. The first measurement (t1) took place during the first weeks of the students’ first semester. The students were tested again at the end of the semester (t2). During the semester, they attended courses and learned about graphs in their respective domains. At t2, the same graph tasks were presented to the students. Study participation was voluntary and was compensated with 20€.

In total, 41 first-year students (matched sample) from the initial study of Klein et al. (2019) participated in the experiment again (Table 1): 20 physics students and 21 economics students. The average age in the sample was 20.27 years, with physics students being slightly younger (19.95 years) than economics students (20.57 years). The grade for higher education entrance qualification also differs systematically between the physics [P] and the economics students [E] [ $t = -2.784, p = 0.009, d = 0.972$ ]; mean (SD)  $P = 1.79$  (0.478); mean (SD)  $E = 2.25$  (0.469)]; 89% of the physics students took an advanced physics course in upper secondary education, whereas only 16% of the economics students attended advanced courses. For an extended description of the sample, see Klein et al. (2019).

Tasks

To assess students’ graph understanding within and across domains, graphs that are regularly used in both domains and are important for learning domain-specific concepts are required. Linear graphs are used extensively in both physics (Klein et al., 2019) and economics (Benedict and Hoag, 2012) and are clearly distinguishable from other forms of graphical representation (e.g., pie charts, Venn diagrams) (Kosslyn, 1999). Although other

graphs are also used in both domains, our study focuses on one single type of graph to avoid distortions caused by the graph type (Strobel et al., 2018).

The study presented here used four isomorphic pairs of line graph tasks (4 from physics and 4 from economics) as they were used by Susac et al. (2018) and Klein et al. (2019) (Figure 1). A  $2 \times 2 \times 2$  (context  $\times$  question  $\times$  concept) design was applied, in which each task belongs either to the domain of physics or economics (context), contains either the graph concept of “slope” or “area” (concept), and requires either a mathematical calculation or purely verbal reasoning from the participants (type of question) (for an example, see Figure 1).

All tasks are presented in a closed-ended format and comprise a question of one or two sentences, a graph, and one correct and up to four incorrect response options. Each graph task also comprises one or two linear curves and other common elements like  $x$ -axis and  $y$ -axis.

Apparatus and ET Analysis

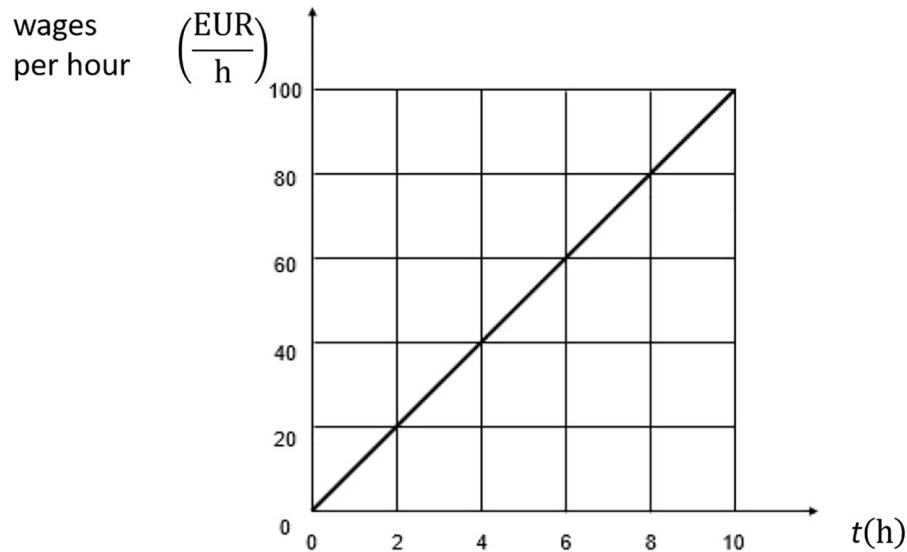
To perform the ET study, the graph comprehension tasks were presented to the students on a 22-inch computer screen ( $1,920 \times 1,080$  pixels). A Tobii Pro X3-120 (120 Hz), which is positioned below the monitor and is not worn by the test taker, was used to record the ET data. The visual angle resolution was below  $0.4^\circ$ . The dwell time (eye movements below an acceleration of  $8,500^\circ/s^2$  and a velocity below  $30^\circ/s$ ) was assessed and used to measure the students’ focus on selected AOIs in the tasks (Figure 2). The AOIs included the task question, the graph itself, axes, and the response options.

After a 9-point calibration process, the eight tasks were presented to the students in a random order, and ensuring that two subsequent tasks were never equal in concept and type of question to avoid students realizing that some tasks only varied in context and realizing that they just need to apply the same task-solving strategy. The order in which the tasks were administered to the students also ensured that isomorphic tasks were never presented one directly after the other. After viewing a task, the test takers had to click and choose one answer from the presented response options using a mouse. Then they had to rate how confident they were that their chosen response option was correct on a six-point Likert scale ranging from very high confidence to very low confidence. By pressing the spacebar, the test takers could proceed to the next task. After the test, each task was

TABLE 1 | Comparison of the postreplication study with the original study by Susac et al. (2018) and the replication study from Klein et al. (2019).

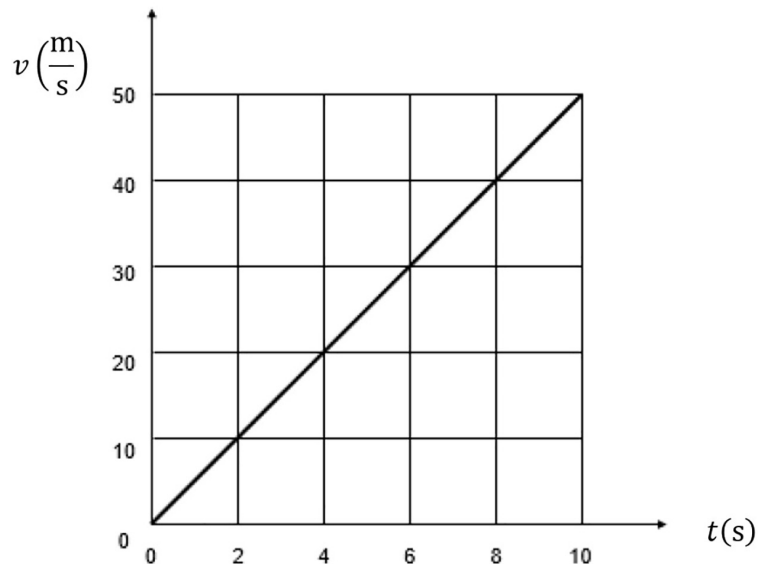
	This study	Klein et al., 2019	Susac et al., 2018
Participants	20 physics students (first year), 21 economics students at t1 and t2	27 physics students (first year), 40 economics students	45 physics students (teacher program, fourth year), 45 psychology students
Materials	Four isomorphic pairs of items about graph slope and area under a curve in the context of physics and economics (finance)		
Apparatus		Tobii X3–120 Hz	SMI RED500 Hz
Additional data		Confidence scores	Student strategies (questionnaire)
Coding scheme		Only direct response (correct or incorrect)	Response and explanation (correction)
Data analysis	ANOVAs to determine the effects of question type, concepts, group, and context on the dependent variables Area of interest (AOIs) question, graph, multiple choice, axis labels, axis tick labels		
Analytic focus	Analysis of student change between t1 and t2	Saccadic direction, attention distribution	Analysis of student strategies

- A** The hourly wage when working in a nuclear power plant increases with longer working hours as shown in the diagram below. How much does an employee earn during the first 8 hours?



- (a) 80 EUR      (b) 100 EUR      (c) 320 EUR      (d) 640 EUR      (e) 1000 EUR

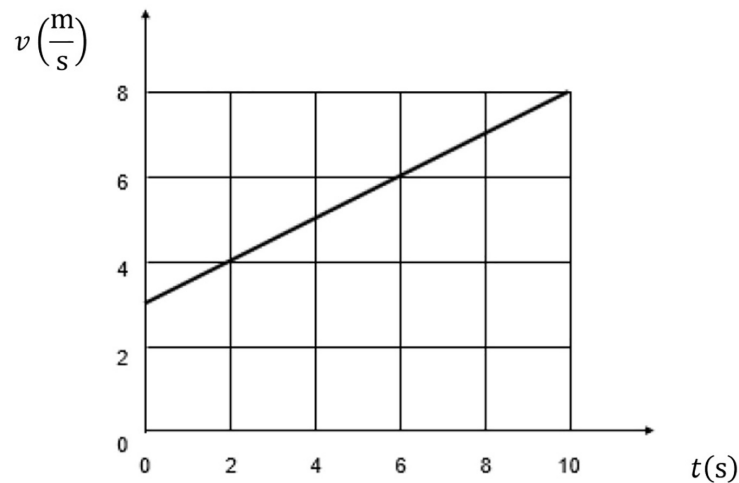
- B** The speed of a Formula 1 car increases as shown in the graph below. That distance does the car travel in the first 8 seconds?



- (a) 40 m      (b) 50 m      (c) 100 m      (d) 160 m      (e) 320 m

**FIGURE 1** | Continued

- C** The diagram shows the temporal course of the speed of a train. Determine the acceleration of the train.



- (a)  $0.5 \text{ m/s}^2$       (b)  $0.8 \text{ m/s}^2$       (c)  $1 \text{ m/s}^2$       (d)  $5 \text{ m/s}^2$       (e)  $8 \text{ m/s}^2$

**FIGURE 1** | Isomorphic task examples: **(A)** quantitative “question type” and the graph concept “area” for economics “context”; **(B)** quantitative “question type” and the graph concept “area” for physics “context”; **(C)** quantitative “question type” and the graph concept “slope” for physics “context” (the “qualitative” concept slope for both contexts can be seen in Klein et al., 2019).

coded with 1 if a student chose the correct response (attractor) and 0 if a student chose one of the distractors (maximum score: 8 points). The confidence rating and the task score sum were linearly transformed into a scale reaching from 0 to 100, with 0 indicating low scores and low confidence and 100 indicating high scores and high confidence.

After completion of all graph tasks (at  $t_1$  and  $t_2$ ), paper-and-pencil questionnaires were administered to collect sociodemographic data (e.g., gender, school education, school leaving grade; for details, see Klein et al., 2019).

## Statistical Approaches

To answer the research questions, several repeated-measures analyses of variance (ANOVAs) were performed, which were also used by Susac et al. (2018) and Klein et al. (2019). This allowed us to systematically explore the relationships between task characteristics (*context*, *concept* and *type of questions*) and examined domains (physics vs. economics) on the basis of the final test scores and to make the comparison of the findings between the three studies transparent. The *measurement point* ( $t_1$ : beginning of first semester or  $t_2$ : end of first semester), the *context*, the *concept*, and the *question* were modeled as within-subject factors, and the domain (physics vs. economics) as intermediate subject factor.

To test the null hypothesis that variance is equal across domains and measurement points, Levene test was used, and

the assumption of homogeneity of variance was met for every ANOVA. Analogous to Susac et al. (2018) and Klein et al. (2019), correlations were calculated using the Bravais–Pearson correlation coefficient.

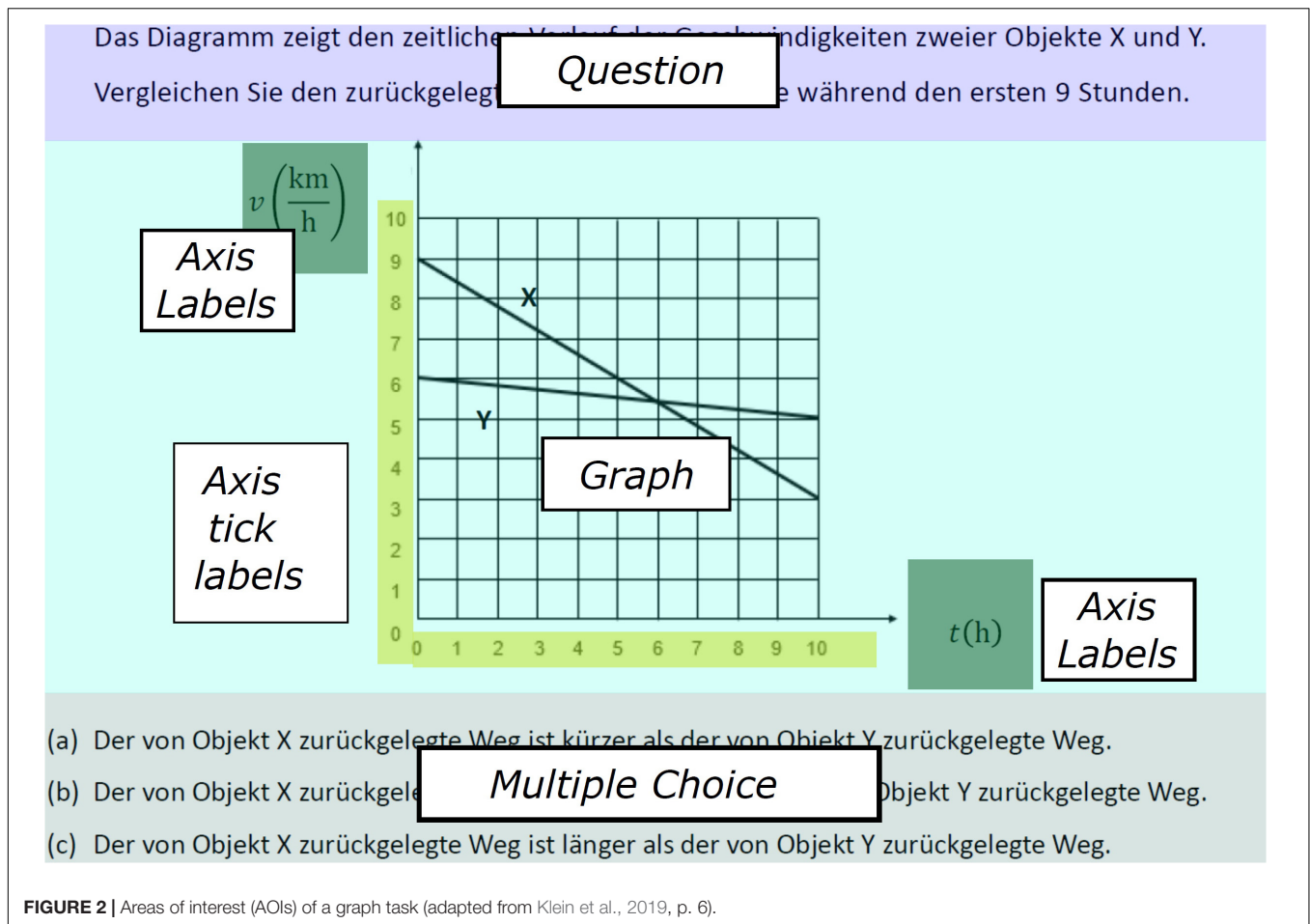
As with the test scores analysis, repeated-measures ANOVAs were performed to analyze the dwell times, taking into account the task characteristics and domains. In addition to the total dwell time during task processing, the dwell times on task-relevant AOIs and on the task questions were analyzed. The total processing time can vary at the second measurement because the test takers are familiar with the type of tasks, recognition effects may occur, and they have attended domain-specific classes in which they learned about graphs in their specific contexts.

## RESULTS

### Changes in Students’ Test Scores Within and Across Domains (H1)

The mean test score of the pretest–posttest sample was ( $60\% \pm 27\%$ ) in  $t_1$  and ( $65\% \pm 25\%$ ) in  $t_2$ , with a change with a small effect size (Cohen, 1988) [ $t(40) = 1.366$ ,  $p = 0.18$ ,  $d = 0.21$ ]. A comparison of the two domains shows that the physics students achieved better results at both measurement points [ $t_1$ : ( $70\% \pm 27\%$ ),  $t_2$ : ( $78\% \pm 18\%$ )] than the economics students [ $t_1$ : ( $49\% \pm 24\%$ ),  $t_2$ : ( $52\% \pm 23\%$ )]. They also





show a comparatively higher increase of about 10% in the test score than the economics students with about 6%. An ANOVA with repeated measurements (t1 and t2) as inner-subject factor and the domain (physics students and economics students) as intermediate subject factor showed that the mean test score difference between the domains is also significantly higher with a large effect size [ $F(1, 39) = 13.355$ ;  $p = 0.001$ ;  $\eta^2_p = 0.255$ ]. However, no significant differences in the increase from t1 to t2 between the two domains, which are mapped by the interaction term (time  $\times$  domain), can be identified [ $F(1, 39) = 0.293$ ;  $p = 0.592$ ;  $\eta^2_p = 0.007$ ]. Thus, students from both domains showed a similar increase in the overall test score.

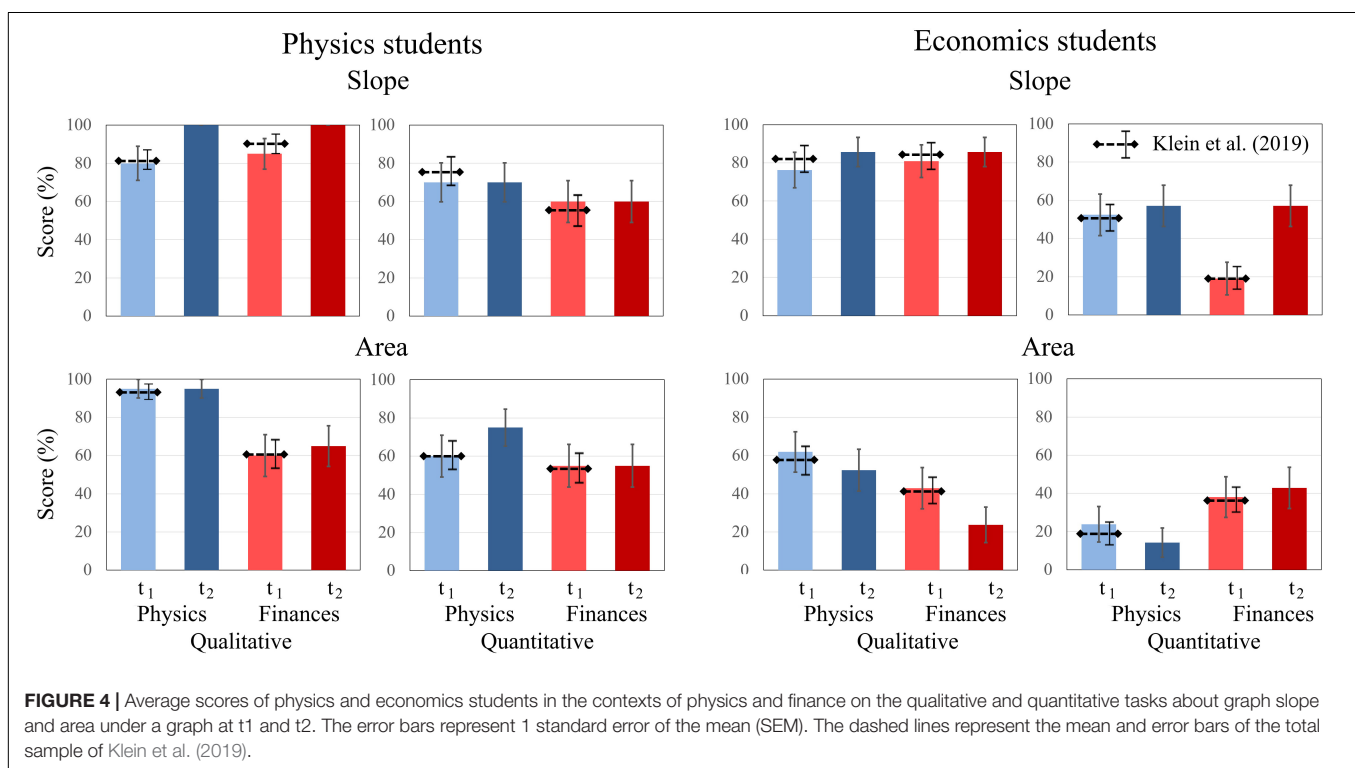
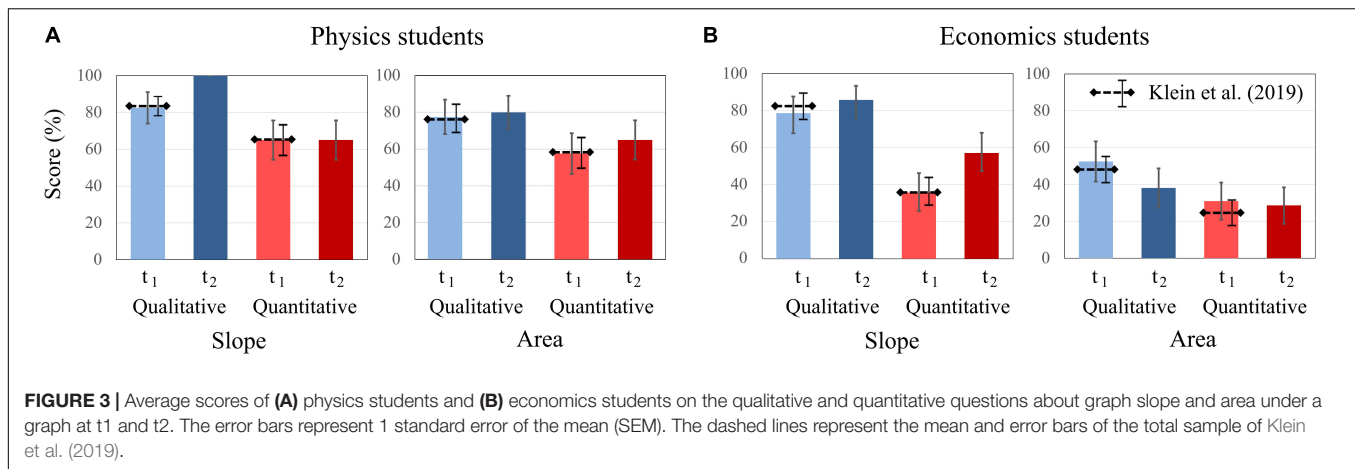
Next, the changes in the test score between the two measurements (t1 and t2) were examined with regard to the question type (qualitative vs. quantitative) and the concept (graph “slope” vs. “area” under the curve). A two-way repeated-measures ANOVA was conducted for each domain. For physics students, a statistically significant main effect was found only for the type of question [ $F(1, 19) = 14.968$ ,  $p = 0.001$ ;  $\eta^2_p = 0.441$ ] and no effect for time [ $F(1, 19) = 1.667$ ,  $p = 0.212$ ,  $\eta^2_p = 0.081$ ] or concept [ $F(1, 19) = 3.449$ ,  $p = 0.079$ ;  $\eta^2_p = 0.154$ ]. The interaction effects were not significant either. For economics students, a significant general time effect was not evident [ $F(1, 20) = 0.373$ ,  $p = 0.548$ ;  $\eta^2_p = 0.018$ ], but significant effects

for question [ $F(1, 20) = 39.174$ ,  $p = 0.000$ ;  $\eta^2_p = 0.662$ ], concept [ $F(1, 20) = 21.774$ ,  $p = 0.000$ ;  $\eta^2_p = 0.521$ ], and the time  $\times$  concept interaction [ $F(1, 20) = 14.440$ ,  $p = 0.001$ ;  $\eta^2_p = 0.419$ ] were found.

Similar to Klein et al. (2019), both physics and economics students scored higher on qualitative than on quantitative tasks. Economics students generally scored higher on tasks that cover the concept of “slope” than on tasks on the concept of “area.” Furthermore, for economics students, there are differences in the changes of the test scores between the two concepts. Economics students’ scores increase on items of “slope” [t1: 57.14%; t2: 71.43%] but decrease on “area” tasks [t1: 41.66%; t2: 33.33%]. Other interaction effects were not significant. For the economics students, the difference between scores on qualitative and quantitative tasks was larger for questions about “slope” from t1 to t2 (Figure 3). “Slope” tasks with quantitative requirements show the largest increase in the scores of economics students.

For physics students, the biggest change was in the test scores of qualitative graphs on “slope” from t1 to t2. In t2, all physics students solved these items correctly.

To compare students from both domains across both contexts, we applied a repeated-measures ANOVA with time and context (physics vs. finance) as a within-subject factor and with the domain (physics vs. economics) as a between-subject factor.



The analysis was performed for each pair of isomorphic tasks (Table 2). Similar to Klein et al. (2019), for qualitative tasks about “slope,” significant differences for time but no other main or interaction effects were found. For quantitative tasks on “slope,” a significant main effect was found only for task context, indicating that across both measurements (t1 and t2) and domains, students generally scored higher on tasks with a physics context than tasks with a finance context. Compared to Klein et al. (2019), students from both domains still solved physics tasks better than finance tasks, although economics students’ scores on quantitative tasks on “slope” in their own domain increased significantly [ $t(20)$ ,  $p = 0.017$ ,  $d = 0.567$ ] (Table 2).

Compared to Klein et al. (2019), physics students had higher scores at both t1 and t2 on qualitative tasks on the “area under

the curve” than economics students. The economics students’ scores on “area” tasks differ from their scores on all other types of task. From t1 to t2, their test scores decreased in the physics context and increased slightly in the finance context, and both for qualitative and quantitative tasks (Figure 4). Thus, significant domain effects for both tasks were found, but no time  $\times$  domain  $\times$  context effect occurred (Table 2).

Overall, physics students scored significantly better on physics tasks than on finance tasks at t1 [ $t(19) = 2.131$ ,  $p = 0.046$ ,  $d = 0.466$ ] and t2 [ $t(19) = 3.040$ ,  $p = 0.007$ ,  $d = 0.68$ ]. The effects for economics students were not significant, although they increased their score on finance tasks (t1: 45%  $\pm$  21%, t2: 52%  $\pm$  28%) more than their score on physics tasks (t1: 54%  $\pm$  30%, t2: 52%  $\pm$  25%).

**TABLE 2 |** Results of the two-way ANOVAs conducted on the students' scores with the time (t1 vs. t2) and the context (physics vs. finance) as within-subject factors and with the domain (physics students vs. economics) as a between-subject factor.

	Time			Domain			Context		
	<i>F</i>	<i>p</i>	$\eta^2_p$	<i>F</i>	<i>p</i>	$\eta^2_p$	<i>F</i>	<i>p</i>	$\eta^2_p$
"Slope" qualitative	5.075	0.030	0.115	1.839	0.183	0.045	0.384	0.539	0.010
"Slope" quantitative	2.518	0.121	0.061	3.592	0.065	0.084	4.393	0.043	0.101
"Area" qualitative	0.955	0.334	0.024	15.678	0.000	0.287	14.158	0.001	0.266
"Area" quantitative	0.133	0.717	0.003	9.030	0.005	0.188	0.556	0.460	0.014
	Time × domain			Time × context			Time × domain × context		
	<i>F</i>	<i>p</i>	$\eta^2_p$	<i>F</i>	<i>p</i>	$\eta^2_p$	<i>F</i>	<i>p</i>	$\eta^2_p$
"Slope" qualitative	0.896	0.350	0.022	0.274	0.604	0.007	0.000	0.990	0.000
"Slope" quantitative	2.518	0.121	0.061	1.448	0.236	0.036	1.448	0.236	0.036
"Area" qualitative	1.938	0.172	0.047	0.049	0.826	0.001	0.503	0.482	0.013
"Area" quantitative	0.495	0.486	0.013	0.001	0.971	0.000	2.266	0.140	0.055
	Domain × context								
	<i>F</i>	<i>p</i>	$\eta^2_p$						
"Slope" qualitative	0.000	0.988	0.000						
"Slope" quantitative	0.275	0.603	0.007						
"Area" qualitative	0.337	0.565	0.009						
"Area" quantitative	8.036	0.007	0.171						

The specification  $F(1,39)$  applies to all *F* values of the following analyses of variance.

## Changes in Students' Confidence Ratings Within and Across Domains (H2)

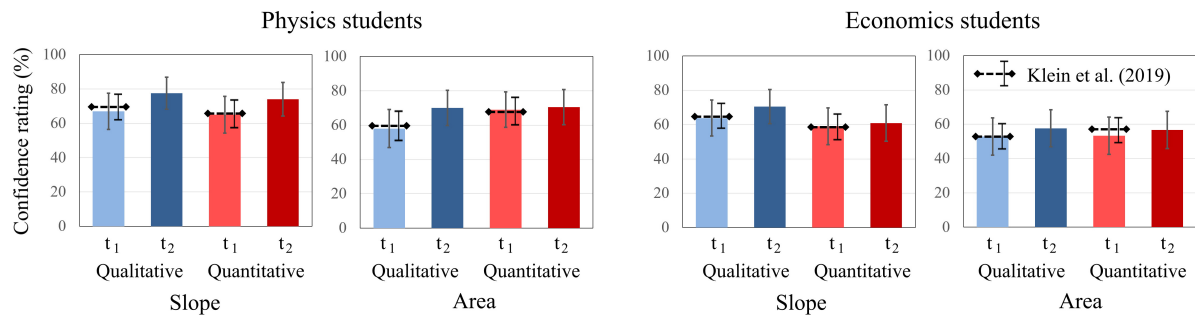
The mean confidence rating and standard deviation were [t1: (61% ± 25%), t2: (67% ± 20%)]. The physics students showed a confidence level of [t1: (65% ± 29%), t2: (73% ± 22%)] and the economics students of [t1: (57% ± 21%), t2: (61% ± 17%)], with no significant differences between the two domains in t1 and t2 ( $p > 0.05$ ). For the physics students, the total test score and the mean confidence level correlated highly in t1, but did not significantly correlate in t2 [t1:  $r(20) = 0.621$ ,  $p < 0.01$ ; t2:  $r(20) = 0.173$ ,  $p > 0.05$ ], whereas for the economics students, there was no significant correlation at either measurement [t1:  $r(21) = 0.198$ ,  $p > 0.05$ ; t2:  $r(21) = -0.297$ ,  $p > 0.05$ ].

To further explore students' confidence ratings, the same analysis procedure was applied as for the test scores. Two-way ANOVAs revealed no significant main effects for the factors time, concept, and type of question for physics students. However, for economics students, the factor concept was significant [ $F(1, 20) = 5.906$ ,  $p < 0.05$ ,  $\eta^2_p = 0.228$ ]. No significant interaction effects between time, concept, and type of question were revealed for either domain. For both question types about graph "slope," students' confidence ratings increased for both domains from t1 to t2, while the increase was more pronounced in physics students. The same applies to both question types about "area" graphs, even though the increase in confidence ratings was weaker for both domains compared to "slope" graphs. For both domains, confidence ratings were higher for "slope" graphs at t1 and t2 compared to "area" graphs (Figure 5).

To analyze the impact of context and time on students' confidence ratings, a repeated-measures ANOVA was run with context and time as the within-subject factors and domain as the between-subject factor for each pair of isomorphic tasks. The results are shown in Table 3.

For qualitative tasks of "slope" and "area" under a curve, significant main effects for time were found. The students' confidence increased from t1 to t2 for all qualitative tasks, but not for quantitative tasks. Furthermore, a significant time × domain × context effect was identified for qualitative tasks on "area" under a curve, showing that physics students' confidence increased over time for each context, whereas economics students' confidence increased over time for finance tasks and decreased over time for physics tasks. All other main and interaction effects were not significant (Table 3).

To investigate the accuracy of students' confidence, the ratings for correct and incorrect responses at each measurement (t1 and t2) were considered. Because of this split of the data across measurements and test scores, and the lack of paired variables (there is only one confidence rating for either a correct or an incorrect response), a repeated-measures analysis was not possible. Hence, all tasks on the "slope" concept and on the "area" concept were aggregated, respectively (Figure 6). For the "slope" concept, the physics students were significantly more confident when responding correctly than when responding incorrectly at t1 but not at t2 [t1:  $t(78) = 2.708$ ,  $p = 0.008$ ; t2:  $t(78) = 1.559$ ,  $p = 0.123$ ]. In contrast, the economics students' confidence was not significantly different between correct and incorrect responses [t1:  $t(82) = 0.362$ ,  $p = 0.718$ ; t2:  $t(82) = 1.369$ ,

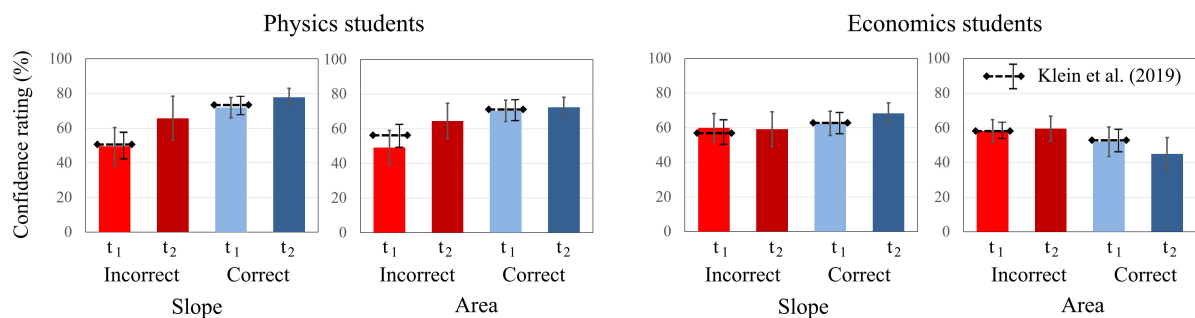


**FIGURE 5 |** Average confidence ratings of physics students and economics students on the qualitative and quantitative tasks about graph slope and area under a graph at t1 and t2. The error bars represent 1 standard error of the mean (SEM). The dashed lines represent the mean and error bars of the total sample of Klein et al. (2019).

**TABLE 3 |** Results of the two-way ANOVAs conducted on the students' confidence ratings with the time (t1 vs. t2) and the context (physics vs. finance) as within-subject factors and with the domain (physics students vs. economics) as a between-subject factor.

	Time			Domain			Context		
	<i>F</i>	<i>p</i>	$\eta_p^2$	<i>F</i>	<i>p</i>	$\eta_p^2$	<i>F</i>	<i>p</i>	$\eta_p^2$
“Slope” qualitative	4.997	0.031	0.114	0.395	0.534	0.010	0.421	0.520	0.011
“Slope” quantitative	1.191	0.282	0.030	2.652	0.111	0.064	0.211	0.649	0.005
“Area” qualitative	4.914	0.033	0.112	1.708	0.199	0.042	1.299	0.261	0.032
“Area” quantitative	0.565	0.457	0.014	3.399	0.073	0.080	0.707	0.406	0.018
	Time × domain			Time × context			Time × domain × context		
	<i>F</i>	<i>p</i>	$\eta_p^2$	<i>F</i>	<i>p</i>	$\eta_p^2$	<i>F</i>	<i>p</i>	$\eta_p^2$
“Slope” qualitative	0.249	0.620	0.006	0.019	0.890	0.000	1.062	0.309	0.027
“Slope” quantitative	0.504	0.482	0.013	1.217	0.277	0.030	0.754	0.390	0.019
“Area” qualitative	0.916	0.344	0.023	3.240	0.080	0.077	12.378	0.001	0.241
“Area” quantitative	0.081	0.777	0.002	1.083	0.305	0.027	0.136	0.715	0.003
	Domain × context								
	<i>F</i>	<i>p</i>	$\eta_p^2$						
“Slope” qualitative	0.714	0.403	0.018						
“Slope” quantitative	1.675	0.203	0.041						
“Area” qualitative	3.046	0.089	0.072						
“Area” quantitative	0.014	0.905	0.000						

The specification  $F(1,39)$  applies to all *F* values of the following analyses of variance.



**FIGURE 6 |** Average confidence ratings of physics and economics students related to correct and incorrect responses on the slope tasks and on the area tasks at t1 and t2. The error bars represent 1 standard error of the mean (SEM). The dashed lines represent the mean and error bars of the total sample of Klein et al. (2019).

$p = 0.175$ ]. For the “area” concept, the results were similar for physics students; whereas they responded correctly with higher confidence at t1, there was no significant difference at t2 [t1:  $t(78) = 2.915$ ,  $p = 0.005$ ; t2:  $t(78) = 1.174$ ,  $p = 0.275$ ]. For economics students, the results were different (**Figure 6**): At t1, they reported a higher confidence in their incorrect responses than their correct responses, although this difference was not significant. At t2, this difference increased, indicating that their confidence was significantly higher for incorrect responses than for correct responses [t2:  $t(82) = -2.810$ ,  $p = 0.006$ ]. Although there is an increase of overall confidence at t2, the self-assessment of students in both domains was less accurate at t2 than at t1.

## Changes in Students’ Dwell Times Within and Across Domains (H3)

### Total Dwell Time

The analysis of students’ eye movements is based on their total dwell time on the tasks before responding and then rating their confidence. The physics students had an average total dwell time of  $412 \pm 86$  s at t1 and  $333 \pm 75$  s at t2. The economics students needed  $461 \pm 172$  s at t1 and  $346 \pm 125$  s at t2 to respond to all tasks.

To compare students’ total dwell time on qualitative and quantitative tasks about graph “slope” and the “area” under a curve, an ANOVA was conducted separately for both domains including time, question, and concept as between factors. For physics students, significant main effects for time [ $F(1, 19) = 13.838$ ;  $p = 0.001$ ;  $\eta^2_p = 0.421$ ] and concept [ $F(1, 19) = 11.291$ ;  $p = 0.003$ ;  $\eta^2_p = 0.373$ ] were found. The factor question type was not significant, but there was a significant interaction effect for question  $\times$  concept [ $F(1, 19) = 11.244$ ;  $p = 0.003$ ;  $\eta^2_p = 0.372$ ]. Physics students spent less time on tasks at t2 and spent more time viewing the “area” tasks than the “slope” tasks (**Figure 7**). The significant interaction effect is similar to Klein et al. (2019), indicating that the question had the opposite effect. Physics students paid more attention to quantitative “slope” tasks and to qualitative “area” tasks.

The effects were similar for economics students. While there were significant main effects for time [ $F(1, 20) = 13.257$ ;  $p = 0.001$ ;  $\eta^2_p = 0.436$ ] and concept [ $F(1, 20) = 9.199$ ;  $p = 0.007$ ;

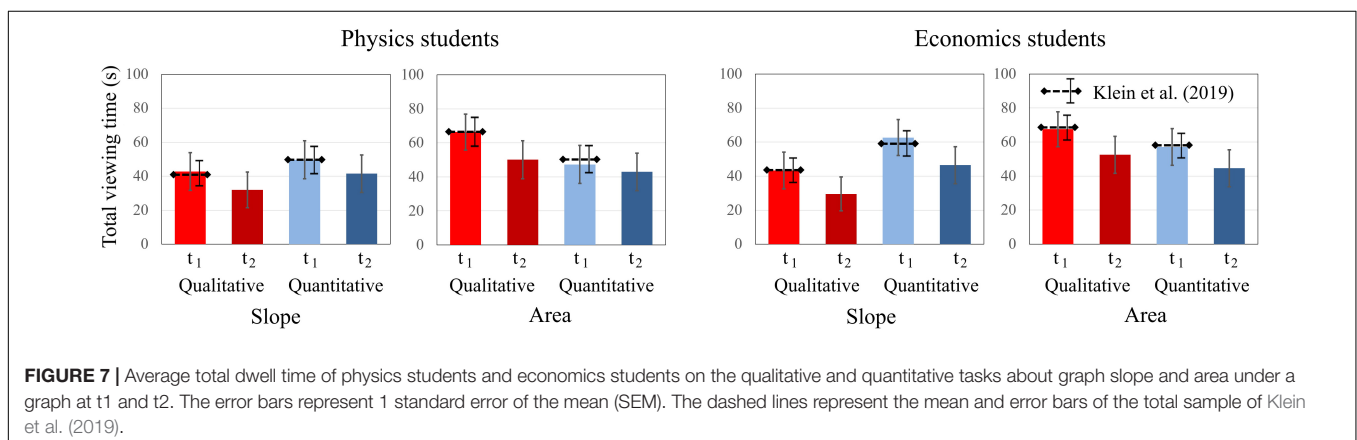
$\eta^2_p = 0.315$ ], the factor question type was not significant. There was a significant interaction effect for question  $\times$  concept [ $F(1, 20) = 13.257$ ;  $p = 0.002$ ;  $\eta^2_p = 0.399$ ]. Economics students also spent less time on tasks at t2 and spent more time viewing the “area” tasks than the “slope” tasks. The significant interaction effect also persists for the economics students. Overall, we found no significant differences between the students’ dwell times at t1 and t2.

To further explore students’ total dwell times, the same analysis was applied as for the test scores and the confidence ratings. The results of a two-way mixed-design ANOVA with the between-subject factor domain and the within-subject factor context for each pair of isomorphic tasks are shown in **Table 4**. The analysis revealed no main effect of context or domain. The interaction domain  $\times$  concept, however, was significant, indicating that physics students needed less time to respond to tasks from the physics context, whereas economics students needed less time to respond to tasks from the finance context. This finding is similar to Klein et al. (2019).

Regarding differences of total dwell time between t1 and t2, significant differences were found for almost each pair of isomorphic tasks. Only the total dwell time in quantitative “area” tasks was slightly below the level of significance ( $p > 0.05$ ). Overall, students needed less time at t2, but as the effect sizes indicate, there were fewer differences between the two measurements for quantitative tasks.

### Dwell Time on Different Areas of Interest (AOI)

In the sample of Klein et al. (2019), no differences were found between physics students and economics students in the defined AOIs (question, graph, and multiple choice) at t1. Compared to Klein et al. (2019), the findings presented here did not differ significantly. Students’ dwell times on the AOIs (question, graph, and multiple choice) were compared between the domains. Six Bonferroni-adjusted  $t$  tests showed no statistical difference between the dwell time of physics and economics students on the AOIs question [t1:  $t(39) = 0.388$ ,  $p = 0.700$ ; t2:  $t(39) = 1.530$ ,  $p = 0.134$ ], graph [t1:  $t(39) = -1.262$ ,  $p = 0.214$ ; t2:  $t(39) = -0.723$ ,  $p = 0.474$ ], and multiple choice [t1:  $t(39) = 0.012$ ,  $p = 0.990$ ; t2:  $t(39) = 0.321$ ,  $p = 0.750$ ]. There was a similar drop of total dwell time from t1 to t2 for students from both domains. In the





**TABLE 4 |** Results of the two-way ANOVAs conducted on the students' dwell times with the time (t1 vs. t2) and the context (physics vs. finance) as within-subject factors and with the domain (physics students vs. economics) as a between-subject factor.

	Time			Domain			Context		
	<i>F</i>	<i>p</i>	$\eta^2_p$	<i>F</i>	<i>p</i>	$\eta^2_p$	<i>F</i>	<i>p</i>	$\eta^2_p$
"Slope" qualitative	22.951	0.000	0.370	0.083	0.775	0.002	0.914	0.345	0.023
"Slope" quantitative	8.626	0.006	0.181	1.203	0.279	0.030	1.752	0.193	0.043
"Area" qualitative	17.616	0.000	0.311	0.128	0.723	0.003	0.163	0.688	0.004
"Area" quantitative	3.414	0.072	0.080	0.858	0.360	0.022	0.080	0.779	0.002
	Time × domain			Time × context			Time × domain × context		
	<i>F</i>	<i>p</i>	$\eta^2_p$	<i>F</i>	<i>p</i>	$\eta^2_p$	<i>F</i>	<i>p</i>	$\eta^2_p$
"Slope" qualitative	0.343	0.562	0.009	0.009	0.925	0.000	9.126	0.004	0.190
"Slope" quantitative	0.954	0.335	0.024	0.302	0.586	0.008	0.566	0.457	0.014
"Area" qualitative	0.027	0.871	0.001	4.301	0.045	0.099	0.316	0.577	0.008
"Area" quantitative	0.797	0.377	0.020	3.217	0.081	0.076	0.856	0.361	0.021
	Domain × context								
	<i>F</i>	<i>p</i>	$\eta^2_p$						
"Slope" qualitative	5.614	0.023	0.126						
"Slope" quantitative	2.102	0.155	0.051						
"Area" qualitative	0.958	0.334	0.024						
"Area" quantitative	0.556	0.460	0.014						

The specification  $F(1,39)$  applies to all *F*-values of the following analyses of variance.

comparison of the two measurements, there were also significant differences in the AOIs of question and graph between t1 and t2. However, no significant differences for the AOI multiple choice between t1 and t2 were found for either physics or economics students (Figure 8).

Next, the total dwell time on the AOI axis labels (adding the dwell times on the *x*-axis and *y*-axis labels) was determined for each item. A two-way mixed-design ANOVA with the between-subject factor domain and the within-subject factors time and context on total dwell time on the AOI axis labels was performed, indicating a significant main effect of time [ $F(1, 39) = 18.196$ ;  $p < 0.001$ ;  $\eta^2_p = 0.318$ ]. In contrast to Klein et al. (2019) and Susac et al. (2018), no interaction effects were found, even when considering the effects only at t2 ( $p > 0.05$ ). Because of the drop of total dwell times, dwell times on axis labels were also not significantly different between students from the two domains.

The dwell times on the axis tick labels were analyzed by a mixed-design ANOVA including time, question, and concept as within-factors for each domain. There was a significant main effect of question type [ $F(1, 19) = 39.752$ ;  $p < 0.001$ ;  $\eta^2_p = 0.677$ ] and a significant interaction effect of time × question × concept [ $F(1, 19) = 4.891$ ;  $p = 0.039$ ;  $\eta^2_p = 0.205$ ] for physics students. Other effects were not significant. Similar to Klein et al. (2019), physics students paid more attention to the axes when responding to quantitative than to qualitative tasks and especially paid more attention to the axes of quantitative "area" tasks at t2 in contrast to the quantitative "slope" tasks (Figure 9). For the economics students, the main effects of the factors concept [ $F(1, 20) = 11.491$ ;  $p = 0.003$ ;  $\eta^2_p = 0.365$ ] and question type

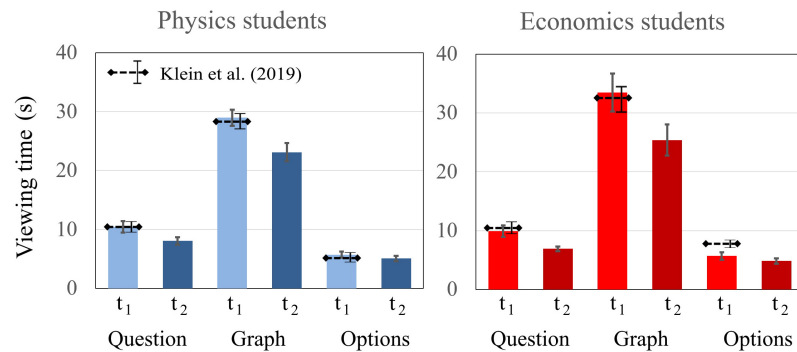
were significant [ $F(1, 20) = 19.976$ ;  $p < 0.001$ ;  $\eta^2_p = 0.500$ ], but the effect for the factor *time* was not. The interaction effects were not significant. Economics students also paid more attention to the axis tick labels of quantitative tasks and to the axis tick labels of tasks about the "area under the curve." All findings were similar to Klein et al. (2019).

## DISCUSSION

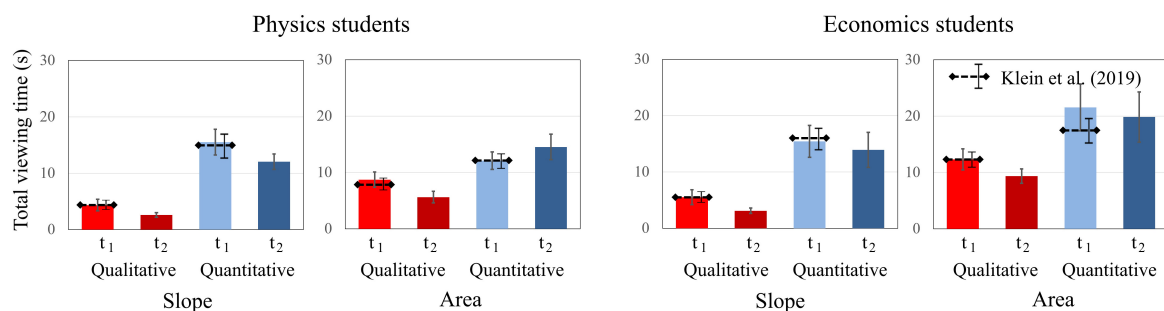
### Changes in Scores Across Contexts and Domains

H1: Physics and economics students solve graph tasks related to the subject they are enrolled in more successfully at the second measurement point than at the first measurement point.

However, the findings indicate differences in the development of graph understanding that are related to the task *context* and the task *concept*. Similar to the results of Klein et al. (2019), the physics students outperformed the economics students in terms of overall test performance and, in particular, achieved higher scores on tasks from the physics context at both t1 and t2. On average, at t2, the physics students also performed better on finance tasks than the economics students. In particular, they achieved higher scores on qualitative tasks on the concept of "slope" in the finance context at t2 than the economics students. On "area" tasks in finance, the scores of physics students remained at a similar level, and the scores of economics students



**FIGURE 8 |** Average fixation time of physics and economics students on the AOIs question, graph, and multiple choice at t1 and t2. The error bars represent 1 standard error of the mean (SEM). The dashed lines represent the mean and error bars of the total sample of Klein et al. (2019).



**FIGURE 9 |** Average total dwell time of physics students and economics students on the AOI axis tick labels for qualitative and quantitative questions about graph slope and area under a graph at t1 and t2. The error bars represent 1 standard error of the mean (SEM). The dashed lines represent the mean and error bars of the total sample of Klein et al. (2019).

increased from t1 to t2, whereas in the physics context, the scores of economics students decreased. These findings indicate that the physics students were more successful in transferring the graph task solution strategies that they had consolidated over the semester to other task contexts.

For the economics students, an increased “transfer effect” of this kind can be seen for the “slope” tasks, as the economics students achieved higher scores on the qualitative and quantitative tasks in the physics context at t2 than at t1. Overall, however, their scores at t2 (and t1) are lower than those of the physics students. The economics students’ scores on “area” tasks in the physics context were lower at t2 than at t1, whereas the physics students achieved higher scores as well as greater graph understanding gains. Whereas at t1 the economics students achieved higher scores in only one of four task pairs in the finance context (Klein et al., 2019), at t2 the opposite became evident for the subsample considered here. Even when taking into account the declining scores in the quantitative “area” tasks in the physics context, at t2 the economics students achieved higher scores in three of four task pairs in the finance context. This indicates that students experience context-specific learning effects that become evident when they expand or transfer their graph solving skills to another context. Similar findings have been reported in previous research, where the transfer of graph understanding over a period of time was different for students from different domains, and

the concepts the students were required to use to solve the tasks also differed (Jerry and Aaron, 2010; Bergey et al., 2015; Miller et al., 2016).

Similar to the results of both reference studies (Susac et al., 2018; Klein et al., 2019), at t2, “area” tasks were solved less successfully than “slope” tasks by students from both domains (physics and economics). The qualitative tasks related to the concept of “slope” were solved more successfully across both domains at t2 than at t1 (80%; Klein et al., 2019), with a correct solution rate of about 92%. The very high solution rate at t2 is in line with the results of our curriculum analyses, because tasks of this kind are an integral part of the curriculum in both domains. “Retest effects” are less likely to occur as the students were not given the solutions to the tasks and more than 3 months had passed between t1 and t2.

For the economics students, an increase in their scores on “slope” tasks from the physics context was also determined at t2, indicating a similar understanding of the representation of this concept in physics and finance graphs. Furthermore, there is a high increase in scores on the quantitative “slope” tasks in the finance context. Fundamental mathematical concepts are taught in economics degree programs right at the beginning of the curriculum, which enables students to understand and analyze subject-related phenomena using these methodological tools (Jensen, 2011; Benedict and Hoag, 2012). Teaching in the domain

of economics in particular places a strong focus on the concept of “slope” (e.g., in the analysis of extremes, cost, and profit trends), which is also generally easier for students to comprehend than the concept of “area under the curve.” Similar findings were reported for first-year students from standardized assessments in higher economics education, which also include tasks that refer to the “slope” concept (e.g., the Test of Understanding in College Economics, Walstad et al., 2007; or the German WiWiKom-Test, Zlatkin-Troitschanskaia et al., 2019), but not to the concept of “area under the curve.”

Differences in scores also occur with respect to the type of question. Both at t1 in the overall sample (Klein et al., 2019) and at t1 and t2 in the subsample examined in this study, qualitative tasks were always solved more successfully than quantitative tasks. Solving a mathematical task appears to require more cognitive resources – which may be measured, for instance, by assessing cognitive load (Gegenfurtner et al., 2011) – than solving a graph task with purely textual requirements (for similar findings, see, e.g., Curcio, 1987; Woolnough, 2000; Benedict and Hoag, 2012; Laging and Voßkamp, 2017; Ceuppens et al., 2019; Shavelson et al., 2019). This finding is also in line with research on mathematical requirements in graph tasks (Curcio, 1987; Planinic et al., 2013; Susac et al., 2018; Wemyss and van Kampen, 2013).

Apart from the “area” tasks, the number of students who had already scored high in the qualitative “slope” tasks at t1 further increased their scores at t2. In the qualitative “area” tasks, higher scores were also identified at t2 for both domains, whereas only physics students succeeded in increasing their scores in the quantitative “area” tasks in the physics context. The finding that the highest score increases or the greatest score decreases in the respective familiar contexts (i.e., “area/finance and “slope/finance for economics students; “area/physics for physics students) occur in the quantitative tasks illustrates that the difficulties the students had with these tasks at the beginning of the semester remained at t2. One possible explanation is that fundamental mathematics courses, which also include graph understanding, are taught in both courses (physics and economics) primarily in the first semesters (for physics, see Küchemann et al., 2019; for economics, see Jensen, 2011).

In summary, despite the discussed differences in terms of *domains*, *task contexts*, *task concepts*, and the *type of question*, H1 cannot be rejected, but more comprehensive research on the graph understanding of students in different domains and contexts is urgently needed.

## Change in Confidence Across Contexts and Domains

H2: Physics and economics students rate their confidence in their solution to tasks related to the subject they are enrolled in more accurately at the second measurement point than at the first measurement point.

A comparison of the two measurements shows that at t2, the confidence rating has only slightly, but not significantly, increased. When looking at the *context*, we did not find any statistically significant effects. Moreover, similar to Klein et al.

(2019), there is no significant difference between the students from both *domains*. With regard to the *task concept*, apart from the “area” task solutions of the economics students, the confidence rating of correct solutions increased at t2. With the exception of economics students’ solutions to the “slope” tasks, incorrect solutions were self-assessed as being correct with more confidence at t2. Already at t1 (Klein et al., 2019), approximately 50% of incorrect solutions were self-assessed as correct, indicating the students’ deficient metacognitive skills. This effect increased to greater than 60% at t2. This finding is also in line with numerous studies across disciplines (Nowell and Alston, 2007; Bell and Volckman, 2011; Guest and Riegler, 2017; Brückner and Zlatkin-Troitschanskaia, 2018).

The increasing confidence in one’s own erroneous solving strategies for graph tasks can be traced back to causes described under the umbrella term “error knowledge,” which includes, for instance, overestimating one’s (task-related) knowledge and skills and deficits in the ability to diagnose errors in the solution process (Kruger and Dunning, 1999). In particular, the latter one can also be caused by didactic priorities. Generally, students are taught to identify possible strategies that will lead them to correct solutions. However, they are less systematically taught to recognize systematic errors in their solution process.

The negative change in the self-assessment of economics students on “area” tasks is particularly remarkable. Compared to all other tasks, both the average correct solution rate and the average correct self-assessment for these tasks decrease significantly. Apparently, there is no recognition effect but a learning effect, so that even wrong task solutions were self-assessed as correct remarkably often. For qualitative tasks on “area under a curve,” physics students’ confidence increased over time for each task context, whereas economics students’ confidence increased for economics tasks and decreased for physics tasks.

In summary, students rated their correctness of responses less accurately at t2. These unexpected findings (e.g., economics students solve “area” tasks less successfully and also rate their solution less accurately at t2) indicate that the students may have developed fundamental misconceptions, which require more in-depth research in further studies. Thus, H2 cannot be confirmed, although there is an increase in confidence rating from t1 to t2 that reflects earlier findings (e.g., Guest and Riegler, 2017).

## Change in Students’ Dwell Times

H3: The dwell time on the tasks and the individual graph components (AOIs) is lower at the second measurement point for students from both domains and in both contexts.

With regard to H3, not only did the total dwell time during task processing decrease significantly at t2, but the students also spent less time reading the tasks. This may be due to a recognition effect or a learning effect in graph understanding, as the scores increased at t2, but familiarity with the tasks increased only slightly. Despite a decrease in total dwell time at t2, students from both domains still spent more time on questions about the “area under the curve” than on questions about graph “slope.” In view of the decreasing or unchanged scores (with the exception



of the “area” quantitative task pair in the physics context), this further indicates the higher cognitive load that “area” tasks elicit in students (Gegenfurtner et al., 2011; Klein et al., 2019).

Furthermore, findings at t2 confirmed the findings at t1 (Klein et al., 2019) that the students’ dwell time is also longer for quantitative tasks than for qualitative tasks if the task is an “area” task. Students from both domains need longer for quantitative “area” tasks than for quantitative “slope” tasks. Thus, in line with Susac et al. (2018) and Klein et al. (2019), students need longer for complex mathematical calculations, such as “area” calculations, than for linear “slope” calculations. Longer dwell times on quantitative tasks can also be attributed to a large extent to the comparably longer dwell time on the axes. While the dwell time on the axes generally decreased at t2, it actually increased for quantitative “area” task processing by physics students. This supports the conclusion that quantitative tasks are more difficult, because comparatively more information has to be extracted from the axes and mentally processed.

Similar to Klein et al. (2019), at t2, students spent the longest time on qualitative “area” tasks. This persistent finding, which is also consistent with the decrease in processing times for all tasks, indicates that estimating the “area” is still more cognitively demanding than determining a “slope,” despite corresponding increases in knowledge. However, the effect of the novelty of such a task was not found. For qualitative tasks, students at t2 from both domains still spent more time on the axis tick labels for the “area” task compared to the “slope” task. It can be assumed that they were looking for further information to estimate the “area” size on the axes. In contrast to Klein et al. (2019), however, no longer dwell times on unfamiliar task contexts were determined at t2 compared to t1. In line with the discussed findings on transferring graph understanding between contexts, this may be due to the fact that by learning how to solve graph problems, students no longer try to decipher the meaning of the axis designations and instead have developed schemes (i.e., heuristics) that enable them to transfer the graph solution strategies from one context to another. Regarding the solving strategies, economics students still needed more time to explore the axis tick labels of qualitative “area” tasks, although they are irrelevant for the solution process. This supports the assumption that economics students use compensatory strategies to respond to these tasks. This is in line with the argument of Beichner (1994) that area estimation stimulates students’ inappropriate use of axis values.

As in Klein et al. (2019) at t1, no overall, differences in dwell times between physics and economics students were found at t2 with regard to the relationship between total dwell time and students’ performance. Intratemporal and intertemporal comparisons between both domains (physics and economics) were conducted over the three areas defined as AOIs (question, graph, multiple choice), indicating no significant differences, as students from both domains spent almost an identical amount of time on the three AOIs at t1 and t2. The significant differences in the students’ scores cannot be explained by a domain-specific change in the time spent on the graph tasks, even for t2. Thus, total dwell time alone does not explain the difference in the performance outcomes between the students from both domains.

However, a comparison of t1 and t2 shows that the time spent on the question and graph decreases significantly, but the time spent on the response options remains almost the same. This indicates that, although the students apparently extract information from the tasks more quickly, they do not have any recognition effects with regard to the task solutions, as they must look at the responses systematically again.

In summary, the effects reported here for the (sub)sample at t2 are not significantly different from those of the entire sample in Klein et al. (2019) at t1. Regarding the drop in total dwell time at t2 and the lower dwell times reported by Susac et al. (2018), there was a correlation between study progress and dwell time and the task *concept* or *question type* across both *domains*. In our study, we thus replicated the findings of both Susac et al. (2018) and Klein et al. (2019) and determined the stability of the time effect, as we found similar effects for t1 and t2. The differences between Susac et al. (2018) and Klein et al. (2019) also persist at t2; there were no significant main effects of *context* at t2. This indicates the stability of the findings over time. Thus, H3 can be confirmed.

## CONCLUSION

### Summary and Future Perspectives

In a postreplication study, based on the two existing studies of Susac et al. (2018) and Klein et al. (2019), using a pretest–posttest measurement, we expanded the analytical research focus to gain initial insights about changes in students’ graph comprehension within and across domains with regard to the theoretically expected (i) time effects (measurements t1 and t2); (ii) domain effects (physics and economics); (iii) question type, concept, and context effects; and possible (vi) interaction effects.

Effects of these kinds could be found at both measurement points. For instance, physics students achieved higher scores than economics students, whereas economics students, at t2 in particular, achieved better results in tasks with a finance *context* than in physics tasks. On average, students from both domains were more likely to correctly solve tasks on the concept “slope” at both measurement points, whereas physics students correctly solved “area” tasks at 67% and “slope” tasks at 75%, and economics students correctly solved “area” tasks at 42% and “slope” tasks at 69%. Furthermore, “slope” tasks were visually processed more quickly than “area” tasks.

Overall, the accuracy of the students’ self-assessment decreased at t2, showing that overestimating incorrect solutions occurs more often than underestimating correct solutions. Further studies are needed with a particular focus on explanations for overestimating incorrect solutions and uncovering possible misconceptions, to form a basis for modified instructional research designs. For example, typical misconceptions could be discussed in classes or short interventions; for instance, digital classroom response systems can be used in larger lectures to gather data on students’ knowledge about a task concept or a solution process. This is especially important because students are increasingly using digital media to construct graphs. However, traditional media are still used in most forms of higher education instruction. Because the present study focuses only on

the understanding and interpretation of graphs and not on haptic construction performance, this problem takes a backseat in the scope of this article (*Limitations and Implications*). However, empirical evidence to examine the difference between paper-based and digital understanding of graphs is also still missing. Meta-analyses would also be desirable for a consolidation of the current findings.

Similar to both referenced studies and at t2 in our study, the total dwell times and the dwell times on the defined AOIs (question, graph, and multiple choice) can hardly predict differences in the scores between the students from both domains. Instead, better predictions can be made by analyzing individual parts of the graph (e.g., axis tick labels). However, in view of the generally faster processing time and on average higher number of correct solutions, more efficient solution strategies and information processing can be assumed for students from both domains at t2. This indicated increase in the efficiency of information processing also shows that, for example, students are less irritated by the axes' labels of unfamiliar domains.

In conclusion, the findings of the postreplication study are mostly consistent with those of the two previous studies. Subsequent studies should now be applied more specifically to the cognitive processes both within and across domains, for instance by expanding the samples and domains to investigate whether existing developments can also be found in other domains.

A more systematic exploration of graph task-relevant aspects could be conducted through expert ratings and additional ET studies with experts to investigate their solving strategies of graph tasks. Combinations of ET with other techniques, for instance, electroencephalography or skin conductance, and in particular verbal data (Leighton et al., 2017) could provide important information, for example, on the extent to which the dwell time on a certain area or the scattering of fixations is more important for solving a graph task or for maintaining existing or transferring solving strategies and emotions during learning in a domain. First insights have been provided by Susac et al. (2018), who retrospectively recorded students' task solving strategies. Computing methods can also provide further evidence as to the extent to which eye movements are linked to information processing (Elling et al., 2012).

Further research should also focus on the consistent decrease in economics students' scores for "area" tasks from foreign contexts, as well as on the high increase in scores for quantitative "slope" tasks at t2, while also controlling for effects of explicit instructional measures in classes, as well as possible learning opportunities outside of university. Considering that the eight isomorphic tasks are tasks that are typically used in textbooks in print or online formats, in lectures, exercises, or in (online) assessments, a more comprehensive analysis of the learning opportunities students have during the semester is required. Studies have shown, for instance, that students increasingly use digital media for their examination preparation (e.g., Wikipedia, see Maurer et al., 2020). In particular, dynamic representations of graphs or the use of graph creation software could promote graph understanding (Stern et al., 2003; Gustafsson and Ryve, 2016; Opfermann et al., 2017). For example, the dynamic hatching of an "area under a curve" with a parallel indication of the calculated

values and a formula display could facilitate understanding in the sense of "learning by examples" (Schalk et al., 2020). However, the extent to which this can have a positive effect on the understanding of certain concepts like "slope" and "area" still needs to be investigated. Further research should also investigate other indicators, such as click rates (Buil et al., 2016; Hunsu et al., 2016) or the use of multiple learning media (Ainsworth, 2006; Mayer, 2009) in the context of experimental studies to provide more precise analyses of information processing and the development of graph comprehension. How students' graph understanding and the identified differences and effects develop over the course of the degree course until their graduation should also be further investigated.

## LIMITATIONS AND IMPLICATIONS

Despite findings that are stable over time and also in line with previous research, these results should be critically discussed in view of the limitations of this study. These limitations concern (i) the construct and the study design, (ii) the sampling, and (iii) the scope of analyses carried out so far.

(i) The study used graph tasks for two types of concepts with linear progressions, "area" and "slope," thus capturing students' understanding only of certain types of graphs (Curcio, 1987). Moreover, the study focus is on the students' internal mental processes rather than on the active construction or drawing of graphs or on communication with third parties. Thus, the focus is on the recognition of trends and areas, i.e., coherent parts of a graph. Graph understanding in terms of the reproduction of individual values, interpolations between graph parts, the extrapolation and prediction of graph progressions, or interpretation in larger contexts (e.g., how an increase in inflation in 5 years will affect the economy) was not captured in this study (Curcio, 1987). Investigating such phenomena requires further task constructions and other study designs.

Furthermore, the test instrument used in this study was limited to eight tasks, which were taken as replications from previous studies; for the same reason, a treatment-control group design (e.g., in which students work on certain tasks or attend selected lectures and courses) was intentionally not used. However, because the findings and the expectations covered by the three studies have for the most part have been confirmed several times, follow-up studies, for instance, in the context of multimedia learning environments, can now be immediately conducted, at least in the investigated domains. Moreover, the present findings are primarily related to digital representations on computer screens. The extent to which extrapolation of other representational formats is possible must also be investigated in further studies.

(ii) Overall, less than two-thirds of the total sample from Klein et al. (2019) could be retested in the study at t2. Nonetheless, compared to other existing ET studies, more than 40 study participants at two measurement points constitute a considerable sample. For future studies, however, larger sample is required to generate a higher generalizability (e.g., investigating correlations of eye movements and scores in different populations), as well

as an expansion at the institutional level, to include more universities, faculties, and students, for instance, to analyze teaching effects. The multilevel structure in which response behavior can vary between and within domains, previous knowledge, and other sociobiographical characteristics should be considered in an expanded sample (Hox et al., 2018). Many non-significant results in the present study can also be traced back to the sample size, so that the differences and correlations were mostly investigated with regard to effect size (see iii).

The calculation of several ANOVAs for the examined factors of graph understanding (task characteristics (context, concept, and type of questions), and domains (physics vs. economics) builds on the studies of Susac et al. (2018) and Klein et al. (2019) and extends the existing analyses by the time factor (t1 vs. t2). Because no comparable findings on the development of graph understanding are available yet, and the comparability between the studies should be ensured, an equally possible comprehensive repeated-measures ANOVA with all within-factors and the domain as between-factor specified in the study was calculated but not presented in this article. The findings of these analyses, taking into account possible inflations of standard error, lead to the same interpretations due to the significance and effect sizes.

In view of low-stakes assessments, deviations in test-taking motivation for larger samples should also be considered in appropriate empirical modeling (Penk et al., 2014). In the present study, this was only possible by means of response time effect modeling (Wise and Kong, 2005). To mitigate the potential supporting effects of test motivation, for instance, in terms of very short and very long dwell times, the study participants were offered monetary compensation, and additional individual surveys were conducted, so that negative hidden mass-group effects on test motivation were as similar as possible across all test takers.

(iii) The present study only regarded dwell times. Qualitative, retrospective interviews (Susac et al., 2018) have shown, however, that students' task-solving strategies differ, and research could be complemented by analyses of saccadic (Klein et al., 2019) or transitional studies of fixation sequences. For example, it is conceivable that students will not only solve "area" tasks better if they look at the graph or the axis tick labels for a longer time, but also if they perform more saccadic eye movements between axis tick labels and graph. Such phenomena could be analyzed more precisely, for instance, by using process and path models.

The "area under the graph" is identified as a crucial concept in graph understanding and its development and should also be researched more intensively, especially among economics students, and treated in a differentiated manner with regard to instructional research designs. Apparently, the two types of

question and the two task concept types are based on different cognitive processes, which are also addressed differently over the semester and thus lead to the changes identified in this study. Potential explanatory factors such as domain-specific prior knowledge might have an effect on these processes (e.g., students with more prior knowledge may use more efficient test-taking strategies) and should also be included in further studies. This can be done, for example, in multilevel linear mixed-effects models (Brückner and Pellegrino, 2016; Strobel et al., 2018), which take into account the structure between subject characteristics, item characteristics and response processes, and the final test scores.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

SB wrote the manuscript and conducted the analyses. OZ-T co-wrote the manuscript and coordinated the analyses. SK, PK, and JK reviewed, corrected, and discussed the article and the analyses, and also introduced the physics perspective into the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

The research in this article was part of the PLATO project funded by the Ministry for Science, Continued Education and Cultural Affairs Rhineland-Palatinate, Germany.

## ACKNOWLEDGMENTS

We would like to thank the two reviewers and the editor who provided constructive feedback and helpful guidance in the revision of this paper.

## REFERENCES

- Ainsworth, S. (2006). DeFT: a conceptual framework for considering learning with multiple representations. *Learn. Instr.* 16, 183–198. doi: 10.1016/j.learninstruc.2006.03.001
- Bartholomé, T., and Bromme, R. (2009). Coherence formation when learning from text and pictures: what kind of support for whom? *J. Educ. Psychol.* 101, 282–293. doi: 10.1037/a0014312
- Becker, S., Gößling, A., Klein, P., and Kuhn, J. (2020a). Investigating dynamic visualizations of multiple representations using mobile video analysis in physics lessons: effects on emotion, cognitive load and conceptual understanding. *Zeitschrift für Didaktik der Naturwissenschaften* 26 (in press).
- Becker, S., Gößling, A., Klein, P., and Kuhn, J. (2020b). Using mobile devices to enhance inquiry-based learning processes. *Learn. Instr.* 69:101350. doi: 10.1016/j.learninstruc.2020.101350

- Beichner, R. J. (1994). Testing student interpretation of kinematics graphs. *Am. J. Phys.* 62, 750–762. doi: 10.1119/1.17449
- Bell, A., and Janvier, C. (1981). The interpretation of graphs representing situations. *Learn. Math.* 2, 34–42.
- Bell, P., and Volckman, D. (2011). Knowledge surveys in general chemistry: confidence, overconfidence, and performance. *J. Chem. Educ.* 88, 1469–1476. doi: 10.1021/ed100328c
- Benedict, M. E., and Hoag, J. (2012). “Factors influencing performance in economics: graphs and quantitative usage,” in *International Handbook on Teaching and Learning in Economics*, eds G. M. Hoyt and K. McGoldrick (Cheltenham: Edward Elgar), 334–340.
- Bergey, B. W., Cromley, J. G., and Newcombe, N. S. (2015). Teaching high school biology students to coordinate text and diagrams: relations with transfer, effort, and spatial skill. *Int. J. Sci. Educ.* 37, 2476–2502. doi: 10.1080/09500693.2015.1082672
- Bollen, L., De Cock, M., Zuza, K., Guisasaola, J., and van Kampen, P. (2016). Generalizing a categorization of students’ interpretations of linear kinematics graphs. *Phys. Rev. Physics Educ. Res.* 12:010108. doi: 10.1103/PhysRevPhysEducRes.12.010108
- Bowen, G. M., and Roth, W.-M. (1998). Lecturing graphing: what features of lectures contribute to student difficulties in learning to interpret graph? *Res. Sci. Educ.* 28, 77–90. doi: 10.1007/bf02461643
- Brand-Gruwel, S., Wopereis, I., and Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Comput. Educ.* 53, 1207–1217. doi: 10.1016/j.compedu.2009.06.004
- Brückner, S., and Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multilevel models to validate an assessment of higher education students’ competency in business and economics. *J. Educ. Measur.* 53, 293–312. doi: 10.1111/jedm.12113
- Brückner, S., and Zlatkin-Troitschanskaia, O. (2018). *Threshold Concepts for Modeling and Assessing Higher Education Students’ Understanding and Learning in Economics*, eds O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, and C. Kuhn (Cham: Springer), 103–121.
- Buil, I., Catalán, S., and Martínez, E. (2016). Do clickers enhance learning? A control-value theory approach. *Comput. Educ.* 103, 170–182. doi: 10.1016/j.compedu.2016.10.009
- Canham, M., and Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learn. Instr.* 20, 155–166. doi: 10.1016/j.learninstruc.2009.02.014
- Ceuppens, S., Bollen, L., Deprez, J., Dehaene, W., and De Cock, M. (2019). 9th grade students’ understanding and strategies when solving  $x(t)$  problems in 1D kinematics and  $y(x)$  problems in mathematics. *Phys. Rev. Phys. Educ. Res.* 15:010101. doi: 10.1103/PhysRevPhysEducRes.15.010101
- Chen, I.-H., Gamble, J. H., Lee, Z.-H., and Fu, Q.-L. (2020). Formative assessment with interactive whiteboards: a one-year longitudinal study of primary students’ mathematical performance. *Comput. Educ.* 150:103833. doi: 10.1016/j.compedu.2020.103833
- Christensen, W. M., and Thompson, J. R. (2012). Investigating graphical representations of slope and derivative without a physics context. *Phys. Rev. ST Phys. Educ. Res.* 8:023101. doi: 10.1103/PhysRevSTPER.8.023101
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. Hillsdale, NJ: Erlbaum Associates.
- Cohn, E., Cohn, S., Balch, D. C., and Bradley, J. (2001). Do graphs promote learning in principles of economics? *J. Econ. Educ.* 32, 299–310. doi: 10.2307/1182879
- Cordova, J. R., Sinatra, G. M., Jones, S. H., Taasoobshirazi, G., and Lombardi, D. (2014). Confidence in prior knowledge, self-efficacy, interest and prior knowledge: influences on conceptual change. *Contemp. Educ. Psychol.* 39, 164–174. doi: 10.1016/j.cedpsych.2014.03.006
- Cowie, B., and Cooper, B. (2017). Exploring the challenge of developing student teacher data literacy. *Assess. Educ.* 24, 147–163. doi: 10.1080/0969594X.2016.1225668
- Cromley, J. G., Bergey, B. W., Fitzhugh, S., Newcombe, N., Wills, T. W., Shipley, T. F., et al. (2013). Effects of three diagram instruction methods on transfer of diagram comprehension skills: the critical role of inference while learning. *Learn. Instr.* 26, 45–58. doi: 10.1016/j.learninstruc.2013.01.003
- Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. *J. Res. Math. Educ.* 18:382. doi: 10.2307/749086
- Elling, S., Lentz, L., and De Jong, M. (2012). Combining concurrent think-aloud protocols and eye-tracking observations: an analysis of verbalizations and silences. *IEEE Trans. Profess. Commun.* 55, 206–220. doi: 10.1109/TPC.2012.2206190
- Freedman, E. G., and Shah, P. (2002). “Toward a model of knowledge-based graph comprehension,” in *Diagrammatic Representation and Inference*, Vol. 2317, eds M. Hegarty, B. Meyer, and N. H. Narayanan (Berlin: Springer-Verlag), 18–30. doi: 10.1007/3-540-46037-3\_3
- Gegenfurtner, A., Lehtinen, E., and Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educ. Psychol. Rev.* 23, 523–552. doi: 10.1007/s10648-011-9174-7
- Guest, J., and Riegler, R. (2017). Learning by doing: do economics students self-evaluation skills improve? *Int. Rev. Econ. Educ.* 24, 50–64. doi: 10.1016/j.iree.2016.10.002
- Gustafsson, P., and Ryve, A. (2016). Developing Design Principles and Task Types for Classroom Response System Tasks in Mathematics: Engineering Mathematical Classroom Discussions. Available online at: <http://urn.kb.se/resolve?urn=urn:nbn:se:mdh:diva-33359> (accessed May 11, 2019).
- Happ, R., Förster, M., Zlatkin-Troitschanskaia, O., and Carstensen, V. (2016). Assessing the previous economic knowledge of beginning students in Germany: implications for teaching economics in basic courses. *Citizenship Soc. Econ. Educ.* 15, 45–57. doi: 10.1177/2047173416646597
- Heublein, U. (2014). Student drop-out from German higher education institutions. *Eur. J. Educ.* 49, 497–513. doi: 10.1111/ejed.12097
- Hochberg, K., Becker, S., Louis, M., Klein, P., and Kuhn, J. (2020). Using smartphones as experimental tools – a follow-up: cognitive effects by video analysis and reduction of cognitive load by multiple representations. *J. Sci. Educ. Technol.* 29, 303–317. doi: 10.1007/s10956-020-09816-w
- Hox, J. J., Moerbeek, M., and van de Schoot, R. (2018). *Multilevel Analysis. Techniques and Applications*, 3rd Edn. New York, NY: Routledge.
- Hunsu, N. J., Adesope, O., and Bayly, D. J. (2016). A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Comp. Educ.* 94, 102–119. doi: 10.1016/j.compedu.2015.11.013
- Ivanjek, L., Planinic, M., Hopf, M., and Susac, A. (2017). “Student difficulties with graphs in different contexts,” in *Cognitive and Affective Aspects in Science Education Research*, eds K. Hahl, K. Juuti, J. Lampiselkä, A. Uitto, and J. Lavonen (Cham: Springer), 167–178. doi: 10.1007/978-3-319-58685-4\_13
- Ivanjek, L., Susac, A., Planinic, M., Andrasevic, A., and Milin-Sipus, Z. (2016). Student reasoning about graphs in different contexts. *Phys. Rev. Phys. Educ. Res.* 12:010106. doi: 10.1103/PhysRevPhysEducRes.12.010106
- Jensen, U. (2011). *Wozu Mathe in den Wirtschaftswissenschaften? Eine Einführung für Studienanfänger [Why Math in Economics? An Introduction for First-Year Students]*. Wiesbaden: Vieweg+Teubner.
- Jerry, T. F. L., and Aaron, C. C. E. (2010). “The impact of augmented reality software with inquiry-based learning on students’ learning of kinematics graph,” in *Proceedings of the 2010 2nd International Conference on Education Technology and Computer*, Vol. 2, Shanghai, V2–V1.
- Just, M. A., and Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychol. Rev.* 87, 329–354. doi: 10.1037/0033-295x.87.4.329
- Klein, P., Küchemann, S., Brückner, S., Zlatkin-Troitschanskaia, O., and Kuhn, J. (2019). Student understanding of graph slope and area under a curve: a replication study comparing first-year physics and economics students. *Phys. Rev. Phys. Educ. Res.* 15, 1–17. doi: 10.1103/PhysRevPhysEducRes.15.020116
- Klein, P., Kuhn, J., Müller, A., and Gröber, S. (2015). “Video analysis exercises in regular introductory mechanics physics courses: effects of conventional methods and possibilities of mobile devices,” in *Multidisciplinary Research on Teaching and Learning*, eds W. Schnotz, A. Kauertz, H. Ludwig, A. Müller, and J. Pretsch (Basingstoke: Palgrave Macmillan), 270–288. doi: 10.1057/9781137467744\_15
- Klein, P., Lichtenberger, A., Küchemann, S., Becker, S., Kekule, M., Viiri, J., et al. (2020). Visual attention while solving the test of understanding graphs in kinematics: an eye-tracking analysis. *Eur. J. Phys.* 41:025701. doi: 10.1088/1361-6404/ab5f51



- Klein, P., Müller, A., and Kuhn, J. (2017). Assessment of representational competence in kinematics. *Phys. Rev. Phys. Educ. Res.* 13:010132. doi: 10.1103/PhysRevPhysEducRes.13.010132
- Kosslyn, S. M. (1999). *Image and Brain: The Resolution of the Imagery Debate*, 4th Edn. Cambridge, MA: MIT Press.
- Kragten, M., Admiraal, W., and Rijlaarsdam, G. (2015). Students' learning activities while studying biological process diagrams. *Int. J. Sci. Educ.* 37, 1915–1937. doi: 10.1080/09500693.2015.1057775
- Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* 77, 1121–1134. doi: 10.1037/0022-3514.77.6.1121
- Küchemann, S., Klein, P., Becker, S., Kumari, N., and Kuhn, J. (2020a). "Classification of students' conceptual understanding in STEM education using their visual attention distributions: a comparison of three machine-learning approaches," in *Proceedings of the CSEDU 2020*, Prague, 36–46.
- Küchemann, S., Klein, P., Fouckhardt, H., Gröber, S., and Kuhn, J. (2020b). Students' understanding of non-inertial frames of reference. *Phys. Rev. Phys. Educ. Res.* 16:010112.
- Küchemann, S., Klein, P., and Kuhn, J. (2019). "Best of Germany: VorleXung: cross-linking recitation sessions and physics lectures using eXperiment-based video-analysis tasks," in *Association for the Advancement of Computing in Education*, ed. EdMedia+ Innovate Learning (Waynesville, NC: AACE), 152–157.
- Laging, A., and Voßkamp, R. (2017). Determinants of maths performance of first-year business administration and economics students. *Int. J. Res. Und. Math. Educ.* 3, 108–142. doi: 10.1007/s40753-016-0048-8
- Leighton, J. P., Tang, W., and Guo, Q. (2017). "Response processes and validity evidence: controlling for emotions in think aloud interviews," in *Understanding and Investigating Response Processes in Validation Research*, eds B. D. Zumbo and A. M. Hubley (Cham: Springer), 137–158.
- Lichti, M., and Roth, J. (2019). Functional thinking: a three-dimensional Construct? *J. Did. Math.* 40, 169–195. doi: 10.1007/s13138-019-00141-3
- Madsen, A. M., Larson, A. M., Loschky, L. C., and Rebello, N. S. (2012). Differences in visual attention between those who correctly and incorrectly answer physics problems. *Phys. Rev. ST Phys. Educ. Res.* 8:10122. doi: 10.1103/PhysRevSTPER.8.10122
- Maurer, M., Schemer, C., Zlatkin-Troitschanskaia, O., and Jitmirski, J. (2020). "Positive and negative media effects on university students' learning: preliminary findings and a research program," in *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*, ed. O. Zlatkin-Troitschanskaia (New York, NY: Springer), 109–119. doi: 10.1007/978-3-030-26578-6\_8
- Mayer, R. E. (2009). *Multimedia Learning*, 2nd Edn. New York, NY: Cambridge University Press.
- McDermott, L. C., Rosenquist, M. L., and van Zee, E. H. (1987). Student difficulties in connecting graphs and physics: examples from kinematics. *Am. J. Phys.* 55, 503–513. doi: 10.1119/1.15104
- Miller, B. W., Cromley, J. G., and Newcombe, N. S. (2016). Improving diagrammatic reasoning in middle school science using conventions of diagrams instruction. *J. Comput.* 32, 374–390. doi: 10.1111/jcal.12143
- Moghavvemi, S., Sulaiman, A., Jaafar, N. I., and Kasem, N. (2018). Social media as a complementary learning tool for teaching and learning: the case of youtube. *Int. J. Manag. Educ.* 16, 37–42. doi: 10.1016/j.ijme.2017.12.001
- Nowell, C., and Alston, R. M. (2007). I thought I got an A! Overconfidence across the economics curriculum. *J. Econ. Educ.* 38, 131–142. doi: 10.3200/JECE.38.2.131-142
- Opfermann, M., Schmeck, A., and Fischer, H. (2017). "Multiple representations in physics and science education—why should we use them?," in *Multiple Representations in Physics Education*, eds D. Treagust, R. Duit, and H. Fischer (Dordrecht: Springer), 1–22. doi: 10.1007/978-3-319-58914-5\_1
- Penk, C., Pöhlmann, C., and Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: an investigation of school-track-specific differences. *Large Scale Assess. Educ.* 2, 1–17. doi: 10.1186/s40536-014-0005-4
- Pinker, S. (1990). "A theory of graph comprehension," in *Artificial Intelligence and the Future of Testing*, ed. R. O. Freedle (London: Routledge), 73–126.
- Planinic, M., Ivanjek, L., Susac, A., and Milin-Sipus, Z. (2013). Comparison of university students' understanding of graphs in different contexts. *Phys. Rev. ST Phys. Educ. Res.* 9:020103. doi: 10.1103/PhysRevSTPER.9.020103
- Planinic, M., Milin-Sipus, Z., Katic, H., Susac, A., and Ivanjek, L. (2012). Comparison of students understanding of line graph slope in physics and mathematics. *Int. J. Sci. Math. Educ.* 10, 1393–1414. doi: 10.1007/s10763-012-9344-1
- Schalk, L., Roelle, J., Saalbach, H., Berthold, K., Stern, E., and Renkl, A. (2020). Providing worked examples for learning multiple principles. *Appl. Cognit. Psychol.* 48:87. doi: 10.1002/acp.3653
- Schlag, J., Zlatkin-Troitschanskaia, O., Kühling-Thees, C., and Brückner, S. (2020). "Influences on the development of economic knowledge over the first academic year," in *Student Learning in German Higher Education: Innovative Measurement Approaches and Research Results*, Bd. 59, eds O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, and C. Lautenbach (Wiesbaden: Springer), 371–399. doi: 10.1007/978-3-658-27886-1\_19
- Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* 13, 90–100. doi: 10.1037/a0015108
- Shah, P., and Hoeffner, J. (2002). Review of graph comprehension research: implications for instruction. *Educ. Psychol. Rev.* 14, 47–69. doi: 10.1023/A:1013180410169
- Shavelson, R. J., Marino, J., Zlatkin-Troitschanskaia, O., and Schmidt, S. (2019). "Reflections on the assessment of quantitative reasoning," in *Shifting Contexts, Stable Core: Advancing Quantitative Literacy in Higher Education*, eds L. Tunstall, G. Karaali, and V. Piercey (Washington, DC: Mathematical Association of America), 163–176.
- Stern, E., Aprea, C., and Ebner, H. G. (2003). Improving cross-content transfer in text processing by means of active graphical representation. *Learn. Instr.* 13, 191–203. doi: 10.1016/S0959-4752(02)00020-8
- Strobel, B., Lindner, M. A., Saß, S., and Köller, O. (2018). Task-irrelevant data impair processing of graph reading tasks: an eye-tracking study. *Learn. Instr.* 55, 139–147. doi: 10.1016/j.learninstruc.2017.10.003
- Susac, A., Bubic, A., Kazotti, E., Planinic, M., and Palmovic, M. (2018). Student understanding of graph slope and area under a graph: a comparison of physics and nonphysics students. *Phys. Rev. Phys. Educ. Res.* 14:020109. doi: 10.1103/PhysRevPhysEducRes.14.020109
- Walstad, W. B., Watts, M., and Rebeck, K. (2007). *Test of Understanding in College Economics. Examiner's Manual*, 4th Edn. New York, NY: National Council on Economic Education.
- Wemyss, T., and van Kampen, P. (2013). Categorization of first-year university students' interpretations of numerical linear distance-time graphs. *Phys. Rev. ST Phys. Educ. Res.* 9:010107. doi: 10.1103/PhysRevSTPER.9.010107
- Wineburg, S., Breakstone, J., McGrew, S., and Ortega, T. (2018). "Why google can't save us: the challenges of our post-Gutenberg moment," in *Positive Learning in the Age of Information (PLATO): A Blessing or a Curse?*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Wiesbaden: Springer), 221–228. doi: 10.1007/978-3-658-19567-0\_13
- Wise, S. L., and Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802\_2
- Woolnough, J. (2000). How do students learn to apply their mathematical knowledge to interpret graphs in physics? *J. Res. Sci. Teach.* 30, 259–267. doi: 10.1007/BF02461633
- Zlatkin-Troitschanskaia, O., Jitmirski, J., Happ, R., Molerov, D., Schlag, J., Kühling-Thees, C., et al. (2019). Validating a test for measuring knowledge and understanding of economics among university students. *German J. Educ. Psychol.* 33, 119–133. doi: 10.1024/1010-0652/a000239

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Brückner, Zlatkin-Troitschanskaia, Küchemann, Klein and Kuhn. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Calibrating the Test of Relational Reasoning: New Information From Oblique Bifactor Models

Denis Federiakin\*

<sup>1</sup> Center for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics, Moscow, Russia

## OPEN ACCESS

### Edited by:

Patricia A. Alexander,  
University of Maryland, Rockville,  
United States

### Reviewed by:

Denis Dumas,  
University of Denver, United States  
Hongyang Zhao,  
University of Maryland, College Park,  
United States

### \*Correspondence:

Denis Federiakin  
dafederiakin@hse.ru

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Psychology

**Received:** 15 May 2020

**Accepted:** 30 July 2020

**Published:** 02 September 2020

### Citation:

Federiakin D (2020) Calibrating  
the Test of Relational Reasoning: New  
Information From Oblique Bifactor  
Models. *Front. Psychol.* 11:2129.  
doi: 10.3389/fpsyg.2020.02129

Relational reasoning (RR) is believed to be an essential construct for studying higher education learning. Relational reasoning is defined as an ability to discern meaningful patterns within any stream of information. Nonetheless, studies of RR are limited by the psychometric structure of the construct. For many instances, the composite nature of RR has been described as a bifactor structure. Bifactor models limit possibilities for studying the inner structure of composite constructs by demanding orthogonality of latent dimensions. Such assumption severely limits the interpretation of the results when it is applied to psychological constructs. However, over the last 10 years, advances in the fields of Rasch measurement led to the development of the oblique bifactor models, which relax the constraints of the orthogonal bifactor models. We show that the oblique bifactor models exhibit model fit, which is superior compared to the orthogonal bifactor model. Then, we discuss their interpretation and demonstrate the advantages of these models for investigating the inner structure of the test of RR. The data are a nationally representative sample of Russian engineering students ( $N = 2,036$ ).

**Keywords:** relational reasoning, the test of relational reasoning, bifactor models, oblique bifactor models, the Extended Testlet Model, the Generalized Subdimensional Model

## INTRODUCTION

Contemporary studies of higher education learning are unthinkable without studies of cognitive processing. Over the past 20 years, educational experiments have advanced our understanding of the intellectual and moral development of students. Moreover, they also have merged educational research with cognitive field (e.g., De Clercq et al., 2013). Researchers more and more tend to explain educational phenomena in terms of information processing and higher-order thinking skills.

Among all higher-order thinking skills, relational reasoning (RR) appears to be one of the most important. Relational reasoning is defined as an ability to discern meaningful patterns within any stream of information (Alexander and The Disciplined Reading and Learning Research Laboratory, 2012; Dumas et al., 2013). The importance of RR is well-established in the educational context; RR has been utilized as a predictive measure in a variety of studies. For example, it can



predict SAT scores both for the verbal section and for the mathematics section (Alexander et al., 2016a). Relational reasoning also demonstrated high levels of predictive validity in the domain of engineering design (Dumas and Schmidt, 2015; Dumas et al., 2016) and medical education (Dumas, 2017). In general, it proved to be a significant predictor of students' ability to produce innovations and solve problems.

As with many other conceptualizations of higher-order thinking skills, RR has been suggested as a composite construct that has many parts. However, some of the most critical manifestations of it are analogy, anomaly, antinomy, and antithesis (Alexander et al., 2016a; Dumas and Alexander, 2016). Each manifestation corresponds to a particular pattern within a set of information. Although researchers can saturate these specific forms of relations with various details of relationships within a set of information elements, these patterns are usually described (Alexander et al., 2016b) as follows:

- Similarity (identifying convergence of change patterns);
- Discrepancy (identifying dissimilarity between one element and all others or finding where the pattern breaks);
- Incompatibility (defining criteria for similarity or dissimilarity and consequently, determining how to classify the elements); and
- Polarity (identifying opposites of continuum and divergence).

However, studies of RR are limited by the psychometric structure of the construct. For many instances, the composite nature of RR has been described as a bifactor structure (Dumas and Alexander, 2016). Although bifactor modeling gained much attention in recent years, its usefulness for practitioners remains somehow restricted by its interpretation and challenges in technical applications (Bonifay et al., 2017). The main problems with it are constraints introduced in the variance–covariance matrix of latent dimensions. This severe assumption is necessary for model identification and avoiding technical difficulties. However, during a recent peak of attention to these models in psychometric literature, several extensions have been proposed to relax this limitation and provide more flexible setups for modeling bifactor structures.

The test of RR (TORR) was designed (Alexander, 2012) and validated (Alexander et al., 2016a) to capture RR and its four manifestations. The TORR was calibrated within classical test theory, item response theory (IRT) and Bayesian networks (Alexander et al., 2016a; Dumas and Alexander, 2016; Grossnickle et al., 2016). Overall, the TORR has good psychometric properties and promising implementations in educational studies. The measure has 32 nonverbal items organized into four 8-item scales that represent the four forms of RR (Figures 1–3 reflect the structure of the TORR under different model assumptions). All items are scored dichotomously and have multiple-choice formats with four response options. Additionally, each TORR scale includes two relatively easy sample items designed to familiarize participants with the content of the tasks.

The authors chose the bifactor structure of the TORR, reflecting the theoretical structure of the construct. An investigation of the TORR's dimensionality argued that a 3PL bifactor model was the best-fitting MIRT model, within which the test was calibrated (Dumas and Alexander, 2016). However, the applied model fixates the correlations of all person-specific parameters at zero, so it is impossible to study the relations between the subcomponents of RR. Therefore, some research questions on RR could not be posed despite being of interest.

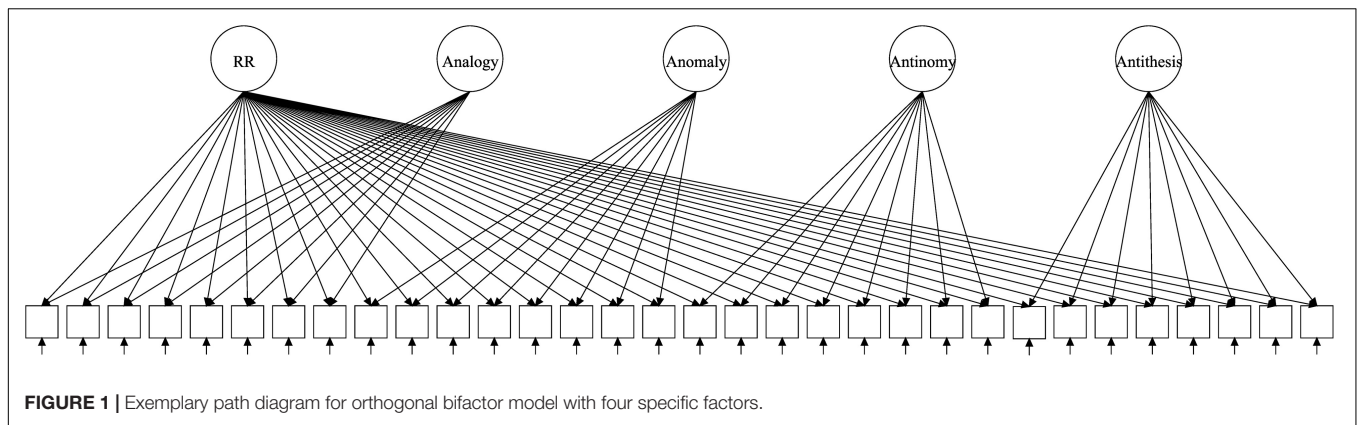
This study aims to enrich the best of our understanding of RR by advancing modeling techniques used to describe the construct. To do so, we apply oblique bifactor models, which impose less strict constraints on the variance–covariance matrix. One of these models is the Extended Testlet Model, which allows specific factors to correlate with the general factor, but forces them to be orthogonal to each other (Paek et al., 2009). Another model is the Generalized Subdimensional Model (GSM) (Brandt and Duckor, 2013), which forces specific factors to be orthogonal to the general factor but allows them to correlate with each other. We discuss the differences in their interpretation and some technical application. Then, we compare the models in terms of their model fit and estimated variance–covariance matrix and review the results obtained using the nonverbal TORR (Alexander et al., 2016a). We conclude this article with a discussion of limitations and possible further research.

The discussed models have been proposed and studied within the paradigm of Rasch measurement. Therefore, all considered models belong to Rasch measurement paradigm to make comparison across them feasible. Because the TORR utilizes dichotomous scoring, we consider only dichotomous versions of the bifactor models. Additionally, all illustrative path diagrams in the description of the models follow the structure of the TORR: 32 dichotomous items divided into four subscales (eight items per subscale).

## Bifactor Models

Bifactor models have a long history in factor analysis (Holzinger and Swineford, 1937; Schmid and Leiman, 1957). Their main feature is that each item loads on the general dimension (we call it “general factor”) and a latent variable defined by a subscale to which an item belongs (we call it “specific factor”). Such structures are useful for modeling composite instruments with non-ignorable local item dependence (LID; Bradlow et al., 1999). Local item dependence implies that item responses are random once values of all latent dimensions are known. As a result of this logic, bifactor IRT model (Bayesian Testlet Model) has been proposed, which attempted to add on latent extra dimensions to make the responses random controlling for them as well as for the general factor.

Nonetheless, such models are overparametrized and cannot be estimated unless the latent dimensions are constrained to be orthogonal (Figure 1). Assumption of total orthogonality of dimensions proposes a problem because it severely restricts the interpretation of the results. Total orthogonality means that specific factors are independent of each other and the general factor. Even if the general factor still can be somehow interpreted as the target dimension of interest, it is “purified” from



components defined by specific factors. However, interpretation of specific factors becomes even more complicated, because they become purified from general RR as well as other components. Further, difficulties in interpretation of such scores met with typically low estimates of their reliability, making the subscores virtually useless (Haberman, 2008).

As a result of this, reasonable setup for bifactor modeling appears to be limited to modeling of LID in educational testing. These (constrained) correlations of specific factors describe correlations of person–testlet interactions (nuisance dimensions) and therefore are not in the focus of interest (e.g., Reise et al., 2010). However, for psychological studies, this remains somewhat questionable assumption because researchers typically expect latent dimensions to correlate (Reise, 2012). A specific example of a consistent application of bifactor models in psychological studies can be an attempt to separate a complex construct from its contexts or situations in which it manifests itself. However, it makes subscores barely useful either way. In the end, as Reise et al. (2010) noted, “researchers view bifactor structures with great suspicion” because of such interpretational difficulty.

A direct example of such approach in Rasch measurement is the original Rasch Testlet Model (Wang and Wilson, 2005). For dichotomous items, Rasch Testlet Model can be represented as

$$g(\pi_{pi}) = \theta_p + \gamma_{p(d)} - \delta_i$$

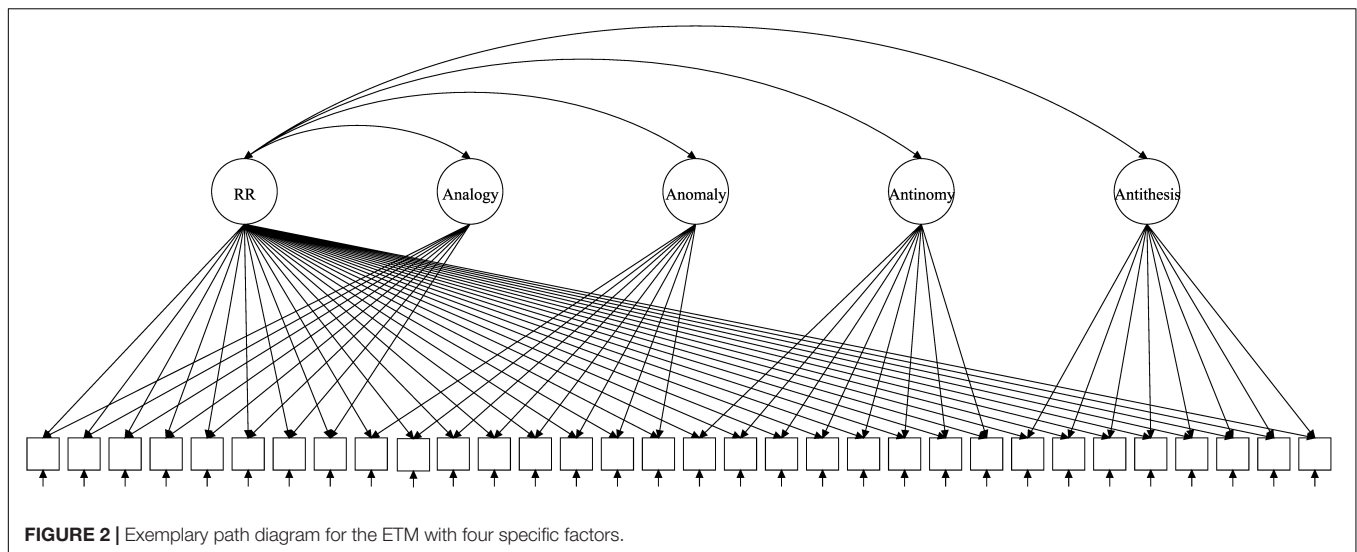
where,  $\pi_{pi}$  is the probability that the person  $p$  gets item  $i$  correctly,  $g(\cdot)$  is a function of choice (in this study, we used inverse logit function),  $\theta_p$  is the ability level of the person  $p$  on the general factor,  $\gamma_{p(d)}$  is an auxiliary ability level of the person  $p$  on the testlet-specific dimension  $d$ , and  $\delta_i$  is the generalized difficulty of the  $i$ th item.

As initially proposed, person parameters are assumed to follow independent normal distributions. Variance of specific factor accumulates the dependency between the items creating the  $d$ th testlet (LID on  $\theta$ ). This parameter varies across persons and remains fixed for all items in a testlet  $d$ ; i.e., it denotes person–testlet interaction. Thus, the probability of a correct response of person  $p$  on item with difficulty  $\delta_i$  depends on the sum of two person-specific parameters:  $\theta_p$  and  $\gamma_{p(d)}$ . As a result of such

decomposition, there are two points to note in interpreting the model. First, even under a low level of the general factor, person  $p$  can perform well for some particular testlet  $d$  if person  $p$  has a relatively high factor score on the corresponding specific factor. Second, the general factor and all specific factors are assumed to be unidimensional.

For the TORR example, orthogonal bifactor model implies that a general factor of RR is abstract, independent of its manifestations (analogy, anomaly, antinomy, and antithesis) and loads items simultaneously with them. However, this assumption is questionable, taking into account the nature of the construct. For example, commonly, researchers conceptualize the search of analogies as a basis for all cognitive functions (e.g., James, 1890; Spearman, 1927; Sternberg, 1977). Regarding four studied manifestations of RR, it means that all of them can be seen as “analogical reasoning plus something extra,” where the subscales differ in additional cognitive operations. Thus, anomaly subscale can be seen as a subscale measuring skill to find what is similar among all elements except one. Antinomy can be seen as a skill to find similarities of an initial element with secondary elements. Then, the correct answer can be determined by exclusion. Antithesis can be seen as a skill to find similarities of an initial element with secondary elements while keeping in mind a rule-implied change and reversing it. So, some elements of analogical reasoning can be found everywhere. Therefore, researchers can expect some nonzero correlations between analogy subscale and all other subscales, which has been established earlier (e.g., Alexander et al., 2016a). At the same time, the orthogonal bifactor model extracts the general factor, which can be severely contaminated by analogical reasoning.

However, such logic can be applied even further, to all other subscales. For example, antinomy subscale can be seen as a search for the anomaly, when the anchor element is presented. In contrast, in anomaly subscale itself, a respondent is required to infer the similarities across elements without the anchor. Antithesis can be seen as a search for multiple anomalies simultaneously, and so on. Therefore, nonzero correlations are expected from all subscales, which is also the case for the correlated factors model without the general factor (Alexander et al., 2016a). As a result of this, the general factor in the orthogonal bifactor model describes nothing more than a



commonality between subscales of the TORR. However, if the generalized ability of RR itself is more than a positive manifold between different types of cognitive operations, the orthogonal bifactor model is not the best choice to describe it.

## The Extended Rasch Testlet Model

As an attempt to overcome limitations of the original bifactor models, Paek et al. (2009) proposed the Extended Rasch Testlet Model (ETM). The key features of this model are correlations of specific factors with the general factor (Figure 2). Consequently, specific factors are purified from each other, but they share some estimated portion of variance with the general factor. Note, that correlations of latent variables can be negative, because items from all subscales define the general factor.

The ETM has the same formulation as the original Rasch Testlet Model and only differs in the assumption applied to the correlations of person-specific parameters. Constraining all covariances between the general factor and specific factors to zero will return a variance–covariance matrix for the original Rasch Testlet Model with the corresponding structure of the testlets. Therefore, the orthogonal Rasch Testlet Model is nested within the ETM. However, the ETM should recover factor scores better than the original Testlet Model because it takes into account the shared variance of person parameters.

It is possible to interpret correlations between specific factors and the general factor as relations between specific subparts of a more general construct and general ability itself controlling for other subparts of the construct. This interpretation follows from the classical interpretation of regression analysis. These correlations may be seen as partial correlations or standardized regression coefficients from a multivariate linear regression model.

For the TORR example, the ETM implies that the general factor of RR preserves correlations with the manifestations of it. Therefore, ETM allows for a tailored test of the hypothesis whether the general factor is just a positive manifold of specific factors or not (Van Der Maas et al., 2006). If the general

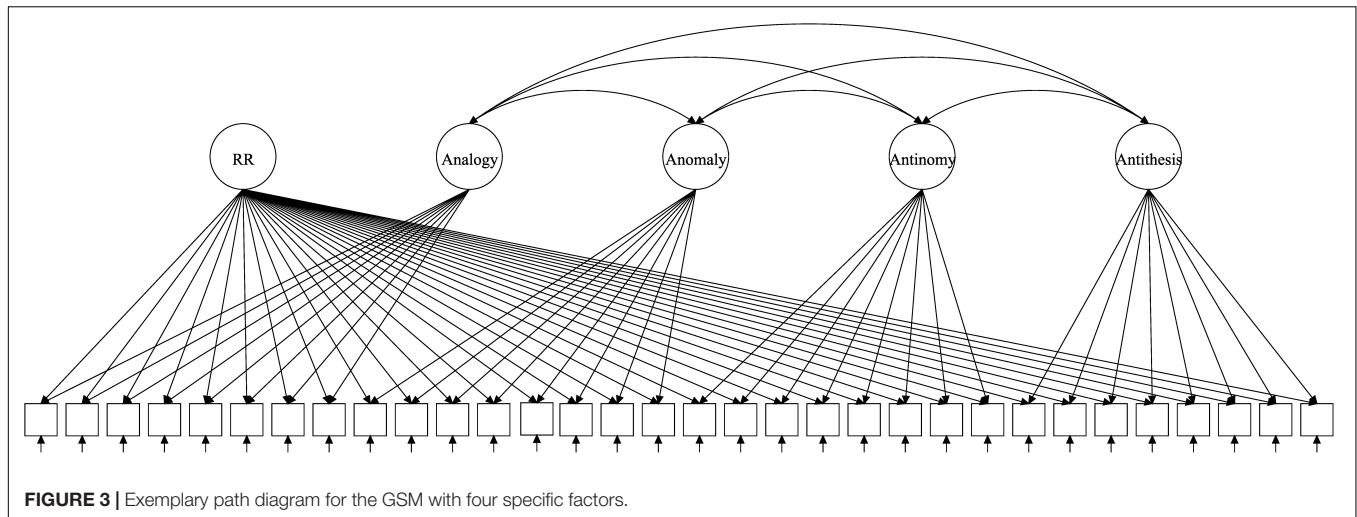
factor of RR preserves nonzero correlations with specific factors of it, then they indeed measure specific manifestations of RR, and the general factor is not an exhaustive descriptor of the latent space of the construct. At the same time, if the correlations of subscales with the general factor become insignificantly different from zero, then the general factor of the orthogonal bifactor model describes nothing more and then a commonality between subscales, and not a specific variable with distinct psychological interpretation. Testing this hypothesis is important because pushing general factor models beyond their limits can lead to the creation of such controversial phenomena, as a general factor of personality (e.g., Revelle and Wilt, 2013).

## The Generalized Subdimensional Model

The GSM (Brandt, 2017) is also a derivative of the original Rasch Testlet Model but in the opposite direction compared to the ETM. Instead of assuming orthogonality between specific factors, it allows them to correlate (Figure 3). Nonetheless, for model identification purposes and to ensure that specific factors represent subscale-specific components of general ability within it, several additional constraints must be made (for details, see Brandt, 2008). They regard to “translation” parameters ( $k_d$ ) weighting the variances of specific factors in order to equalize them: the sum of squares of translation parameters is constrained to be equal to the number of specific factors ( $D$ , for details, see Brandt and Duckor, 2013). The GSM can be described (Robitzsch et al., 2020) as

$$g(\pi_{pi}) = k_d (\theta_p + \gamma_{p(d)} - \delta_i).$$

Note that the GSM requires skipping one of the specific factors to avoid overconstraining (Brandt, 2008). This is achieved by defining the skipped specific factor as a negative sum of all remaining specific factors. Because one of the specific dimensions is excluded from the calibration, it is necessary to recalibrate the model



**FIGURE 3 |** Exemplary path diagram for the GSM with four specific factors.

with alternative reparameterizations at least three times to gather the full variance–covariance matrix of the dimensions, e.g.,

- (1) Excluding the last  $D$ th dimension to recover all covariances between all dimensions but covariances with dimension  $D$ ,
- (2) Excluding dimension  $D-1$  to recover all covariances of dimension  $D$  but the covariance of dimension  $D$  with dimension  $D-1$ , and
- (3) Excluding dimension  $D-2$  to recover the covariance of dimensions  $D$  and  $D-1$ .

A direct interpretation of this model assumes that specific factors are not purified from each other, but they are allowed to correlate freely (even negatively). Therefore, this model describes how specific factors relate to each other after the general factor is extracted. Brandt and Duckor (2013) recommended interpreting the general factor as a shared variance of dimensions from a truly multidimensional construct.

Within the context of TORR, this model describes differences in commonalities between the subscales. After the general RR is extracted, this model reveals how similar or how different the used subscales are and what is the degree of shared cognitive processing that they provoke. The correlations close to zero will mean that the subscales are virtually independent controlling for the general RR, and *vice versa*. Note that these relations are not the same as with correlated factors model, where the general factor is distributed across subscales, causing possible positive correlations. GSM explicitly models “residual” correlations between subscales, which are not described by the general factor.

When comparing the ETM and the GSM, it is important to distinguish their purposes: they are meant to answer different research questions in terms of studying the internal structure of composite constructs. These two models complement each other in terms of their focus of interest. Usage of them in a directly competitive manner fits only for deciding which model orders respondents better by the general factor. Note, however, that

they extract different factor structures. This happens because of differences in constraints imposed on the variance–covariance matrix. While orthogonal testlet models and the GSM describe general RR, which is independent of its manifestations, the ETM describes general RR, which is correlated to them. Moreover, the ETM and orthogonal testlet models describe specific factors that are independent of each other. In contrast, the GSM describes specific factors that share some portion of variance with each other.

Roughly all of these models are special cases of the multidimensional random coefficients multinomial logit model (MRCMLM; Adams et al., 1997). Therefore, the TAM package for R software (Robitzsch et al., 2020) can be used to calibrate these models. Although the GSM itself is not a special case of MRCMLM (Brandt, 2017), its predecessor—the Rasch model with subdimensions (Brandt, 2008)—is. Therefore, all discussed models can be calibrated with TAM package, using the same algorithms for likelihood estimation. The parameters were estimated with the quasi–Monte-Carlo algorithm implemented in the TAM package, which proved to be efficient in the presence of high-dimensional latent ability space (Wu et al., 2007). To estimate reliability, we used expected *a posteriori* (EAP) estimates of factor scores (Bock and Mislevy, 1982) because of their flexibility in complex multidimensional setup. Moreover, EAP uses distributional information from the variance–covariance matrix to increase the precision of the estimates.

To demonstrate the advantages of oblique bifactor models in terms of global model fit, we analyzed absolute and relative model fit indices. To estimate the absolute global fit, we used root mean square error of approximation (RMSEA; Steiger, 1990) and standardized root mean square residual (SRMSR; Hu and Bentler, 1999) according to the recommendations given by Shi et al. (2020). Root mean square residual can be interpreted as an unstandardized measure of the distance between the data-generating model and the hypothesized model. Standardized root mean square residual possesses a straightforward interpretation: it is just on average of correlation residuals. As a result of



this, models with lower values of these indices are preferable. We also used comparative fit index (CFI; Bentler, 1990) as an additional measure of incremental model fit. In contrast to RMSEA, CFI is commonly interpreted as a measure of the distance between the hypothesized model and the baseline model, where all the variables are uncorrelated. Therefore, models with higher CFI values are preferable. Note, however, that despite conventional “rules of thumb” derived in factor analytical approach, there are no strict cutoff criteria for IRT models (e.g., Maydeu-Olivares, 2013; Savalei, 2018; Xia and Yang, 2019). Consequently, we cannot definitively conclude that some or all models fit or do not fit the data. Additionally, we compared the relative fit of the models with the Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). These indices allow for comparison of model fit across nonnested models, introducing a penalty for extra parameters (AIC) with respect to sample size (BIC). Lower values of these indices imply a better global model-data fit accounting for model complexity.

## DATA

The data used for this study is a part of a larger project, called the Super Test Project, led by researchers at Stanford University in collaboration with ETS and researchers from various countries including China and Russia. The overall purpose of this project is to examine learning outcomes and institutional- and individual-level factors related to them for electrical engineering and computer science students across multiple countries. To this end, the research team also collected a wealth of contextual survey data from students, faculty, and administrators.

As a part of the Super Test Project, the TORR was administered to Russian electrical engineering and computer science students. We randomly included 34 Russian universities in a nationally representative sample of engineering students. The testing was conducted in November–December 2016 among students graduating in 2017 (when they were in the middle of their fourth year of studying) and in April 2017 among students graduating in 2019 (when they were at the end of their second year). The testing was conducted in a computer-based format. Students had 60 min to complete the TORR. The data cleaning procedure included the deletion of all response profiles with 50% or less of the responses on any subscale. Consequently, 76 profiles were deleted from the database (approximately 3.6%). We compared correlations between subscales in raw scores before and after deletion of the profiles, to prove that the deleted responses did not bias the subsequent analyses. The change in correlations was less than 0.001. The final sample size is 2,036 students.

## RESULTS

The results of the global model fit analysis are reported in **Table 1** (note that the deviance statistic in the GSM

**TABLE 1** | Results of the model comparison.

Statistics	Models		
	Testlet model	ETM	GSM
$\chi^2$ statistics for the baseline model		10,980.342	
Degrees of freedom for $\chi^2$ statistics		496	
Sample size		2,036	
Number of free parameters	37	41	42
Degrees of freedom for $\chi^2$ statistics	491	487	486
$\chi^2$ statistics	3,290.570	2,315.269	2,299.300
RMSEA	0.053	0.043	0.043
CFI	0.733	0.826	0.827
SRMSR	0.058	0.051	0.050
Deviance	82,009.13	81,677.74*	81,661.43
AIC	82,083.13	81,759.74	81,745.43
BIC	82,291.52	81,990.65	81,981.97

\*Likelihood ratio test reveals that ETM fits significantly better than Rasch Testlet model (critical  $\chi^2$  value is 9.49 for 4 degrees of freedom on  $p < 0.05$  significance level; empirical  $\chi^2$  value is 331.39).

**TABLE 2** | Internal structure of country-specific relational reasoning construct from orthogonal Rasch Testlet model.

Scale	Variance	EAP reliability
General RR	0.63	0.60
Analogy	0.75	0.39
Anomaly	0.47	0.22
Antinomy	0.61	0.32
Antithesis	0.83	0.46

is averaged over its four reparameterizations). As **Table 1** suggests, both the ETM and the GSM fit data better, indicating that oblique bifactor models provide a better description of RR than orthogonal bifactor model. In other words, correlations of latent dimensions should not be ignored while studying RR.

The results from Rasch Testlet Model are presented in **Table 2**. The results indicate that the sample appears to be rather homogeneous in terms of the ability distribution. Relatively small variances of the latent abilities can explain the low reliability of estimates. Variance of specific factors from this model measures a degree of local dependence (Wang and Wilson, 2005). Therefore, it is notable that analogy and antithesis subscales possess more specific variance (LID) than the entire general factor.

The results from the ETM are presented in **Table 3**. The results suggest that the variance of three components of RR lowered compared to their estimates from orthogonal Rasch Testlet Model (analogy, anomaly, and antithesis). However, the variance of the fourth component (antinomy) increased. Notably, the variance of the general RR did not change across the models, but its reliability increased. We emphasize that the interpretation of factors differs across these models because of the difference in the modeled structures.



**TABLE 3** | Internal structure of country-specific relational reasoning construct from the ETM.

Scale	Correlation with general RR	Variance	EAP reliability
General RR	–	0.63	0.67
Analogy	0.27*	0.34	0.45
Anomaly	–0.09*	0.17	0.12
Antinomy	–0.70*	0.76	0.46
Antithesis	–0.02	0.63	0.38

\* $p < 0.001$ .**TABLE 4** | Internal structure of country-specific relational reasoning construct from the GSM.

Scale	Scales			Variance	EAP reliability
	Analogy	Anomaly	Antinomy		
General RR	–	–	–	0.57	0.75
Analogy				0.24	0.39
Anomaly	0.32*			0.25	0.35
Antinomy	–0.55*	–0.70*		0.86	0.44
Antithesis	–0.21*	–0.03	–0.53*	0.36	0.45

\* $p < 0.001$ .

The results from the ETM suggest that the better engineering students perform in general on RR, the worse they are at defining criteria to distinguish continuums (antinomy scale). However, this exact subscale describes a measure of the ability to identify compromises between different solutions (Dumas and Schmidt, 2015). This may be a sign of potential difficulties in future engineering performance for students. At the same time, positive relations between the overall reasoning and analogical reasoning have been identified in several previous studies (Carpenter et al., 1990) and demonstrated here. However, relations among other forms of RR and general RR itself are negative or insignificant, suggesting that these parts of RR do not relate to it in any way that cannot be explained by other subscales (that is, controlling for other subscales).

The last portion of the results came from the GSM (Table 4). Note that these results are averaged across four recalibrations of the model (skipping every specific factor from calibration). However, the maximum difference between the same parameter across different recalibrations is less than 0.02. The results suggest that this model provides overall the most balanced and reliable estimates of a general RR general as well as its specific factors. That is, although variances of latent variables are not the biggest across the three considered models, the reliability of them appears to be optimal. Notably, the general RR returns the highest reliability under the GSM structure along with shrinking its variance. However, the variance of antinomy subscale reaches its peak in this model, implying that this scale measures cognitive skill distinct from general RR. Patterns of correlations of latent variables support this conclusion.

These relationships may indicate how students achieve a score on general RR. The abilities to find anomalies and analogies are positively correlated. It is possible to conclude that these

abilities share, to some extent, the same cognitive processing: to define which elements are to be excluded, one should define what is similar among other elements. Interestingly, scores on the anomaly subscale do not depend on scores on the antithesis subscale: the ability to define an outlying sign of a breaking pattern does not relate to the ability to find the opposite pattern.

## DISCUSSION

Relational reasoning is believed to be an essential construct for studying higher education learning. Nature of RR reflects the ability of an individual to capture complex relations between patterns within the stream of information. Accordingly, RR can be conceptualized in a multitude of forms, based on the content of information (e.g., professional knowledge or common sense), its type (verbal, numerical, graphical), complexity of relations (e.g., number of analyzed rules), or kind of relations (such as resemblance or divergence). The analyzed TORR conceptualizes it in four types of relations connecting abstract geometric patterns: analogy (similarity), anomaly (discrepancy), antinomy (incompatibility), and antithesis (polarity; Alexander et al., 2016a; Dumas and Alexander, 2016). Many studies proved its predictive power and importance, and the TORR itself has been shown to exhibit good psychometric properties.

However, studying the nature of RR has been limited by the traditions of psychometric modeling. Because RR itself has a composite nature, researchers applied bifactor models to describe it. As a result of this, extracted factor scores do not correlate with each other because of technical necessity. For the case of the TORR, this means that scores on the analogy subscale are not related to general RR; nor are they related to any other subscale. However, analogical reasoning is regarded as the basis of cognitive processing (Gust et al., 2008). Therefore, at least this subscale should be correlated with general RR as well as with other subscales.

Bifactor modeling techniques require severe constraints to be forced on relations of latent variables: they are assumed to be orthogonal. As a result of this, their interpretation becomes sophisticated and barely useful for practitioners (Bonifay et al., 2017). That is, interpretation of specific factors implies that they do not contain any information, described by the general factor; nor do they contain information described by other specific factors. Consequently, the domain of bifactor models usually is limited by the separation of the general factor from contexts of its manifestations. Primordial example of this is modeling LID, caused by shared stimuli of items (DeMars, 2006). Within this example, subscores do not possess any meaningful interpretation from the beginning and are extracted only to reach local independence of items on person parameters. This is, clearly, not the case for composite psychological constructs, where components have meaningful interpretation and cannot be expected to be orthogonal.

Oblique bifactor models can be considered to overcome these limitations. These models allow relaxing the assumption of total orthogonality traditionally required for bifactor modeling. The set of these models includes (but is not limited to) (1) the ETM

(Paek et al., 2009) and (2) the GSM (Brandt, 2017). While the ETM allows specific factors to correlate with the general factor but not with each other, the GSM allows them to correlate with each other but not with the general factor. As a result of this, these models extract general factors that differ in interpretation and psychological meaning but allow researchers to study the inner structure of psychological constructs. However, these models do not exhaust the set of oblique bifactor models; e.g., one can conceive models with zero constraints on the sum of some or all values in the variance–covariance matrix (e.g., Robitzsch et al., 2020). Nonetheless, the interpretation of such models is next to impossible because it is next to impossible to have theoretical expectations of this kind. It appears such models can only be used to improve model fit in the case when the orthogonal bifactor model exhibits inappropriate model fit. Despite that, further investigation of oblique bifactor models appears to be promising. Such further research include other constraints on the variance–covariance matrix (including nonzero constraints on the sum of its values) and using strong priors about variance–covariance values in the Bayesian paradigm.

For the TORR example, the GSM is the best-fitting model. This means that after extraction of the general RR subscales preserve some relations between each other. Also, these correlations are more important than correlations of the subscales with the general RR. This means that the manifestations of the RR differ more significantly in their relation to each other, whereas their relation to the RR is more homogeneous. Moreover, the assumption of their orthogonality leads to misspecification of the measurement model. Combining results of the ETM and the GSM, several conclusions arise. First, cognitive processing of analogies is the basis of RR, as well as other intellectual activities (Carpenter et al., 1990; Gust et al., 2008). Second, students of engineering programs can increase their total RR scores by having higher scores of one of analogy and anomaly, antinomy, or antithesis abilities. Because this indicates, to some extent, mutually exclusive groups of cognitive abilities, a possible investigation of these results may be directed profiling of cognitive abilities. Third, the most outlying manifestation of RR is antinomy. It correlates negatively to negligibly with other components of RR and the general RR itself. More in-depth investigation of this cognitive process is of great interest.

Unfortunately, the TORR subscores from oblique bifactor models appear to be unreliable, as well as from orthogonal bifactor model. Although this may not be the case for other instruments, this is a natural result for bifactor modeling (Haberman and Sinharay, 2010). However, for some purposes, it is required to have specific subscores with reliable estimates. There are several ways to do so. One of them is recalibrating data within correlated factors model and defactor ignoring model fit indices. This approach is unpopular in the statistical literature, although it fits to willingness to not restrict interpretation to a single model (Organization for Economic Co-operation and Development, 2005; Brandt et al., 2014). Another approach is the application of the composite model, which combines reflective and formative approaches within a single model (Wilson and Gochyyev, 2020). However, this model is more or less equivalent to the correlated factors model and therefore describes the same

relations between subscales. While bifactor models *extract* the general factor from the subscales, the composite model *distributes* it across them in the same manner as models without general factor do. As a result of this, it provides high estimates of reliability for subscores.

Several significant limitations cannot be ignored. In this study, we did not discuss the TORR comparability across various demographics groups, for two reasons. The first reason is regarding the graphical nature of the test and therefore the plausible assumption for item comparability. Second, previous studies revealed decent item-level cross-demographics comparability of the TORR in terms of race and gender (Dumas, 2016; Dumas and Alexander, 2018). However, those demographic groups were sampled inside the United States. Therefore, cross-national comparability of the TORR remains unknown. Nevertheless, studying cross-national comparability in terms of item behavior is possible using modifications of the orthogonal bifactor model that allow for the decomposition of differential item functioning into testlet-based and item-based components (Paek and Fukuhara, 2015; Fukuhara and Paek, 2016). Applications of this approach to enhanced bifactor models and changes in their interpretation are of interest. Nonetheless, since the topic of comparability lies beyond the scope of this article, we did not test it. Another limitation concerns the interpretation of subscores and their relations. Although they can be described in terms of original names of the subscales, further theoretical and, probably, experimental study of subscales purified from general RR and subscales purified from each other is required. We also did not consider higher-order model. Even this model is nested within the same class of hierarchical models as bifactor models (Yung et al., 1999; Rijmen, 2010), they reflect latent structures, which can be analytically inferred from the correlated factors model without general factor. Therefore, the second-order models are vulnerable to the positive manifold effect. Moreover, they do not imply the use of specific factor scores, which makes them less useful for practitioners.

Probably, the most significant limitation of this study concerns the application of only Rasch-type models. The used oblique bifactor models were proposed and studied only within Rasch modeling approach. This guarantees that these models return unbiased estimates. Moreover, Rasch modeling setup provides numerical stability, which is desirable for such heavily parametrized models as oblique bifactor models. However, the counterparts of the described models can be conceived within 2PL (Birnbaum, 1968) and, probably, other IRT models. Rasch modeling imposes strict assumptions regarding item discrimination parameters. On the one hand, it guarantees that the probability of solving an easier item is always (on any level of ability) higher than the probability of solving a harder item. This allows for a straightforward interpretation of parameters and facilitates the development of the continuum of observed behavior. On the other hand, it implicates that all items share an equal portion of variance with corresponding latent variable. This assumption may not be as feasible for psychological constructs as it is for educational constructs. Therefore, replication of this study under IRT models with more parameters per item is of interest.

Given, of course, that oblique bifactor models are as well-behaved under those IRT models as under Rasch modeling framework.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and

institutional requirements. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

DF conducted the analyses and wrote the manuscript.

## ACKNOWLEDGMENTS

Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.

## REFERENCES

- Adams, R. J., Wilson, M., and Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *App. Psychol. Meas.* 21, 1–23. doi: 10.1177/0146621697211001
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 716–723. doi: 10.1109/TAC.1974.1100705
- Alexander, P. A. (2012). *The Test of Relational Reasoning*. College Park, MD: Disciplined Reading and Learning Research Laboratory.
- Alexander, P. A., Dumas, D. G., Grossnickle, E. M., List, A., and Firetto, C. M. (2016a). Measuring relational reasoning. *J. Exp. Educ.* 84, 119–151. doi: 10.1080/00220973.2014.963216
- Alexander, P. A., Jablansky, S., Singer, L. M., and Dumas, D. G. (2016b). Relational reasoning: what we know and why it matters. *Policy Insights Behav. Brain Sci.* 3, 36–44. doi: 10.1177/2372732215622029
- Alexander, P. A., and The Disciplined Reading, and Learning Research Laboratory [DRLRL] (2012). Reading into the future: competence for the 21st century. *Educ. Psychol.* 47, 259–280. doi: 10.1080/00461520.2012.722511
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol. Bull.* 107, 238–246. doi: 10.1037/0033-2909.107.2.238
- Birnbaum, A. L. (1968). “Some latent trait models and their use in inferring an examinee’s ability,” in *Statistical Theories of Mental Test Scores*, eds F. M. Lord and M. R. Novick (Reading, MA: Addison-Wesley).
- Bock, R. D., and Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Appl. Psychol. Meas.* 6, 431–444. doi: 10.1177/014662168200600405
- Bonifay, W., Lane, S. P., and Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clin. Psychol. Sci.* 5, 184–186. doi: 10.1177/2167702616657069
- Bradlow, E. T., Wainer, H., and Wang, X. (1999). A bayesian random effects model for testlets. *Psychometrika* 64, 153–168. doi: 10.1007/BF02294533
- Brandt, S. (2008). “Estimation of a Rasch model including subdimensions,” in *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments*, Vol. 1, eds M. von Davier and D. Hastedt (Princeton, NJ: IEA-ETS Research Institute), 51–70.
- Brandt, S. (2017). Concurrent unidimensional and multidimensional calibration within item response theory. *Pensamiento Educativo. Revista de Investigación Educativa Latinoamericana* 54, 1–16. doi: 10.7764/PEL.54.2.2017.4
- Brandt, S., and Duckor, B. (2013). Increasing unidimensional measurement precision using a multidimensional item response model approach. *Psychol. Asses. Modeling* 55, 148–161.
- Brandt, S., Duckor, B., and Wilson, M. (2014). “A utility-based validation study for the dimensionality of the performance assessment for California teachers,” in *Presented in the Annual Conference of the American Educational Research Association (AERA)*, Philadelphia, PA.
- Carpenter, P. A., Just, M. A., and Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychol. Rev.* 97, 404–431. doi: 10.1037/0033-295X.97.3.404
- De Clercq, M., Galand, B., and Frenay, M. (2013). Chicken or the egg: longitudinal analysis of the causal dilemma between goal orientation, self-regulation and cognitive processing strategies in higher education. *Stud. Educ. Eval.* 39, 4–13. doi: 10.1016/j.stueduc.2012.10.003
- DeMars, C. E. (2006). Application of the Bi-Factor multidimensional item response theory model to Testlet-Based tests. *J. Educ. Meas.* 43, 145–168. doi: 10.1111/j.1745-3984.2006.00010.x
- Dumas, D. G. (2016). *Seeking Cultural Fairness in a measure of Relational Reasoning*. dissertation, University of Maryland, College Park, MD, doi: 10.13016/M2T48H
- Dumas, D. G. (2017). Relational reasoning in science, medicine, and engineering. *Educ. Psychol. Rev.* 29, 73–95. doi: 10.1007/s10648-016-9370-6
- Dumas, D. G., and Alexander, P. A. (2016). Calibration of the test of relational reasoning. *Psychol. Assess.* 28, 1303–1318. doi: 10.1037/pas0000267
- Dumas, D. G., and Alexander, P. A. (2018). Assessing differential item functioning on the test of relational reasoning. *Front. Educ.* 3:14. doi: 10.3389/educ.2018.00014
- Dumas, D. G., Alexander, P. A., and Grossnickle, E. M. (2013). Relational reasoning and its manifestations in the educational context: a systematic review of the literature. *Educ. Psychol. Rev.* 25, 391–427. doi: 10.1007/s10648-013-9224-4
- Dumas, D. G., and Schmidt, L. (2015). Relational reasoning as predictor for engineering ideation success using TRIZ. *J. Eng. Des.* 26, 74–88. doi: 10.1080/09544828.2015.1020287
- Dumas, D. G., Schmidt, L. C., and Alexander, P. A. (2016). Predicting creative problem solving in engineering design. *Think. Skills Creat.* 21, 50–66. doi: 10.1016/j.tsc.2016.05.002
- Fukuhara, H., and Paek, I. (2016). Exploring the utility of logistic mixed modeling approaches to simultaneously investigate item and testlet DIF on testlet-based data. *J. Appl. Meas.* 17, 79–90.
- Grossnickle, E. M., Dumas, D. G., Alexander, P. A., and Baggetta, P. (2016). Individual differences in the process of relational reasoning. *Learn. Instr.* 42, 141–159. doi: 10.1016/j.learninstruc.2016.01.013
- Gust, H., Krumnack, U., Kühnberger, K. U., and Schwering, A. (2008). Analogical reasoning: a core of cognition. *Künstliche Intelligenz* 1, 8–12.
- Haberman, S. J. (2008). When can subscores have value? *J. Educ. Behav. Stat.* 33, 204–229. doi: 10.3102/1076998607302636
- Haberman, S. J., and Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika* 75, 209–227. doi: 10.1007/s11336-010-9158-4
- Holzinger, K. J., and Swineford, F. (1937). The bi-factor method. *Psychometrika* 2, 41–54. doi: 10.1007/BF02287965
- Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Modeling* 6, 1–55. doi: 10.1080/10705519909540118
- James, W. (1890). *The Principles of Psychology*. New York, NY: Holt.
- Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement* 11, 71–101. doi: 10.1080/15366367.2013.831680
- Organization for Economic Co-operation, and Development [OECD] (2005). *PISA 2003 Technical Report*. Paris: OECD.
- Paek, I., and Fukuhara, H. (2015). Estimating a DIF decomposition model using a random-weights linear logistic test model approach. *Behav. Res. Methods* 47, 890–901. doi: 10.3758/s13428-014-0512-9

- Paek, I., Yon, H., Wilson, M., and Kang, T. (2009). Random parameter structure and the testlet model: extension of the rasch testlet model. *J. Appl. Meas.* 10, 394–407.
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivar. Behav. Res.* 47, 667–696. doi: 10.1080/00273171.2012.715555
- Reise, S. P., Moore, T. M., and Haviland, M. G. (2010). Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *J. Pers. Assess.* 92, 544–559. doi: 10.1080/00223891.2010.496477
- Revelle, W., and Wilt, J. (2013). The general factor of personality: a general critique. *J. Res. Personal.* 47, 493–504. doi: 10.1016/j.jrp.2013.04.012
- Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *J. Educ. Meas.* 47, 361–372. doi: 10.1111/j.1745-3984.2010.00118.x
- Robitzsch, A., Kiefer, T., and Wu, M. (2020). *Package TAM: Test Analysis Modules. R Package Version 3.5–19*.
- Savalei, V. (2018). On the computation of the RMSEA and CFI from the mean-and-variance corrected test statistic with nonnormal data in SEM. *Multivar. Behav. Res.* 53, 419–429. doi: 10.1080/00273171.2018.1455142
- Schmid, J., and Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika* 22, 53–61. doi: 10.1007/BF02289209
- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136
- Shi, D., Maydeu-Olivares, A., and Rosseel, Y. (2020). Assessing fit in ordinal factor analysis models: SRMR vs. RMSEA. *Struct. Equ. Modeling* 27, 1–15. doi: 10.1080/10705511.2019.1611434
- Spearman, C. (1927). *The Abilities of Man: Their Nature and Measurement*. New York, NY: Macmillan.
- Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivar. Behav. Res.* 25, 173–180. doi: 10.1207/s15327906mbr2502\_4
- Sternberg, R. J. (1977). *Intelligence, Information Processing, and Analogical Reasoning: The Componential Analysis of Human Abilities*. Mahwah, NJ: Erlbaum.
- Van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., and Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychol. Rev.* 113, 842–861. doi: 10.1037/0033-295X.113.4.842
- Wang, W. C., and Wilson, M. (2005). The Rasch testlet model. *Appl. Psychol. Meas.* 29, 126–149. doi: 10.1177/0146621604271053
- Wilson, M., and Gochyyev, P. (2020). Having your cake and eating it too: multiple dimensions and a composite. *Measurement* 151:107247. doi: 10.1016/j.measurement.2019.107247
- Wu, M. L., Adams, R. J., Wilson, M., and Haldane, S. A. (2007). *ConQuest, ACER Generalised Item Response Modeling Software*. Camberwell: Australian Council for Educational Research.
- Xia, Y., and Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: the story they tell depends on the estimation methods. *Behav. Res. Methods* 51, 409–428. doi: 10.3758/s13428-018-1055-2
- Yung, Y. F., Thissen, D., and McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika* 64, 113–128. doi: 10.1007/BF02294531

**Conflict of Interest:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Federiakin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Performance Assessment of Critical Thinking: Conceptualization, Design, and Implementation

Henry I. Braun<sup>1\*</sup>, Richard J. Shavelson<sup>2</sup>, Olga Zlatkin-Troitschanskaia<sup>3</sup> and Katrina Borowiec<sup>1</sup>

<sup>1</sup> Lynch School of Education and Human Development, Boston College, Chestnut Hill, MA, United States, <sup>2</sup> Graduate School of Education, Stanford University, Stanford, CA, United States, <sup>3</sup> Department of Business and Economics Education, Johannes Gutenberg University, Mainz, Germany

## OPEN ACCESS

### Edited by:

Isabel Benítez,  
University of Granada, Spain

### Reviewed by:

Anders Jönsson,  
Kristianstad University, Sweden  
Katrina Roohr,  
Educational Testing Service,  
United States

### \*Correspondence:

Henry I. Braun  
braunh@bc.edu

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Education

**Received:** 30 May 2020

**Accepted:** 04 August 2020

**Published:** 08 September 2020

### Citation:

Braun HI, Shavelson RJ,  
Zlatkin-Troitschanskaia O and  
Borowiec K (2020) Performance  
Assessment of Critical Thinking:  
Conceptualization, Design,  
and Implementation.  
Front. Educ. 5:156.  
doi: 10.3389/feduc.2020.00156

Enhancing students' critical thinking (CT) skills is an essential goal of higher education. This article presents a systematic approach to conceptualizing and measuring CT. CT generally comprises the following mental processes: identifying, evaluating, and analyzing a problem; interpreting information; synthesizing evidence; and reporting a conclusion. We further posit that CT also involves dealing with dilemmas involving ambiguity or conflicts among principles and contradictory information. We argue that performance assessment provides the most realistic—and most credible—approach to measuring CT. From this conceptualization and construct definition, we describe one possible framework for building performance assessments of CT with attention to extended performance tasks within the assessment system. The framework is a product of an ongoing, collaborative effort, the *International Performance Assessment of Learning* (iPAL). The framework comprises four main aspects: (1) The storyline describes a carefully curated version of a complex, real-world situation. (2) The challenge frames the task to be accomplished (3). A portfolio of documents in a range of formats is drawn from multiple sources chosen to have specific characteristics. (4) The scoring rubric comprises a set of scales each linked to a facet of the construct. We discuss a number of use cases, as well as the challenges that arise with the use and valid interpretation of performance assessments. The final section presents elements of the iPAL research program that involve various refinements and extensions of the assessment framework, a number of empirical studies, along with linkages to current work in online reading and information processing.

**Keywords:** critical thinking, performance assessment, assessment framework, scoring rubric, evidence-centered design, 21st century skills, higher education

## INTRODUCTION

In their mission statements, most colleges declare that a principal goal is to develop students' higher-order cognitive skills such as critical thinking (CT) and reasoning (e.g., Shavelson, 2010; Hyytinen et al., 2019). The importance of CT is echoed by business leaders (Association of American Colleges and Universities [AACU], 2018), as well as by college faculty (for curricular analyses in Germany, see e.g., Zlatkin-Troitschanskaia et al., 2018). Indeed, in the 2019 administration of the Faculty Survey of Student Engagement (FSSE), 93% of faculty



reported that they “very much” or “quite a bit” structure their courses to support student development with respect to thinking critically and analytically. In a listing of 21st century skills, CT was the most highly ranked among FSSE respondents (Indiana University, 2019). Nevertheless, there is considerable evidence that many college students do not develop these skills to a satisfactory standard (Arum and Roksa, 2011; Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019). This state of affairs represents a serious challenge to higher education – and to society at large.

In view of the importance of CT, as well as evidence of substantial variation in its development during college, its proper measurement is essential to tracking progress in skill development and to providing useful feedback to both teachers and learners. Feedback can help focus students’ attention on key skill areas in need of improvement, and provide insight to teachers on choices of pedagogical strategies and time allocation. Moreover, comparative studies at the program and institutional level can inform higher education leaders and policy makers.

The conceptualization and definition of CT presented here is closely related to models of information processing and online reasoning, the skills that are the focus of this special issue. These two skills are especially germane to the learning environments that college students experience today when much of their academic work is done online. Ideally, students should be capable of more than naïve Internet search, followed by copy-and-paste (e.g., McGrew et al., 2017); rather, for example, they should be able to critically evaluate both sources of evidence and the quality of the evidence itself in light of a given purpose (Leu et al., 2020).

In this paper, we present a systematic approach to conceptualizing CT. From that conceptualization and construct definition, we present one possible framework for building *performance assessments* of CT with particular attention to extended performance tasks within the test environment. The penultimate section discusses some of the challenges that arise with the use and valid interpretation of performance assessment scores. We conclude the paper with a section on future perspectives in an emerging field of research – the iPAL program.

## CONCEPTUAL FOUNDATIONS, DEFINITION AND MEASUREMENT OF CRITICAL THINKING

In this section, we briefly review the concept of CT and its definition. In accordance with the principles of evidence-centered design (ECD; Mislevy et al., 2003), the conceptualization drives the measurement of the construct; that is, implementation of ECD directly links aspects of the assessment framework to specific facets of the construct. We then argue that performance assessments designed in accordance with such an assessment framework provide the most realistic—and most credible—approach to measuring CT. The section concludes with a sketch of an approach to CT measurement grounded in *performance assessment*.

## Concept and Definition of Critical Thinking

Taxonomies of 21st century skills (Pellegrino and Hilton, 2012) abound, and it is neither surprising that CT appears in most taxonomies of learning, nor that there are many different approaches to defining and operationalizing the construct of CT. There is, however, general agreement that CT is a multifaceted construct (Liu et al., 2014). Liu et al. (2014) identified five key facets of CT: (i) evaluating evidence and the use of evidence; (ii) analyzing arguments; (iii) understanding implications and consequences; (iv) developing sound arguments; and (v) understanding causation and explanation.

There is empirical support for these facets from college faculty. A 2016–2017 survey conducted by the Higher Education Research Institute (HERI) at the University of California, Los Angeles found that a substantial majority of faculty respondents “frequently” encouraged students to: (i) evaluate the quality or reliability of the information they receive; (ii) recognize biases that affect their thinking; (iii) analyze multiple sources of information before coming to a conclusion; and (iv) support their opinions with a logical argument (Stolzenberg et al., 2019).

There is general agreement that CT involves the following mental processes: identifying, evaluating, and analyzing a problem; interpreting information; synthesizing evidence; and reporting a conclusion (e.g., Erwin and Sebrell, 2003; Kosslyn and Nelson, 2017; Shavelson et al., 2018). We further suggest that CT includes dealing with dilemmas of ambiguity or conflict among principles and contradictory information (Oser and Biedermann, 2020).

Importantly, Oser and Biedermann (2020) posit that CT can be manifested at three levels. The first level, *Critical Analysis*, is the most complex of the three levels. Critical Analysis requires both knowledge in a specific discipline (conceptual) and procedural analytical (deduction, inclusion, etc.) knowledge. The second level is *Critical Reflection*, which involves more generic skills “. . . necessary for every responsible member of a society” (p. 90). It is “a basic attitude that must be taken into consideration if (new) information is questioned to be true or false, reliable or not reliable, moral or immoral etc.” (p. 90). To engage in Critical Reflection, one needs not only apply analytic reasoning, but also adopt a reflective stance toward the political, social, and other consequences of choosing a course of action. It also involves analyzing the potential motives of various actors involved in the dilemma of interest. The third level, *Critical Alertness*, involves questioning one’s own or others’ thinking from a skeptical point of view.

Wheeler and Haertel (1993) categorized higher-order skills, such as CT, into two types: (i) when solving problems and making decisions in professional and everyday life, for instance, related to civic affairs and the environment; and (ii) in situations where various mental processes (e.g., comparing, evaluating, and justifying) are developed through formal instruction, usually in a discipline. Hence, in both settings, individuals must confront situations that typically involve a problematic event, contradictory information, and possibly conflicting principles. Indeed, there is an ongoing debate concerning whether CT

should be evaluated using generic or discipline-based assessments (Nagel et al., 2020). Whether CT skills are conceptualized as generic or discipline-specific has implications for how they are assessed and how they are incorporated into the classroom.

In the iPAL project, CT is characterized as a multifaceted construct that comprises conceptualizing, analyzing, drawing inferences or synthesizing information, evaluating claims, and applying the results of these reasoning processes to various purposes (e.g., solve a problem, decide on a course of action, find an answer to a given question or reach a conclusion) (Shavelson et al., 2019). In the course of carrying out a CT task, an individual typically engages in activities such as specifying or clarifying a problem; deciding what information is relevant to the problem; evaluating the trustworthiness of information; avoiding judgmental errors based on “fast thinking”; avoiding biases and stereotypes; recognizing different perspectives and how they can reframe a situation; considering the consequences of alternative courses of actions; and communicating clearly and concisely decisions and actions. The order in which activities are carried out can vary among individuals and the processes can be non-linear and reciprocal.

In this article, we focus on generic CT skills. The importance of these skills derives not only from their utility in academic and professional settings, but also the many situations involving challenging moral and ethical issues – often framed in terms of conflicting principles and/or interests – to which individuals have to apply these skills (Kegan, 1994; Tessier-Lavigne, 2020). Conflicts and dilemmas are ubiquitous in the contexts in which adults find themselves: work, family, civil society. Moreover, to remain viable in the global economic environment – one characterized by increased competition and advances in second generation artificial intelligence (AI) – today’s college students will need to continually develop and leverage their CT skills. Ideally, colleges offer a supportive environment in which students can develop and practice effective approaches to reasoning about and acting in learning, professional and everyday situations.

## Measurement of Critical Thinking

Critical thinking is a multifaceted construct that poses many challenges to those who would develop relevant and valid assessments. For those interested in current approaches to the measurement of CT that are not the focus of this paper, consult Zlatkin-Troitschanskaia et al. (2018).

In this paper, we have singled out *performance assessment* as it offers important advantages to measuring CT. Extant tests of CT typically employ response formats such as forced-choice or short-answer, and scenario-based tasks (for an overview, see Liu et al., 2014). They all suffer from moderate to severe construct underrepresentation; that is, they fail to capture important facets of the CT construct such as perspective taking and communication. High fidelity performance tasks are viewed as more authentic in that they provide a problem context and require responses that are more similar to what individuals confront in the real world than what is offered by traditional multiple-choice items (Messick, 1994; Braun, 2019). This greater verisimilitude promises higher levels of construct representation and lower levels of construct-irrelevant variance.

Such performance tasks have the capacity to measure facets of CT that are imperfectly assessed, if at all, using traditional assessments (Lane and Stone, 2006; Braun, 2019; Shavelson et al., 2019). However, these assertions must be empirically validated, and the measures should be subjected to psychometric analyses. Evidence of the reliability, validity, and interpretative challenges of performance assessment (PA) are extensively detailed in Davey et al. (2015).

We adopt the following definition of performance assessment:

A performance assessment (sometimes called a work sample when assessing job performance) ... is an activity or set of activities that requires test takers, either individually or in groups, to generate products or performances in response to a complex, most often real-world task. These products and performances provide observable evidence bearing on test takers’ knowledge, skills, and abilities—their competencies—in completing the assessment (Davey et al., 2015, p. 10).

A performance assessment typically includes an extended performance task and short constructed-response and selected-response (i.e., multiple-choice) tasks (for examples, see Zlatkin-Troitschanskaia and Shavelson, 2019). In this paper, we refer to both individual performance- and constructed-response tasks as performance tasks (PT) (For an example, see **Table 1** in section “iPAL Assessment Framework”).

## AN APPROACH TO PERFORMANCE ASSESSMENT OF CRITICAL THINKING: THE IPAL PROGRAM

The approach to CT presented here is the result of ongoing work undertaken by the International Performance Assessment of Learning collaborative (iPAL<sup>1</sup>). iPAL is an international consortium of volunteers, primarily from academia, who have come together to address the dearth in higher education of research and practice in measuring CT with performance tasks (Shavelson et al., 2018). In this section, we present iPAL’s assessment framework as the basis of measuring CT, with examples along the way.

### iPAL Background

The iPAL assessment framework builds on the Council of Aid to Education’s Collegiate Learning Assessment (CLA). The CLA was designed to measure cross-disciplinary, generic competencies, such as CT, analytic reasoning, problem solving, and written communication (Klein et al., 2007; Shavelson, 2010). Ideally, each PA contained an extended PT (e.g., examining a range of evidential materials related to the crash of an aircraft) and two short PT’s: one in which students either critique an argument or provide a solution in response to a real-world societal issue.

Motivated by considerations of adequate reliability, in 2012, the CLA was later modified to create the CLA+. The CLA+ includes two subtests: a PT and a 25-item Selected Response

<sup>1</sup><https://www.ipal-rd.com/>

**TABLE 1** | The iPAL assessment framework.

Aspect	Description	Refugee crisis exemplar
Storyline	The storyline describes a curated version of a real-world situation.	With regional economic, health, crime and political challenges, there is an increasing demand for migrant entry into the country of Dorado in Central America. The question of whether it is safe to increase immigration (and add to the number of Reception Centers) has come before the country's Homeland Commission. A related question is whether Reception Centers have become local "hotspots" for crime.
Challenge	The challenge frames the tasks the respondent must carry out based on the dilemma or problem (potentially including moral or ethical aspects) presented in the <i>storyline</i> . The challenge should be sufficiently complex so that its resolution requires the respondent: (i) To apply multiple aspects of reasoning and judgment, and (ii) To consider the trade-offs that occur when adopting one potential solution over another – or deciding among competing principles.	(1a) Enumerate the pros and cons, if any, for accepting more refugees. (1b) Identify the documents and evidence in them to justify the list of pros and cons. (2a) Elaborate and recommend a concrete course of action: stem the flow of refugees at the border, control the flow of refugees (perhaps admitting certain types only like doctors and scientists), or take in a quota decided upon by the inter-governmental agreements. (2b) Identify the documents and evidence in them that lead to the recommendation. (3) Provide a set of recommendations on how the country can address challenges of the poor conditions to which refugees are now exposed, as well as dealing with crime rates in or near Reception Centers. (4) Suggest what additional information, if any, you would like to have to increase your confidence in the recommendation.
Documents	The storyline is augmented by a portfolio of documents in a range of formats (e.g., government reports, newspaper articles, web blogs, YouTube videos). Documents are collected or developed purposively to represent different sources of information and multiple perspectives. They vary with respect to the trustworthiness of the information; the relevance of the information; and the extent to which the information provided provokes the respondent to make judgmental errors or show bias.	(1) A letter from the Director of the Valparaiso Metropolitan Reception Center titled "Need for Reception Centers in Crisis Situations." (2) Three tables displaying crime statistics and demographic data provided by the Doradian Bureau of Statistics, presented separately for El Doradians and "foreigners." (3) An interview regarding the integration of migrants with a professor who is an expert on migration. (4) A newspaper article titled "Crimes committed by foreigners are on the rise." (5) An excerpt from a 2016 government report titled "Immigration and security: current status and future predictions." (6) Excerpt from the United Nation's "Universal Declaration of Human Rights." (7) Extract from an OECD Migration Report.
Scoring rubric	The scoring rubric comprises six dimensions. The first three dimensions involve comparing, evaluating, and justifying the characteristics of the information provided in the document collection regarding: (1) <b>Trustworthiness</b> of the information—dealing primarily with the information source, its context, its (hidden) motivation, and its potential conflicts with other evidence. (2) <b>Relevance</b> of the information as it pertains to the problem in the storyline. (3) <b>Bias</b> in information due to susceptibility to bias or proneness to use faulty heuristics in judgment and decision-making. The last three dimensions pertain to tacit and explicit response processes: (4) Analysis of different <b>perspectives</b> at play, addresses questions about the source of (hidden) motivation, control, expertise, and legitimacy (Mejia et al., 2019). (5) Demonstrating an <b>openness to the consequences</b> of prioritizing certain perspectives in the source provided—including any course of action suggested by the materials. (6) Formulating and communicating a <b>coherent argument</b> for the position taken, drawing from the five dimensions above.	<b>Refugee Crisis: Trustworthiness, Relevance, Bias, and Ethical Considerations in Documents</b> <b>Document 1:</b> A letter from the Director of the private reception center (both relevant and irrelevant, <i>baseline heuristic</i> ). <b>Document 2:</b> Doradian Bureau of Statistics – Crime statistics (relevant, <i>representative and baseline heuristics</i> ). <b>Document 3:</b> An interview with a professor who is an expert on immigration (relevant/focuses on the key factors influencing on the success of integration). <b>Document 4:</b> Newspaper story (irrelevant, <i>biased/fake news</i> ). <b>Document 5:</b> Government report (relevant). <b>Document 6:</b> The United Nations, The Universal Declaration of Human Rights (relevant). <b>Document 7:</b> A graph/table from an OECD report with data bearing on increase in refugees and non-refugees and crime (irrelevant, <i>biased</i> ).

Question (SRQ) section. The PT presents a document or problem statement and an assignment based on that document which elicits an open-ended response. The CLA+ added the SRQ section (which is not linked substantively to the PT scenario) to increase the number of student responses to obtain more reliable estimates of performance at the student-level than could be achieved with a single PT (Zahner, 2013; Davey et al., 2015).

## iPAL Assessment Framework Methodological Foundations

The iPAL framework evolved from the Collegiate Learning Assessment developed by Klein et al. (2007). It was also informed by the results from the AHELO pilot study (Organisation for Economic Co-operation and Development [OECD], 2012, 2013), as well as the KoKoHs research program in Germany

(for an overview see, Zlatkin-Troitschanskaia et al., 2017, 2020). The ongoing refinement of the iPAL framework has been guided in part by the principles of Evidence Centered Design (ECD) (Mislevy et al., 2003; Mislevy and Haertel, 2006; Haertel and Fujii, 2017).

In educational measurement, an assessment framework plays a critical intermediary role between the theoretical formulation of the construct and the development of the assessment instrument containing tasks (or items) intended to elicit evidence with respect to that construct (Mislevy et al., 2003). Builders of the assessment framework draw on the construct theory and operationalize it in a way that provides explicit guidance to PT's developers. Thus, the framework should reflect the relevant facets of the construct, where relevance is determined by substantive theory or an appropriate alternative such as behavioral samples from real-world situations of interest (criterion-sampling; McClelland, 1973), as well as the intended use(s) (for an example, see Shavelson et al., 2019). By following the requirements and guidelines embodied in the framework, instrument developers strengthen the claim of construct validity for the instrument (Messick, 1994).

An assessment framework can be specified at different levels of granularity: an assessment battery ("omnibus" assessment, for an example see below), a single performance task, or a specific component of an assessment (Shavelson, 2010; Davey et al., 2015). In the iPAL program, a performance assessment comprises one or more extended performance tasks and additional selected-response and short constructed-response items. The focus of the framework specified below is on a single PT intended to elicit evidence with respect to some facets of CT, such as the evaluation of the trustworthiness of the documents provided and the capacity to address conflicts of principles.

From the ECD perspective, an assessment is an instrument for generating information to support an evidentiary argument and, therefore, the intended inferences (claims) must guide each stage of the design process. The construct of interest is operationalized through the *Student Model*, which represents the target knowledge, skills, and abilities, as well as the relationships among them. The student model should also make explicit the assumptions regarding student competencies in foundational skills or content knowledge. The *Task Model* specifies the features of the problems or items posed to the respondent, with the goal of eliciting the evidence desired. The assessment framework also describes the collection of *task models* comprising the instrument, with considerations of construct validity, various psychometric characteristics (e.g., reliability) and practical constraints (e.g., testing time and cost). The student model provides grounds for evidence of validity, especially cognitive validity; namely, that the students are thinking critically in responding to the task(s).

In the present context, the target construct (CT) is the competence of individuals to think critically, which entails solving complex, real-world problems, and clearly communicating their conclusions or recommendations for action based on trustworthy, relevant and unbiased information. The situations, drawn from actual events, are challenging and may arise in many possible settings. In contrast to more

reductionist approaches to assessment development, the iPAL approach and framework rests on the assumption that properly addressing these situational demands requires the application of a constellation of CT skills appropriate to the particular task presented (e.g., Shavelson, 2010, 2013). For a PT, the assessment framework must also specify the rubric by which the responses will be evaluated. The rubric must be properly linked to the target construct so that the resulting score profile constitutes evidence that is both relevant and interpretable in terms of the student model (for an example, see Zlatkin-Troitschanskaia et al., 2019).

## iPAL Task Framework

The iPAL 'omnibus' framework comprises four main aspects: A *storyline*, a *challenge*, a *document library*, and a *scoring rubric*. **Table 1** displays these aspects, brief descriptions of each, and the corresponding examples drawn from an iPAL performance assessment (Version adapted from original in Hyytinen and Toom, 2019). *Storylines* are drawn from various domains; for example, the worlds of business, public policy, civics, medicine, and family. They often involve moral and/or ethical considerations. Deriving an appropriate storyline from a real-world situation requires careful consideration of which features are to be kept *in toto*, which adapted for purposes of the assessment, and which to be discarded. Framing the *challenge* demands care in wording so that there is minimal ambiguity in what is required of the respondent. The difficulty of the *challenge* depends, in large part, on the nature and extent of the information provided in the *document library*, the amount of scaffolding included, as well as the scope of the required response. The amount of information and the scope of the challenge should be commensurate with the amount of time available. As is evident from the table, the characteristics of the documents in the library are intended to elicit responses related to facets of CT. For example, with regard to bias, the information provided is intended to play to judgmental errors due to fast thinking and/or motivational reasoning. Ideally, the situation should accommodate multiple solutions of varying degrees of merit.

The dimensions of the *scoring rubric* are derived from the *Task Model* and *Student Model* (Mislevy et al., 2003) and signal which features are to be extracted from the response and indicate how they are to be evaluated. There should be a direct link between the evaluation of the evidence and the claims that are made with respect to the key features of the *task model* and *student model*. More specifically, the *task model* specifies the various manipulations embodied in the PA and so informs scoring, while the *student model* specifies the capacities students employ in more or less effectively responding to the tasks. The score scales for each of the five facets of CT (see section "Concept and Definition of Critical Thinking") can be specified using appropriate behavioral anchors (for examples, see Zlatkin-Troitschanskaia and Shavelson, 2019). Of particular importance is the evaluation of the response with respect to the last dimension of the scoring rubric; namely, the overall coherence and persuasiveness of the argument, building on the explicit or implicit characteristics related to the first five dimensions. The scoring process must be monitored carefully to



ensure that (trained) raters are judging each response based on the same types of features and evaluation criteria (Braun, 2019) as indicated by interrater agreement coefficients.

The scoring rubric of the iPAL omnibus framework can be modified for specific tasks (Lane and Stone, 2006). This generic rubric helps ensure consistency across rubrics for different storylines. For example, Zlatkin-Troitschanskaia et al. (2019, p. 473) used the following scoring scheme:

Based on our construct definition of CT and its four dimensions: (D1-Info) recognizing and evaluating information, (D2-Decision) recognizing and evaluating arguments and making decisions, (D3-Conseq) recognizing and evaluating the consequences of decisions, and (D4-Writing), we developed a corresponding analytic dimensional scoring ... The students' performance is evaluated along the four dimensions, which in turn are subdivided into a total of 23 indicators as (sub)categories of CT ... For each dimension, we sought detailed evidence in students' responses for the indicators and scored them on a six-point Likert-type scale. In order to reduce judgment distortions, an elaborate procedure of 'behaviorally anchored rating scales' (Smith and Kendall, 1963) was applied by assigning concrete behavioral expectations to certain scale points (Bernardin et al., 1976). To this end, we defined the scale levels by short descriptions of typical behavior and anchored them with concrete examples. ... We trained four raters in 1 day using a specially developed training course to evaluate students' performance along the 23 indicators clustered into four dimensions (for a description of the rater training, see Klotzer, 2018).

Shavelson et al. (2019) examined the interrater agreement of the scoring scheme developed by Zlatkin-Troitschanskaia et al. (2019) and "found that with 23 items and 2 raters the generalizability ("reliability") coefficient for total scores to be 0.74 (with 4 raters, 0.84)" (Shavelson et al., 2019, p. 15). In the study by Zlatkin-Troitschanskaia et al. (2019, p. 478) three score profiles were identified (low-, middle-, and high-performer) for students. Proper interpretation of such profiles requires care. For example, there may be multiple possible explanations for low scores such as poor CT skills, a lack of a disposition to engage with the challenge, or the two attributes jointly. These alternative explanations for student performance can potentially pose a threat to the evidentiary argument. In this case, auxiliary information may be available to aid in resolving the ambiguity. For example, student responses to selected- and short-constructed-response items in the PA can provide relevant information about the levels of the different skills possessed by the student. When sufficient data are available, the scores can be modeled statistically and/or qualitatively in such a way as to bring them to bear on the technical quality or interpretability of the claims of the assessment: reliability, validity, and utility evidence (Davey et al., 2015; Zlatkin-Troitschanskaia et al., 2019). These kinds of concerns are less critical when PT's are used in classroom settings. The instructor can draw on other sources of evidence, including direct discussion with the student.

## Use of iPAL Performance Assessments in Educational Practice: Evidence From Preliminary Validation Studies

The assessment framework described here supports the development of a PT in a general setting. Many modifications are possible and, indeed, desirable. If the PT is to be more deeply embedded in a certain discipline (e.g., economics, law, or medicine), for example, then the framework must specify characteristics of the narrative and the complementary documents as to the breadth and depth of disciplinary knowledge that is represented.

At present, preliminary field trials employing the omnibus framework (i.e., a full set of documents) indicated that 60 min was generally an inadequate amount of time for students to engage with the full set of complementary documents and to craft a complete response to the challenge (for an example, see Shavelson et al., 2019). Accordingly, it would be helpful to develop modified frameworks for PT's that require substantially less time. For an example, see a short performance assessment of civic online reasoning, requiring response times from 10 to 50 min (Wineburg et al., 2016). Such assessment frameworks could be derived from the omnibus framework by focusing on a reduced number of facets of CT, and specifying the characteristics of the complementary documents to be included – or, perhaps, choices among sets of documents. In principle, one could build a 'family' of PT's, each using the same (or nearly the same) storyline and a subset of the full collection of complementary documents.

Paul and Elder (2007) argue that the goal of CT assessments should be to provide faculty with important information about how well their instruction supports the development of students' CT. In that spirit, the full family of PT's could represent all facets of the construct while affording instructors and students more specific insights on strengths and weaknesses with respect to particular facets of CT. Moreover, the framework should be expanded to include the design of a set of short answer and/or multiple choice items to accompany the PT. Ideally, these additional items would be based on the same narrative as the PT to collect more nuanced information on students' precursor skills such as reading comprehension, while enhancing the overall reliability of the assessment. Areas where students are under-prepared could be addressed before, or even in parallel with the development of the focal CT skills. The parallel approach follows the co-requisite model of developmental education. In other settings (e.g., for summative assessment), these complementary items would be administered after the PT to augment the evidence in relation to the various claims. The full PT taking 90 min or more could serve as a capstone assessment.

As we transition from simply delivering paper-based assessments by computer to taking full advantage of the affordances of a digital platform, we should learn from the hard-won lessons of the past so that we can make swifter progress with fewer missteps. In that regard, we must take validity as the touchstone – assessment design, development and deployment must all be tightly linked to the operational definition of the CT construct. Considerations of reliability and practicality come into play with various use cases that highlight different purposes for the assessment (for future perspectives, see next section).



The iPAL assessment framework represents a feasible compromise between commercial, standardized assessments of CT (e.g., Liu et al., 2014), on the one hand, and, on the other, freedom for individual faculty to develop assessment tasks according to idiosyncratic models. It imposes a degree of standardization on *both* task development and scoring, while still allowing some flexibility for faculty to tailor the assessment to meet their unique needs. In so doing, it addresses a key weakness of the AAC&U's VALUE initiative<sup>2</sup> (retrieved 5/7/2020) that has achieved wide acceptance among United States colleges.

The VALUE initiative has produced generic scoring rubrics for 15 domains including CT, problem-solving and written communication. A rubric for a particular skill domain (e.g., critical thinking) has five to six dimensions with four ordered performance levels for each dimension (1 = lowest, 4 = highest). The performance levels are accompanied by language that is intended to clearly differentiate among levels.<sup>3</sup> Faculty are asked to submit student work products from a senior level course that is intended to yield evidence with respect to student learning outcomes in a particular domain and that, they believe, can elicit performances at the highest level. The collection of work products is then graded by faculty from other institutions who have been trained to apply the rubrics.

A principal difficulty is that there is neither a common framework to guide the design of the challenge, nor any control on task complexity and difficulty. Consequently, there is substantial heterogeneity in the quality and evidential value of the submitted responses. This also causes difficulties with task scoring and inter-rater reliability. Shavelson et al. (2009) discuss some of the problems arising with non-standardized collections of student work.

In this context, one advantage of the iPAL framework is that it can provide valuable guidance and an explicit structure for faculty in developing performance tasks for both instruction and formative assessment. When faculty design assessments, their focus is typically on content coverage rather than other potentially important characteristics, such as the degree of construct representation and the adequacy of their scoring procedures (Braun, 2019).

## CONCLUDING REFLECTIONS

### Challenges to Interpretation and Implementation

Performance tasks such as those generated by iPAL are attractive instruments for assessing CT skills (e.g., Shavelson, 2010; Shavelson et al., 2019). The attraction mainly rests on the assumption that elaborated PT's are more authentic (direct) and more completely capture facets of the target construct (i.e., possess greater construct representation) than the widely used selected-response tests. However, as Messick (1994) noted

authenticity is a "promissory note" that must be redeemed with empirical research. In practice, there are trade-offs among authenticity, construct validity, and psychometric quality such as reliability (Davey et al., 2015).

One reason for Messick (1994) caution is that authenticity does not guarantee construct validity. The latter must be established by drawing on multiple sources of evidence (American Educational Research Association et al., 2014). Following the ECD principles in designing and developing the PT, as well as the associated scoring rubrics, constitutes an important type of evidence. Further, as Leighton (2019) argues, response process data ("cognitive validity") is needed to validate claims regarding the cognitive complexity of PT's. Relevant data can be obtained through cognitive laboratory studies involving methods such as think aloud protocols or eye-tracking. Although time-consuming and expensive, such studies can yield not only evidence of validity, but also valuable information to guide refinements of the PT.

Going forward, iPAL PT's must be subjected to validation studies as recommended in the *Standards for Psychological and Educational Testing* by American Educational Research Association et al. (2014). With a particular focus on the criterion "relationships to other variables," a framework should include assumptions about the theoretically expected relationships among the indicators assessed by the PT, as well as the indicators' relationships to external variables such as intelligence or prior (task-relevant) knowledge.

Complementing the necessity of evaluating construct validity, there is the need to consider potential sources of construct-irrelevant variance (CIV). One pertains to student motivation, which is typically greater when the stakes are higher. If students are not motivated, then their performance is likely to be impacted by factors unrelated to their (construct-relevant) ability (Lane and Stone, 2006; Braun et al., 2011; Shavelson, 2013). Differential motivation across groups can also bias comparisons. Student motivation might be enhanced if the PT is administered in the context of a course with the promise of generating useful feedback on students' skill profiles.

Construct-irrelevant variance can also occur when students are not equally prepared for the format of the PT or fully appreciate the response requirements. This source of CIV could be alleviated by providing students with practice PT's. Finally, the use of novel forms of documentation, such as those from the Internet, can potentially introduce CIV due to differential familiarity with forms of representation or contents. Interestingly, this suggests that there may be a conflict between enhancing construct representation and reducing CIV.

Another potential source of CIV is related to response evaluation. Even with training, human raters can vary in accuracy and usage of the full score range. In addition, raters may attend to features of responses that are unrelated to the target construct, such as the length of the students' responses or the frequency of grammatical errors (Lane and Stone, 2006). Some of these sources of variance could be addressed in an online environment, where word processing software could alert students to potential grammatical and spelling errors before they submit their final work product.

<sup>2</sup><https://www.aacu.org/value>

<sup>3</sup>When test results are reported by means of substantively defined categories, the scoring is termed "criterion-referenced". This is, in contrast to results, reported as percentiles; such scoring is termed "norm-referenced".

Performance tasks generally take longer to administer and are more costly than traditional assessments, making it more difficult to reliably measure student performance (Messick, 1994; Davey et al., 2015). Indeed, it is well known that more than one performance task is needed to obtain high reliability (Shavelson, 2013). This is due to both student-task interactions and variability in scoring. Sources of student-task interactions are differential familiarity with the topic (Hyytinen and Toom, 2019) and differential motivation to engage with the task. The level of reliability required, however, depends on the context of use. For use in formative assessment as part of an instructional program, reliability can be lower than use for summative purposes. In the former case, other types of evidence are generally available to support interpretation and guide pedagogical decisions. Further studies are needed to obtain estimates of reliability in typical instructional settings.

With sufficient data, more sophisticated psychometric analyses become possible. One challenge is that the assumption of unidimensionality required for many psychometric models might be untenable for performance tasks (Davey et al., 2015). Davey et al. (2015) provide the example of a mathematics assessment that requires students to demonstrate not only their mathematics skills but also their written communication skills. Although the iPAL framework does not explicitly address students' reading comprehension and organization skills, students will likely need to call on these abilities to accomplish the task. Moreover, as the operational definition of CT makes evident, the student must not only deploy several skills in responding to the challenge of the PT, but also carry out component tasks in sequence. The former requirement strongly indicates the need for a multi-dimensional IRT model, while the latter suggests that the usual assumption of local item independence may well be problematic (Lane and Stone, 2006). At the same time, the analytic scoring rubric should facilitate the use of latent class analysis to partition data from large groups into meaningful categories (Zlatkin-Troitschanskaia et al., 2019).

## Future Perspectives

Although the iPAL consortium has made substantial progress in the assessment of CT, much remains to be done. Further refinement of existing PT's and their adaptation to different languages and cultures must continue. To this point, there are a number of examples: The refugee crisis PT (cited in **Table 1**) was translated and adapted from Finnish to US English and then to Colombian Spanish. A PT concerning kidney transplants was translated and adapted from German to US English. Finally, two PT's based on 'legacy admissions' to US colleges were translated and adapted to Colombian Spanish.

With respect to data collection, there is a need for sufficient data to support psychometric analysis of student responses, especially the relationships among the different components of the scoring rubric, as this would inform both task development and response evaluation (Zlatkin-Troitschanskaia et al., 2019). In addition, more intensive study of response processes through cognitive laboratories and the like are needed to strengthen the

evidential argument for construct validity (Leighton, 2019). We are currently conducting empirical studies, collecting data on both iPAL PT's and other measures of CT. These studies will provide evidence of convergent and discriminant validity.

At the same time, efforts should be directed at further development to support different ways CT PT's might be used—i.e., use cases—especially those that call for formative use of PT's. Incorporating formative assessment into courses can plausibly be expected to improve students' competency acquisition (Zlatkin-Troitschanskaia et al., 2017). With suitable choices of storylines, appropriate combinations of (modified) PT's, supplemented by short-answer and multiple-choice items, could be interwoven into ordinary classroom activities. The supplementary items may be completely separate from the PT's (as is the case with the CLA+), loosely coupled with the PT's (as in drawing on the same storyline), or tightly linked to the PT's (as in requiring elaboration of certain components of the response to the PT).

As an alternative to such integration, stand-alone modules could be embedded in courses to yield evidence of students' generic CT skills. Core curriculum courses or general education courses offer ideal settings for embedding performance assessments. If these assessments were administered to a representative sample of students in each cohort over their years in college, the results would yield important information on the development of CT skills at a population level. For another example, these PA's could be used to assess the competence profiles of students entering Bachelor's or graduate-level programs as a basis for more targeted instructional support.

Thus, in considering different use cases for the assessment of CT, it is evident that several modifications of the iPAL omnibus assessment framework are needed. As noted earlier, assessments built according to this framework are demanding with respect to the extensive preliminary work required by a task and the time required to properly complete it. Thus, it would be helpful to have modified versions of the framework, focusing on one or two facets of the CT construct and calling for a smaller number of supplementary documents. The challenge to the student should be suitably reduced.

Some members of the iPAL collaborative have developed PT's that are embedded in disciplines such as engineering, law and education (Crump et al., 2019; for teacher education examples, see Jeschke et al., 2019). These are proving to be of great interest to various stakeholders and further development is likely. Consequently, it is essential that an appropriate assessment framework be established and implemented. It is both a conceptual and an empirical question as to whether a single framework can guide development in different domains.

## Performance Assessment in Online Learning Environment

Over the last 15 years, increasing amounts of time in both college and work are spent using computers and other electronic devices. This has led to formulation of models for the *new literacies* that attempt to capture some key characteristics of these activities. A prominent example is a model proposed by Leu et al. (2020). The model frames online reading as a process of

problem-based inquiry that calls on five practices to occur during online research and comprehension:

1. Reading to identify important questions,
2. Reading to locate information,
3. Reading to critically evaluate information,
4. Reading to synthesize online information, and
5. Reading and writing to communicate online information.

The parallels with the iPAL definition of CT are evident and suggest there may be benefits to closer links between these two lines of research. For example, a report by Leu et al. (2014) describes empirical studies comparing assessments of online reading using either open-ended or multiple-choice response formats.

The iPAL consortium has begun to take advantage of the affordances of the online environment (for examples, see Schmidt et al. and Nagel et al. in this special issue). Most obviously, Supplementary Materials can now include archival photographs, audio recordings, or videos. Additional tasks might include the online search for relevant documents, though this would add considerably to the time demands. This online search could occur within a simulated Internet environment, as is the case for the IEA's ePIRLS assessment (Mullis et al., 2017).

The prospect of having access to a wealth of materials that can add to task authenticity is exciting. Yet it can also add ambiguity and information overload. Increased authenticity, then, should be weighed against validity concerns and the time required to absorb the content in these materials. Modifications of the design framework and extensive empirical testing will be required to decide on appropriate trade-offs. A related possibility is to employ some of these materials in short-answer (or even selected-response) items that supplement the main PT. Response formats could include highlighting text or using a drag-and-drop menu to construct a response. Students' responses could be automatically scored, thereby containing costs. With

automated scoring, feedback to students and faculty, including suggestions for next steps in strengthening CT skills, could also be provided without adding to faculty workload. Therefore, taking advantage of the online environment to incorporate new types of supplementary documents should be a high priority and, perhaps, to introduce new response formats as well. Finally, further investigation of the overlap between this formulation of CT and the characterization of online reading promulgated by Leu et al. (2020) is a promising direction to pursue.

## DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

## AUTHOR CONTRIBUTIONS

HB wrote the article. RS, OZ-T, and KB were involved in the preparation and revision of the article and co-wrote the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was funded in part by the Spencer Foundation (Grant No. #201700123).

## ACKNOWLEDGMENTS

We would like to thank all the researchers who have participated in the iPAL program.

## REFERENCES

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, D.C: American Educational Research Association.
- Arum, R., and Roksa, J. (2011). *Academically Adrift: Limited Learning on College Campuses*. Chicago, IL: University of Chicago Press.
- Association of American Colleges and Universities (n.d.). *VALUE: What is value?*. Available online at: <https://www.aacu.org/value> (accessed May 7, 2020).
- Association of American Colleges and Universities [AACU] (2018). *Fulfilling the American Dream: Liberal Education and the Future of Work*. Available online at: <https://www.aacu.org/research/2018-future-of-work> (accessed May 1, 2020).
- Braun, H. (2019). Performance assessment and standardization in higher education: a problematic conjunction? *Br. J. Educ. Psychol.* 89, 429–440. doi: 10.1111/bjep.12274
- Braun, H. I., Kirsch, I., and Yamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th grade NAEP reading assessment. *Teach. Coll. Rec.* 113, 2309–2344.
- Crump, N., Sepulveda, C., Fajardo, A., and Aguilera, A. (2019). Systematization of performance tests in critical thinking: an interdisciplinary construction experience. *Rev. Estud. Educ.* 2, 17–47.
- Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., and Wise, L. (2015). *Psychometric Considerations for the Next Generation of Performance Assessment*. Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service.
- Erwin, T. D., and Sebrell, K. W. (2003). Assessment of critical thinking: ETS's tasks in critical thinking. *J. Gen. Educ.* 52, 50–70. doi: 10.1353/jge.2003.0019
- Haertel, G. D., and Fujii, R. (2017). "Evidence-centered design and postsecondary assessment," in *Handbook on Measurement, Assessment, and Evaluation in Higher Education*, 2nd Edn, eds C. Secolsky and D. B. Denison (Abingdon: Routledge), 313–339. doi: 10.4324/9781315709307-26
- Hyttinen, H., and Toom, A. (2019). Developing a performance assessment task in the Finnish higher education context: conceptual and empirical insights. *Br. J. Educ. Psychol.* 89, 551–563. doi: 10.1111/bjep.12283
- Hyttinen, H., Toom, A., and Shavelson, R. J. (2019). "Enhancing scientific thinking through the development of critical thinking in higher education," in *Redefining Scientific Thinking for Higher Education: Higher-Order Thinking, Evidence-Based Reasoning and Research Skills*, eds M. Murtonen and K. Ballou (London: Palgrave MacMillan).
- Indiana University (2019). *FSSE 2019 Frequencies: FSSE 2019 Aggregate*. Available online at: [http://fsse.indiana.edu/pdf/FSSE\\_IR\\_2019/summary\\_tables/FSSE19\\_Frequencies\\_\(FSSE\\_2019\).pdf](http://fsse.indiana.edu/pdf/FSSE_IR_2019/summary_tables/FSSE19_Frequencies_(FSSE_2019).pdf) (accessed May 1, 2020).
- Jeschke, C., Kuhn, C., Lindmeier, A., Zlatkin-Troitschanskaia, O., Saas, H., and Heinze, A. (2019). Performance assessment to investigate the domain specificity of instructional skills among pre-service and in-service teachers of mathematics and economics. *Br. J. Educ. Psychol.* 89, 538–550. doi: 10.1111/bjep.12277
- Kegan, R. (1994). *In Over Our Heads: The Mental Demands of Modern Life*. Cambridge, MA: Harvard University Press.

- Klein, S., Benjamin, R., Shavelson, R., and Bolus, R. (2007). The collegiate learning assessment: facts and fantasies. *Eval. Rev.* 31, 415–439. doi: 10.1177/0193841x07303318
- Kosslyn, S. M., and Nelson, B. (2017). *Building the Intentional University: Minerva and the Future of Higher Education*. Cambridge, MA: The MIT Press.
- Lane, S., and Stone, C. A. (2006). “Performance assessment,” in *Educational Measurement*, 4th Edn, ed. R. L. Brennan (Lanham, MA: Rowman & Littlefield Publishers), 387–432.
- Leighton, J. P. (2019). The risk–return trade-off: performance assessments and cognitive validation of inferences. *Br. J. Educ. Psychol.* 89, 441–455. doi: 10.1111/bjep.12271
- Leu, D. J., Kiili, C., Forzani, E., Zawilinski, L., McVerry, J. G., and O’Byrne, W. I. (2020). “The new literacies of online research and comprehension,” in *The Concise Encyclopedia of Applied Linguistics*, ed. C. A. Chapelle (Oxford: Wiley-Blackwell), 844–852.
- Leu, D. J., Kulikowich, J. M., Kennedy, C., and Maykel, C. (2014). “The ORCA Project: designing technology-based assessments for online research,” in *Paper Presented at the American Educational Research Annual Meeting*, Philadelphia, PA.
- Liu, O. L., Frankel, L., and Roehr, K. C. (2014). Assessing critical thinking in higher education: current state and directions for next-generation assessments. *ETS Res. Rep. Ser.* 1, 1–23. doi: 10.1002/ets2.12009
- McClelland, D. C. (1973). Testing for competence rather than for “intelligence.” *Am. Psychol.* 28, 1–14. doi: 10.1037/h0034092
- McGrew, S., Ortega, T., Breakstone, J., and Wineburg, S. (2017). The challenge that’s bigger than fake news: civic reasoning in a social media environment. *Am. Educ.* 4, 4–9, 39.
- Mejía, A., Mariño, J. P., and Molina, A. (2019). Incorporating perspective analysis into critical thinking performance assessments. *Br. J. Educ. Psychol.* 89, 456–467. doi: 10.1111/bjep.12297
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educ. Res.* 23, 13–23. doi: 10.3102/0013189x023002013
- Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Res. Rep. Ser.* 2003, i–29. doi: 10.1002/j.2333-8504.2003.tb01908.x
- Mislevy, R. J., and Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educ. Meas. Issues Pract.* 25, 6–20. doi: 10.1111/j.1745-3992.2006.00075.x
- Mullis, I. V. S., Martin, M. O., Foy, P., and Hooper, M. (2017). *ePIRLS 2016 International Results in Online Informational Reading*. Available online at: <http://timssandpirls.bc.edu/pirls2016/international-results/> (accessed May 1, 2020).
- Nagel, M.-T., Zlatkin-Troitschanskaia, O., Schmidt, S., and Beck, K. (2020). “Performance assessment of generic and domain-specific skills in higher education economics,” in *Student Learning in German Higher Education*, eds O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, and C. Lautenbach (Berlin: Springer), 281–299. doi: 10.1007/978-3-658-27886-1\_14
- Organisation for Economic Co-operation and Development [OECD] (2012). *AHELO: Feasibility Study Report*, Vol. 1. Paris: OECD. Design and implementation.
- Organisation for Economic Co-operation and Development [OECD] (2013). *AHELO: Feasibility Study Report*, Vol. 2. Paris: OECD. Data analysis and national experiences.
- Oser, F. K., and Biedermann, H. (2020). “A three-level model for critical thinking: critical alertness, critical reflection, and critical analysis,” in *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*, ed. O. Zlatkin-Troitschanskaia (Cham: Springer), 89–106. doi: 10.1007/978-3-030-26578-6\_7
- Paul, R., and Elder, L. (2007). Consequential validity: using assessment to drive instruction. *Found. Crit. Think.* 29, 31–40.
- Pellegrino, J. W., and Hilton, M. L. (eds) (2012). *Education for life and work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington DC: National Academies Press.
- Shavelson, R. (2010). *Measuring College Learning Responsibly: Accountability in a New Era*. Redwood City, CA: Stanford University Press.
- Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educ. Psychol.* 48, 73–86. doi: 10.1080/00461520.2013.779483
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. P. (2019). Assessment of university students’ critical thinking: next generation performance assessment. *Int. J. Test.* 19, 337–362. doi: 10.1080/15305058.2018.1543309
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., and Marino, J. P. (2018). “International performance assessment of learning in higher education (iPAL): research and development,” in *Assessment of Learning Outcomes in Higher Education: Cross-National Comparisons and Perspectives*, eds O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, and C. Kuhn (Berlin: Springer), 193–214. doi: 10.1007/978-3-319-74338-7\_10
- Shavelson, R. J., Klein, S., and Benjamin, R. (2009). The limitations of portfolios. *Inside Higher Educ.* Available online at: <https://www.insidehighered.com/views/2009/10/16/limitations-portfolios>
- Stolzenberg, E. B., Eagan, M. K., Zimmerman, H. B., Berdan Lozano, J., Cesar-Davis, N. M., Aragon, M. C., et al. (2019). *Undergraduate Teaching Faculty: The HERI Faculty Survey 2016–2017*. Los Angeles, CA: UCLA.
- Tessier-Lavigne, M. (2020). *Putting Ethics at the Heart of Innovation*. Stanford, CA: Stanford Magazine.
- Wheeler, P., and Haertel, G. D. (1993). *Resource Handbook on Performance Assessment and Measurement: A Tool for Students, Practitioners, and Policymakers*. Palm Coast, FL: Owl Press.
- Wineburg, S., McGrew, S., Breakstone, J., and Ortega, T. (2016). *Evaluating Information: The Cornerstone of Civic Online Reasoning. Executive Summary*. Stanford, CA: Stanford History Education Group.
- Zahner, D. (2013). *Reliability and Validity-CL+.* Council for Aid to Education. Available online at: <https://pdfs.semanticscholar.org/91ae/8edfac44bce3bed37d8c9091da01d6db3776.pdf>.
- Zlatkin-Troitschanskaia, O., and Shavelson, R. J. (2019). Performance assessment of student learning in higher education [Special issue]. *Br. J. Educ. Psychol.* 89, i–iv, 413–563.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Lautenbach, C., Molerov, D., Toepper, M., and Brückner, S. (2017). *Modeling and Measuring Competencies in Higher Education: Approaches to Challenges in Higher Education Policy and Practice*. Berlin: Springer VS.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Toepper, M., and Lautenbach, C. (eds) (2020). *Student Learning in German Higher Education: Innovative Measurement Approaches and Research Results*. Wiesbaden: Springer.
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., and Pant, H. A. (2018). “Assessment of learning outcomes in higher education: international comparisons and perspectives,” in *Handbook on Measurement, Assessment, and Evaluation in Higher Education*, 2nd Edn, eds C. Secolsky and D. B. Denison (Abingdon: Routledge), 686–697.
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., and Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *Br. J. Educ. Psychol.* 89, 468–484. doi: 10.1111/bjep.12286

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Braun, Shavelson, Zlatkin-Troitschanskaia and Borowiec. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# The Role of Students' Beliefs When Critically Reasoning From Multiple Contradictory Sources of Information in Performance Assessments

Olga Zlatkin-Troitschanskaia<sup>1\*</sup>, Klaus Beck<sup>1</sup>, Jennifer Fischer<sup>1</sup>, Dominik Braunheim<sup>1</sup>, Susanne Schmidt<sup>1</sup> and Richard J. Shavelson<sup>2</sup>

<sup>1</sup> Department of Business and Economics Education, Johannes Gutenberg University Mainz, Mainz, Germany, <sup>2</sup> Graduate School of Education, Stanford University, Palo Alto, CA, United States

## OPEN ACCESS

### Edited by:

Lawrence Jun Zhang,  
The University of Auckland,  
New Zealand

### Reviewed by:

Steve Oswald,  
Université de Fribourg, Switzerland  
Rainer Bromme,  
University of Münster, Germany

### \*Correspondence:

Olga Zlatkin-Troitschanskaia  
troitschanskaia@uni-mainz.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Psychology

**Received:** 26 May 2020

**Accepted:** 04 August 2020

**Published:** 11 September 2020

### Citation:

Zlatkin-Troitschanskaia O, Beck K,  
Fischer J, Braunheim D, Schmidt S  
and Shavelson RJ (2020) The Role  
of Students' Beliefs When Critically  
Reasoning From Multiple  
Contradictory Sources of Information  
in Performance Assessments.  
Front. Psychol. 11:2192.  
doi: 10.3389/fpsyg.2020.02192

Critical reasoning (CR) when confronted with contradictory information from multiple sources is a crucial ability in a knowledge-based society and digital world. Using information without critically reflecting on the content and its quality may lead to the acceptance of information based on unwarranted claims. Previous personal beliefs are assumed to play a decisive role when it comes to critically differentiating between assertions and claims and warranted knowledge and facts. The role of generic epistemic beliefs on critical stance and attitude in reflectively dealing with information is well researched. Relatively few studies however, have been conducted on the influence of *domain-specific beliefs*, i.e., *beliefs in relation to specific content encountered in a piece of information or task, on the reasoning process*, and on how these beliefs may affect decision-making processes. This study focuses on students' *task- and topic-related beliefs* that may influence their reasoning when dealing with multiple and partly contradictory sources of information. To validly assess CR among university students, we used a newly developed computer-based *performance assessment* in which the students were confronted with an authentic task which contains theoretically defined psychological stimuli for measuring CR. To investigate the particular role of task- and topic-related beliefs on CR, a purposeful sample of 30 university students took part in a performance assessment and then were interviewed immediately afterward. In the semi-structured cognitive interviews, the participants' task-related beliefs were assessed. Based on qualitative analyses of the interview transcripts, three distinct *profiles of decision-making* among students have been identified. More specifically, the different types of students' beliefs and attitudes derived from the cognitive interview data suggest their influence on information processing, reasoning approaches and decision-making. The results indicated that the students' beliefs had an influence on their selection, critical evaluation and use of information as well as on their reasoning processes and final decisions.

**Keywords:** critical reasoning, multiple source use, reasoning profiles, performance assessment, domain-specific beliefs, decision-making, cognitive interview protocols, criteria-driven online search



## RESEARCH BACKGROUND AND STUDY OBJECTIVES

Critical reasoning (CR) when confronted with contradictory information from multiple sources is a crucial ability in a knowledge-based society and digital world (Brooks, 2016; Newman and Beetham, 2017; Wineburg and McGrew, 2017). The Internet presents a flood of complex, potentially conflicting, and competing information on one and the same issue. To build a dependable and coherent knowledge base and to develop sophisticated (domain-specific and generic) attitudes and an analytical, reflective stance, students must be able to select and critically evaluate, analyze, synthesize, and integrate incoherent, fragmented, and biased information.

Students' mental CR strategies may likely be insufficient for what is demanded for understanding heterogeneous information and, what is more, for effective and productive participation in a complex information environment (for a meta-study, see Huber and Kuncel, 2016, for university students, see McGrew et al., 2018; Wineburg et al., 2018; Hahnel et al., 2019; Münchow et al., 2019; Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019b). As a coping strategy, they may choose to reduce complexity by various means, for instance, by using cognitive heuristics, preferring simplified forms of information presentation, or relying on sources without verification, which can be exploited for manipulation (Walther and Burkell, 2002; Metzger, 2007; Horstmann et al., 2009; Metzger et al., 2010).

In addition, certain information representations may be (sub)consciously preferred not for their informational but for their entertainment value, their elicitation of certain affects, or their engagement properties (Maurer et al., 2018, 2020). Based on students' previous media experience, knowledge, and expectations, they may have learned to assume that certain types of media representations are more trustworthy than others (McGrew et al., 2017). They may follow a heuristic that similar media representations offer similarly reliable evidence, without considering the communicative context, communicator's intentions, and possibilities of becoming a victim of manipulation. This is particularly the case when it comes to online information channels (Metzger et al., 2010; Ciampaglia, 2018).

As some current studies indicate, students who habitually avoid information that contradicts their beliefs may easily miss important content and fall prey to biased information [see Section "State of Research on Beliefs and Their Impact on (Online) Information Processing"]. Using information without critically reflecting on the content and its quality may lead to the acceptance of information based on unwarranted claims. Deficits in due critical evaluation arise most likely because of shallow processing or insufficient reasoning and may occur subconsciously (Stanovich, 2003, 2016).

Insufficient reasoning can be amplified when information on a topic is distorted or counterfactual and when students do not recognize biased or false information and use it to build knowledge. As a result, learners may neglect complex, academically warranted knowledge and rely more on lower-quality information that is consistent with their beliefs and biases and that is easier to comprehend (Hahnel et al., 2019;

Schoor et al., 2019). The internalization of this biased information may subsequently affect learning by acting to inhibit or distort more advanced information processing and knowledge acquisition (List and Alexander, 2017, 2018).

Theoretically, previous personal beliefs are assumed to play a very decisive role when it comes to critically differentiating between assertions and claims on the one hand and warranted knowledge and facts on the other hand. Rather, the role of generic epistemic beliefs on critical stance and attitude in reflectively dealing with information is well researched [see Section "State of Research on Beliefs and Their Impact on (Online) Information Processing"]. Relatively few studies have been conducted on the influence of *domain-specific beliefs*, i.e., *beliefs in relation to specific content encountered in a piece of information or task*, on the reasoning process. Beliefs of this kind are usually measured using psychological scales, but without insight into the reasoning process and how these beliefs may affect the information-processing and decision-making processes [for an overview of current research, see Section "State of Research on Beliefs and Their Impact on (Online) Information Processing"].

With our study, we aim to contribute to this research desideratum. This study focuses on students' *task- and topic-related beliefs* that may influence their reasoning when dealing with multiple and partly contradictory sources of information. To validly assess CR among university students, we used a newly developed computer-based performance assessment of learning in which the students are confronted with an authentic task which contains theoretically defined psychological stimuli for measuring CR (for details, see Section "Assessment Frameworks for Measuring Critical Reasoning") in accordance with our construct definition (see Section "Critically Reasoning from Multiple Sources of Information"; Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019b).

To investigate the particular role of task- and topic-related beliefs on CR, a purposeful sample of 30 university students from the overall sample of the overarching German iPAL study took part in a performance assessment and then were interviewed immediately afterward (for details, see Sections "A Study on Performance Assessments of Higher Education Students' CR" and "Materials and Methods"). In the semi-structured cognitive interviews, the task-related beliefs of the participants were elicited and then qualitatively analyzed (see Section "Cognitive Interviews and Qualitative Analyses"). The cognitive interview transcripts were examined in order to address the two overarching questions (i) *how do students' beliefs influence their selection, evaluation and use of information as well as their subsequent reasoning and decision-making?* and (ii) *how do students' beliefs change as they progress through the task and encounter multiple new information sources along the way*. Based on qualitative analyses (Strauss and Corbin, 1998; for details, see "Materials and Methods"), different profiles of participants have been identified, which can be distinguished by different personal characteristics such as the level of prior knowledge.

In this paper, we present our theoretical and empirical analyses to address these two questions (see Section "Results"). The study results – despite the necessary limitations (see Section "Limitations and Implication for Future Research") – lead to a valuable specification of theoretical assumptions for further

empirical research in this highly relevant but under-researched field (see Section “Summary and Interpretation of Results”).

## STATE OF RESEARCH ON BELIEFS AND THEIR IMPACT ON (ONLINE) INFORMATION PROCESSING

For a systematic analysis of the state of research, we conducted a criteria-driven online search. Based on expert interviews, we determined a set of keywords and conducted the search on the ERIC database and Google Scholar for the period 2009–2020. The stepwise search using keywords related to online information processing and critical reasoning among university students resulted in 56 eligible studies. The review of the abstracts showed that students' beliefs were assessed and analyzed in 15 studies. The essential results of these studies are summarized in the following overview (see **Table 1**).

About half of these 15 studies focus explicitly on the relation between beliefs and (online) information processing (see **Table 1**), while critical reasoning was only implicitly addressed. Despite this narrow research focus, all studies indicate a clear connection between *epistemic beliefs* and the approach to (online) information processing, especially regarding *judgment* of information sources and their content. Well-developed and more advanced epistemic beliefs positively influenced the quality of students' information processing.

Ulyshen et al. (2015) provided one of the few studies specifically investigating the relation between *general epistemic beliefs* and *Internet search behavior*. Using participants' think-aloud protocols they investigated the impact of students' task-related epistemic beliefs on their information processing. The results indicate a positive impact of well-developed epistemic beliefs on evaluating the quality and credibility of information.

Chiu et al. (2013) used a questionnaire to investigate the relation between university students' *Internet-specific epistemic beliefs* and Internet search behavior. The authors identified four dimensions of beliefs: *certainty*, *simplicity*, *source*, and *justification of Internet-based knowledge*. The results indicate a positive association between Internet searching and *justification*, and a negative association with *simplicity* and *source*. In a follow-up study, Chiu et al. (2015) examined gender differences in students' Internet-specific epistemic beliefs, indicating a gender gap in *certainty* and *simplicity*, and revealing more perceived *uncertainty* and *complexity* among females compared to males.

Mason et al. (2010) specifically focused on students' *Internet search* when working on *academic tasks dealing with a controversial topic*, and in relation to *epistemic metacognition*, which was defined as students' ability to spontaneously reflect on the accessed information. In addition, they examined the relationship between personal characteristics and prior topic-related knowledge. Test participants were asked to *think aloud during their Internet search*. Qualitative and quantitative analyses revealed diverse epistemic metacognitions among all study participants, but to different extents and levels. No correlation between epistemic beliefs and prior knowledge was

identified. Overall, two patterns of epistemic metacognition were determined that significantly affected students' Internet search. Students who spontaneously generated more sophisticated reflections about the sources and the information provided outperformed students who were at a lower epistemic level.

In an experimental study with an intervention-control group design (the intervention aiming at improving *medicine-specific epistemic beliefs*), Kienhues et al. (2011) focused on the relationship between *processing conflicting versus consistent (medical) information on the Internet and topic-related and medicine-specific epistemic beliefs*. The intervention groups differed in both their topic-related and medicine-specific epistemic beliefs, and were more advanced compared to the control group.

van Strien et al. (2016) examined the influence of attitude strength on the processing and evaluation of sources of information on the Web. In an eye-tracking study, university students received information from pre-selected websites from different sources on a controversial topic. Participants who felt strongly about the topic did not consider websites with attitude-inconsistent content for as long and did not rate the credibility of this information as highly as students with less strongly established prior attitudes. The participants with strong prior attitudes also included more attitude-consistent information in an essay task than participants with weaker prior attitudes. Thus, differences in prior attitudes bias the evaluation and processing of information in different ways. Even though students were not fully biased during initial information processing, they were so when evaluating the information and presenting it in an essay task.

Similar biases were found in a study by Bråten et al. (2011), who examined how undergraduates judged the trustworthiness of information sources on a controversial topic. Students judged information differently depending on the *sources* (e.g., textbooks, official documents, newspapers). In addition, students with limited *topic-specific knowledge* were inclined to trust less trustworthy sources. Lucassen and Schraagen (2011) show similar results in terms of relation with *domain-specific knowledge and source expertise*.

Following the assumption that students spontaneously engage in *epistemic cognition* when processing conflicting scientific information, van Strien et al. (2012b) examined how this epistemic cognition is related to students' actual beliefs. In addition, the interplay of students' epistemic beliefs and prior attitudes when encountering conflicting and partly attitude-inconsistent information on a controversial socio-scientific topic was studied using think-aloud methods. The results indicated that students' difficulties in adequately evaluating diverse and conflicting information do not correlate with their *prior epistemic beliefs*. These beliefs might be developing from *naïve to sophisticated*, i.e., from absolutism to multiplism to evaluativism (which were measured using a test developed by van Strien et al., 2012a).

van Strien et al. (2014) investigated the effects of prior attitudes on how students deal with conflicting information in multiple texts, indicating that students with strong prior attitudes were significantly more likely to write essays that were biased toward

**TABLE 1 |** Overview of recent studies on beliefs and their impact on information processing.

Authors	Study	Focus of analysis	Study design	Sample
(1) Bråten et al., 2011	Trust and mistrust when students read multiple information sources about climate change	Evaluation of source information	Demographic information sheet, Topic knowledge questionnaire, texts and trustworthiness questionnaire	128 undergraduate students (80.2% female) from a university in southeast Norway
(2) Chiu et al., 2013	Internet-specific epistemic beliefs and self-regulated learning in online searches for academic information	Internet-specific epistemic beliefs	Exploratory factor analysis, confirmatory analysis and the hypothesized model	748 male and female university students in Taiwan
(3) Chiu et al., 2015	Testing measurement invariance and latent mean differences across gender groups in college students' Internet-specific epistemic beliefs.	Internet-specific epistemic beliefs	Internet-specific epistemic beliefs questionnaire	735 male and female university students in Taiwan
(4) Hsu et al., 2014	Epistemic beliefs, online search strategies, and behavioral patterns while exploring socioscientific issues	Scientific-epistemic beliefs (SEB)	SEB questionnaire, Online Information Searching Strategies Inventory, screen-capture videos, sequential analysis	42 undergraduate and graduate students in Taiwan
(5) Johnson et al., 2016	Students' approaches to the evaluation of digital information: Insights from their trust judgments	Trust judgements	55 5-point Likert-scale statements, questions about their disposition to trust and their health status	531 1st-year and 3rd year-students
(6) Kahne and Bowyer, 2017	Educating for democracy in a partisan age: confronting the challenges of motivated reasoning and misinformation	Digital media and motivated reasoning	Survey about students' online activity and political participation	2101 participants (17–25 years old)
(7) Kammerer and Gerjets, 2012	Effects of search interface and Internet-specific epistemic beliefs on source evaluations during Web search for medical information	Internet-specific epistemic beliefs	Web search, source evaluations, search interface design, eye-tracking	80 university students from different majors at a German university
(8) Kienhues et al., 2011	Dealing with conflicting or consistent medical information on the web	Epistemic beliefs	Pretest–posttest experimental design: an intervention group (website with conflicting contents), another intervention group (website with consistent contents) and a no-intervention group (control group, no web search)	100 mostly (84%) female students attending a German university
(9) Lucassen and Schraagen, 2011	Factual accuracy and trust in information: The role of expertise	Trust judgements; source, semantic, surface-model	Online questionnaire; novice–expert-design	657 participants (recruited in different online forums)
(10) Mason et al., 2010	Searching the Web to learn about a controversial topic: are students epistemically active?	Epistemic metacognition	Online information searching; think-aloud procedure during the search; two measures for verbal and visuospatial memory capacity; writing an essay	46 students from an university in northern Italy
(11) Mason et al., 2014	Reading information about a scientific phenomenon on webpages varying for reliability: an eye-movement analysis	Source evaluation; eye movements; Internet reading	Eye-tracking; prior knowledge questions, complete the Connotative Aspects of Epistemological Beliefs and then read the 4 webpages to get information for writing a report	49 female undergraduate students from the faculty of psychology, public university in north-eastern Italy
(12) van Strien et al., 2012b	Do prior attitudes influence epistemic cognition while reading conflicting information?	Epistemic cognition	Reading a number of pre-selected texts on climate change; thinking aloud; writing a short essay	98 students from a Dutch school for pre-university education; 25 students in the follow-up study

*(Continued)*

TABLE 1 | Continued

Authors	Study	Focus of Analysis	Study Design	Sample
(13) van Strien et al., 2014	Dealing with conflicting information from multiple nonlinear texts: Effects of prior attitudes	Prior attitudes; evaluating conflicting information	Reading and writing task (essays were scored on the perspective taken and the origin of information)	61 students in pre-university education in the Netherlands
(14) van Strien et al., 2016	How attitude strength biases information processing and evaluation on the Internet	Prior attitudes; attitude strength; source evaluation	Online questionnaire; eye-tracking; writing an essay; computer-based questionnaire	79 students (56 female) from a German university
(15) Ulyshen et al., 2015	Understanding the connection between epistemic beliefs and Internet searching	Epistemic beliefs	An ill-structured task using the Google search engine; the revised Cognitive Flexibility Inventory, prior content knowledge test, verbal comprehension test, complexity of learning strategies (think aloud procedure) & retrospective interview	53 undergraduate students from a Midwestern University

their prior attitudes. Moreover, students with strong attitudes took explicit stances and used large proportions of information not presented in the sources in their essays, while students with neutral attitudes wrote syntheses and used more information from the given documents.

To gain a deeper insight into the role of experience in the evaluation phase of the information search process and into the development of beliefs influencing the evaluation of information, Johnson et al. (2016) found significant differences between first-year vs. third-year undergraduates regarding the factors that influence their judgment of the trustworthiness of online information. The results indicate that the more advanced students were not only more sophisticated in evaluating information sources but also more aware in terms of making use of the evaluation criteria.

Likewise, Hsu et al. (2014) examined how students' different levels of development of their scientific-epistemic beliefs impact their online information searching strategies and behaviors. They divided undergraduates and graduates into two groups depending on whether they employed a naïve or sophisticated strategy. They measured students' self-perceived online searching strategies and video recorded their search behaviors. Students with higher-quality scientific epistemic beliefs showed more advanced online searching strategies and demonstrated a rather meta-cognitive search pattern.

Mason et al. (2014) studied whether *topic-specific prior knowledge and epistemological beliefs* influence visual behavior when reading verbally and graphically presented information on webpages. They found that readers made a presumably implicit evaluation of the sources they were confronted with. University students with more elaborated topic-specific epistemic beliefs spent more time on graphics in the context of more reliable sources and increased their knowledge of the topic.

The study of Kahne and Bowyer (2017) is of particular interest for our analysis, as they took policy positions into consideration, an aspect which plays an important role in the task scenario we administered to our test participants (see section

"Research Questions"). In their survey of young adults, which is representative for the U.S., they asked participants to judge the veracity of simulated online postings. Controlling for political knowledge and media literacy, their main finding was that the alignment of statements with prior policy beliefs is more decisive for the evaluation of information quality than their accuracy.

Summing up, from the findings reported in recent literature, we register several commonalities in respect to the relation between *beliefs and the evaluation of internet-based information*. First, information as such and especially information encountered on the Internet was generally recognized and processed on the basis of beliefs and attitudes. Initially, students were always inclined to consider information trustworthy that corresponds with their own (prior) knowledge, whereas they tended to neglect conflicting information. Other biasing factors were prior beliefs (attitudes), which were of comparatively greater impact on the ascription of quality of information in terms of credibility, reliability, plausibility, or trustworthiness. Students appear to be liable to believe and to use information sources in line with their previous convictions, i.e., to avoid "cognitive dissonances" (Festinger, 1962). In addition, the impact of these factors is moderated by their strength (i.e., attitude strength). All in all, well-developed and more advanced (domain-specific) prior knowledge and epistemic beliefs seem to positively influence the quality of students' Internet searches and (online) information processing.

## RESEARCH QUESTIONS

In the studies we referenced above, the question of whether (prior) beliefs and attitudes are personal traits or states and to what extent they may *change* remains open. We do not yet know whether (prior) beliefs and attitudes will change during the information acquisition process, and if so, under what circumstances. Our study aims to shed some light on the answers to these questions.



More specifically, based on the analyses of the current state of international research (see Section “State of Research on Beliefs and Their Impact on (Online) Information Processing”), we developed an analytical framework for our study as presented in **Figure 1**, and specify the following research questions (RQs):

(I) The Relationship between Beliefs and Decision-Making

RQ1: *Students' beliefs at the beginning of task processing*

- *Do the students indicate that they held certain beliefs before they began the performance assessment?*
- *Is it possible to identify distinct types based on these beliefs?*

RQ2: *The relationship between students' beliefs and their reasoning process as well as their final decision (written task response)*

- *At which point in time during task processing did the students make their decision?*
- *Do the students' beliefs affect their decision-making process?*
- *Is it possible to identify distinct profiles of decision-making?*
- *Which reasoning approaches become evident that may influence the decision-making of the participants?*

(II) Change of Beliefs While Solving the Task.

RQ3: *Interaction between students' beliefs and the processing of the given information (in the task)*

- *Do the students' beliefs change as they progress through the task and encounter multiple new information sources along the way (which could indicate that the processed information influences the students' beliefs)? If so, to what extent is this reflected in their final decision (written task response)?*

## CONCEPTUAL AND METHODOLOGICAL BACKGROUND

### Critically Reasoning From Multiple Sources of Information

Students' skills in judging (online) information are of central importance to avoid the acquisition of erroneous domain-specific and generic knowledge (Murray and Pérez, 2014; Brooks, 2016). The abilities involved in finding, accessing, selecting, critically evaluating, and applying information from the Internet and from various media are crucial to learning in a globalized digital information society (Pellegrino, 2017; List and Alexander, 2019). Students need *critical reasoning* (CR) skills to judge the quality of the information sources and content they access inside as well as outside of higher education (Harrison and Luckett, 2019). Students need CR to recognize easily available biased and counterfactual information, withstand manipulation attempts (Wineburg and McGrew, 2017; McGrew et al., 2018), and avoid generating erroneous domain-specific and generic knowledge or arguments.<sup>1</sup>

<sup>1</sup> In contrast to other concepts related to *critical thinking*, critical online reasoning (COR) is explicitly limited to the online information environment and includes the specific ability of “online information acquisition”. While there is currently no

In our study, we follow the definition of CR and its facets as described in Zlatkin-Troitschanskaia et al. (2019b). CR is defined as students' (I.) *identification, evaluation, and integration of data sources*; (II.) *recognition and use of evidence*; (III.) *reasoning based on evidence, and synthesis*; (IV) *(causal and moral) recognition of consequences of decision-making, which ultimately lead to* (V) *the use of appropriate communicative action*. The performance assessments used in this study to measure CR (see next section) are based on these five theoretically driven central facets of this definition of CR. Students' ability to critically reason from multiple sources of information as a specific representation of CR was measured within this assessment framework.

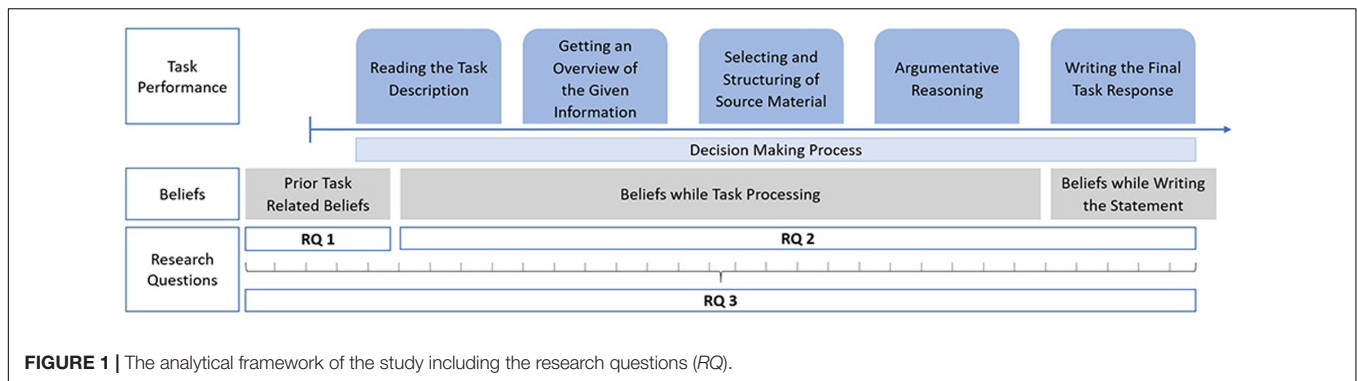
### Assessment Frameworks for Measuring Critical Reasoning

Valid measurement of CR skills is an important component of a program of research on how CR can be effectively promoted in higher education. Moreover, as part of a validity argument, CR's relation to other related constructs needs to be examined. Based on existing psychological learning models (Mislevy, 2018; Pellegrino, 2020), analyses of this kind can provide a significant contribution to developing appropriate explanatory approaches to CR. Despite the urgency of this topic for higher education (Harrison and Luckett, 2019), theoretically sophisticated CR learning and performance assessment tools have so far been developed by only a few projects internationally (for an overview, see Zlatkin-Troitschanskaia et al., 2018a).

Multidimensional and multifaceted (meta)cognitive higher-order (procedural) skills, such as CR, can be validly measured with closed-format tests to a limited extent, as selected-response items fall at the lower end of the ‘lifelike fidelity scale.’ Multiple-choice tests predominantly assess declarative and conceptual/factual knowledge (e.g., Braun, 2019). As Liu et al. (2014) and Oser and Biedermann (2020) documented, there are several closed-format tests for assessing CR (or related constructs). One main shortcoming of tests of this kind is their limited face validity, ecological validity, or content validity (Davey et al., 2015). They usually demonstrate (extremely) strong correlations with tests focused on general intellectual ability [e.g., intelligence tests or the Scholastic Aptitude Test (SAT)], but they tend to fail to measure more specific procedural skills regarding the use and the evaluation of information sources used for learning in higher education. Well-established CR assessments have been based on standard-setting research (Facione, 1990; Facione et al., 1998), but have used multiple-choice formats and brief situational contexts and have assessed generic minimal

unified definition of COR, there are numerous definitions of its related construct critical thinking (CT) that include and describe different dimensions or levels. For instance, Oser and Biedermann (2020) distinguish between CT as alertness, CT as immediate reflection, and CT as analysis. Facione (2004, p. 9) describes CT as “inference, explanation, interpretation, evaluation, analysis, self-regulation” (for further definitions of CT, see Moore, 2013; Beck, 2020). As Brookfield (1987) emphasizes, “Being a critical thinker involves more than cognitive activities such as logical reasoning or scrutinizing arguments for assertions unsupported by empirical evidence. Thinking critically involves us *recognizing the assumptions underlying our beliefs and behaviors*”.





**FIGURE 1 |** The analytical framework of the study including the research questions (RQ).

inferencing abilities.<sup>2</sup> Despite the broad use of this test type in educational assessment, it remains unclear to what extent these tasks are ecologically valid and whether students can transfer the measured abilities to more authentic and complex requirement situations.

At the other end of the assessment spectrum are traditional essay prompts with open responses and rubric scoring. Their suitability for assessing CR based on multiple sources of information in particular, may be limited by challenges in objective scoring and the brevity of the prompt (Zlatkin-Troitschanskaia et al., 2019b). While ecological validity in particular is especially limited in standardized tests (Braun, 2019), CR can be more adequately measured through *performance assessments* that simulate the complex environment students find themselves in 'in everyday life,' and provide an addition to standardized measures, as they are better suited to reflect current contexts and learning conditions inside and outside of higher education (Oliveri and Mislevy, 2019; Shavelson et al., 2019).

So far, to measure university students' performance on concrete, real-world tasks and to tap their critical thinking skills, the Council for Aid to Education (CAE) has developed the Collegiate Learning Assessment (CLA) (Klein et al., 2007), which was also used in the Assessment of Higher Education Learning Outcomes study, and has launched a refined performance test on CT, the CLA+. The assessment contains an hour-long performance task and a half-hour set of multiple-choice items so as to produce reliable individual student scores (Zahner, 2013). The CLA+ is available internationally for various countries (Wolf et al., 2015). It has been used in the United States and was also adapted and used in Finland<sup>3</sup>, Italy, and the United Kingdom (Zahner and Ciolfi, 2018), and has undergone preliminary validation for Germany (Zlatkin-Troitschanskaia et al., 2018b). This computer-delivered assessment consists of a performance task where students are confronted with a complex scenario. Additionally, they are presented with a collection of documents with further information and data that should be used to properly evaluate the case and decide on a course of action. The test has an open-ended response format and is complemented by 25 selected-response questions on separate item stems. According to

Wolf et al. (2015), the assessment measures the following student abilities: Problem-solving and analysis, writing effectiveness, writing mechanics, reasoning scientifically and quantitatively, reading critically and evaluatively, and critiquing an argument.

Other assessments that were recently developed for higher education, such as HEIghten by ETS<sup>4</sup> or The Cap Critical Reasoning test, can be considered knowledge-based analytical-thinking, multiple-choice tests<sup>5</sup> and do not encompass any performance tasks (for an overview, see Zlatkin-Troitschanskaia et al., 2018a).

## The iPAL Study on Performance Assessments of Higher Education Students' CR

In iPAL (international Performance Assessment of Learning), an international consortium focuses on the development and testing of performance assessments as the next generation of student learning outcome measurements (Shavelson et al., 2019). The researchers address the question how performance assessments can enhance targeted student learning beyond rote memorization of facts and actively foster students' acquisition of 21st century skills (including CR). The subproject presented here is designed to measure higher education students' CR by simulating real-life decision-making and judgment situations (Shavelson et al., 2019).

The German iPAL subproject follows a criterion-sampling measurement approach to assessing students' CR. Criterion-sampled performance assessment tasks present real-world decision-making and judgment situations that students may face in academic and professional domains as well as in public and private life. Test takers are assigned a role in an authentic holistic scenario and are given additional documents and links to Internet sources related to the topic of the task (presented in different print and online formats) to be judged in respect to their varying degrees of trustworthiness and relevance. The skillset tapped by these tasks comprises skills necessary to critically reason from multiple sources of information, i.e., to critically select and evaluate (online) sources and information, and to use them to make and justify an evidence-based conclusive decision.

In the German iPAL study, we developed a performance assessment with a case scenario (renewable energy) [comprising

<sup>2</sup>One well-known test of this kind is the Watson-Glaser Critical Thinking Appraisal (2002), which comprises tasks on inferences, recognition of assumptions, deduction, interpretation, and evaluation of arguments (Watson and Glaser, 2002).

<sup>3</sup>[https://ktl.jyu.fi/fi/hankkeet/kappas/copy\\_of\\_lyhyesti](https://ktl.jyu.fi/fi/hankkeet/kappas/copy_of_lyhyesti)

<sup>4</sup>[www.ets.org/heighten/about/critical\\_thinking/](http://www.ets.org/heighten/about/critical_thinking/)

<sup>5</sup><http://practice.cappassessments.com>

22 (ir)relevant, (un)reliable and partly conflicting pieces of information]. This newly developed computer-based performance assessment was comprehensively validated according to the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], 2014; see Zlatkin-Troitschanskaia et al., 2019b; Nagel et al., 2020). Validity evidence was gathered (i) by evaluating the test-takers' responses to the performance assessment (for details, see Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019b), (ii) a semi-structured cognitive interview, and (iii) an additional questionnaire on the students' personal characteristics such as prior knowledge, general intellectual ability, and media use behavior (for details, see Nagel et al., 2020).

In the following, we focus on the validation work conducted in (ii) cognitive interviews and present the analyses of transcripts of these cognitive interviews and corresponding results. To strengthen our validity argument (in the sense of Messick, 1994; Kane, 2013; Mislevy, 2018), we additionally refer to the particular findings from (i) to demonstrate how students' beliefs and reasoning processes as identified in the cognitive interviews are related to their task performance (written final response on the case presented in the task).

## MATERIALS AND METHODS

In this section, we first describe the entire study, including the performance task and the other assessments applied, before presenting the sub-study of the cognitive interviews and its results.

### Instruments

#### The Performance Task

To assess students' CR and their ability to critically reason from multiple sources of information, the German iPAL study developed and tested the "Wind Turbine" performance task. This computer-based assessment consists of a realistic short-frame scenario that describes a particular situation and requests a recommendation for a decision based on information provided in an accompanying document library (including 22 snippets and sources of information of different types; e.g., Wikipedia articles, videos, public reports, official statistics). These information sources, on which the students are to base their decision recommendation, vary in their relevance to the task topic and in the trustworthiness of their contents (for detailed descriptions of the performance task, see Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019b).

In this case scenario, test takers were assigned the role of a member of the municipal council of a small town confronted with the opportunity to build a wind farm on its grounds. They were asked to review the information sources provided in the task and, based on the evidence, to write a policy recommendation for a course of action, i.e., to recommend to the city council whether or not to permit the construction of the wind turbines in its agricultural countryside (for more details, see Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019b).

### Accompanying Assessments

To control for task- and topic-related prior knowledge, we used a short version of the WiWiKom test, which was comprehensively validated in the representative nation-wide WiWiKom study as an indicator of knowledge in economics and social sciences (Zlatkin-Troitschanskaia et al., 2019a). As two indicators of general cognitive ability, the scale "Choosing figures" of the Intelligence Structure Test (IST-2000 R, Liepmann et al., 2007) as well as the grade of university entrance qualification were used in the present study (for details, see Nagel et al., 2020). The participants' levels of interest in the task topic and case scenario (renewable energy) as well as their test motivation were also assessed in this study using two five-point-Likert-type scales (validated in the previous studies cited).

Furthermore, socio-demographic information and personal characteristics (e.g., scales on 'media use,' 'need of evaluation,' 'information overload'; for details, see Nagel et al., 2020) expected to affect test performance were collected. Indicators of relevant expertise in the context of solving the performance task, such as completed commercial or vocational training, were also surveyed, as they might also influence task performance.

### Study Design and Validation

To test and validate the "Wind Turbine" task in accordance with the Standards of Educational and Psychological Testing (American Educational Research Association [AERA], 2014), assessments were conducted with a total of 95 undergraduate and graduate students from different study domains (e.g., business and economics, teacher education, psychology) at a German university (Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019b).

The students worked on the task in a controlled laboratory on computers configured specifically for this assessment and had no access to other resources to solve the task. The study was carried out in small groups on several dates under the supervision of experienced test administrators.

The total test time for the performance task was 60 min. The time limit put the participants under pressure, which led to them not having enough time to study all given sources intensively. Instead, it required them to decide which sources and contents to select and review more thoroughly, considering their relevance, validity, and trustworthiness. After the performance task (and a short break), the participants were asked to work on the accompanying assessments. The participants received an incentive of €20, and were offered an individual feedback on their test results.

Test performance was scored using a 6-point Likert-type anchored rating scheme based on the CR definition with 5 dimensions and 23 subdimensions (Section "Sample and Data"; for details on scoring, see Zlatkin-Troitschanskaia et al., 2019b). The individual task responses, i.e., short essays, were randomly assigned to two of four trained raters, and the written responses were evaluated according to this newly developed and validated scoring scheme with behavioral anchors for each sub-category. Two raters independently evaluated the participants' task responses, and a sufficient interrater agreement

was determined (Cohen's  $\kappa > 0.80$ ;  $p = 0.000$ , for the overall test score).

In terms of psychometric diagnostics, the student response when solving the performance task (well-founded written final decision) was interpreted as a manifestation of their latent (meta)cognitions. The students' task performance, i.e., their written responses, were regarded as valid indicators of the students' ability to critically reason from multiple and contradictory sources of information (in the sense of the construct definition, see Section "Critically Reasoning from Multiple Sources of Information"). The theoretically hypothesized multidimensional internal structure of this construct was supported empirically using confirmatory factor analysis (CFA) (Zlatkin-Troitschanskaia et al., 2019b).

As theoretically expected, analyses of the task performance did not identify any significant domain-specific effects among students from different study domains (Nagel et al., 2020). This also holds true for prior knowledge from previous vocational training, which showed no significant effect on the test results. As the performance task was developed to measure generic CR skills, this finding indicates that as expected, the assessment is not domain-specific. However, with regard to theories in learning and expertise research, it could be assumed that domain-specific expertise, though acquired within a certain domain, can be transferred to generic problems or tasks (Alexander, 2004). In this respect, these results indicate that students may have deficits in their (meta)cognitive abilities that would enable them to transfer their prior knowledge and skills to the new context encountered in the performance task. Overall, the results from these validation studies provide evidence of the technical quality of the developed performance task and provided evidence as to the test's construct validity and reliability (Zlatkin-Troitschanskaia et al., 2019b; Nagel et al., 2020).

## Cognitive Interviews and Qualitative Analyses

The analyses of test performance *per se* do not permit us to draw valid conclusions on students' underlying response processes when performing this task. In accordance with our construct and test definition, we expect that while working on the performance task, the test participants selected and evaluated the given information with the goal of finding relevant and reliable/trustworthy information for their evidence-based reasoning, decision-making, and final conclusion in the written response. To investigate how the test participants dealt with multiple contradictory sources, to what extent they integrated and evaluated given information during their reasoning and decision-making process, and which individual factors influenced their response processes, a semi-structured cognitive interview with stimulated recalls was conducted immediately after the performance assessment with a subsample of participants (see next section). The participants were first shown a screen displaying all 22 documents included in the document library one after the other. The students reflected and commented on, for instance, whether they considered the source in question and the information given therein to be relevant and/or credible (and why), and whether they used or ignored this source and information (and why). Particular

attention was paid to controlling whether the students were aware of their task topic-related beliefs and attitudes and if so, whether they were aware to which extent these influenced their critical reasoning while processing the task, for example resulting in selective inclusion of the given information. The interviews took approximately 80 minutes and were recorded for later transcription.

The interview questions included, for instance, whether a test participant held task topic-related beliefs about wind power, the environment, etc. prior to the performance assessment, and to what extent previous experience, individual knowledge, attitudes and beliefs relevant to the task topic influenced the students' information selection, evaluation and decision-making. In particular, the students reported at which point in time during the task processing they made their decision whether or not to suggest to the municipal council to allow the building of the wind farm (for instance, indicating that many students made their decision before they had even read the given information; see Section "Results"). The cognitive interviews also indicated that the performance task with its task prompt is clear and suitable for the objectives of the presented study.

The evaluation of the data from the cognitive interviews in the German iPAL project was carried out using the software MaxQDA. Based on the cognitive interview protocols, a differentiated category system was developed and validated. More specifically, the qualitative analysis of the cognitive interviews was guided by grounded- and data-driven theory for developing a coding scheme (Strauss and Corbin, 1998).

The iterative process of coding consists of (1) open, (2) axial (Strauss and Corbin, 1990) and (3) selective coding. At first, an open coding was used to access the data that did not yet follow any schematics. In the subsequent step, first categories were identified, such as the students' beliefs at the beginning of the task or at which time point during the task processing the decision was made. Then, all interviews were analyzed and coded based on the defined categories. The coding scheme highlighted the points of reference regarding different information sources used by the students in their interviews. It linked the use of different sources to the students' reasoning processes while reaching their decision, making it possible to derive and generalize response patterns. Coding development was complemented by an analytical approach of constantly comparing phenomena within the dataset (minimum and maximum contrast between the phenomena). Selected codes with a focus on student beliefs are presented in this paper (see Section "Results").

The classification of the participants into the three profiles described in Section "Results" was based on a combination of criteria from the category system. These were primarily: (1) at what point during the assessment they made their decision (reported in the cognitive interviews), (2) their decision-making process (pros and cons; intuitively, based on original opinion; based on a specific source, etc.) and (3) the strength of their beliefs (strong personal beliefs primarily about nature conservation/animal welfare, etc. and general personal identification with the task topic). As the participants were classified into profiles based on a combination of these three categories, participants classified into different profiles may share certain attributes (e.g., listing pros and cons).



## Sample and Data

The semi-structured interviews were conducted with a purposefully selected subsample<sup>6</sup> (which is part of the overall sample used in the German iPAL study) of 30 university students from one German university and from different courses of study. With this subsample, we aimed to include students from all study domains represented in the main sample in the cognitive interviews as well. Accordingly, the subsample consists of about 50% students of economics education, while the other half comprises students from other study domains (economics, psychology, and geoscience). Another important criterion for purposeful sampling was to include as wide a range as possible in terms of participants' prior study experience and other personal characteristics that may influence students' task topic-related beliefs, attitudes and knowledge, and their task performance. Accordingly, the subsample consists of students from both undergraduate and graduate programs and in different study semesters. To gain first indications of the possible impact of knowledge and skills achieved during academic studies, we focused on more experienced students. The average duration of studies to date among subsample participants was therefore 5.1 semesters, indicating that the students were fairly advanced in their respective study programs. Additionally, the university entrance qualification (with an average grade of 2.1; range from 1.0 = best to 6.0 = worst;  $n = 25^{*7}$ ). To control for the possible impact of pre-university education and practical experience, we included students with completed vocational education and training (11 students had completed an apprenticeship before beginning their academic studies;  $n = 29^{*7}$ ). The average interviewee's age was 24 years; 21 students were female – these proportions were similar to those in the overall sample.

Despite this purposeful sampling procedure, as participation in this study was voluntary, our sample cannot be considered representative. However, no significant deviations from the entire student population described in Nagel et al. (2020) were found with regard to the socio-biographical characteristics (e.g., gender and age).

## RESULTS

### Prior Findings on Test Performance and Additional Assessments

The students achieved an average intelligence-test (IST) score of 17.18 points (out of a maximum score of 40 points;  $n = 28^{*7}$ ) and an average economics knowledge test score of 10.46 points (out of a maximum score of 15 points;  $n = 23^{*7}$ ). Only four students stated that they had previous practical experience with Wind Turbines. Most students reported a low to medium level of task topic-related previous knowledge while a high level of knowledge on wind turbines was very rare ( $n = 1$ ).

<sup>6</sup>The criteria for selection from the overall sample and inclusion were the participants' socio-biographical and educational characteristics to ensure: (1) gender balance, (2) age distribution, (3) course of study representation (all), (4) study year/progress (advanced students), and (5) prior education (e.g., completed vocational training).

<sup>7</sup>The deviation from the total sample size (30 participants) is due to missing values.

The mean performance on the task was 3.52 points with 6 points being the highest possible score (for the scoring, see Zlatkin-Troitschanskaia et al., 2019b). The median number of information sources students used in their written statements was 7 (out of 22 information sources given in the task). The written argument-based statements within the scope of the task processing differed in length, which was on average 426 words, with a maximum of 866 words and a minimum of 68 words, indicating a high deviation ( $SD = 196$  words) within the sample.

In the following, these results were used as external criteria to demonstrate how the following results from the analyses of the cognitive interviews correspond to these results of the quantitative analyses of the test scores.

## The Relationship Between Beliefs and Decision-Making

### RQ1: Students' Beliefs at the Beginning of Task Processing

In the cognitive interviews, the students were asked whether they had been aware of their task topic-related beliefs prior to working on the task and if so, whether they were aware that their personal beliefs may have influenced their decision in the performance task and how they believed this influence may have shaped their response. Most participants ( $n = 23$ ) stated that *they had already held certain beliefs on the task topic before beginning the task*. For instance, one participant stated: “[...] I think I would have recommended this from the beginning because this is also a topic I hear about in the media from time to time, so that I already have a personal opinion about wind power and energy” (ID15). In response to the question whether his personal beliefs had influenced his response, interviewee ID7 stated: “Sure, because then I did not even look at the controversial sources at all and, that is... for example, if I believed that the bats from source 21 were extremely important, then of course I would have looked at the source.” Seven participants ( $n = 7$ ) reported that they did not have any prior beliefs about the topic of the performance assessment.

In terms of distinct types based on the reported beliefs, both groups of students – those who indicated prior beliefs and those who did not – can be further distinguished into two subgroups each (i) depending on the students' *positive or negative stance toward wind turbines*, which vary considerably in stance strength and (ii) which can also be linked to a *more economics-focused or a more ecologically oriented reasoning perspective* (see Section “The Relationship between Beliefs and Decision-Making”).

### R2: The Relationship Between Students' Beliefs, Their Decision-Making Process and Their Final Decision

#### *Time of the decision-making and types of decision (intuition-based vs. evidence-based)*

In the cognitive interviews, the students reported *at which point in time while working on the performance task and processing the information they made their decision* as to whether renewable energies should be promoted or not in the given case (see Table 2). About one third of the students ( $n = 8$ ) made their decision *at the beginning of the task, after having read the*



scenario, even though they had not yet read or considered the given information at all or only very briefly.

Another group of students ( $n = 8$ ) used the given sources and made their decision mostly after (more or less thoroughly) looking through the information provided. For instance, when asked when he had decided in favor of or against the construction of wind turbines, interviewee (ID 1) stated “yes [...], I actually knew from the beginning when I went through this [task] what direction my statement would go in.” Interviewee (ID 23) made his decision while working on the audiovisual information: “So after I watched the videos [...] I changed my opinion.”

In contrast, other students made their recommendations after having reviewed the information material and after weighing up the pros and cons ( $n = 9$ ), as it is the case with, for example, interviewee (ID 7): “Interviewer: So that means that you first read all the sources and all the arguments? Participant: First the pros and cons, and only then I had a feeling.” This finding indicates that for some students the creation of pro and contra lists was an important step in their decision-making process. However, not all of those who made their decision comparably late in the task-solving process stated that they had done so based on weighing up the pros and cons: five students indicated that they made a late but still intuitive decision.

Overall, with regard to the time of decision-making, four types can be distinguished among the participants (Table 2), which differ in terms of intuition-based vs. evidence-based decision-making as well as the extent to which the given information and pros/cons were considered or ignored.

Students who made a late intuition-based decision ( $n = 5$ ) performed worse, with an average test score of 3.25. Students who made their decision at the end of the task based on weighing pros and cons ( $n = 9$ ) performed better compared to all other participants, with an average test score of 3.83. Noticeably, there were hardly any differences in task performance between the students who decided intuitively at the beginning of the task and the students who weighed up pros and cons and decided at the end of the task.

## The Relationship Between Students' Beliefs and the Decision-Making Process: Profiles of Decision-Making

Regarding the question to what extent students' were aware that their beliefs impact their decision-making process and whether distinct profiles of decision-making can be determined, among the 30 study participants, we identified students who indicated that their previous beliefs played a major, minor or no role

in their decision-making process. Combined with the time at which they made their decision, we distinguished three profiles of decision-making:

**Profile 1 “determined”:** Participants who ignored the given information and made their decision solely based on their individual beliefs, almost immediately after having read the task ( $n = 7$ ). For example, (ID5) stated: “I wouldn't have made a recommendation that goes against my gut instinct. For example, I think that even if the sources had been chosen in such a way that they would have given me a negative impression, I am not sure whether that would have caused me to change my initial positive stance. I simply couldn't just ignore my background knowledge and my personal attitude when giving my recommendation at the end.”

**Profile 2 “deliberative”:** Participants who decided contrary to their task topic-related beliefs, and changed their decision after having read the information provided in the task ( $n = 11$ ), as well as participants who stated that they held certain beliefs at the beginning of the task but weighed up pros and cons while processing the task and made their decision based on these considerations ( $n = 5$ ). The two cases were merged into one profile as students in both cases stated that they held certain beliefs but made their decision based on the pros and cons of the evidence rather than on those beliefs.

There were some differences within this decision-making profile. For instance, some students switched between being in favor of or against the construction of wind turbines while working on the task: “So basically I'm for it and then while I was writing this I just started to waver, you have to list the negative things and then I doubted it for a moment but then I finally decided in favor at the end.” (ID 8); other students changed their prior opinions by reflecting on their own beliefs in the context of the given information: “At the beginning I would have said yes [impact of belief on decision-making]. But then I tried to be as unbiased as possible, or rather to be subjective in my role as a member of the council. And then I kind of abandoned my [initial] decision and my personal belief.” (ID 17).

**Profile 3 “open minded”:** Participants who did not state any prior beliefs, took note of the provided information, and made their decision after considering pros and cons ( $n = 7$ ). Interviewee (ID7) stated that he had had no prior beliefs before starting the task, and that he made his decision after considering the given information and making a pro and con list: “No, I couldn't decide at the beginning, it just happened toward the end of the argumentation. Well, I was not for or against it from the beginning. I just did not know how to decide.”

Since the participants were classified into profiles based on a combination of the scoring categories (see Section “Cognitive Interviews and Qualitative Analyses”), participants classified into different profiles may share some attributes (e.g., listing pros and cons) and there may be some overlaps between the profiles.

## The Relationship Between the Decision-Making Profiles and Task Performance (Test Scores) as Well as the Results of Additional Assessments

Noticeably, the participants in profile 3 on average achieved higher test scores than the other two profiles (Table 3). Students who based their decision on their beliefs (profile 1) performed

**TABLE 2 |** Time of decision-making and type of decision.

Time of the decision-making	Frequency	Average test score
Early, <u>intuitive</u> decision (before source evaluation)	8	3.76
Decision after reading (selected) sources (multiple) times)	8	3.11
Late, <u>intuitive</u> decision (after source evaluation)	5	3.25
Late decision based on pro-/con-arguments	9	3.83

**TABLE 3 |** Means of task performance of different profiles.

Impact of students' beliefs	Frequency	Average test score	SD
<b>Profile 1 “determined”:</b> Decision based on firm beliefs prior to processing the task	7	2.95	1.10
<b>Profile 2 “deliberative”:</b> Decision made after reflection on beliefs	16	3.38	0.59
<b>Profile 3 “open minded”:</b> Decision made after a neutral approach, primarily reflecting on the source material	7	4.01	0.44

**TABLE 4 |** Characteristics of the three decision-making profiles.

Profiles	Profile 1 “determined” (n = 7)	Profile 2 “deliberative” (n = 16)	Profile 3 “open minded” (n = 7)	Total Sample (n = 30)
Number of Information Sources Used	5	8.25	7.85	7.5
Length of Written Response	506	370	473	426
Gender	Female: 7 Male: 0	Female: 8 Male: 8	Female: 6 Male: 1	Female: 21 Male: 9
Age (n = 29)	21.17	24.75	24.43	23.93
Degree	Bachelor: 6 Master: 1	Bachelor: 14 Master: 2	Bachelor: 6 Master: 1	Bachelor: 26 Master: 4
Vocational Training (n = 29)	Yes: 2 No: 4	Yes: 5 No: 11	Yes: 4 No: 3	Yes: 11 No: 18
Intelligence Test Score (n = 28)	16.67	17.40	17.14	17.18
Economic Knowledge Test Score (n = 23)	10.20	10.65	10.20	10.35
University Entry Qualification Grade (n = 29)	1.87	2.18	2.17	2.11

worse compared to other participants (profile 3). In terms of the average test score, the deviation between these two profiles (1 and 3) was more than 1 point.

Upon further characterizing the three profiles, we found additional differences between the groups of students in terms of the number of information sources used and the number of words in the final recommendation statements, which differ greatly (Table 4). Compared to students who made a decision based on their beliefs (profile 1), the average number of information sources used was 3.25 points higher for students who changed their beliefs (profile 2) and 2.86 points higher for students of profile 3. The mean number of words in the written final recommendation statements also varied heavily. Remarkably, the responses of “deliberative” students (profile 2) were the shortest with an average of 370 words. “Determined” students (profile 1) who did not change their beliefs wrote on average 33 more words than “open minded” students (profile 3) with a mean of 473 words.

With regard to personal characteristics, there were no significant differences in the intelligence test scores for the three profiles (Table 4). The same was true for performance in the economics knowledge test, with results ranging from 10.20 to 10.65 points (on a 15-point scale).

Students who made their decision based on evidence and pros/cons, despite their beliefs or without considering previous task-related beliefs, tended to be older (profile 2: 3.58 years older on average; profile 3: 3.26 years older on average) than “determined” (profile 1) students. There were no significant differences in terms of gender, pre-university education (vocational training or university entry qualification grade) or degree course, which does not indicate any substantial influence of prior education on the response processes.

## Task-Topic Related Attitudes and Their Relationship to Reasoning Processes and Decision-Making

Another approach to identifying certain beliefs and their possible relationship to information processing and critical reasoning was to analyze students' task-topic related attitudes and their impact on reasoning approaches when solving the performance task. In this respect, the reasoning lines identified in the cognitive interviews (as well as in the written task responses) can be categorized as follows:

(1) The first category differentiates between primarily *economics-focused* or *ecologically oriented* reasoning lines. Twelve students' recommendations had a primary economical focus in their reasoning, while 18 students relied more on ecological aspects and sources presenting an ecological perspective.

Remarkably, students who were in favor of building wind turbines tended to choose an economics-focused reasoning line, while students against the construction chose an ecologically focused perspective (Table 5). An example for an economical reasoning line can be seen in the following statements: “*The trade tax to be paid by the operator could be sensibly invested in the modernization of facilities, the infrastructure of the place and the marketing of the local recreation area. This source of income seems to be important for the community, especially in the future, against the background of an increasingly dwindling agriculture*” (ID 13); “*In my opinion, the offer should be accepted, as the positive aspects outweigh the negative ones and, in general, the construction of wind turbines would mean a macroeconomic, long-term benefit for the community. In addition, it is an investment in infrastructure.*” (ID 25). In contrast, an example for an ecological line of reasoning and their relationship to information processing and decision-making can be seen in the following statement (ID 26): “*That caused me to have*

**TABLE 5 |** Economic- and ecological-focused reasoning and decision against or in favor of wind turbines at the end of the tasks.

Reasoning Approach Frequency/(Average test score) <i>n</i> = 29	In favor of building wind turbines at the end of the task	Against building wind turbines at the end of the task
Economic-focused reasoning	9 (3.68)	3 (3.82)
Ecological-focused reasoning	6 (3.95)	11 (3.17)

*fewer choices, and I had already had the notion in mind that wind turbines are good and nuclear power plants are bad, which is why I said from the very outset that yes, no matter in which form, more renewable energy should be produced and, well, that's why I said all along that that would be the most sensible result in my opinion, without any of those arguments."*

(2) The students' decision-making processes and (final) recommendations can also be categorized in terms of the extent to which the specific situation described in the task was considered. While half of the participants took the *task-specific perspective* of the local council and the current situation of the city into account (*n* = 15), other students choose a more general approach in making a recommendation for or against wind turbines (*n* = 15).

One example for considering of local conditions can be found in the statement of participant (ID 7): *"I consider the construction of the wind turbines in the north of the municipality to be an incalculable risk, as the tertiary sector and especially the tourism that goes along with it represent an important source of income for the town. I think it makes much more sense to locate the wind turbines in the west. Farmers who live there, such as Mr. Anders and Mr. Bender, should welcome an additional source of income besides agriculture, so that they should agree to the construction of the wind turbines."* A more general approach is expressed in the statement of participant (ID 16): *"The fact that wind energy is initially a clean and environmentally friendly way of generating energy speaks for the installation of wind turbines. In addition, there are also economic reasons for this, as good money can be made from the rent that incurs when a wind turbine is installed. [...]"*

While the majority of students who took the task-specific current situation of the city into account tended to express a negative attitude about wind turbines, students who took a more general reasoning approach were rather in favor of building wind turbines (Table 6).

### The Relationship Between the Reasoning Approaches and Task Performance (Test Scores)

In terms of task performance, no significant differences were found between the students with different reasoning approaches, although students who chose economics-focused reasoning achieved slightly higher performances than the other students. When taking into account the positive vs. negative stance toward wind turbines at the end of the

**TABLE 6 |** Perspective of reasoning (local council included or not) and decision against or in favor of wind turbines at the end of the tasks.

Reasoning Approach Frequency/(Average test score) <i>n</i> = 29	In favor of building wind turbines at the end of the task	Against building wind turbines at the end of the task
Included perspective of local council in the decision regarding the construction of wind turbines	5 (3.97)	10 (3.37)
Made a general decision regarding the construction of wind turbines	10 (3.70)	4 (3.16)

task, however, the difference in task performance of students with ecological-focused reasoning is about 0.8 points, whereas the difference in the group with economic-focused approach is only 0.1 point.

### Change of Beliefs While Solving the Task RQ3: Interaction Between Students' Beliefs and Processing of the Given Information

Looking at the time of decision-making, we found that some students changed their opinion about the construction of wind turbines (once or several times) while processing and working on the task, while others did not. While 14 interviewees reported that they did not change their opinion about the wind turbines over the course of their task solving, 12 interviewees changed their opinion after processing of information given in the task (Table 7). Four participants claimed that they had not been initially disposed either way. Both groups of students—those who changed their opinion and those who did not—can each be further distinguished into two subgroups depending on their positive or negative stance toward wind turbines, which vary considerably in size. Within the group with no change of opinion, participants who had voted against the construction on wind power plants at the beginning of the tasks and remained negative (*n* = 3) can be distinguished from participants who had a positive stance toward wind turbines before and after completing the task (*n* = 12). We can also differentiate between students who have changed their opinion during working on the task. Some students initially had negative attitudes toward wind turbines, but changed their opinion during the task processing and in the end voted in favor wind turbines (*n* = 2). The same applies to participants who were in favor of constructing wind turbines at the beginning, but ultimately spoke out against wind turbines (*n* = 9).

### The Relationship Between a Change of Students' Beliefs and Task Performance (Test Scores)

There was hardly any significant difference in the test score of the two groups, although students who did not change their opinion performed slightly better than students who changed their opinion: the difference in task performance was about 0.7 points.

TABLE 7 | Change of opinion.

Change of opinion about wind turbines:	Frequency	Average test score
Change of opinion	12	3.17
No change of opinion	14	3.86
(No statement at the beginning and/or end)	4	3.24
<b>Subdimensions: Change of opinion about wind turbines: Student. . .</b>		
. . . maintained a positive stance towards the construction of wind turbines	12	3.83
. . . maintained a negative stance towards the construction of wind turbines	3	3.96
. . . changed from positive to negative stance towards the construction of wind turbines	9	3.12
. . . changed from negative to positive stance towards the construction of wind turbines	2	3.39

## DISCUSSION AND CONCLUSION

### Summary and Interpretation of Results

The data from the cognitive interviews on the students' beliefs, information processing and reasoning processes make a valuable contribution to explaining the students' CR abilities and the complex interplay between their underlying thought processes and task topic-related beliefs. In the interviews, most participants expressed that they were aware of holding certain beliefs at the beginning of task processing (RQ1). The results of the qualitative analysis of the cognitive interview protocols indicated that the students' task topic-related beliefs had an influence on their selection, critical evaluation and use of information as well as on their reasoning process and final decision (RQ2). As an additional decisive contribution to existing research [see Section "State of Research on Beliefs and Their Impact on (Online) Information Processing"], we provide initial evidence that some students' task topic-related beliefs changed over the course of task processing, indicating that the processed information (recognized and reflected evidence and pros/cons) influenced the students' beliefs to varying degrees (RQ3).

Overall, the evidence from this qualitative analysis suggests a complex reciprocal and changeable relationship between students' task topic-related beliefs, their processing of new (confirm or deviant) information and their decision-making based on both beliefs and evidence.

More specifically, the types of beliefs and attitudes derived from the cognitive interview data suggest their influence on information processing, reasoning approaches and decision-making. In particular, the students who already had strong task topic-related beliefs at the beginning regarded these as decisive while solving the task. For instance, students who had already made a decision based on their beliefs at the beginning of the task cited fewer sources in their written response (final decision).

Overall, the *selection, evaluation, and use of information* while working on the task were influenced, in particular, by the participants' *task topic-related beliefs* (RQ2). By contrast, hardly any differences became evident in terms of students' relevant knowledge. However, the majority of the participants had only little prior knowledge of the subject, i.e., a large amount of the information in the task was new to them. Though most students had a positive or negative stance toward renewable energy in general, their personal beliefs concerning wind energy in particular did not appear to be very firm and well-founded. The few test participants who had already dealt with the subject

area in more detail appeared to have more solid personal beliefs about wind energy (RQ1). Furthermore, there were no differences in terms of students' general interest in the topic. However, two reasoning lines – more ecologically oriented vs. economics-focused approaches – became evident, which appear to influence students' decision-making processes and final decision.

Remarkably, the students who had more elaborated beliefs prior to processing the task were more likely to come to a decision that contradicted their personal beliefs. For instance, the information on the negative effects of wind turbines on the health of people and animals living in the vicinity of a wind farm (noise emission, bird strike, infrasound) was particularly relevant for these participants when making their decision; they were more astonished by this information than the students who had hardly any prior knowledge about the subject and no well-developed beliefs (RQ2).

Most students started selecting information right away after obtaining an initial overview of the sources presented in the task. The participants' subsequent evaluation of the given information with regard to the *reliability, validity, objectivity, and trustworthiness* of the respective sources (as stated in the interviews) does not appear to have had much of an influence on their selection and use of information. In contrast, the participants evaluated the *relevance* of the sources differently, whereby a large number of the sources that were evaluated as relevant were used to inform their decisions and help them formulate their written recommendations. For instance, in the interviews, the majority of students rated Wikipedia as a less reliable source (of course the exact details vary, but in general, it received rather negative ratings), as Wikipedia pages can potentially be edited by any Internet user. However, the choice as to whether or not to use information from Wikipedia sources was primarily made on the basis of the *content* of these sources ("do I want to address bird mortality or not?"). In contrast, when it came to the evaluation of the public-service broadcaster videos, a large number of participants assessed these videos as trustworthy despite not having watched them, as they considered this source to be particularly reliable.

Overall, in the cognitive interviews it became evident that the students mostly selected and evaluated (or ignored) new information depending on media or source type (i.e., whether they believed that certain types of media and presented sources are relevant and reliable) but not on the particular content/evidence. This finding is in line with previous research reported in the Section "State of Research on Beliefs and Their



Impact on (Online) Information Processing” and stresses the importance of epistemic beliefs regarding information sources, which was not a focus of this study and requires further investigations in the particular context of online reasoning (for limitations, see the next section). In addition, this result points to a demand for more observational studies that capture in detail what documents, what parts of these documents, and which content the participants read and comprehend while solving the task.

Although participants used different *sources* in their statements, most of the students did not compile the information provided to them and weigh the evidence (pros/cons), but rather selected information related to their own beliefs, indicating biased selection, evaluation and use of information (for the confirmation bias, see Mercier and Sperber, 2009, 2011; Metzger et al., 2010; Metzger and Flanagin, 2015). A (repeated) critical examination of the information and evidence provided did not take place.

Linking the results from the qualitative analyses of the cognitive interviews with task performance further suggests a confirmation bias in reasoning, showing that students who only made their decision based on their beliefs (profile 1) had the worst test scores on average. This was also reflected in the number of sources used. They wrote the longest statements but based on the lowest number of used *sources*, without sufficiently reflecting on the available information and evidence. This finding is also supported by the lower performance of students who tended to overemphasize a single source while neglecting all contradicting source information (for the authority bias, see Metzger et al., 2010; Metzger and Flanagin, 2015). Overall, the finding from the qualitative analyses that often no sufficient critical reasoning took place in the decision-making process and that the decision was based on beliefs (and bias) was also reflected in the students' statements.

In contrast, the students with no early inclination (profile 3) approached more source material neutrally and decided on the incorporation of the information and evidence individually, outperforming the other students in terms of task score. Their statements were less belief-driven since they addressed the specific task scenario and prioritized the town's needs and restrictions over their personal stance on renewable energy.

As the students only had limited time (60 min) to respond, time pressure also played an important role and forced them to gather relevant information as quickly as possible. If the participants selected the information they intended to read more precisely, worked with it and then used it in their decision-making at an early stage (*right at the beginning*), quickly (*without deliberative thinking*), and consistently (*without changing their minds*), the issue of time pressure apparently did not have much of an effect on their task-solving efforts. The cognitive interviews indicate that for some students, however, selecting suitable information was a major challenge while working on the task (indicating the higher cognitive load; Sweller, 1988). These participants often opted to use internal sources as opposed to external sources, indicating that they mostly focused on the information that was available within the task document itself and disregarded the hyperlinks. The majority of participants did

not watch the two videos (completely) due to time issues. This aspect also points to some limitations of our study (see next section).

## Limitations and Implication for Future Research

Though the study provides some important insights into the complex reciprocal relationship between students' beliefs and their reasoning and decision-making process, some limitations (besides those related to the sample, see Section “Sample and Data”) must be critically noted, which indicate some perspectives for further research.

While the results of the qualitative analyses pertaining to RQs 1&2 allow for some clear statements about students' beliefs and their influence on critical reasoning, the findings pertaining to RQ3 regarding changes in beliefs are still limited. First, in our study, we can only derive conclusions about task topic-related beliefs. These need to be distinguished from general personal (e.g., epistemic) beliefs, which were not analyzed in our study. In prior research, general beliefs usually were seen as a trait that does not change during the course of solving a task. However, measuring epistemic beliefs is considered challenging from a conceptual as well as a methodological perspective, and requires further research (van Strien et al., 2012a).

Second, based on the cognitive interview protocols, a clear distinction between a change in task topic-related beliefs and a change of overall opinion could also only be made to a limited extent. Although some students clearly stated that they had beliefs prior to processing the task that influenced their information processing and decision-making, and they had changed their opinion, we cannot conclude, on the basis of the interviews, whether this *change of opinion was due to a change in their underlying beliefs*. It is also questionable whether students were able to clearly distinguish between their belief, their attitude toward the task topic, and their opinion, and to express this difference in the interviews. This limitation results in an important follow-up for further research: Is a change of opinion accompanied by a change of task topic-related beliefs?

Though the results of both assessed scales on students' interest in the task topic and students' test motivation showed (very) high levels among all participants in this study, we noticed some differences in the way students approached the cognitive interviews. While some students were very communicative and talked a lot about their beliefs and task processing, other students gave short answers. Consequently, the cognitive interview protocols vary substantially in length and detail. The results of the qualitative analyses must therefore be viewed critically in terms of this data limitation. For instance, it could not be ruled out that students who did not express that they had topic-related beliefs prior to processing the task may not have deliberately reflected on this interview question or simply not have wanted to share this information (e.g., due to a bias of social desirability). Despite the use of a standardized guide in the semi-structured interviews, the comparability of the cognitive interview protocols may be limited in this regard.

The task topic may also be not without bias in this respect, since renewable energy can be generally framed in a positive

light. For this reason, it can also not be ruled out that students' responses to the task and their answers to the interview questions were biased in terms of social desirability. However, the fact that some students in our sample were both initially and ultimately against the construction of the wind turbines ( $n = 3$ ) may contradict this assumption.

In addition, though (i) the task prompt to write an evidence-based statement regarding the decision for the community should have been clear and strong enough to indicate that a discussion of the evidence (pros-cons) made available in the task was required, and (ii) (very) high levels of assessed interest in the task topic and overall test motivation among participants were determined, a difference among participants in terms of (metacognitively) engaging their critical reasoning skills when solving the performance task can still not be ruled out. Based on prior research, however, it can be assumed that the activation of critical reasoning abilities requires metacognitive skills (e.g., Brand-Gruwel and Stadler, 2011). Therefore, further understanding of students' (metacognitive) engaging (and other influences) during the decision-making process is required to help identify certain patterns in task processing strategies for this type of performance assessment and to further improve computer-based simulations in terms of their ecological validity and reliability to ensure more authentic assessment (for a critical discussion, see also Mercier and Sperber, 2011).

In this context, it is remarkable that the group of students who were aware of the influence of their beliefs – despite the task prompt asking them to include the given information and evidence in their decision-making process – decided to use only information that supported their beliefs (profile 1). These students had already recognized at the beginning of the task processing that their beliefs would have a decisive influence on their decision. If we transfer this finding to other real-life situations, in particular the everyday use of online sources in Internet searching, further research is required as to whether students, when searching for sources and in the context of their university education, also specifically focus on sources and information that confirm their beliefs. In this respect, the identified reasoning profile 1 may lead to an acquisition of biased (domain-specific) knowledge. In contrast, the “open minded” profile 3 approached more information neutrally, outperforming the other students in terms of the scored quality of written statements.

In this context, it is also important to focus on those students who claimed to have certain beliefs on the topic before starting the task but still reviewed all the information given and even partly decided against their beliefs after having regarded all information (“deliberative” profile 2). This profile should be analyzed more in-depth, especially taking into consideration both additional underlying cognitive and non-cognitive student characteristics as well as specific learning opportunities that this group might have had to develop this deliberative reasoning approach. Here, further questions arise: Why did students choose this approach and decide against their beliefs? What personal or contextual factors may have played a decisive role?

The complex relationship between prior knowledge and beliefs also requires further in-depth investigation. Ho et al.

(2008) found that task topic-related beliefs interact with the amount and quality of topic-relevant knowledge, whereby the topic-related beliefs may have a stronger impact on decision-making than knowledge. Analogously, the results of our study suggest that in general, no matter how experienced a student is in a topic or how much previous knowledge they had, certain beliefs seemed to be influential and predominant. However, to what extent the beliefs influence students in their approach to a task topic and which aspects were particularly crucial for students to be influenced by their beliefs (e.g., strength of beliefs or additional personal characteristics) must also be analyzed in further research (for an overview, see Brand-Gruwel and Stadler, 2011).

In addition, looking at the differences in the students' reported reasoning processes, we can conclude that diverse students' beliefs and attitudes, which were related to the task context and topic to a very different extent (e.g., in the area of sustainability), had an influence on the students' decision-making and final decision. Based on the data from the cognitive interview protocols, however, we were not able to analyze the complex relationship between beliefs, reasoning approaches and *lines of argumentation*. Though critical reasoning is indeed related to aspects of argumentative skills, this latter aspect was not the focus of our study (as described in Section “Conceptual and Methodological Background”) and requires further investigation in several regards. Particular investigation of argumentative skills would require a substantial change and further development of the experimental and assessment setting. For instance, there are several performance tests available that specifically focus on measuring argumentative skills (e.g., Argument Structure Test, Münchow et al., 2020; Argument Judgment Test, Münchow et al., 2019) and are suitable for discriminant validation of CR assessments, which should be investigated in a follow-up research. In addition, comprehensive qualitative analyses of both the argumentative importance of the material on the one hand as well as (i) arguments (more or less reflective or intuitive, Mercier and Sperber, 2011) used by students in their responses and (ii) (new) arguments created by the students themselves based on given arguments in the provided information on the other hand need to be conducted in further studies, and explicitly linked to students' critical reasoning ability and performance.

Finally, the method of cognitive interviews also has certain limitations in terms of understanding and explaining students' reasoning processes during task-solving, for instance due to a bias of social desirability (e.g., Kahne and Bowyer, 2017) as mentioned above or limited mental recall capacities. However, one central focus of the presented study lies on the investigation of self-awareness of one's own beliefs, i.e., whether the students were aware of their beliefs and whether they were aware if their beliefs influencing their perception, evaluation, selection and use of the given information. Hence, cognitive interviews were necessary to gain indications regarding the students (critical) reflection on their thought processes involved in solving the task, i.e., writing a statement. Especially any conclusions about self-awareness regarding one's beliefs and their relation to decision-making can best be reached by means of stimulated recalls in cognitive

interviews, which has been shown to approximate think-aloud methods in the study settings where participants cannot think aloud while processing the task (as in this computer-based test environment).

Follow-up research observing these limitations and implications would provide a better understanding of successful CR and a more significant basis for developing targeted instructional interventions in order to promote students' CR skills in dealing with new more or less trustworthy or contradictory information.

## DATA AVAILABILITY STATEMENT

The datasets presented in this article will be made available by the authors, without undue reservation, to any qualified researcher. Requests to access the datasets should be directed to troitschanskaia@uni-mainz.de.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## REFERENCES

- Alexander, P. A. (2004). "A model of domain learning: Reinterpreting expertise as a multidimensional, multistage process," in *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development*, eds D. Y. Dai and R. J. Sternberg (New Jersey: Lawrence Erlbaum Associates), 273–298.
- American Educational Research Association [AERA], American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Beck, K. (2020). "On the relationship between "Education" and "Critical Thinking"," in *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*, ed. O. Zlatkin-Troitschanskaia (New York, NY: Springer), 73–87.
- Brand-Gruvel, S., and Stadler, M. (2011). Solving information-based problems: evaluating sources and information. *Learn. Instr.* 21, 175–179. doi: 10.1016/j.learninstruc.2010.02.008
- Bråten, I., Strømso, I. H., and Salmerón, L. (2011). Trust and mistrust when students read multiple information sources about climate change. *Learn. Instr.* 21, 180–192. doi: 10.1016/j.learninstruc.2010.02.002
- Braun, H. (2019). Performance assessment and standardization in higher education: a problematic conjunction? *Br. J. Educ. Psychol.* 89, 429–440. doi: 10.1111/bjep.12274
- Brookfield, S. D. (1987). *Developing Critical Thinkers: Challenging Adults to Explore Alternative Ways of Thinking and Acting*. San Francisco, CA: Jossey-Bass.
- Brooks, C. (2016). *ECAR Study of Students and Information Technology*. Louisville KY: ECAR.
- Chiu, Y.-L., Liang, Y.-C., and Tsai, C.-C. (2013). Internet-specific epistemic beliefs and self-regulated learning in online academic information searching. *Metacogn. Learn.* 8, 235–260. doi: 10.1007/s11409-013-9103-x

## AUTHOR CONTRIBUTIONS

OZ-T provided the idea for the study, co-developed the assessment, supervised the analyses, and co-wrote the manuscript. KB co-developed the assessment, supervised the analyses, and was involved in preparing and reviewing the manuscript. JF and DB conducted the analyses, and were involved in preparing the manuscript. SS was involved in the data collection and in the analyses. RS was involved in the development of the performance assessment and in preparing the manuscript. All the authors contributed to the article and approved the submitted version.

## FUNDING

This study is part of the PLATO project, which is funded by the German federal state of Rhineland-Palatinate.

## ACKNOWLEDGMENTS

We would like to thank the two reviewers and the editor who provided constructive feedback and helpful guidance in the revision of this manuscript. We would also like to thank all students from the Johannes Gutenberg University Mainz who participated in this study as well as the raters who evaluated the written responses.

- Chiu, Y.-L., Tsai, C.-C., and Liang, J.-C. (2015). Testing measurement invariance and latent mean differences across gender groups in college students' internet-specific epistemic beliefs. *Austr. J. Educ. Technol.* 31, 486–499. doi: 10.14742/ajet.1437
- Ciampaglia, G. L. (2018). "The digital misinformation pipeline," in *Positive Learning in the Age of Information*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Wiesbaden: Springer), 413–421. doi: 10.1007/978-3-658-19567-0\_25
- Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., and Wise, L. (2015). *Psychometric Considerations for the Next Generation of Performance Assessment*. Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service.
- Facione, P. A. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction—The Delphi Report*. Millbrae, CA: Academic Press.
- Facione, P. A. (2004). *Critical Thinking: What It Is and Why It Counts*. Millbrae, CA: Academic Press.
- Facione, P. A., Facione, N. C., and Giancarlo, C. A. F. (1998). *The California Critical Thinking Disposition Inventory (CA)*. Cambridge, MA: Academic Press.
- Festinger, L. (1962). Cognitive dissonance. *Sci. Am.* 207, 93–107. doi: 10.1038/scientificamerican1062-93
- Hahnel, C., Kroehne, U., Goldhammer, F., Schoor, C., Mahlow, N., and Artelt, C. (2019). Validating process variables of sourcing in an assessment of multiple document comprehension. *Br. J. Educ. Psychol.* 89, 524–537. doi: 10.1111/bjep.12278
- Harrison, N., and Luckett, K. (2019). Experts, knowledge and criticality in the age of 'alternative facts': re-examining the contribution of higher education. *Teach. Higher Educ.* 24, 259–271. doi: 10.1080/13562517.2019.1578577
- Ho, S., Bossard, D., and Scheufele, D. A. (2008). Effects of value predispositions, mass media and knowledge on public attitudes toward embryonic stem cell research. *Int. J. Public Opin. Res.* 20, 171–192. doi: 10.1093/ijpor/edn017

- Horstmann, N., Ahlgrimm, A., and Glockner, A. (2009). How distinct are intuition and deliberation? An eye tracking analysis of instruction-induced decision modes. *Judgm. Decis. Mak.* 4, 335–354.
- Hsu, C.-H., Tsai, M.-J., Hou, H.-T., and Tsai, C.-C. (2014). Epistemic beliefs, online search strategies, and behavioral patterns while exploring socioscientific issues. *J. Sci. Educ. Technol.* 23, 471–480. doi: 10.1007/s10956-013-9477-1
- Huber, C. R., and Kuncel, N. R. (2016). Does college teach critical thinking? A meta-analysis. *Rev. Educ. Res.* 86, 431–468. doi: 10.3102/0034654315605917
- Johnson, F., Shaffi, L., and Rowley, J. (2016). Students' approaches to the evaluation of digital information: insights from their trust judgments. *Br. J. Educ. Technol.* 47, 1243–1258. doi: 10.1111/bjet.12306
- Kahne, J., and Bowyer, B. (2017). Educating for democracy in a partisan age: confronting the challenges of motivated reasoning and misinformation. *Am. Educ. Res. J.* 54, 3–34. doi: 10.3102/0002831216679817
- Kammerer, Y., and Gerjets, P. (2012). Effects of search interface and Internet-specific epistemic beliefs on source evaluations during Web search for medical information: an eye-tracking study. *Behav. Inform. Technol.* 31, 83–97. doi: 10.1080/0144929X.2011.599040
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50, 1–73. doi: 10.1111/jedm.12000
- Kienhues, D., Stadler, M., and Bromme, R. (2011). Dealing with conflicting or suspect medical information on the web: when expert information breeds laypersons' doubts about experts. *Learn. Instr.* 21, 193–204. doi: 10.1016/j.learninstruc.2010.02.004
- Klein, S., Benjamin, R., Shavelson, R., and Bolus, R. (2007). The collegiate learning assessment: facts and fantasies. *Eval. Rev.* 31, 415–439. doi: 10.1177/0193841X07303318
- Liepmann, D., Beauducel, A., Brocke, B., and Amthauer, R. (2007). *I-S-T 2000 R. Intelligenz-Struktur-Test 2000 R*, 2nd Edn. Göttingen: Hogrefe.
- List, A., and Alexander, P. A. (2017). Cognitive affective engagement model of multiple source use. *Educ. Psychol.* 52, 182–199. doi: 10.1080/00461520.2017.1329014
- List, A., and Alexander, P. A. (2018). "Cold and warm perspectives on the cognitive affective engagement model of multiple source use," in *Handbook of Multiple Source Use*, eds J. L. G. Braasch, I. Bråten, and M. T. McCrudden (New York, NY: Routledge), 34–54. doi: 10.4324/9781315627496-3
- List, A., and Alexander, P. A. (2019). Toward an integrated framework of multiple text use. *Educ. Psychol.* 54, 20–39. doi: 10.1080/00461520.2018.1505514
- Liu, O. L., Frankel, L., and Roehr, K. C. (2014). Assessing critical thinking in higher education: current state and directions for next-generation assessments. *ETS Res. Rep.* 2014, 1–23. doi: 10.1002/ets2.12009
- Lucassen, T., and Schraagen, J. M. (2011). Factual accuracy and trust in information: the role of expertise. *J. Am. Soc. Inform. Sci. Technol.* 62, 1232–1242. doi: 10.1002/asi.21545
- Mason, L., Boldrin, A., and Ariasi, N. (2010). Searching the Web to learn about a controversial topic: are students epistemically active? *Instr. Sci.* 38, 607–633. doi: 10.1007/s11251-008-9089-y
- Mason, L., Pluchino, P., and Ariasi, N. (2014). Reading information about a scientific phenomenon on webpages varying for reliability: an eye-movement analysis. *Educ. Technol. Res. Dev.* 62, 663–685. doi: 10.1007/s11423-014-9356-3
- Maurer, M., Quiring, O., and Schemer, C. (2018). "Media effects on positive and negative learning," in *Positive Learning in the Age of Information (PLATO) – A Blessing or a Curse?*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Wiesbaden: Springer), 197–208. doi: 10.1007/978-3-030-26578-6
- Maurer, M., Schemer, C., Zlatkin-Troitschanskaia, O., and Jitomirski, J. (2020). "Positive and negative media effects on university students' learning: preliminary findings and a research program," in *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*, ed. O. Zlatkin-Troitschanskaia (New York, NY: Springer), 109–119. doi: 10.1007/978-3-030-26578-6\_8
- McGrew, S., Breakstone, J., Ortega, T., Smith, M., and Wineburg, S. (2018). Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory Res. Soc. Educ.* 46, 165–193. doi: 10.1080/00933104.2017.1416320
- McGrew, S., Ortega, T., Breakstone, J., and Wineburg, S. (2017). The challenge that's bigger than fake news: civic reasoning in a social media environment. *Am. Edu.* 41, 4–9.
- Mercier, H., and Sperber, D. (2009). "Intuitive and reflective inferences," in *In Two Minds: Dual Processes and Beyond*, eds J. Evans and K. Frankish (Oxford: Oxford University Press), 149–170. doi: 10.1093/acprof:oso/9780199230167.003.0007
- Mercier, H., and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* 34, 57–74. doi: 10.1017/S0140525X100009
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Edu. Res.* 23, 13–23. doi: 10.3102/0013189X023002013
- Metzger, M. J. (2007). Making sense of credibility on the web: models for evaluating online information and recommendations for future research. *J. Am. Soc. Inform. Sci. Technol.* 58, 2078–2091. doi: 10.1002/asi.20672
- Metzger, M. J., and Flanagin, A. J. (2015). Credibility and trust of information in online environments: the use of cognitive heuristics. *J. Prag.* 59, 210–220. doi: 10.1016/j.pragma.2013.07.012
- Metzger, M. J., Flanagin, A. J., and Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *J. Commun.* 60, 413–439. doi: 10.1111/j.1460-2466.2010.01488.x
- Mislevy, R. J. (2018). *Socio-Cognitive Foundations of Educational Measurement*. New York, NY: Routledge.
- Moore, T. (2013). Critical thinking: seven definitions in search of a concept. *Stud. Higher Educ.* 38, 506–522. doi: 10.1080/03075079.2011.586995
- Münchow, H., Richter, T., and Schmid, S. (2020). "What does it take to deal with academic literature?," in *Student Learning in German Higher Education: Innovative Measurement Approaches and Research Results*, Vol. 2, eds O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, and C. Lautenbach (Wiesbaden: Springer), 241–260. doi: 10.1007/978-3-658-27886-1\_12
- Münchow, H., Richter, T., von der Mühlen, S., and Schmid, S. (2019). The ability to evaluate arguments in scientific texts: measurement, cognitive processes, nomological network, and relevance for academic success at the university. *Br. J. Educ. Psychol.* 89, 501–523. doi: 10.1111/bjep.12298
- Murray, M. C., and Pérez, J. (2014). Unraveling the digital literacy paradox: how higher education fails at the fourth literacy. *Issues in Inform. Sci. Inform. Technol.* 11, 85–100. doi: 10.28945/1982
- Nagel, M.-T., Zlatkin-Troitschanskaia, O., Schmidt, S., and Beck, K. (2020). "Performance assessment of generic and domain-specific skills in higher education economics," in *Student Learning in German Higher Education: Innovative Measurement Approaches and Research Results*, O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, and C. Lautenbach (Wiesbaden: Springer), 281–299. doi: 10.1007/978-3-658-27886-1\_14
- Newman, T., and Beetham, H. (2017). *Student Digital Experience Tracker 2017: the Voice of 22,000 UK Learners*. Bristol: JISC.
- Oliveri, M. E., and Mislevy, R. J. (2019). Introduction to "Challenges and Opportunities in the Design of 'Next-Generation Assessments of 21st Century Skills'" Special Issue. *Int. J. Test.* 19, 97–102. doi: 10.1080/15305058.2019.1608551
- Oser, F., and Biedermann, H. (2020). "A three-level model for critical thinking: critical alertness, critical reflection, and critical analysis," in *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*, ed. O. Zlatkin-Troitschanskaia (New York, NY: Springer), 89–106. doi: 10.1007/978-3-030-26578-6\_7
- Pellegrino, J. (2017). "Teaching, learning and assessing 21st century skills," in *Pedagogical Knowledge and the Changing Nature of the Teaching Profession*, ed. S. Guerriero (Paris: OECD Publishing), doi: 10.1787/9789264270695-12-en
- Pellegrino, J. W. (2020). Sciences of learning and development: some thoughts from the learning sciences. *Appl. Dev. Sci.* 24, 48–56. doi: 10.1080/10888691.2017.1421427
- Schoor, C., Melzner, N., and Artelt, C. (2019). The effect of the wording of multiple documents on learning. *Zeitschrift für Pädagogische Psychologie* 33, 223–240. doi: 10.1024/1010-0652/a000246
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. (2019). Assessment of university students' critical thinking: next generation performance assessment. *Int. J. Test.* 19, 337–362. doi: 10.1080/15305058.2018.1543309
- Stanovich, K. E. (2003). "The fundamental computational biases of human cognition: heuristics that (sometimes) impair decision making and problem



- solving,” in *The Psychology of Problem Solving*, eds J. E. Davidson and R. J. Sternberg (New York, NY: Cambridge University Press), 291–342. doi: 10.1017/cbo9780511615771.011
- Stanovich, K. E. (2016). *The Rationality Quotient: Toward a Test of Rational Thinking*. 1st Edn. Cambridge, MA: MIT Press.
- Strauss, A., and Corbin, J. (1990). *Basic of Grounded Theory Methods*. Beverly Hills, CA: Sage.
- Strauss, A., and Corbin, J. (1998). *Basics of Qualitative Research. Techniques and Procedures for Developing Grounded Theory*, 2nd Edn. Thousand Oaks, CA: Sage Publications.
- Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285. doi: 10.1207/s15516709cog1202\_4
- Ulyshen, T. Z., Koehler, M. J., and Gao, F. (2015). Understanding the connection between epistemic beliefs and internet searching. *J. Educ. Comput. Res.* 53, 345–383. doi: 10.1177/0735633115599604
- van Strien, J. L. H., Brand-Gruwel, S., and Boshuizen, H. P. A. (2014). Dealing with conflicting information from multiple nonlinear texts: effects of prior attitudes. *Comput. Hum. Behav.* 32, 101–111. doi: 10.1016/j.chb.2013.11.021
- van Strien, J. L. H., Bijker, M., Brand-Gruwel, S., and Boshuizen, H. P. A. (2012a). “Measuring sophistication of epistemic beliefs using rasch analysis,” in *The Future of Learning: Proceedings of the 10th International Conference of the Learning Sciences (ICLS 2012) Volume 2. Short Papers, Symposia, and Abstracts*, eds J. van Aalst, K. Thompson, M. J. Jacobson, and P. Reimann (Sydney: International Society of the Learning Sciences), 197–201.
- van Strien, J. L. H., Brand-Gruwel, S., and Boshuizen, H. P. A. (2012b). *Do Prior Attitudes Influence Epistemic Cognition While Reading Conflicting Information? Poster Presented at the Biannual Meeting of the EARLI Special Interest Group Comprehension of Text and Graphics in August 2016. Grenoble (France)*. Available at: [https://www.researchgate.net/publication/254848942\\_Do\\_prior\\_attitudes\\_influence\\_epistemic\\_cognition\\_while\\_reading\\_conflicting\\_information](https://www.researchgate.net/publication/254848942_Do_prior_attitudes_influence_epistemic_cognition_while_reading_conflicting_information) (accessed May 16, 2020).
- van Strien, J. L. H., Kammerer, Y., Brand-Gruwel, S., and Boshuizen, H. P. A. (2016). How attitude strength biases information processing and evaluation on the web. *Comput. Hum. Behav.* 60, 245–252. doi: 10.1016/j.chb.2016.02.057
- Walthen, C. N., and Burkell, J. (2002). Believe it or not: factors influencing credibility on the web. *J. Am. Soc. Infor. Sci. Technol.* 53, 134–144. doi: 10.1002/asi.10016
- Watson, G., and Glaser, E. (2002). *Watson-Glaser Critical Thinking Appraisal – UK Edition*. London: Pearson Assessment.
- Wineburg, S., and McGrew, S. (2017). “Lateral reading: Reading less and learning more when evaluating digital information,” in *Stanford History Education Group Working Paper No. 2017-A1*. Available at: <https://ssrn.com/abstract=3048994> (accessed May 16, 2020).
- Wineburg, S., Smith, M., and Breakstone, J. (2018). What is learned in college history classes? *J. Am. History* 104, 983–993. doi: 10.1093/jahist/jax434
- Wolf, R., Zahner, D., and Benjamin, R. (2015). Methodological challenges in international comparative post-secondary assessment programs: lessons learned and the road ahead. *Stud. Higher Educ.* 40, 471–481. doi: 10.1080/03075079.2015.1004239
- Zahner, D. (2013). *Reliability and Validity-CLA+*. New York, NY: CAE.
- Zahner, D., and Cioffi, A. (2018). “International comparison of a performance-based assessment in higher education,” in *Assessment of Learning Outcomes in Higher Education. Methodology of Educational Measurement and Assessment*, eds O. Zlatkin-Troitschanskaia, M. Toepper, H. Pant, C. Lautenbach, and C. Kuhn (Wiesbaden: Springer), doi: 10.1007/978-3-319-74338-7
- Zlatkin-Troitschanskaia, O., Jitomirski, J., Happ, R., Molerov, D., Schlax, J., Kühling-Thees, C., et al. (2019a). Validating a test for measuring knowledge and understanding of economics among university students. *Zeitschrift für Pädagogische Psychologie*, 33, 119–133.
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., and Beck, K. (2019b). On the complementarity of holistic and analytic approaches to performance assessment scoring. *Br. J. Educ. Psychol.* 89, 468–484. doi: 10.1111/bjep.12286
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., and Pant, H. A. (2018a). “Assessment of Learning outcomes in higher education: international comparisons and perspectives,” in *Handbook on Measurement, Assessment and Evaluation in Higher Education*, C. Secolsky and B. Denison (New York, NY: Routledge).
- Zlatkin-Troitschanskaia, O., Toepper, M., Molerov, D., Buske, R., Brückner, S., and Pant, H. A. (2018b). “Adapting and validating the collegiate learning assessment to measure generic academic skills of students in germany – implications for international assessment studies in higher education,” in *Assessment of Learning Outcomes in Higher Education*, eds O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, C. Lautenbach, and C. Kuhn (Wiesbaden: Springer), 245–266.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Zlatkin-Troitschanskaia, Beck, Fischer, Braunheim, Schmidt and Shavelson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Learning to Fly Through Informational Turbulence: Critical Thinking and the Case of the Minimum Wage

Gerhard Minnameier\* and Rico Hermkes\*

Chair of Business Ethics and Business Education, Faculty of Economics and Business Administration, Goethe University Frankfurt am Main, Frankfurt, Germany

## OPEN ACCESS

### Edited by:

Olga Zlatkin-Troitschanskaia,  
Johannes Gutenberg University  
Mainz, Germany

### Reviewed by:

Henry Braun,  
Boston College, United States  
Lucia Mason,  
University of Padua, Italy

### \*Correspondence:

Gerhard Minnameier  
minnameier@econ.uni-frankfurt.de  
Rico Hermkes  
hermkes@econ.uni-frankfurt.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Education

**Received:** 15 June 2020

**Accepted:** 03 September 2020

**Published:** 24 September 2020

### Citation:

Minnameier G and Hermkes R  
(2020)  
Learning to Fly Through Informational  
Turbulence: Critical Thinking  
and the Case of the Minimum Wage.  
Front. Educ. 5:573020.  
doi: 10.3389/feduc.2020.573020

The paper addresses online reasoning and information processing with respect to a much debated issue: the pros and cons of the minimum wage. Like with all controversial issues, one can easily remain in a self-reinforcing bubble, once one has taken sides, and immunize oneself against criticism. Paradoxically, the more information we have at our disposal, the easier this gets (Roetzel, 2019). The only (and possibly universal) antidote seems to be “critical thinking” (Ennis, 1987, 2011). However, critical thinking is a very broad concept, purported to include diverse kinds of information processing, and it is also thought to be content-specific. Therefore, we aim at addressing both understanding of content knowledge and reasoning processes. We pursue three goals with this paper: First, we conduct a conceptual analysis of the learning content and of reasoning patterns for and against the minimum wage. Second, we explicate an inferential framework that can be applied for processes of critical thinking. Third, teaching strategies are discussed to support reasoning processes and to promote critical thinking skills.

**Keywords:** critical thinking, inferential processes, abduction, argumentation, online reasoning, multiple-document comprehension, cognitive conflict

## INTRODUCTION

In the digital age, online reasoning and the processing of information from multiple sources with contradictory viewpoints are of outstanding importance. Instead of immunizing oneself against criticism and other viewpoints, one should on the one hand be open-minded. On the other hand, one should not blindly follow the opinions of others and resist manipulative information and campaigns. Critical thinking addresses this ability and is regarded as one of the key twenty first century learning and innovation skills (OECD). Oser (2018) sums it up pointedly by stating that “(c)ritical thinking is seen as a means for guarding against fake news in its psychological and emotional dimension” (p. 368). However, critical thinking is a very broad concept, purported to include diverse kinds of (generic) information processing activities, and it is also thought to be content-specific (see Ennis, 1987, 2011; Tarchi and Mason, 2020).

How do critical thinking skills manifest themselves *in situ*? As Dormann et al. (2018) state, there is a consensus on the processes involved in critical thinking. The authors write:

The consensus list of critical thinking skills “[is]” comprised of six skills (sub-skills in parentheses): interpretation (categorization, decoding significance, clarifying meaning), analysis (examining ideas, detecting arguments, analyzing arguments), evaluation (assessing claims, assessing arguments), inference (querying evidence, conjecturing alternatives, drawing conclusions), explanation (stating results, justifying procedures, presenting arguments), and self-regulation (self-examination, self-correction) (p. 329f).

Consensus is fine. However, this structured list also reveals that there is still a need to further disentangle and differentiate the processes at work (as they obviously shade into each other and lack clear boundaries). Moreover, it shows that critical thinking cannot be limited to purely logical-analytical processes. According to Hitchcock (2017), “some critical thinking, but not all, is logical analysis of argument. In thinking critically, we not only want to find out, if a single piece of reasoning or argument is good or bad. We also want to know more about its context and see it in a broader framework of alternative choices, ways or options. We want to trace the best path toward our understanding of a problem and make the best decision about it. We also look at the extent to which all our judgments and decisions are supported by evidence while examining as well the quality of this evidence” (p. 484). In this respect, an essential property of critical thinking is the embedding of arguments in an overall context.

Shavelson et al. (2019) differentiate between two types of contexts: (i) everyday life contexts, in which decision making- or problem solving-processes take place and (ii) argumentative contexts, in which one’s own positions are to be developed and justified. Such argumentative contexts are largely formed by other people. For this reason, critical thinking is often closely associated with multiple-text-comprehension (see Stadler and Bromme, 2013; Richter and Maier, 2017; Tarchi and Mason, 2020).

A suitable starting point for a systematization is offered by da Silva Almeida and Rodrigues Franco (2011). They state, that critical thinking “is a multifaceted cognitive construct, with an inductive, deductive and creative nature” (p. 179). On the one hand, this aims at inferences as suitable “candidates” for processes of critical thinking. On the other hand, it also goes beyond a logical-analytical framework, as Hitchcock puts it. The problem here is that “logic” is equated with deduction. However, inferences should be understood in the context of the rather new and wider program of the naturalization of logic, which is not limited to classical (deductive) logical approaches, but also includes processes of practical reasoning and eco-logical judgments in real-world contexts (Gabbay and Woods, 2005; Magnani, 2009, 2018; Minnameier, 2019). In particular, it includes abduction and induction and how they, together with deduction, shape reasoning processes in general. In this sense, logic counts as the theory of right reasoning. Understood in this way and in

this wider sense, inferential reasoning is what critical thinking is mainly about (apart from non-cognitive factors that concern, e.g., self-regulation).

This is particularly important, because not only are the nature and the amount of information processed constantly changing, but so is the world, too. Consequently, the need to explain world phenomena is also evolving continuously. Finally, this also accounts for the truth of propositions about the world. As Magnani puts it, an abductive “inferential problem can be enhanced by the emergence of new information in a temporal dimension that favors the restarting of the inferential process itself” (p. 12). In this respect, as the title suggests, one can speak of “informational turbulences,” through which critical thinking has to maneuver.

The paper is structured in three parts. In the first part, we conduct a conceptual analysis of the learning content in the case of the minimum wage. In the second part, we explicate an inferential framework that can be applied for assessing processes of critical thinking. We reconstruct typical reasoning patterns for and against the minimum wage and identify problems of reasoning. In the third part, teaching strategies are discussed to promote critical thinking.

## THE CASE OF THE MINIMUM WAGE

### Content Analysis

Minimum wages exist in many countries around the world, especially in Anglo-American and European countries, but not everywhere, and they have always been, and still are, highly debated. Opponents think it undermines the market mechanism, increases poverty and unemployment among low-skilled workers and threatens businesses that cannot afford the higher costs induced by the minimum wage. Conversely, supporters believe that it increases the standard of living for those workers, reduces inequality and poverty and therefore brings about more justice in income.<sup>1</sup> One of the last countries having introduced the minimum wage so far is Germany, where it was enforced in 2015, starting with a rather high level of €8.50, which was gradually increased to currently €9.35 and is about to be increased further to €9.80 in January 2021.

For its supporters, the minimum wage is hailed as a kind of universal antidote to all evils of global capitalism, which affects poorly skilled people in developed countries, because their jobs or job opportunities move toward emerging (or rather emerged) economies, in particular in Asia, while the well-educated in upper income segments keep on benefitting. In modern Western societies, the widening gap between rich and poor is perceived as a case of injustice and as a huge social and economic problem, which in some sense it certainly is. Unskilled work can be done by anyone, and if labor is cheaper in other parts of the world, those jobs move away from affluent, high-wage to low-wage countries, leaving the not so well-off in the rich countries behind.

<sup>1</sup><https://worldpopulationreview.com/countries/minimum-wage-by-country/>

This is a fact. However, it is also a normal and natural process in a market economy. And even though the social problem cannot and should not be denied, low-wages indicate a gap between supply and demand of unskilled or low-skilled work, which cannot and should not be denied either. It could be understood as a sign that especially young people, who have to think about how to make a living in their future working-lives, have to orient themselves toward job opportunities as well as the skills and education that are needed to be able to grasp those opportunities.

As far as the long-term unemployed are concerned, they certainly need support. Hence, it is not the question, whether society has to do something about their situation, but *how* this should be done. Establishing a minimum wage is one way, subsidizing low income from labor with additional transfers is another. The latter strategy has been favored mostly by economists, the former by the German trade union federation (DGB)<sup>2</sup> in their campaign for the minimum wage, in which they championed ten key arguments<sup>3</sup> of which three are of particular interest:

- Minimum wages would prevent “wage poverty” and make sure that workers do not depend on additional subsidies.
- Minimum wages would relieve the federal budget, because it is the duty of companies to provide high enough wages, not the duty of the government to support workers.
- Minimum wages would ensure justice by stopping the downward spiral of wages.

These arguments concern both positive and normative aspects. From a *normative* point of view, it is argued that poverty should not be understood in terms of total income, but in terms of earned wages. According to this (new) concept of “wage poverty,” transfer payments do not count; it is rather the earned wages as such that should get workers above the culturally agreed minimum livelihood. If someone works fulltime and delivers decent work, he or she should earn a decent wage (in accordance also with the third argument).

This line of reasoning is problematic insofar as in a market economy, prices are not meant to be just. They are meant to be efficient (while the issues of justice and efficiency are systematically decoupled). The concept of a “just price” relates to ancient and scholastic conceptions of the economy and simply does not fit into the modern notion of markets, where prices basically have a steering function. They should indicate where to move productive resources. High prices indicate scarcity of the respective goods and services, while low prices indicate overabundance. Fairness, for its part, is provided in two ways: first by setting rules against exploitation, child labor and so forth that apply to all and are built into the market order, second by redistributing income through taxation and subsidies.

As to the *positive* analysis, it is a market-economic truism, that defining a lower bound for prices (here: wages) reduce demand (here: for low-skilled labor), unless price elasticity is zero or

close to zero. In principle, demand is therefore negatively affected by minimum prices above the market price. This is true, even if in reality negative employment effects may not show up or may not be significant. High growth rates may compensate the negative effect, and if the elasticity of demand is (close to) zero, a minimum wage may merely set a new reference point without any negative employment effect. Furthermore, minimum wages might not have detrimental effects, if low wages only result from a low bargaining power of the workers and where jobs cannot be displaced (as e.g., in the case of waiters and hairdressers).

Hence, the crucial question is not, whether a person is in favor of or opposed to minimum wages, but how she construes the pros and cons and, in particular, whether and to what extent they. . .

- misconstrue the causal effects of minimum prices in general, with the possibly paradoxical result that the creation of much needed jobs in the low-wage sector is disincentivized.
- Confound means and ends, because the minimum wage is meant to solve the *justice problem* of wealth distribution and the allocation of income within the market framework, while markets are meant to solve the *efficiency problem* of wealth creation and the allocation of factors of production.

The first fallacy gets the causal relations wrong, the second is an example of a category error. Both errors are possibly fatal, because they could take us to jump out of the frying pan into the fire, i.e., to create more harm out of good intentions. As for Germany, the number of workers who have to be subsidized in addition to their wage income has only been slightly reduced, because only 3 per cent of those who receive the minimum wage are fulltime working singles. And while the hourly wages have been augmented, working hours have been diminished so that a 14 percent increase in terms of hourly wages in the relevant group results in only a 4 percent increase in monthly wages. While employment remained largely stable or even increased slightly between 2016 and 2017 in industries not affected by the minimum wages, it was reduced significantly in those affected by the minimum wage (Mindestlohn-Kommission, 2018). By and large the effects so far have been only moderate, but as we know from recent research, this is partly due to substantial non-compliance. Out of a total of roughly 4 million workers who are eligible for the minimum wage, 750,000 are paid less than the minimum wage (see Caliendo et al., 2019). Based on calculations of the true hourly wage that is paid, the German Institute for Economic Research (Fedorets et al., 2020) even reports that we end up with 2.4 million workers who are paid below the minimum wage in their main occupations,<sup>4</sup> even though they concede that wage inequality has declined in Germany since 2006.

## Thinking Critically About the Minimum Wage

As one can easily see, the task critical thinkers with an economics background face is manifold. Concerning the status quo, they first

<sup>2</sup>The abbreviation stands for “Deutscher Gewerkschaftsbund.”

<sup>3</sup><https://www.dgb.de/schwerpunkt/mindestlohn/hintergrund/argumente>

<sup>4</sup>If sideline jobs are included, the number rises to 3.8 million employees paid below the minimum wage (Fedorets et al., 2020).



have to reconstruct, analyze, and evaluate the reasoning that had led to the introduction of the minimum wage. Four decisive parts can be extracted in this respect:

1. Identifying the underlying problem and its epistemic domain: positive (explanatory), prescriptive (instrumental), or normative (ethical).
2. Understanding the minimum wage as a possible solution to this problem.
3. Deriving consequences based on background knowledge or assumptions.
4. Gathering evidence, weighing pros and cons, and judging whether the minimum wage is an acceptable solution or not.

(1) It should be clear that the basic problem at issue is that wages are perceived to be too low from an ethical point of view. While it is questionable whether the problem persists in the light of additional transfer payments by the government, at least for the DGB the fundamental issue is wage poverty, which is held to be unjust. At least for the DGB this is taken for granted, so that they do not see an *ethical* problem of how to determine what is just, but an *instrumental* one of how to *implement* justice, as they understand it. Ennis (1996) calls this step identifying the focus. Jenicek and Hitchcock (2005) refer to this step as problem identification and analysis (see Hitchcock, 2017).

(2) Setting a minimum wage is the straightforward answer to this problem. While there might be other possible solutions (including the pre-existing one of topping up wages by governmental subsidies), wages lower than the minimum wage will certainly be pushed up, as long as employers comply with the rule.

(3) Apart from rising hourly wages, critical thinkers might infer that employers could try to formally reduce working hours (while the overall workload for each worker remains essentially the same), or that they might reduce jobs as a reaction to higher costs. Conversely, they might reckon that if margins for employers are sufficiently high, no jobs would be lost, and that only the overall surplus would be divided differently between workers and employers.

(4) Based on the aspects taken into account and the evidence gathered in their respect, beliefs have to be formed or updated, which ultimately take reasoners to their final conclusion. Accordingly, they might speak out for or against the minimum wage, or they might remain indifferent.

Whatever they think after this analysis, they will either remain with or encounter a new problem, when they are exposed to the report of the minimum wage commission. *Advocates* of the minimum wage must face the (possible) problem of *inefficiency* owing to non-compliance and job losses. Hence, they will find themselves in a situation where the problem of implementing “just wages” is either not solved (owing to various forms of non-compliance) or that a new *technological* problem emerges, namely that of creating new jobs or preventing job losses. Opponents may find that this problem is automatically solved, if the minimum wage is withdrawn. However, they face a problem of injustice and see themselves in a situation, where they have to be able to offer a solution to solve the original problem in a different way (other

than the minimum wage). The identification and specification of this problem is the crucial step that has to be taken by both advocates and opponents of the minimum wage.

## AN INFERENTIAL FRAMEWORK FOR PROCESSES OF CRITICAL THINKING

### Reasoning-Based Dimensions of Critical Thinking

The four parts of CT explained in section “Thinking Critically About the Minimum Wage” correspond by and large to the reasoning-related dimensions described in Zlatkin-Troitschanskaia et al. (2019): *Recognizing and evaluating arguments and making decisions*<sup>5</sup> and *recognizing and evaluating the consequences of decisions*. An inferential reconstruction can further clarify the meaning and different aspects of processes like “decision making.” In addition to the reasoning-related dimensions, there are also dimensions of critical thinking, which concerns the research and evaluation of information from multiple sources and the examination of sources [e.g., *recognizing and evaluating information* (Zlatkin-Troitschanskaia et al., 2019) or *location, evaluation and integration/synthesis of information* (see Leu et al., 2014)]. The skills related to these dimensions are particularly relevant if the task involves information from online sources, which do not have to undergo an editorial process and where the sources of the information as well as the distributors of the information may be unknown.

In the present article, however, we focus only on reasoning-related dimensions, since the conceptualization and assessment of the reasoning-related dimension of CT requires further clarification, which can be delivered by the inferential framework (as we try to show in the following subsections). First, the inferential taxonomy provides a unified framework by which CT-relevant cognitive activities (like interpretation, analysis, inference, evaluation, explanation etc.; see Dormann et al., 2018) can be subsumed under specific inferences. Second, distinctive skill categories can be defined. For instance, the category “inference” (sensu Dormann et al., 2018), which includes *querying evidence*, *conjecturing alternatives*, *drawing conclusions*, can be specified: *Querying evidence* is part and parcel of induction, *conjecturing alternatives* concerns abduction,<sup>6</sup> while *drawing conclusions* is a particular step in all kinds of inferences (see section “Inferential Processes”). In effect, the individual skills can be localized in the dynamics of an inferential cycle, which allows the explication of inaccurate cognitive activities as errors of reasoning. Finally, validity conditions for each inference can be specified. In the case of induction, the validity criterion also depends on the domain of reasoning. Thus, validity conditions can also be differentiated by domain (positive, prescriptive, or normative). As a consequence, the assessment of CT can be based

<sup>5</sup>Note, that Zlatkin-Troitschanskaia et al. (2019) merged dimension “recognizing and evaluating information” and dimension “recognizing and evaluating arguments and making decisions” into one dimension after statistical analyses.

<sup>6</sup>da Silva Almeida and Rodrigues Franco (2011) denote this as the creative facet (see section “Introduction”).

on the inferential processes individuals undergo and the extent to which their conclusions are warranted.

The inferential framework applies to individuals' reasoning processes such as the derivation of implications and consequences or the development of sound arguments (subsumed as synthetic dimensions of CT by Liu et al., 2014) as well as the comprehension and evaluation of arguments inferred by others (analytical dimensions of CT; *ibid.*).

## The Inferential Triad

As opposed to the "consensus list of critical thinking skills," which was reported in the introduction, we believe that *all* critical thinking is inferential in some sense and takes place in the context of certain inferences. It has to be taken into account, of course, that Dormann et al. (2018) understanding of "inferences" is narrower than the present one and only pertains to evidential reasoning. Conversely, the inferential theory of learning can be applied to all issues of analysis, evaluation, and problem solving, and integrate these issues systematically within an overall context of inferential reasoning. As a consequence, different aspects of critical thinking are then understood and addressed as different aspects of inferential reasoning as such.

In the following section, we present this inferential framework and explicate the mental processes underlying critical thinking in detail. This yields a view of critical thinking processes that is both differentiated and integrated, and that allows us to account for *correct* and *incorrect* reasoning with respect to specific reasoning processes. It also allows us to distinguish different kinds of fallacies as well as specific clues for constructive support. And it provides general guidelines for the initiation of reasoning processes and the promotion of critical thinking skills.

The inferential theory of learning rests on C. S. Peirce's pragmatist theory of knowledge creation and the inferential processes he distinguishes. In particular, Peirce has introduced the concept of abduction to modern epistemology and philosophy of science and created a whole theory of how knowledge is acquired from the first perception of a problem to its solution. The following model of these inferences was first suggested in Minnameier (2004). It shows that the three inferences of abduction, deduction, and induction form a recursive triad (Figure 1).

The triad begins at  $t_0$  with the surprising problem, and matches very well with Peirce's description in (CP 5.171).<sup>7</sup> The surprising facts then call for an explanation, and abduction (at least in explanatory abduction) describes the process of developing *potential* explanations, which are subsequently examined by deduction and induction.

What is particular about this model is that the inductive arrow points back to the starting point, rather than to the

theory (which is either accepted or rejected as the outcome of induction). The reason is that induction cannot "prove" a theory, even if the evidence is decisive. Therefore, the acceptance of a theory is interpreted in the sense of projecting the content of the theory onto its cases – the original surprising one, the tested ones, and the untested ones at present, in the past, and in the future. This implies that any theory is implicitly evaluated each time it is applied.

To be sure, the reasoning process as such starts even one step before abduction, so as to produce surprise (or a cognitive state equivalent to surprise). The "surprise" in this model comes as a "negative induction," which reveals that something expected is not actually the case. In other words, the theory and our current observations decohere. This decoherence is what generates a new abductive problem, and a successful abduction is understood as re-establishing coherence.

With regard to critical thinking the focus is on identifying problems and errors of reasoning, rather than the creative search for solutions that has just been mentioned in the preceding paragraph. However, the whole triad is important for critical thinking, because in order to reflect on minimum wages, critical thinkers have to comprehend the original problem minimum wages are meant to solve, understand how they should function, be able to derive deductive consequences based on their background knowledge and to evaluate that pros and cons based on what we know or on future empirical research to be carried out.

## Inferential Processes

Inferences, whether abductive, deductive, or inductive, are cognitive processes with a definite beginning and a definite end. If we take, e.g., deduction as the most common inference, it starts from putting together a couple of premises, from which one then tries to derive necessary consequences, i.e., results that are implicit in the premises. Once we identify candidates for such results we have to judge, whether they really follow from the premises (because the moment of finding a result as such is a spontaneous event). If the result is judged valid, the process is terminated. Thus, the bare-bone structure of any inference consists of three distinctive steps that Peirce calls "colligation," "observation," and "judgment" (CP 2.442–444 [c. 1893]).<sup>8</sup>

<sup>8</sup>"The first step of inference usually consists in bringing together certain propositions which we believe to be true, but which, supposing the inference to be a new one, we have hitherto not considered together, or not as united in the same way. This step is called *colligation*. The compound assertion resulting from colligation is a *conjunctive proposition* . . . Colligation is a very important part of reasoning, calling for genius perhaps more than any other part of the process" (CP 2.442).

"An inference, then, may have but a single premiss, or several premisses may be united by colligation. In the latter case, they form, when colligated, one conjunctive proposition. But even if there be but one premiss, the icon of that proposition is always more or less complex. The next step of inference to be considered consists in the contemplation of that complex icon, the fixation of the attention upon a certain feature of it, and the obliteration of the rest of it, so as to produce a new icon" (CP 2.443). "It thus appears that all knowledge comes to us by observation" (CP 2.444).

"Whenever one thing suggests another, both are together in the mind for an instant. In the present case, this conjunction is specially (*sic!*) interesting, and

<sup>7</sup>"Abduction is the process of forming an explanatory hypothesis. It is the only logical operation which introduces any new idea; for induction does nothing but determine a value, and deduction merely evolves the necessary consequences of a pure hypothesis. Deduction proves that something *must* be; Induction shows that something *actually* is operative; Abduction merely suggests that something *may* be. Its only justification is that from its suggestion deduction can draw a prediction which can be tested by induction, and that, if we are ever to learn anything or to understand phenomena at all, it must be by abduction that this is to be brought about" (CP 5.171 [1903]).

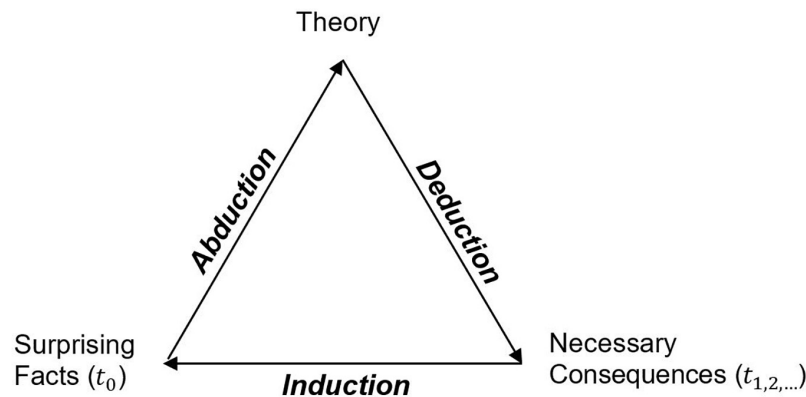


FIGURE 1 | The inferential triad.

Accordingly, any inference starts from the *colligation* of certain premises, from which conclusions are to be drawn. These premises are observed so as to produce some result and answer a specific question, which is the target of the inference. However, since every such result first springs to our minds spontaneously in the process of *observation*, it has to be followed by a *judgment* to make it a conclusion and in order to prevent spurious thoughts and mistakes. Such judgments are not to be misunderstood as meta-reflections or rule-following, but merely as satisfying the mind that the premises sanction the conclusion. Again, for abduction this means that the conclusion removes the incoherence inherent to the premises. If it does, the abduction is valid. A deductive inference is valid, if the conclusion is implied by the premises. Finally, validity of an inductive inference means that not only the empirical tests confirm the underlying theory, but that possible rival theories can also be rejected. As to this general understanding of inference (see also Stalnaker, 1987; Woods, 2013, 2017; Valaris, 2017; Hofmann, 2019).

Every inference must have a target, for without such a target it would be useless to engage in making the inference anyway. The target of abduction is to remove epistemic surprise and produce coherence. The target of deduction is to determine what follows necessarily from certain premises or assumptions by revealing their implications. The target of induction is to determine whether the underlying theory or action plan it to be accepted or rejected, albeit perhaps preliminary. The latter is also known as the “inference to the best explanation” (IBE) (Lipton, 2004; see also Minnameier, 2004, 2017, 2019; Yu and Zenker, 2018).

Induction, or IBE, leads to projecting the content of the theory onto all relevant cases (past, present, or future). And this is true in the positive case, where we accept the theory as well as in the negative case, where we just project its negation onto

the cases (meaning that what the theory assumes is not true of those cases). Inductive judgments are always preliminary in the sense that every application of the theory might just as well challenge the theory as it might corroborate it. This is the essence of pragmatism. However, inductive judgments might also be preliminary in the sense of a tentative decision, especially when we are under pressure to act and have to make choices based on weak evidence.

## Domains of Reasoning

Extending Peirce’s analysis of explanatory reasoning, we can conceive of inferential reasoning in non-explanatory domains (Gabbay and Woods, 2005), in particular in the domains of *prudential* (or instrumental, strategic or technological)<sup>9</sup> reasoning and *ethical* reasoning (Minnameier, 2017). The latter two forms might both be called normative, but in two different respects. *Prudential* reasoning concerns the question of how to reach a certain goal effectively. It looks for suitable means to reach specified ends. Conversely, *ethical* reasoning focuses on determining what would be suitable ends to pursue. Quite often we conflate these two distinct forms within the concept of *normative* statements as opposed to *positive* statements. The positive/normative-distinction is so commonplace, today, that we often fail to clearly separate instrumental reasoning from ethical reasoning. However, this is of vital importance with respect to the present subject matter, i.e., minimum wages, where both aspects are central, yet have to be kept strictly separate.

To our knowledge, the idea that the simple and strict positive/normative dichotomy is erroneous goes back to Putnam (2002). One of his main claims is “that the activity of justifying factual claims presupposes value judgments” (p. 137). In particular, every acceptance of a theory – by way of induction, as we might say – hinges on criteria for evaluation that we

in its turn suggests that the one necessarily involves the other. A few mental experiments. . . satisfy the mind that the one icon would at all times involve the other. . . Hence the mind is not only led from believing the premiss to judge the conclusion true, but it further attaches to this judgment another – that *every* proposition *like* the premiss, that is having an icon like it, *would* involve, and compel acceptance of, a proposition related to it as the conclusion then drawn is related to that premiss. [This is the third step of inference.]” (CP 2.444; emphasis).

<sup>9</sup>The concept of “prudence” is a technical term in philosophy. In particular, it is used to distinguish “moral” or “ethical” judgments from “prudential” ones (cf. e.g., Crisp, 2018). Both suggest courses or action, but the latter ones relate only to the interest of the agent. In everyday language they are also called “instrumental” or “strategic.” “Technological” can also be used as a synonym, if its meaning is not restricted to the use of some kind of machinery, but covers means-end reasoning and optimization in general.

use (whether we champion simplicity, coherence, predictive validity, or whatsoever, which are all *values* (cf. *ibid.*, p. 142). Whichever criterion is chosen, we end up assigning a truth-*value* to the theory we *evaluate*. And within the frame of reference of positive – or call it “descriptive” or “explanatory” – reasoning and research we follow the regulative idea of truth.

Accordingly, we can distinguish three domains of reasoning that we could label

- (i) positive/explanatory,
- (ii) prescriptive/instrumental, and
- (iii) normative/ethical

While positive (or explanatory) reasoning is guided by the regulative idea of truth, prescriptive (or instrumental) reasoning is guided by the regulative idea of “efficiency” or “effectiveness”,<sup>10</sup> and normative (or ethical) reasoning by that of “justice” (or the “good life,” in general).

With respect to the problem at hand it has already been pointed out that the minimum wage can be analyzed and evaluated in terms justice as well as in terms of efficiency, and that critical thinkers would have to consider both independently. However, as far as we know, they do not seem to differentiate much. A study comparing economists’ and laypeople’s evaluation of labor market policies (Haferkamp et al., 2009) has revealed that most laypeople favor policies like the establishment of a general minimum wage and consider them both fair (more than 75%) and efficient (more than 50%), while almost all people with an economic background regard them as unfair and inefficient. Hence, a crucial element of critical thinking with respect to the minimum wage is to distinguish between the aspect of normativity (i.e., what one wants to achieve) and implementation (i.e., how it can best be achieved). Justice relates to the former, efficiency to the latter, because efficiency in terms of growth and economic prosperity provides the basis for a normatively motivated redistribution of wealth or resources.

## The Minimum Wage Task and Analysis

In our study, we aim at describing critical thinking in terms of cognitive skills, as they are captured by the inferential learning theory. While this also encompasses forms of tacit knowing, like e.g., intuitive decision-making (see Hermkes, 2016; Minnameier, 2019), we concentrate on deliberative thinking in present study. Accordingly, we do not focus on issues like biased information and framing effects, but try to find out, how students actually think and reason based on the information presented to them. In particular, the strategy is to present mutually incompatible accounts of a certain subject matter, which in our case consists in the pros and cons of a general minimum wage. Critical thinkers

in our sense have to understand, analyze, and evaluate conflicting views on this particular issue.

The participants are Bachelor and Master students of economics. Bachelor students should already have successfully completed the introductory courses in economics (3rd semester and beyond). When designing the task, we are guided by the investigation of Zlatkin-Troitschanskaia et al. (2019): The minimum wage task is designed to be presented on a computer in a controlled setting. The stimulus material includes the following documents: (1) ten arguments pro minimum wage of the DGB, (2) a reply to the arguments, comprising ten counter-arguments, (3) report of the German commission on the minimum wage (Mindestlohn-Kommission, 2018), (4) chapters from a standard textbook on economics, including the following contents: Market forces of supply and demand, elasticity, efficiency of markets, labor market theory (pdf documents). Moreover, students are allowed to search for information on the internet. The response format is a written statement. The argumentation serves as a data basis for the rating. The task is processed as follows: First, students are presented the ten arguments of the DGB in favor of the minimum wage and asked to evaluate the claims critically. This analysis probes their critical thinking in terms of distinguishing domains (justice versus efficiency).

Next, students are confronted with the 2018 report of the German commission on the minimum wage (Mindestlohn-Kommission, 2018) and allowed to search the internet for additional information. Their task at this stage is to determine what speaks in favor of the minimum wage and what against. Inasmuch as students are critical thinkers, they should not just decide for or against and then prop up and preserve their view (in the sense of immunizing it against counter-arguments), but should rather be capable of addressing the fundamental tradeoff between justice and efficiency that is at the heart of the debate.

The advantage of this tradeoff-constellation is not only that critical thinkers can prove how they not just immunize their own view, but take up valid critical aspects advanced by their opponents. Therefore, it does not really matter, whether an individual is for or against the minimum wage, because the situation for critical thinking would roughly be the same, just that the problem be inverse with respect to the two sides of the tradeoff:

- Those in favor of the minimum wage mainly focus on the justice aspect. However, they have to see and face the problem of inefficiency, that the minimum wage generally crowds out jobs (where more would be needed), or might entail non-compliance if it cannot be fully enforced.
- Those against the minimum wage focus on the efficiency aspect and the functioning of the market mechanism, but have to acknowledge and address the challenge of injustice.

In both cases, performances can be reconstructed as an inferential cycle. The inferential processes can be assessed according to the “reasoning”-related dimensions of CT (recognizing and evaluating arguments and making decisions; recognizing and evaluating the consequences of decisions). According to the four-part model of critical thinking explicated in section “Thinking

<sup>10</sup>There is a terminological problem, because “efficiency” has different uses. Here it is meant in the sense of means-end relationships where efficiency refers to optimality in terms of getting the most out of a specific input or reaching a pre-specified output with the minimum input. However, economists also use “efficiency” in terms of the best outcome for a group of people (as for instance in welfare economics). In this latter sense, “efficiency” belongs to normative reasoning in the context of the regulative idea of the good life (for a group of people), although not in the sense of justice. Hence, the “good life” is the broader notion, referring to either individuals, a group of individuals or from a transindividual point of view, where only the latter concerns “justice.”



Critically About the Minimum Wage,” students’ reasoning includes

- (1) identifying the domain in which the problem to be solved is located,
- (2) understanding to what extent a minimum wage can be a possible solution,
- (3) seeing consequences for the labor market or related to social justice,
- (4) evaluating them in terms of justice and efficiency.

From an inferential point of view, both can be understood as abductively inferred conclusions, which represent solutions for a technological problem. The only difference is that the former try to find a technological solution to implement their idea of just wages, while the latter try to solve a problem of efficiency with respect to jobs and the economy in general.

From their respective theoretical points of view, both will have to consider economic theory as background knowledge and deduce consequences from them (based on what they know and what they find in the materials). Both advocates and opponents will finally have to address the above-mentioned difficulties in the inductive stage and acknowledge that there clearly are drawbacks that have to be addressed.

Regarding the processes of inferential reasoning explained in section “Inferential Processes,” the abductive, deductive and inductive inferences can be described in more detail. Students have to *colligate* the content of the respective problem, in this case the justice problem, and understand the minimum wage as a solution to this problem (judgment step in abduction). From here, the critical evaluation of the minimum wage starts. In their deductive examination, they have to *colligate* not only the result of abduction (the problem and the minimum wage as solution), but also relevant information from the materials and from their previously acquired background knowledge. From this they have to derive – by way of *observation* and *judgment* – the various scenarios that follow from the respective assumptions. If given in the materials, they just have to comprehend these argumentations. The inductive part consists of *colligating* these deductive consequences as well as available empirical data or established facts and evaluating the minimum wage based on the evidence. *Observation* relates to taking notice of all the pros and cons and the probabilities of respective events or outcomes. In particular, it includes attending to the trade-off between justice and efficiency in the light of the evidence available. Finally, they have to *judge* whether to accept or reject the (general) minimum wage, or whether and why it has to remain an open question at the current state of affairs.

## TEACHING STRATEGIES TO PROMOTE CRITICAL THINKING – THE EVOCATION OF COGNITIVE CONFLICTS

The inferential framework can be used as a foundation for the assessment of critical thinking skills because it reveals the cognitive deep structures that underlie the processes of

argumentation and the students’ engagement with the arguments of others. In addition to its relevance for the assessment of critical thinking, knowledge of cognitive deep structures can also be helpful when focusing on questions about appropriate teaching strategies. Abrami et al. (2008) review the effectiveness of instructional interventions to promote critical thinking skills. 117 studies with approximately 20,000 participants were examined for this purpose. As the authors summarize, critical thinking does not develop in an incidental manner alone, but requires appropriate teaching strategies and instructional methods. This accounts for the acquisition of critical thinking skills as well as for triggering of critical thinking *in situ*.

A suitable strategy to trigger critical thinking is the evocation of cognitive conflicts as explained above with respect to the minimum wage issue. Moreover, cognitive conflicts can serve to counteract the immunization of one’s own position and can be an occasion to focus on the positions of other parties and to appreciate their arguments. This strategy can be placed in the context of constructivist learning theories, which consider disturbances to be the driving force for learning processes. Its appropriateness becomes obvious when one considers the explanations of critical thinking in the VALUE rubrics of the AAC&U.<sup>11</sup> Critical thinking is defined there as “the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion.”<sup>12</sup>

From a constructivist point of view, it can be argued that conceptual changes and the questioning of incoming “data” need an occasion. In Piagetian terms, what is needed is a disequilibrium between *assimilation* and *accommodation*. Of course, dispositions such as open-mindedness or inquisitiveness (see Facione, 2000) can play a role and explain why some are generally more alert and inquisitive than others. Nevertheless, it would be erroneous to expect that each and every content presented would be generally doubted and called into question. For this to happen, events are needed that have the capacity to trigger critical thinking, which it does if it entails a disequilibrium in the Piagetian sense. This is where cognitive conflicts come into play.

Whether critical thinking skills are applied in specific situations depends – apart from the motivation to use them<sup>13</sup> – on background knowledge and prior beliefs about the topic (Tarchi and Mason, 2020). Prior beliefs<sup>14</sup> moderate the effect of critical thinking skills on the quality of arguments in the context of multiple text comprehension. Belief consistent information has a higher probability of being colligated, which may result (i) in a biased situational (mental) model in favor of the existing beliefs (see Maier and Richter, 2013) and (ii) in poor judgments. This is referred to as the “belief-consistency effect.” In the case of minimum wages, students may, for example,

<sup>11</sup><https://www.aacu.org/value/rubrics/critical-thinking>

<sup>12</sup>In inferential terminology, “comprehensive exploration” includes the abductive and deductive inferences, while “accepting an opinion or conclusion” refers to the final inductive judgment.

<sup>13</sup>For the motivational component of critical thinking see Facione et al. (1997), da Silva Almeida and Rodrigues Franco (2011); for the effects of motivation on students’ performance see Liu et al. (2015), Braun (2019).

<sup>14</sup>E.g., about effects of vaccination.

believe that the government is responsible for regulating markets when they cause (perceived) injustice. But they can also be convinced that regulations are detrimental interventions in markets that undermine the functioning of the market economy. As a consequence, depending on what beliefs are held, arguments for (or against) the minimum wage could either be appreciated and considered in one's reasoning or ignored and neglected. An explanatory model for the belief-consistency effect in the context of multiple text comprehension was presented by Richter and Maier (2017). As Richter and Maier put it, two processing steps take place in multiple text comprehension, which are associated with intuitive and deliberative thinking. The reference to intuitive and deliberative thinking is noteworthy, because it puts the model in the context of dual-process theories (see Kahneman, 2011). Dual-process theories are based on the assumption that there are two ways of thinking: a fast, intuitive and resource-saving way of thinking, and a slower, deliberative and more resource-intensive way of thinking.<sup>15</sup> As a consequence, the empirical findings can be interpreted as instantiations of general patterns of human thinking.

Bearing in mind that individuals tend to focus on belief-consistent information and that one cannot expect that each and every content presented would be cast into doubt, the question arises of how students can be “stimulated” to think critically when the subject area actually requires critical thinking. Or to be more specific: How can students be encouraged to include information and appreciate arguments that do not fit their current point of view? A first step to answer this question – the evoking of cognitive conflicts to trigger critical thinking – will be outlined in the following. In this context, the concept of epistemic vigilance can be taken into account (see Stadtler et al., 2015). Vigilance, however, should not only be understood in the context of reasoning as a trait, but rather as a state of conflict within the intuitive system *in situ*. Emerging conflicts can trigger subsequent reasoning activities. The occurrence of such activities marks the transition to deliberative reasoning (system 2), which is guided by strategic objectives (or in the case of multiple-text-comprehension by reading goals such as defending one's own position or reading out of epistemic curiosity; see Richter and Maier, 2017, p. 152).

How can cognitive conflicts arise? Based on the distinction between intuition and deliberation, it can be said that intuitive judgments are sometimes hasty and biased (for a well-known example see “the bat-and-ball” problem; Kahneman, 2011). System 2 could intervene to correct the biases produced by system 1. But there is also another way in which system 1 itself can deal with such biased judgments. This is due to the fact that conflicts can arise between prior beliefs and background knowledge, on the one hand, and intuitive judgments on the other hand. Both, the processing of background knowledge and intuitive decision-making are carried out non-deliberatively in system 1. As Trémolière and De Neys (2014) put it, “such a conflict with our background knowledge will decrease the appeal

of the substituted response and might thereby actually help people to reason better” (p. 487).

However, how can individuals notice such conflicts in the first place? A possible explanation given by Trémolière and De Neys (2014) is that the conflict leads to disfluency. The disturbed fluency serves as a signal, shifting the salience from the intuitively processed content to the conflict. As research on human intuition indicates, processing in system 1 and epistemic feelings are strongly intertwined (Schwartz, 1990; Koriati, 2000; McDermott, 2004; Proust, 2015). For example, feelings of fluency inform the individual whether cognitive processing is flowing or stagnating. Moreover, the monitoring function of epistemic feelings not only relates to the intuitive processes themselves, but also to the results of these processes (e.g., feelings of rightness, coherence, or uncertainty). As Proust (2015) states, the feelings’ “valence and intensity tell the agent whether she should accept or reject a cognitive outcome” (p. 6).

The findings of Maier and Richter (2013) as well as those of Trémolière and De Neys (2014) point to the same “mechanism” in system 1. However, there is a main difference between the studies. Richter and Maier (2017) focus on the negative influence of background knowledge (and topic-related beliefs). According to the text-belief consistency effect, the “background” impedes a more elaborate text comprehension. In contrast, Trémolière and De Neys (2014) focus on the case where background knowledge plays a positive role in reasoning by preventing biased judgments. The latter is of particular interest for our task concerning the minimum wage in two respects:

- (1) Economic content knowledge matters. Cognitive conflicts, which can trigger critical thinking processes, do not only arise when students begin to reflect on intuitive judgments, but already occur in the course of intuitive processing. However, to cause such conflicts, background knowledge is required. Since the processes take place within system 1, tacit knowledge (acquired e.g., through participation in a community of practice) can also be part of the knowledge base.
- (2) If an uncritical and one-sided reception of a party's point of view occurs, or if that party just tries to persuade the addressee, then the evocation of cognitive conflicts can help to “stimulate” critical thinking. According to the inferential approach, such a judgment can be the result of an inductive inference at the end of the inferential cycle. A cognitive conflict should therefore be understood as instigating and motivating a new cycle.

It is worth mentioning that external “stimulation” is only one way to induce cognitive conflicts. The occurrence of cognitive conflicts can also be caused intrinsically. In our case of the minimum wage, this occurs when students, having argued either for the minimum wage as a solution to the justice problem or against it because of the efficiency problem, acknowledge<sup>16</sup> that there are relevant drawbacks that have to be addressed. With this problem in mind, there would be a reason for students to

<sup>15</sup>The various approaches differ in particular in the conceptualization of the interaction between the two ways of thinking. A prominent theory is the interventionist approach of Kahneman (2011). System-1 thinking is assumed to be the default mode, system 2 intervenes especially when cognitive conflicts occur.

<sup>16</sup>Note that “acknowledge” can also be understood in terms of system-1 processes and does not necessarily have to aim at a deliberative process.

initiate another inferential cycle and find an integrated solution (as outlined in section “The Case of the Minimum Wage”). But of course, there is no certainty that this will happen. Richter and Maier (2017) point out that the occurrence of a conflict does not necessarily mean that a new cycle is initiated. Students may perceive the conflict, but they can still ignore it. This leads to the question of how salient a conflict has to be for the students to start engaging it its solution. This can currently be regarded as an open empirical question.

## CONCLUSION

Focusing on the example of the minimum wage, we have elaborated on how to engage critically with a controversial and much debated topic. An inferential framework was presented that enables us to reconstruct the cognitive deep structure of reasoning processes when arguing for or against minimum wages. Thus, the foundation has been laid for empirical studies to follow, in which students’ critical thinking skills can be assessed.

## REFERENCES

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamin, R., et al. (2008). Instructional interventions affecting critical thinking skills and dispositions: a stage 1 Meta-analysis. *Rev. Educ. Res.* 78, 1102–1134. doi: 10.3102/0034654308326084
- Braun, H. (2019). Performance assessment and standardization in higher education: a problematic conjunction? *Br. J. Educ. Psychol.* 89, 429–440. doi: 10.1111/bjep.12274
- Caliendo, M., Schröder, C., and Wittbrodt, L. (2019). The causal effects of the minimum wage introduction in Germany – An overview. *Ger. Econ. Rev.* 20, 257–292. doi: 10.1111/geer.12191
- Crisp, R. (2018). “Prudential and moral reasons,” in *The Oxford Handbook of Reasons and Normativity*, ed. D. Star (Oxford: Oxford University Press), 800–820. doi: 10.1093/oxfordhb/9780199657889.013.35
- da Silva Almeida, L., and Rodrigues Franco, A. H. (2011). Critical thinking: its relevance for education in a shifting society. *Rev. Psicol.* 29, 175–195.
- Dormann, C., Demerouti, E., and Bakker, A. (2018). “A model of positive and negative learning. Learning demands and resources, learning engagement, critical thinking, and fake news detection,” in *Positive Learning in the Age of Information: A Blessing or a Curse?*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Dordrecht: Springer), 315–346. doi: 10.1007/978-3-658-19567-0\_19
- Ennis, R. (1987). “A taxonomy of critical thinking abilities and dispositions,” in *Teaching Thinking Skills: Theory and Practice*, eds J. Baron and R. Sternberg (New York: W. H. Freeman), 9–26.
- Ennis, R. (1996). *Critical Thinking*. Upper Saddle River: Prentice-Hall.
- Ennis, R. (2011). Critical thinking. *Refl. Perspect.* 26, 4–18. doi: 10.5840/inquiryctnews20112613
- Facione, P. A. (2000). The disposition toward critical thinking: its character, measurement, and relationship to critical thinking skill. *Informal Log.* 20, 61–84. doi: 10.22329/il.v20i1.2254
- Facione, P. A., Facione, N. C., and Giancarlo, C. A. (1997). “The motivation to think in working and learning. in Preparing Competent College Graduates,” in *Setting New and Higher Expectations for Student Learning*, ed. E. Jones (San Francisco, CA: Jossey-Bass Publishers), 67–79. doi: 10.1002/he.36919969608
- Fedorets, A., Grabka, M. M., Schröder, C., and Seebauer, J. (2020). Lohnungleichheit in Deutschland sinkt. *DIW Wochenbericht* 87, 92–97. doi: 10.18723/diw\_wb:2020-7-1
- Gabbay, D. M., and Woods, J. (2005). *A Practical Logic of Cognitive Systems. Volume 2: The Reach of Abduction – Insight and Trial*. Amsterdam: Elsevier. doi: 10.1016/S1874-5075(05)80020-8
- Depending on the reasoning dynamics (e.g., the appearance of reasoning errors) and existing student dispositions, instructional interventions can take place. One teaching strategy to encourage critical thinking is the evocation of cognitive conflicts. With the reconstructed deep structure of the students’ cognitive processes, “cognitive activation” can be geared to each student’s mindset and in an adaptive way. The effectiveness of such interventions, e.g., against immunization or students’ proneness to persuasive agitation, can thus be investigated in more detail.
- ## AUTHOR CONTRIBUTIONS
- GM has presented and analyzed the case of the minimum wage, delivered the inferential approach and the inferential reconstruction of the task and instances of critical thinking in dealing with the case. RH has contributed the teaching strategies to promote critical thinking and parts of the inferential reconstruction of the problem. Both authors contributed to the article and approved the submitted version.
- Haferkamp, A., Fetschenhauer, D., Belschak, F., and Enste, D. (2009). Efficiency versus fairness: the evaluation of labor market policies by economists and laypeople. *J. Econ. Psychol.* 30, 527–539. doi: 10.1016/j.joep.2009.03.010
- Hermkes, R. (2016). “Perception, abduction, and tacit inference,” in *Model-Based Reasoning in Science and Technology – Logical, Epistemological, and Cognitive Issues*, eds L. Magnani and C. Casadio (Heidelberg: Springer), 399–418. doi: 10.1007/978-3-319-38983-7\_22
- Hitchcock, D. (2017). “Critical thinking as an educational ideal,” in *On Reasoning and Argument. Essays in Informal Logic and on Critical Thinking*, ed. D. Hitchcock (Cham: Springer Nature Switzerland), 477–497. doi: 10.1007/978-3-319-53562-3\_30
- Hofmann, F. (2019). How to know one’s experiences transparently. *Philos. Stud.* 176, 1305–1324. doi: 10.1007/s11098-018-1064-0
- Jenicek, M., and Hitchcock, D. (2005). *Evidence-Based Practice: Logic and Critical Thinking in Medicine*. Chicago: AMA Press.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.
- Koriat, A. (2000). The feeling of knowing: some metatheoretical implications for consciousness and control. *Cons. Cogn.* 9, 149–171. doi: 10.1006/ccog.2000.0433
- Leu, D. J., Kiili, C., Forzani, E., Zawilinski, L., McVerry, J. G., and O’Byrne, W. I. (2014). The new literacies of online research and comprehension: rethinking the reading achievement gap. *Read. Res. Q.* 50, 37–59. doi: 10.1002/rrq.85
- Lipton, P. (2004). *Inference to the Best Explanation*, 2nd Edn. London: Routledge. doi: 10.4324/9780203470855
- Liu, O. L., Frankel, L., and Roohr, K. C. (2014). Assessing critical thinking in higher education: current state and directions for next-generation assessments. *ETS Res. Rep. Ser.* 1, 1–23. doi: 10.1002/ets2.12009
- Liu, O. L., Rios, J. A., and Borden, V. (2015). The effects of motivational instruction on college students’ performance on low-stakes assessments. *Educ. Assess.* 30, 79–94. doi: 10.1080/10627197.2015.1028618
- Magnani, L. (2009). *Abductive Cognition. The Epistemological and Eco-Cognitive Dimension of Hypothetical Reasoning*. Berlin: Springer. doi: 10.1007/978-3-642-03631-6
- Magnani, L. (2018). The urgent need of a naturalized logic. *Philosophies* 3:44. doi: 10.3390/philosophies3040044
- Maier, J., and Richter, T. (2013). Text-belief consistency effects in the comprehension of multiple texts with conflicting information. *Cogn. Instr.* 31, 151–175. doi: 10.1080/07370008.2013.769997
- McDermott, R. (2004). The feeling of rationality: the meaning of neuroscientific advances for political science. *Perspect. Polit.* 2, 691–706. doi: 10.1017/S1537592704040459

- Mindestlohn-Kommission (2018). *Zweiter Bericht zu den Auswirkungen des gesetzlichen Mindestlohns: Bericht der Mindestlohnkommission an die Bundesregierung nach §9 Abs. 4 Mindestlohngesetz*. Available online at: <https://www.mindestlohn-kommission.de/DE/Bericht/pdf/Bericht2018.html?nn=7081728> (accessed July 17, 2020).
- Minnameier, G. (2004). Peirce-Suit of Truth – Why inference to the best explanation and abduction ought not to be confused. *Erkenntnis* 60, 75–105. doi: 10.1023/B:ERKE.0000005162.52052.7f
- Minnameier, G. (2017). “Forms of abduction and an inferential taxonomy,” in *Springer Handbook of Model-Based Reasoning*, eds L. Magnani and T. Bertolotti (Dordrecht: Springer), 175–195. doi: 10.1007/978-3-319-30526-4\_8
- Minnameier, G. (2019). “Re-orienting the logic of abduction and the naturalization of logic,” in *Natural Arguments: A Tribute to John Woods*, eds D. Gabbay, L. Magnani, W. Park, and A. V. Pietarinen (London: College Publications), 353–374.
- Oser, F. (2018). “Positive learning through negative learning. The wonderful burden of PLATO,” in *Positive Learning in the Age of Information: A Blessing or a Curse?*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Dordrecht: Springer), 363–370. doi: 10.1007/978-3-658-19567-0\_21
- Proust, J. (2015). *The Representational Structure of Feelings*. Available online at: <https://open-mind.net/papers/the-representational-structure-of-feelings> (accessed July 17, 2020).
- Putnam, H. (2002). *The Collops of the Fact/Value Dichotomy and Other Essays*. Cambridge, MA: Harvard University Press.
- Richter, T., and Maier, J. (2017). Comprehension of multiple documents with conflicting information: a two-step model of validation. *Educ. Psychol.* 52, 148–166. doi: 10.1080/00461520.2017.1322968
- Roetzel, P. G. (2019). Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Bus. Res.* 12, 479–522. doi: 10.1007/s40685-018-0069-z
- Schwartz, N. (1990). “Feelings as information: Informational and motivational functions of affective states,” in *Handbook of Motivation and Cognition: Foundations of Social Behavior*, Vol. 2, eds E. T. Higgins and R. M. Sorrentino (New York, NY: Guilford), 527–651.
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. P. (2019). Assessment of university students’ critical thinking: next generation performance assessment. *Int. J. Test* 19, 337–362. doi: 10.1080/15305058.2018.1543309
- Stadtler, M., and Bromme, R. (2013). Multiple document comprehension: an approach to public understanding of science. *Cogn. Instr.* 31, 122–129. doi: 10.1080/07370008.2013.771106
- Stadtler, M., Paul, J., Globoschütz, S., and Bromme, R. (2015). *Watch out! - An Instruction Raising Students’ Epistemic Vigilance Augments Their Sourcing Activities. Mind, Technology, and Society*. Austin: Cognitive Science Society, 2278–2283.
- Stalnaker, R. C. (1987). *Inquiry*. Cambridge, MA: A Bradford Book.
- Tarchi, C., and Mason, L. (2020). Effects of critical thinking on multiple-document comprehension. *Eur. J. Psychol. Educ.* 35, 289–313. doi: 10.1007/s10212-019-00426-8
- Trémolière, B., and De Neys, W. (2014). When intuitions are helpful: Prior beliefs can support reasoning in the bat-and-ball problem. *J. Cogn. Psychol.* 26, 486–490. doi: 10.1080/20445911.2014.899238
- Valaris, M. (2017). What reasoning might be. *Synthese* 194, 2007–2024. doi: 10.1007/s11229-016-1034-z
- Woods, J. (2013). *Errors of Reasoning: Naturalizing the Logic of Inference*. London: College Publications.
- Woods, J. (2017). “Reorienting the logic of abduction,” in *Handbook of Model-Based Science*, eds L. Magnani and T. Bertolotti (Dordrecht: Springer), 137–150. doi: 10.1007/978-3-319-30526-4\_6
- Yu, S., and Zenker, F. (2018). Peirce knew why abduction isn’t IBE: a scheme and critical questions for abductive argument. *Argumentation* 32, 569–587. doi: 10.1007/s10503-017-9443-9
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., and Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *Br. J. Educ. Psychol.* 89, 468–484. doi: 10.1111/bjep.12286

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Minnameier and Hermkes. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Strategy Use in Learning From Multiple Texts: An Investigation of the Integrative Framework of Learning From Multiple Texts

Alexandra List<sup>1\*</sup> and Patricia A. Alexander<sup>2</sup>

<sup>1</sup> Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University, State College, PA, United States, <sup>2</sup> Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, United States

## OPEN ACCESS

### Edited by:

Lawrence Jun Zhang,  
The University of Auckland,  
New Zealand

### Reviewed by:

Hossein Bozorgian,  
University of Mazandaran, Iran  
Tiefu Shaun Zhang,  
University of Electronic Science  
and Technology of China, China

### \*Correspondence:

Alexandra List  
azl261@psu.edu

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Education

**Received:** 30 June 2020

**Accepted:** 22 September 2020

**Published:** 22 October 2020

### Citation:

List A and Alexander PA (2020)  
Strategy Use in Learning From  
Multiple Texts: An Investigation of the  
Integrative Framework of Learning  
From Multiple Texts.  
Front. Educ. 5:578062.  
doi: 10.3389/feduc.2020.578062

This study examined undergraduates' strategy use when learning about a complex and controversial topic (i.e., mass incarceration in the United States) based on information presented across multiple texts. Guided by the Integrated Framework of Learning from Multiple Texts, this study directed students to engage in one of three forms of strategy use while learning from multiple texts. In particular, students were asked to identify relevant and important information in texts (i.e., intratextual processing), to form relations or connections across texts (i.e., intertextual processing), or to identify easy or difficult to understand information in texts (i.e., metacognitive processing). In addition to receiving task instructions directing them to engage in these modes of processing, students were also provided with a highlighting tool to scaffold their strategy use (e.g., by allowing important and relevant information to be marked in green, in the intratextual processing condition). This highlighting tool also enabled researchers to collect log data of students' manifest strategy use. Students were found to demonstrate differential patterns of strategy use in accordance with their assigned processing conditions. Moreover, students' use of strategies directed toward multiple texts was found to predict multiple text task performance.

**Keywords:** multiple text comprehension, multiple text processing, strategic processing, integration- and synthesis- oriented strategies, metacognition

## INTRODUCTION

This study examined whether prompting students to engage in different types of processing when learning from multiple texts impacted their strategy use and task performance. The multiple text task used in this study required learners to understand and write about a complex and controversial topic (i.e., mass incarceration in the United States) based on information presented across multiple texts. This task was designed to represent the types of academic assignments that undergraduate students are frequently asked to complete (Hendley, 2012; Datig, 2016; Weston-Sementelli et al., 2018). It was also devised to address the type of social issues, discussed in the popular press, that students may be driven to research or to learn more about on their own (Bazelon, 2019; Uhrmacher, 2020).

Similar tasks have been employed in prior research examining students' learning from multiple texts (e.g., Wiley et al., 2009; Barzilai and Weinstock, 2015). This body of research has established that students need a variety of sophisticated skills and strategies to learn about complex and controversial topics from multiple texts. Such sophisticated strategies include being able to

identify relevant content in texts (Potocki et al., 2017; McCrudden, 2018); synthesize and connect information introduced across disparate texts (Kobayashi, 2009; List et al., 2019b); and, make metacognitive judgments regarding comprehension quality and adequacy of task performance (e.g., Stadtler and Bromme, 2008; Wang and List, 2019).

Despite the need for students to demonstrate sophisticated and erudite strategy use when learning from multiple texts, relatively few studies have examined the nature of such strategy use during task performance (Wolfe and Goldman, 2005; Anmarkrud et al., 2014; Du and List, 2020). Those studies that are the exception have generally employed real-time methods like think-alouds to capture students' strategic processing. In those studies, students' more sophisticated text processing (e.g., use of evaluative and cross-textual linking strategies) has been associated with better task performance, particularly as assessed through writing (Goldman et al., 2012; Anmarkrud et al., 2014). This positive association notwithstanding, students' strategy use when learning from multiple texts has rarely been experimentally manipulated. In this study, by comparison, we expressly altered the directions that students received regarding how to process an assigned set of four multiple texts, in order to examine the influence of such task manipulations on students' strategy use and task performance. Specifically, we examined whether prompting students to (a) attend to relevant or important information in individual texts (i.e., *intratextual processing*); (b) relate or connect content across texts (i.e., *intertextual processing*); or (c) monitor the ease or difficulty of their understanding (i.e., *metacognitive processing*) impacted their demonstrated strategy use and task performance.

## Integrated Framework of Learning From Multiple Texts

The design of this study was guided by the Integrated Framework of Learning from Multiple Texts (IF-MT; List and Alexander, 2019). Synthesizing much of the literature on students' multiple text learning, the IF-MT depicts such learning as unfolding over three stages, preparation, execution, and production. In the *preparation* stage, students analyze the task guiding multiple text use based on its various objective characteristics (e.g., the topic or domain) and their subjective perceptions of these characteristics (e.g., students' topic or domain interest). Students' task analysis and subjective perceptions result in their adoption of a default stance or guiding orientation toward task completion. In adopting a default stance, students make decisions about their investment in and strategic approach toward task completion.

In the *execution* stage of the IF-MT, students engage in strategic processing consistent with the default stances they adopted in the preparation stage. Three categories of strategic processing characterize students' interactions with multiple texts and predict students' accomplishment of various learning outcomes. These three modes of strategic processing are behavioral, cognitive, and metacognitive. *Behavioral strategies* reflect students' observable interactions with multiple texts, including text access and navigation. *Cognitive strategies* are defined as the internal operations or mental processes that

students perform during reading. Finally, *metacognitive strategies* represent students' efforts to monitor and regulate their own understanding during reading (i.e., comprehension monitoring), to appraise text quality (i.e., epistemic monitoring), and to judge the extent of their learning (i.e., monitoring of task outcomes).

In the present study, we targeted students' cognitive and metacognitive strategy use. The cognitive strategies we examined included those involved in intratextual comprehension (i.e., understanding individual texts) and in intertextual integration (i.e., cross-textual linking). *Intratextual strategies*, like prior knowledge activation and elaboration, reflect the cognitive processes that students intentionally use before, during, and after reading. These strategies have been found to support single text comprehension in prior work (Woloshyn et al., 1994; McNamara, 2004; Dinsmore and Alexander, 2012; Parkinson and Dinsmore, 2018). By comparison, *intertextual strategies*, including organization and corroboration, involve students' formation of cross-textual links in the service of developing an integrated and coherent representation of a central topic or issue discussed across multiple texts (Kobayashi, 2009; Bråten and Strømso, 2011; Hagen et al., 2014). Behavioral strategy use was not examined in this study as we were focused on capturing the covert (i.e., cognitive and metacognitive) processes that students engaged during multiple text use. These have been examined to a more limited extent in prior work, as compared to behavioral strategies that are easier to capture, for instance, via students' notes or log data (Hagen et al., 2014; List and Alexander, 2017). Although students' behavioral strategy use was not the target of this investigation, students were nevertheless asked to use behavioral strategies (i.e., highlighting) to support the mode of cognitive or metacognitive processing that they were instructed to deploy when learning from multiple texts.

Although a variety of strategies are identified as important in the IF-MT, their differential impact on multiple text learning has yet to be established. Therefore, in this study, we set out to determine the extent to which providing students with various processing directives in the preparation stage of the IF-MT influenced their strategy use in the execution stage and, ultimately, their formation of various cognitive (mental) and external (i.e., tangible) outcomes in the production stage. *Cognitive outcomes* are defined in the IF-MT as the mental models or representations of complex topics or issues that students construct based on information introduced across multiple texts. Tangible or *external outcomes* reflect the physical products (e.g., written responses) that students compose based on the cognitive outcomes they generate. In part, these external outcomes are what allow learning from multiple texts to be evaluated and assessed. External outcomes are considered separately from their underlying cognitive bases in the production stage of the IF-MT. This is done to underscore that the external products that students develop are typically only selective or stylized representations of all the information that students may internalize and cognitively represent when processing multiple texts. For instance, when students write a summary based on multiple texts, they may only include main ideas in the external responses that they compose, while retaining additional salient details in their cognitive representations.

In this investigation, we assessed students' cognitive representations of multiple texts and their correspondent external outcomes. In particular, two types of outcome measures were employed. First, we analyzed the quality of students' research reports. Research reports were the external outcome that students were asked to produce based on the multiple texts they processed in this study. Second, we examined the quality of students' responses to open-ended questions designed to probe the richness of students' cognitive representations of the overlapping topics or issues discussed across four carefully crafted texts. By assessing these two outcome measures, in conjunction with the mode of text processing that students were asked to adopt (i.e., intratextual, intertextual, or metacognitive), we sought to achieve a deeper understanding of the relation between students' manifest strategic processing and task performance when learning from multiple texts.

That is, the design of this study mirrored each phased of the IF-MT. In the preparation phase, students were instructed to engage in one of three modes of multiple text processing (i.e., intratextual, intertextual, or metacognitive) while consulting multiple texts to compose a research report. It was our expectation that these strategy use directives would be incorporated into students' task analysis and planning of task completion. In the execution phase, we expected students to engage in the intratextual, intertextual, or metacognitive processing of multiple texts, in accordance with their assigned task condition. We supported such strategy use by providing students with a highlighting tool, customized to their assigned condition. Finally, in the production phase, we assessed both students' cognitive representations of the information introduced across texts (i.e., via the open-ended questions) and students' manifest performance on the assigned multiple text task (i.e., composing a research report).

## Strategy Use When Learning From Multiple Texts

The strategic processes that students engaged during the execution stage of the IF-MT were of particular interest in this study. Indeed, there has been a substantive and growing body of work documenting the various strategies that students may use when learning from multiple texts (e.g., Wineburg, 1991; Daher and Kiewra, 2016; Brante and Strømsø, 2018). In a recent taxonomy, the *Comprehensive Strategy Framework* (CSF), List (2020) suggests that these strategies may be differentiated according to two primary dimensions. That is, the strategies that students use when learning from multiple texts vary in their *functions* (i.e., goals for strategy deployment) and in their *referents* (i.e., informational foci).

Based in Cho et al.'s (2018) work, three possible strategy functions are identified in the CSF. According to Cho et al. (2018), when learning about complex topics using multiple texts, students engage in constructive-integrative, critical-analytic, and metacognitive-reflective processing. *Constructive-integrative processing* refers to students' efforts to make meaning or to develop a single, coherent cognitive representation of information presented across multiple texts. *Critical-analytic processing* encompasses students' efforts to establish source trustworthiness or information veracity during multiple text

learning. Finally, *metacognitive-reflective processing* captures students' efforts to deploy, monitor, and regulate their use of constructive-integrative and critical-analytic processing strategies, including metacognition and self-regulation.

List (2020) points out that each of the aforementioned functions may be directed toward at least three possible strategy referents or informational targets: (a) a single text, (b) multiple texts, or (c) learners' prior knowledge and beliefs. For instance, when engaging in constructive-integrative processing during multiple text reading, students may elaborate the information introduced in a single text based on information included in that same text (i.e., single text referent), information explained in another text (i.e., multiple text referent), or based on their own experiences (i.e., prior knowledge and beliefs referent). In this case, students' constructive-integrative processing may be thought of as both uniform in function and distinct in referent, with students' efforts at meaning-making extended across single texts, multiple texts, and their own prior knowledge. Crossing the three strategy functions identified by Cho et al. (2018) with the three strategy referents from the IF-MT, List (2020) charts the landscape of students' potential strategy use when learning from multiple texts.

This function by referent mapping of strategy engagement has been observed in prior work. For instance, in a think-aloud study, Anmarkrud et al. (2014) investigated students' use of linking strategies (i.e., strategies connecting two or more texts) as a distinct strategy referent. They found linking to be disproportionately distributed across students' constructive-integrative (47.1%), critical-analytic (36.3%), and metacognitive-reflective (16.7%) processing of multiple texts. Similarly, Wolfe and Goldman (2005) found that the elaborative strategies that students reported using differed according to whether these were associated with learners' referencing of a single text, of multiple texts, or of their earlier generated think-aloud comments. In this study, we similarly investigate differences in students' strategy use across the three different types of strategy referents identified in the CSF (List, 2020). In doing so, we build on prior work that has only documented the nature of students' strategy use by explicitly directing students to engage in different modes of strategic processing when learning from multiple texts. Thus, in this study, we explicitly directed students to engage in intratextual, intertextual, or metacognitive processing during a multiple text task.

## Task Assignment When Learning From Multiple Texts

Task instructions, or the directives that students receive prior to reading, have repeatedly been found to play an important role in students' learning from multiple texts (Le Bigot and Rouet, 2007; Gil et al., 2010a,b; McCarthy and Goldman, 2015; List et al., 2019a). Task instructions specify the types of external products that students may be asked to produce from multiple texts and direct students' attention and strategy engagement toward particular content in texts (McCrudden and Schraw, 2007; McCrudden et al., 2011). Nevertheless, to date, only the first of these task instruction functions has been well-investigated. That is, students asked to produce different external products based on multiple texts have been found to differ in the quality of

their performance, with this association explained by features of students' strategy use (Wiley and Voss, 1999; Cerdán and Vidal-Abarca, 2008; Stadler and Bromme, 2008; Kobayashi, 2009; McCrudden and Sparks, 2014). In this study, rather than varying the types of task products that students were asked to produce we instead varied the modes of processing that students were asked to engage during multiple text use. We did this by directing students to engage in intratextual, intertextual, or metacognitive strategy use when learning from multiple texts and by facilitating such strategy use by asking students to highlight and explain information in texts that were consistent with their assigned task condition. For instance, students in the intertextual processing strategy condition received task instructions and a highlighting tool, with three different color options, to aid them in identifying similar, different, and otherwise related information introduced across four multiple texts. Students in the metacognitive processing strategy condition were instructed to identify content that was easy or difficult to understand and were provided with two highlighter options to aid them in doing so (i.e., a red highlighter to mark difficult to understand information and a green highlighter to mark easy to understand information). For this study, we posed the following research questions:

- (1) Are there differences in strategy use among students prompted to engage in the intratextual, intertextual, or metacognitive processing of multiple texts?

We expected the nature of students' strategy use across conditions to differ in both function and referent. In particular, we expect students in the intratextual condition to exhibit the greatest degree of constructive-integrative strategy use directed toward single texts. We expected students in the intertextual condition to manifest the greatest degree of constructive-integrative strategy use directed toward multiple texts. Finally, we expected students in the metacognitive processing condition to exhibit the most metacognitive-reflective strategy use, across referents.

- (2) Are there differences in writing performance, citation use, and responses to open-ended integration questions among students prompted to engage in the intratextual, intertextual, or metacognitive processing of multiple texts? Due to the important role that integration or cross-textual connection formation plays in students' learning from multiple texts (Britt et al., 1999; Perfetti et al., 1999), we expected students assigned to the intertextual processing condition to demonstrate the greatest degree of task performance. Then, based on Stadler and Bromme's (2008) work, we expected students in the metacognitive processing condition to outperform students in the intratextual condition, across the outcome measures examined.
- (3) Is there an association between students' multiple text strategy use and their external outcomes (i.e., research report writing quality, citation use, and responses to open-ended integration questions) when learning from multiple texts? We expected students' greater engagement in strategy use directed toward multiple texts to be associated with

research report writing quality and with open-ended integration performance.

## MATERIALS AND METHODS

### Participants

Participants were 71 undergraduate students enrolled at a large university in the Mid-Atlantic region of the United States (age:  $M = 20.59$ ,  $SD = 1.98$ ; female: 57.89%,  $n = 33$ ; male: 42.11%,  $n = 24$ ). Students participating identified their race/ethnicity as White (42.59%,  $n = 23$ ), Black/African-American (7.41%,  $n = 4$ ), Asian (29.63%,  $n = 16$ ), Hispanic/Latino (12.96%,  $n = 7$ ), or as representing more than one racial or ethnic group (7.41%  $n = 4$ ). Students represented a variety of class standings—freshman: 8.77% ( $n = 5$ ); sophomores: 36.84% ( $n = 21$ ); juniors: 21.05% ( $n = 12$ ); seniors: 33.33% ( $n = 19$ ). Ten students (14.08% of the sample) only completed the individual difference measures and did not complete the multiple text task, reducing our analysis sample to 61.

### Procedures

This study proceeded in three main phases. First, students were asked to complete assessments of prior topic knowledge and topic interest. Then, students were asked to complete a multiple text task, including a reading and a writing phase. Students were randomly assigned to intratextual, intertextual, or metacognitive processing task conditions, as they did so. Finally, students were asked to respond to a post-task assessment. Students completed the study online, via the Qualtrics platform, at a time and location of their choosing. The study took students approximately 1 h to complete.

### Overview of Study Measures

Three types of measures were collected in this study. First, individual difference measures were collected to use as controls. Second, process measures of students' multiple text use were gathered. These data were collected using log indicators, namely students' text highlights and associated explanations. Third, performance data were collected. Process and performance data were used to answer the focal research questions in this study.

### Individual Difference Measures

Two individual difference factors, found to be associated with multiple text task performance in prior work, were assessed in this investigation (i.e., prior topic knowledge and topic interest, Bråten et al., 2014).

### Prior Topic Knowledge

Prior topic knowledge was assessed via a term identification measure. In particular, students were asked to define eight terms, relevant to the task (i.e., mass incarceration) and taken directly from the experimental texts (i.e., cash bail, mandatory minimums, mass incarceration, misdemeanor, over-policing, parole, probation, and recidivism). Students were instructed to write N/A if they were unfamiliar with a particular term. Students' definitions for each term were scored as 1 (correct) or 0 (incorrect or N/A). For instance, one student defined probation



as: “a system set up to prevent incarceration and allow for some giving back to the community from the offender,” which received a score of 1. Another student wrote that probation was: “the conditional release of a convict into society,” which was scored as a 0, since this was the definition of parole. Cohen’s kappa inter-rater agreement for students’ prior knowledge was 0.90.

### Topic Interest

Students were asked to rate their interest in each of five topics, related to mass incarceration (i.e., criminal justice, criminology, public policy, social issues, social justice). Students’ interest in each topic was rated on a seven-point scale from *not at all interested* to *very interested*. Cronbach’s alpha reliability for the five-item measure was 0.84. Students’ mean interest ratings were 4.16 ( $SD = 1.24$ ), indicating that, on average, students were at least moderately interested in this study.

## Multiple Text Task

### Topic

Mass incarceration in the United States was selected as the topic of this study for several reasons. First, it represented a complex and multifaceted topic. Understanding mass incarceration required students to make sense of a number of difficult and interrelated concepts, including cash bail and mandatory minimums in sentencing. Second, mass incarceration was a topic about which conflicting, but comparably valid, points of view could be introduced. For example, while some experts consider parole to be an effective antidote to mass incarceration, others contend that parole increases recidivism by prolonging individuals’ contact with the carceral system. Third, mass incarceration constitutes an important societal issue addressed with some frequency in the popular press (Bazelon, 2019; Uhrmacher, 2020). Therefore, we expected students in this study to have some familiarity with this topic. Finally, mass incarceration was a topic about which much data were publicly available and readily accessible, facilitating our construction of texts that included relevant statistical information in support of various issues introduced.

### Texts

Four texts were constructed for the purpose of this study. These were developed to be parallel in structure and overlapping in content, such that key issues were commonly introduced across texts, albeit from different perspectives. Structurally, each text consisted of three paragraphs, each introducing a key issue related to mass incarceration in the United States. Each key issue was supported by one piece of relevant statistical information, attributed to an embedded source cited in-text, such that there were three pieces of statistical data, and associated sources, included in each text. In terms of content, the texts were designed to include some complementary information (i.e., that agreed with information in another text) and some conflicting information (i.e., that disagreed with information in another text). For instance, two texts agreed that the United States incarcerated more individuals and a greater proportion of individuals than any other country in the world, while two texts expressed conflicting views. One of those conflicting texts suggested that the War on Drugs was responsible for increases

in mass incarceration in the United States, whereas the other contended that only a minority of criminal convictions were for drug-related crimes.

All texts were created to appear trustworthy in nature by attributing them to faculty at prestigious post-secondary institutions in the United States. Texts were further presented as feature articles published in well-respected press outlets (e.g., *Economist*, *Atlantic Magazine*). Texts ranged from 253 to 258 words in length. The Flesch-Kincaid grade level readabilities ranged from 14.9 to 16.6, indicating that texts were appropriate for use with an undergraduate audience. Texts were presented in a random order and students could navigate backward and forward across texts. Information about study texts is summarized in **Table 1**.

## Task Conditions

Students’ assignment to task condition had two phases. First, students received task instructions, consistent with their experimental condition, prior to completing the multiple text task. Second, students were provided with external supports (i.e., customized highlighting tools) to support their strategy engagement during processing.

### Task Instructions

All students received the following task instructions prior to reading: *We will ask you to read four texts to write a research report about mass incarceration in the United States. Your research report should connect information presented across texts and cite your sources.* This general set of instructions was followed by one of three specific task directives, asking students to engaged in intratextual, intertextual, or metacognitive processing while learning from multiple texts. Students were randomly assigned to receive one of these three specific task directives.

### External Supports

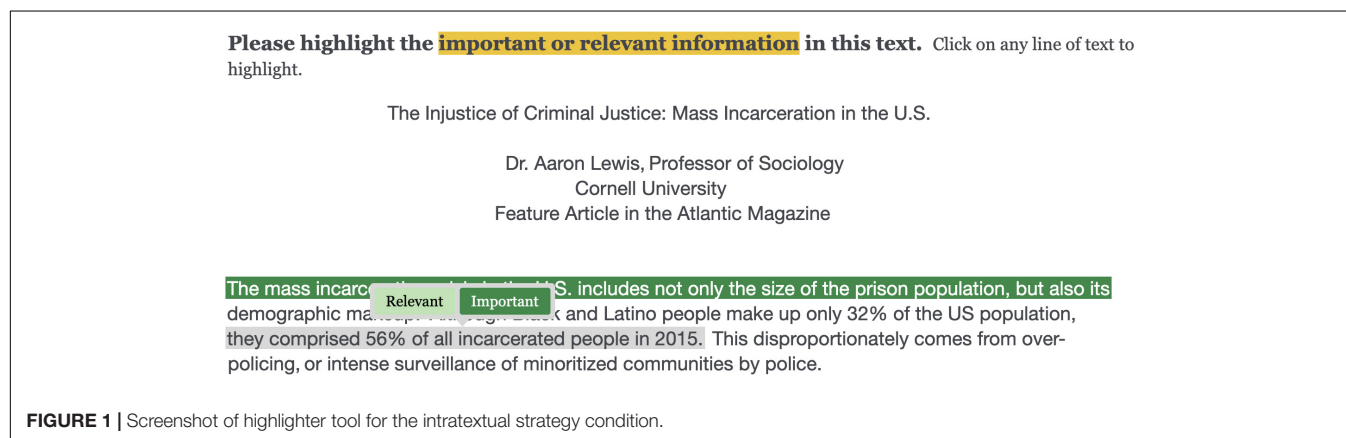
Additionally, students were asked to highlight information in the four study texts in accordance with the processing directives they received. Students were also asked to explain their highlights in a text-box provided for this purpose. Highlights and associated explanations were used both as a physical scaffold to support students’ strategy use, in accordance with their assigned strategy condition, and as a log-data indicator of what information students had attended to during reading and what type of processing was facilitated. While students across conditions may have highlighted the same sentence in text, the highlighter color and students’ associated explanations were used to determine what type of processing each instance of highlighting represented. For instance, students marking the same information may have done so in making a judgment of information relevance (i.e., engaging in intratextual processing), in forming of a cross-textual connection (i.e., reflecting intertextual processing), or in determining a sentence’s comprehension ease (i.e., corresponding to metacognitive processing). **Figure 1** includes a screenshot of the highlighting tool available to students in association with each strategy condition.

Before viewing and highlighting any of the four experimental texts, students were introduced to a practice text that they could highlight according to their assigned strategy condition. The goal

**TABLE 1** | Overview of study texts.

Title	Author and affiliation	Source	Words	Grade level <sup>1</sup>
Understanding the History of Mass Incarceration (Text 1)	Dr. John Williams, Professor of American History, Princeton University	<i>Economist</i>	253	14.8
U.S. is Unique in the World in terms of Mass Incarceration (Text 2)	Dr. Sam Campbell, Professor of Criminology, Dartmouth University	<i>Foreign Policy Review</i>	255	15.6
Facts and Myths about Mass Incarceration in the United States (Text 3)	Dr. Mark Miller, Professor of Public Policy, University of Pennsylvania	<i>New Yorker</i>	254	15.7
The Injustice of Criminal Justice: Mass Incarceration in the United States (Text 4)	Dr. Aaron Lewis, Professor of Sociology, Cornell University	<i>Atlantic Magazine</i>	258	14.9

<sup>1</sup> Flesch-Kincaid Grade-Level Readability.



of this practice text was both to familiarize students with the use of the highlighting tool and to define the construct of mass incarceration for readers. During this practice exercise students were also shown how to navigate forward and backward in the texts they read. This forward and backward navigation was specifically introduced to foster students' potential connection formation across texts.

### Intratextual Processing

In the intratextual condition, students were asked to identify the important or relevant information included within each study text. In particular, students were told: *As you read, we will also ask you to highlight any information that you consider to be relevant or important in each text.* Students were further provided with two highlighting colors allowing relevant (light green) and important (dark green) information to be differentially marked (see **Figure 1**). In this case, prompting students to attend to relevant and important information during reading was viewed as fostering an intratextual strategic approach in that students were prompted to attend to the (relevant and important) information included *within each text*. The intratextual processing condition served as a control or comparison group to which students' more intertextually- or metacognitively-focused multiple text processing could be compared.

### Intertextual Processing

The intertextual strategy condition asked students to identify connections or relations across texts: *As you read, we will also ask you to highlight any information that you consider to be*

*related or connected across texts.* Students in this condition were provided with three highlighting colors to mark similar (green), different (red), or otherwise related (blue) content across texts. See **Figure 2**. This condition was intended to direct students' attention toward the connections or links that could be formed across multiple texts.

### Metacognitive Processing

Metacognitive strategy use was elicited by asking students to highlight the easy or difficult to understand information in each text: *As you read, we will also ask you to highlight any information that is easy or difficult for your to understanding in each text.* "Easy to understand" content was highlighted in green and "difficult to understand" content was marked in red. See **Figure 3**. Prompting students to identify text-based information as either easy or difficult to understand was expected to cue students' engagement in comprehension monitoring during reading.

## Processing Measures

### Text Highlighting

Students' highlights and highlight explanations were coded in terms of their quantity and content. Quantitatively, the number of sentences students highlighted, across texts, was totaled. Inter-rater agreement for the number of highlights in students' responses was 100%, based on 23 responses scored (32.39% of the sample).

With regard to content, students' highlight explanations were coded per List's (2020) *Comprehensive Strategy Framework*,

according to their strategy functions (i.e., constructive-integrative, critical-analytic, or metacognitive-reflective processing) and referents (i.e., a single text, multiple texts, or students' prior knowledge/beliefs). This  $3 \times 3$  taxonomy resulted in students receiving nine separate scores to capture their reported strategy use across all four study texts. For example, students received three separate scores reflecting their engagement in constructive-integrative processing directed toward a single text, toward multiple texts, and toward their own prior knowledge and beliefs. As an example, one student explained the information they highlighted as follows: "The article is easy to understand...The author gives detail and support to his or her thesis very well." This explanation was coded as reflecting metacognitive-reflective and critical-analytic processing, both directed toward a single text. Another student explained one of their highlights as follows: "This text expresses the failure of the community correction programs, similar to the previous text," with this explanation coded as reflecting constructive-integrative processing directed toward multiple texts. It is important to note that while students' highlights and explanations could correspond to their assigned task condition, this was not always the case. For instance, students directed to engage in intratextual processing often identified important and relevant information in texts, however, students assigned to the intertextual processing condition also, at times, explained their highlights as reflecting relevance determinations (see **Table 2** for additional coding examples). Inter-rated agreement for strategy coding was 80.19%, based on our coding of all student responses.

## Research Report

Following the reading phase, students were asked to confirm that they were ready to compose their research reports on the topic of mass incarceration in the United States, with students able to return to the four study texts if they wanted. In composing their research reports, students were asked to "identify connections across the texts you read" and to include citations in their writing.

The research reports students composed were scored using a six-point rubric. The rubric was designed to award points for (a) the number of key issues related to mass incarceration

that students discussed, (b) the extent to which key issues were explained or elaborated in students' writing, and (c) the extent to which key issues were described in an integrative fashion, based on information introduced across two or more texts. Students' responses were assigned a 1 when they introduced a single issue discussed anywhere in the study texts and a 2 when a single issue was not only introduced, but also discussed in an elaborated fashion, with associated evidence, examples, or explanations introduced. Responses assigned a 3 discussed multiple issues introduced within the study texts, while a 4 was assigned to responses that both discussed multiple issues and elaborated on at least two of these, with additional evidence or explanations provided. Finally, responses assigned a 5 contained one instance of intertextual integration, or discussed one issue related to mass incarceration, based on information introduced across two or more texts. Responses assigned a 6 included the integrated discussion of two or more issues based on information introduced across multiple texts. See **Table 3** for a rubric with sample responses. Inter-rater agreement for the scores assigned to students' written responses was Cohen's  $\kappa = 0.75$  (exact agreement: 78.94%). The number of unique citations included in students' research reports was also totaled. Exact agreement for the number of citations included in students' responses was 92.45%.

## Post-task Assessment

Although the rubric used to score research reports was designed to capture both the breadth (i.e., number of issues discussed) and depth (i.e., elaborated and integrated discussion of issues) of students' understanding of mass incarceration, we were interested in probing this understanding further. As such, students were asked to respond to a series of open-ended questions designed to assess their integrative understanding of various key issues discussed across the four study texts. That is, while students could choose whether or not to write about the controversial issue of the War on Drugs in the research reports that they composed, students' understanding of this issue was directly assessed in the open-ended questions that students were asked to answer. In particular, students were asked to: *Think about the four texts*

**TABLE 2** | Examples of strategy explanation codings.

Functions	Referents		
	Single text	Multiple texts	Prior knowledge/beliefs
Constructive-Integrative processing	"I highlighted information that I thought were key points in the reading"	"This piece was very similar to the definitions that were previously stated. This was seen through the consistent discussion of words such as parole and mandatory minimums."	"I also highlighted things that may not be completely familiar to me, therefore pointing it out from the other information I read."
Critical-Analytic processing	"Citations make stuff seem credible"	"This text talks about how parole leads to re-incarceration, however the next text states that twice as many people are on parole/probation than incarcerated."	"I am not familiar with the Equal Justice Initiative, but. The Equal Justice Initiative found misdemeanors to make up 80% of all arrests in 2017, but these arrests are made to maintain law and order."
Metacognitive-Reflective processing	"I do not understand the red highlighted points."	"I also highlighted the information about the mandatory minimums on drug offenses because it helps me better understand the argument in the first reading"	"Mass incarceration and pardon are two new words for me. Therefore, this sentence is difficult for me to understand."

**TABLE 3 |** Rubric for scoring students' research reports.

Score	Description	Example	Frequency
1	Single, specific issue introduced	All four texts discussed the issue the US have surrounding the topic of mass incarceration. They all talked about how mass incarceration is being implemented and why they were created. Additionally, the reasoning for these mass incarcerations are due to drug crimes in which many police officers target minorities.	8.77% ( <i>n</i> = 5)
2	Single, specific issue introduced and elaborated	When I see those data about mass incarceration for the first time. I was shocked by those numbers. However, the US government always states that they will treat people equally no matter the race. But one of the factors which contribute to mass incarceration is cultural background. I used to learn CCJS 100, and one of the lectures talked about that blacks are more likely to commit crimes than whites. Did they really do something bad? Or just some people have a bias on them. . .	1.75% ( <i>n</i> = 1)
3	Multiple issues introduced	The United States holds the greatest number of people incarcerated, compared to other countries around the world. The United States found that arrests related to drug use have increased 10 times. Increasing parole and probation have been considered to help monitor these issues, but this might not be the most effective solution. To continue, Black and Latino people make up a small portion of the United States' population but they make up a vast percentage of people incarcerated, which indicates over-policing.	14.04% ( <i>n</i> = 8)
4	Multiple issues introduced and elaborated	Mass incarceration in the United States is a large issue that should be addressed. Many contributing factors have to do with this issue. Some of these factors include over-policing, over-use of the parole system, and over-emphasis on minority communities and not the population as a whole. . . One of the readings stated that the United States assigns the longest punishments compared to all other countries for the same crimes. Additionally, more arrests and convictions are made against people in the Latino or African American communities, compared to other individuals. An interesting point made in one of the readings is that all people, white or black, engage in the same amount of drug-activity and crime. I believe that if less parole opportunities were granted for individuals who may not be able to comply with all the rules and regulations, the recidivism rate would likely decline. If prisoners were forced to finish out their sentences and not receive any special treatment or early release, they will likely integrate themselves back into society more effectively compared to going back into society while still paying the price for your crime. The policing system is obviously flawed and could use improvements in several areas.	15.79% ( <i>n</i> = 9)
5	Multiple issues introduced, elaborated, and single instance of integration	. . . Many researchers have been looking into the reasoning for this recurring problem, and why the trend has been increasing over the past years rapidly. <b>Dr. John Williams</b> sees this problem and points out that in 2018, over 2.3 million people were in U.S prisons. He says that, "Those incarcerated for drugs increasing from 40,000 in 1980 to over 400,000 in 2017." A big problem encouraging this increase is all of the prisoners being brought in during this war on drugs. <u><i>Not only did Dr. Williams see this problem, but so did Dr. Sam Campbell stating</i></u> , "Analysis from the United Nations Office on Drugs and Crime found that the U.S has less than 5% of the world's population, but almost 25% of the world's prisoners." Drugs are a huge factor for why the prisons in the U.S are so overcrowded now, but that is not the only reason. Dr. Aaron Lewis found that misdemeanors make up 80% of all arrests in 2017. Another factor for why the jails/prisons are so crowded is because they are locking people up who don't necessarily need to be locked up.	21.05% ( <i>n</i> = 12)
6	Multiple issues introduced, elaborated, and multiple issues of integration	<b>Dr. John Williams</b> suggests that the war on drug plays a large role in mass incarceration due to the major increase in imprisonment of drug-related crimes (Williams). His main argument points to how mandatory minimums are enforced for even possession of drugs, which then ultimately leads to the mass imprisonment of many people for an extended amount of time and for menial crimes (Williams). <u><i>Dr. Aaron Lewis presents a related factor toward the overall root of mass incarceration</i></u> . He brings up the idea of mass incarceration being inherently racist due to the hyperfocus on those who are of the minority in the US, Blacks and Hispanics (Lewis). . . . To address solutions to the issue of mass incarceration, <b>Dr. Williams</b> proposes favoring for probations and paroles (Williams). He claims that it is more cost efficient and promotes community corrections (Williams). <u><i>However, Dr. Lewis, Dr. Miller and Dr. Campbell note that utilizing more paroles is not the most effective solution</i></u> and that around half of those on parole do not succeed (Lewis) due to them being sent back for breaking a minor violation (Campbell) or being unable to pay certain fees (Lewis)	33.33% ( <i>n</i> = 19)

*Instances of integration are underlined and italicized; Citations are bolded.*

you read. Please summarize what the texts said about each of these main points. Please be sure to think about the information presented across all four texts in the summaries you compose. Students were then asked to summarize information related to each of four key issues discussed across multiple texts (i.e., the number of incarcerated individuals in the U.S., the War on Drugs, the "tough on crime" culture in the U.S., and the advantages and disadvantages of community corrections). Students' responses to

each open-ended question were assigned a score of 0, 1, or 2, according to whether these were incorrect or incomplete (0), reflected information only from a single text (1), or considered information provided in more than one study text (2). Students' scores were totaled across all four of the key issues that they were asked to summarize, based on information introduced across multiple texts. Sample responses are included in **Table 4**. Exact agreement for students' open-ended response scores was 88.32%.



## RESULTS

### Research Question 1: Differences in Strategy Use by Task Condition

Our first research question examined differences in students' manifest strategy use across task conditions. A series of one-way ANOVAs were conducted, with alpha adjusted to 0.007 for multiple comparisons (i.e.,  $\alpha = 0.05/8$ ). Because so few students exhibited critical analytic processing directed toward their prior knowledge, this aspect of strategy use was not analyzed. Descriptive information for strategy use across conditions is presented in **Table 5**.

To start, students' use of constructive-integrative strategies directed at multiple texts differed significantly across task conditions [ $F(2,56) = 12.48$ ,  $p < 0.001$ ,  $\eta^2 = 0.31$ ], indicating a large effect. *Post hoc* analyses, using Tukey's HSD, determined that students assigned to the intertextual condition employed significantly more constructive-integrative strategies directed at multiple texts ( $M = 3.60$ ,  $SD = 3.98$ ) than students assigned to either the intratextual ( $M = 0.42$ ,  $SD = 0.84$ ) or the metacognitive ( $M = 0.25$ ,  $SD = 0.44$ ) strategy conditions,  $ps < 0.001$ .

Moreover, students' use of metacognitive-regulatory strategies directed at single texts differed by task condition [ $F(2,56) = 21.71$ ,

$p < 0.001$ ,  $\eta^2 = 0.44$ ], reflecting a large effect. *Post hoc* analyses using Tukey's HSD found students assigned to the metacognitive strategy condition to use significantly more metacognitive-reflective strategies directed at single texts ( $M = 4.05$ ,  $SD = 3.73$ ) than students assigned to either the intratextual ( $M = 0.16$ ,  $SD = 0.50$ ) or the intertextual ( $M = 0.00$ ,  $SD = 0.00$ ) strategy conditions,  $ps < 0.001$ . Likewise, students' metacognitive-reflective strategy directed toward their prior knowledge differed across conditions,  $F(2,56) = 6.61$ ,  $p < 0.007$ ,  $\eta^2 = 0.19$ . In particular, this approach to strategic processing was only manifest by students in the metacognitive processing condition ( $M = 0.55$ ,  $SD = 0.94$ ),  $ps < 0.01$ .

No other strategy categories were found to significantly differ across conditions,  $ps > 0.02$ . The amount of information that students highlighted also did not differ by task condition,  $p = 0.52$ .

### Research Question 2: Performance Differences by Task Condition

For the second research question, we used three one-way ANOVAs to examine whether students' quality of research report writing, citation use, or responses to open-ended questions tapping integration differed by task condition. However, writing

**TABLE 4 |** Sample open-ended responses.

	Inaccurate/incomplete (0)	Summary based on a single text (1)	Summary based on multiple texts (2)
Number of people in U.S. prisons	A lot	Is the most in the world, accounts for a quarter of the world's prison population	2.3 million, makes up 25% of the world's imprisoned
Probation and parole and mass incarceration	Probation and parole are monitored from an officer	Over 40% of those on probation and parole re-offend and are sent back to prison.	Probation and parole may lessen the financial burden of mass incarceration; however, it overall will not decrease the amount of people in jail because these practices often lead to recidivism.

**TABLE 5 |** Descriptives.

	Intratextual processing	Intertextual processing	Metacognitive processing	Total
<b>Strategic processing</b>				
Total highlights	19.95 (8.21)	16.48 (7.97)	17.95 (11.59)	18.06 (9.43)
CI-ST	5.68 (4.10)	2.50 (3.10)	3.35 (3.22)	3.81 (3.68)
CI-MT	0.42 (0.84)	3.60 (3.98)	0.25 (0.44)	1.44 (2.81)
CI-PK	0.53 (1.31)	0.45 (1.57)	0.60 (1.23)	0.53 (1.36)
CA-ST	0.74 (1.41)	1.35 (2.16)	0.55 (0.94)	0.88 (1.60)
CA-MT	0.05 (0.23)	0.50 (0.89)	0.05 (0.22)	0.20 (0.58)
CA-PK	0.00 (0.00)	0.00 (0.00)	0.05 (0.22)	0.02 (0.13)
MR-ST	0.16 (0.50)	0.00 (0.00)	4.05 (3.73)	1.42 (2.87)
MR-MT	0.11 (0.32)	0.00 (0.00)	0.45 (0.89)	0.19 (0.57)
MR-PK	0.00 (0.00)	0.00 (0.00)	0.55 (0.94)	0.19 (0.60)
<b>Performance</b>				
Research report quality	4.00 (1.81)	4.53 (1.87)	4.15 (1.90)	4.23 (1.84)
Number of citations	0.83 (1.42)	1.53 (1.87)	1.65 (1.53)	1.35 (1.63)
Open-ended responses	3.94 (2.24)	3.79 (2.37)	3.63 (1.61)	3.79 (2.06)

CI is constructive-integrative processing; CA is critical-analytic processing; MR is metacognitive-reflective processing; ST is single text; MT is multiple texts; PK is prior knowledge and beliefs.

quality, citation use, and open-ended response scores did not differ by condition ( $ps > 0.26$ ).

### Research Question 3. Association Between Processing Differences and Task Performance

For our third research question, we examined the role of strategy use in students' performance on the external outcomes examined in this study (i.e., the quality of students' research reports, citation use, and responses to open-ended integration questions). For each model, prior topic knowledge was controlled for in Step 1. Experimental condition, indicator coded relative to the intratextual strategy use condition, was entered in Step 2, and the total volume of information that students highlighted, across texts, as well as students' manifest strategy use, were entered as predictors in Step 3. Because of the volume of strategy types examined in this study, and because our interest was specifically focused on students' cross-textual linking or integration-focused strategy use, only variables reflecting students' multiple text-directed strategy use, including constructive-integrative, critical-analytic, and metacognitive-reflective processing, were included in Step 3. **Table 6** includes correlations among key variables.

#### Research Report Scores

The model predicting students' research report writing quality was not significant,  $p = 0.34$ .

#### Citations

The model predicting the number of citations included in students' written responses was not significant,  $p = 0.07$ .

#### Open-Ended Integration

The model predicting students' open-ended integration scores was overall significant [ $F(7,45) = 2.72, p < 0.05, R^2_{\text{adj}} = 0.19$ ] corresponding to a medium effect. However, only students' engagement in critical-analytic strategy use directed toward multiple texts was an individually significant predictor in the model ( $\beta = 0.43, p < 0.01$ ). See **Table 7** for a model summary.

## DISCUSSION

Guided by the IF-MT, this study examined whether directing students to engage in intratextual, intertextual, or metacognitive processing in the preparation stage of multiple text learning, resulted in variable strategy use during execution, and in differences in production, or in learners' task performance. Two key findings emerged in this study. First, students' manifest strategy use was found to differ in association with the processing directives that they received prior to reading. Second, students' engagement in constructive-integrative processing directed at multiple texts was found to predict open-ended integration performance, one of the outcome measures examined in this study. We discuss each of these main findings, in turn. As a whole, results from this study align with theoretical insights from the IF-MT. In particular, using an innovative methodological approach, we establish that (a) modes of strategic processing can be instantiated via task instructions, (b) students direct strategic processing toward a variety of referents when learning from multiple texts, and (c) strategy use is associated with integration performance.

### Differences in Processing by Strategy Condition

In this study, we asked students to engage in intratextual, intertextual, or metacognitive processing when completing a multiple text task. We then tracked such processing, or students' manifest strategy use, by asking learners to highlight particular information in texts, in accordance with their strategy condition, as well as to explain their highlights. When the association between assigned mode of processing and manifest strategy use was examined, students assigned to the intertextual processing condition were found to use more constructive-integrative strategies directed at multiple texts than students asked to engage in intratextual or metacognitive processing. Likewise, directing students to engage in metacognitive strategy use during reading resulted in their significantly higher deployment of

**TABLE 6 |** Correlation among key indicators.

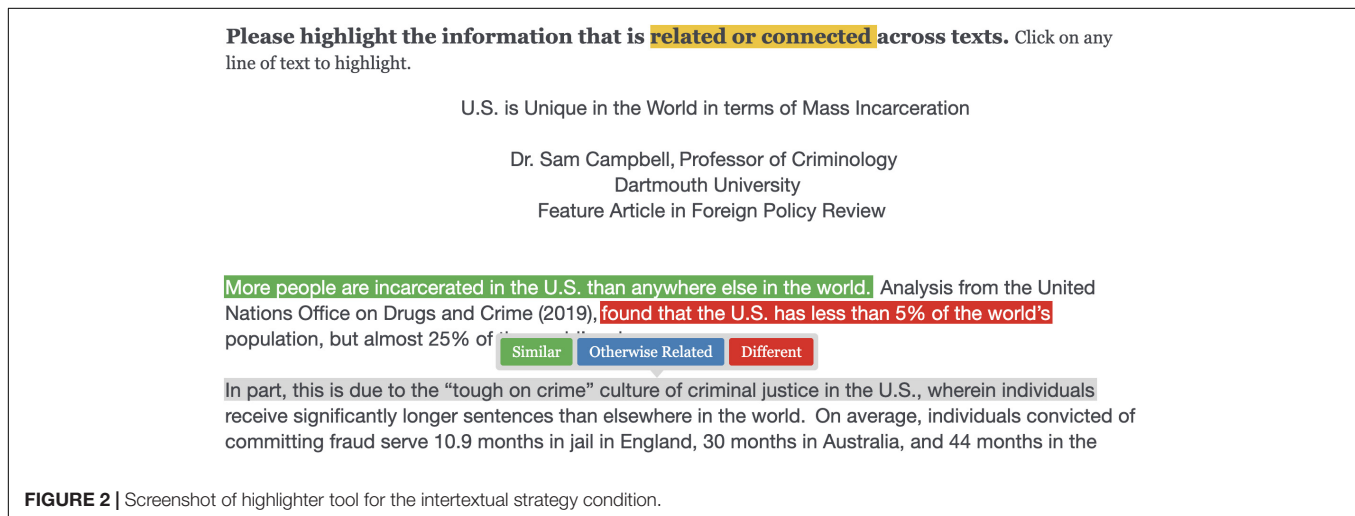
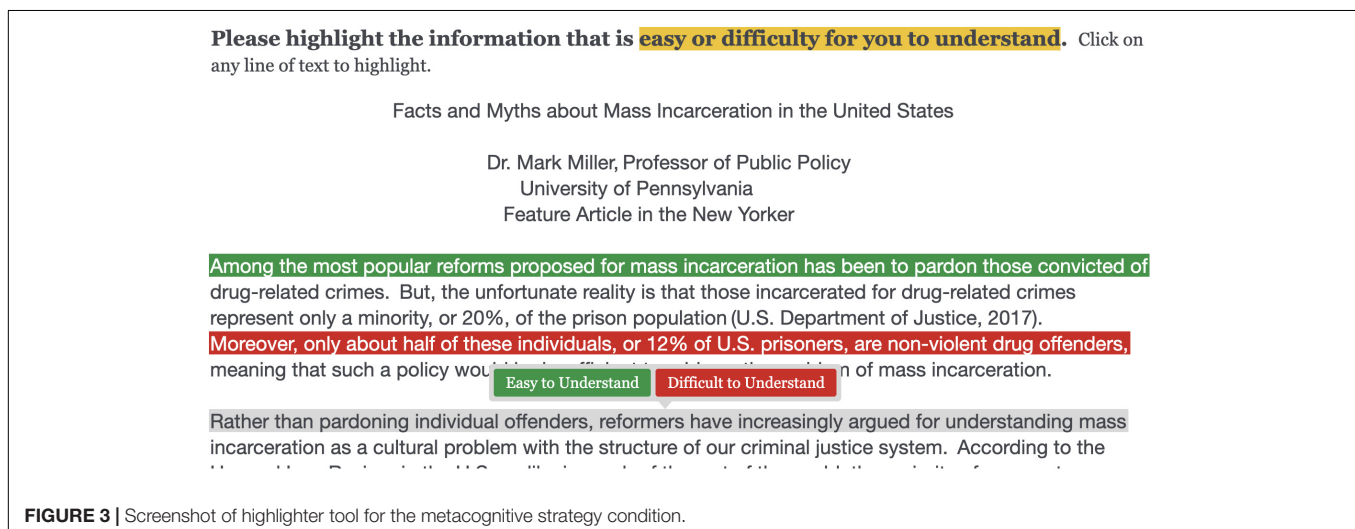
	1	2	3	4	5	6	7	8	9	10	11
(1) No. H	1.00										
(2) CI-ST	0.24	1.00									
(3) CI-MT	-0.14	-0.28*	1.00								
(4) CI-PK	0.06	0.07	-0.06	1.00							
(5) CA-ST	0.28*	-0.16	-0.01	-0.19	1.00						
(6) CA-MT	-0.08	-0.18	0.41***	0.10	0.18	1.00					
(7) MR-ST	0.06	-0.25	-0.21	-0.04	-0.15	-0.10	1.00				
(8) MR-MT	0.32*	-0.04	-0.15	-0.06	0.01	-0.06	0.53***	1.00			
(9) MR-PK	-0.10	-0.03	-0.16	0.13	-0.14	-0.11	0.39**	0.25	1.00		
(10) RR Quality	0.20	0.11	0.15	0.04	0.14	0.20	-0.10	-0.16	-0.12	1.00	
(11) Cites	-0.10	-0.07	0.27*	-0.07	0.14	0.37**	0.26	0.03	0.09	0.50***	1.00
(12) Open-ended	0.15	0.07	0.25	0.18	0.07	0.45***	-0.03	0.12	-0.03	0.51***	0.41**

CI is constructive-integrative processing; CA is critical-analytic processing; MR is metacognitive-reflective processing; ST is single text; MT is multiple texts; PK is prior knowledge and beliefs; No H. is the number of highlights; RR Quality is the quality of students' research reports. \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ .

**TABLE 7 |** Model summary predicting open-ended integration performance.

Predictors	<i>B</i>	<i>SE(B)</i>	$\beta$	<i>p</i>
<b>Step 1: Control</b>				
Prior knowledge	0.12	0.15	0.11	0.42
<b>Step 2: Strategy condition</b>				
Intertextual processing	−0.96	0.76	−0.22	0.21
Metacognitive processing	−0.22	0.63	−0.05	0.73
<b>Step 3: Strategy use</b>				
Total number of highlights	0.03	0.04	0.10	0.48
Constructive-integrative processing directed at multiple texts	0.16	0.11	0.23	0.16
Critical-analytic processing directed at multiple texts	1.45	0.49	0.43	0.005
Metacognitive-reflective processing directed at multiple texts	0.78	0.71	0.14	0.28

$F(7,45) = 2.72$ ,  $p < 0.05$ ,  $R^2_{adj} = 0.19$ . All coefficients based on last step of the model.

**FIGURE 2 |** Screenshot of highlighter tool for the intertextual strategy condition.**FIGURE 3 |** Screenshot of highlighter tool for the metacognitive strategy condition.

metacognitive-regulatory strategies direct at both single texts and at their prior knowledge or beliefs, as compared to students in the other two conditions. We draw four key conclusions based on these results.

First, as suggested by the IF-MT, the preparation and execution stages of multiple text learning were, indeed, found to be linked in this study. Prior work on learning from multiple texts, has long found task assignments asking students to

produce various types of external outcomes to be associated with differences in task performance (Wiley and Voss, 1999; Gil et al., 2010a,b; List et al., 2019a). Here we demonstrate that task assignments can further be used to specify a desired mode of processing for students to engage during reading. Instructing students to engage in particular types of processing during task completion may be a particularly effective approach to use in instances when students have inaccurate or incomplete schema for what strategies different tasks may require (Wiley et al., 2018; List et al., 2019a). Instructing students to employ particular forms of processing may also increase the frequency with which students engage in deep-level strategy use (e.g., evaluation, metacognition), with such strategy use rarely spontaneously reported (Gerjets et al., 2011; Du and List, 2020).

Indeed, in this study we were encouraged to find that asking students to engage in intertextual processing resulted in their increased strategy use directed toward multiple texts. This constitutes a key contribution of this study. In effect, while prior work has recognized the importance of students' engagement in intertextual processing, students have been found to manifest such processing to varying extents and often only in accordance with the degree of support for such processing provided by the study design (e.g., Britt and Sommer, 2004; Cerdán and Vidal-Abarca, 2008). In this study, we found a rather large volume of processing to be directed toward multiple texts, with such processing including constructive-integrative, critical-analytic, and metacognitive-reflective modes of strategy use. This suggests that the provision of task instructions to cue processing, in addition to various other physical scaffolds (e.g., highlighting; backward/forward navigation across texts), may increase students' engagement in intertextual processing during reading.

As a third point, we found it fruitful to compare the relative prevalence of the various forms of strategic processing that students exhibited in this study to those documented in prior work. In examining similar categories of strategic processing, Anmarkrud et al. (2014) found students' linking strategies (i.e., directed toward multiple texts) to most commonly reflect efforts to identify and learn important information (47.1%) and to evaluate sources and information (36.3%), with these categories generally corresponding to constructive-integrative and critical-analytic processing, respectively. In this study, too, strategies directed at multiple texts were most commonly constructive-integrative in nature, with critical-analytic and metacognitive-reflective strategies directed toward multiple texts to a considerably more limited extent. In this study, these somewhat reduced rates of critical-analytic and metacognitive-reflective strategy use, directed at multiple texts may be partly explained by the task directives we employed. That is, because students assigned to the intertextual processing strategy condition were directed to focus on multiple texts and to identify the connections among these, it seems logical that constructive-integrative strategy use dominated other strategy functions.

Among students directing any strategy functions toward multiple texts (52.54%,  $n = 31$ ), 70.24% of the multiple-text directed strategies used were focused on construction-integration, with only 19.89% of such strategies focused on metacognition-reflection and 9.87% of these focused on critical-analytic processing. This suggests that when a particular approach to processing is cued, learners' focus on such processing may come at the cost of a broader or more varied repertoire of strategy use. Alternatively, particularly in the case of critical-analytic processing, such processing may have been particularly limited both because it was not explicitly cued in any of the task conditions and because all of the texts used in this study were designed to be trustworthy in nature. Nevertheless, we were heartened by some students' efforts to corroborate and compare information across texts, as demonstrated in responses such as: "This highlight shows a different statistic that only 20% of crimes are drug related," reflecting critical-analytic processing or efforts to corroborate statistical information across texts.

When examining strategy use across conditions, a number of additional patterns emerged. For one, the majority of students' strategy functions were directed toward single texts and constructive-integrative processing. This dominance is understandable given that, at its heart, this task involved students trying to learn about a complex and controversial topic, based on information presented within a series of individual, albeit conceptually connected, texts. It therefore follows that strategies aimed, fundamentally, at constructing meaning dominated students' learning. Likewise, it seems logical that strategies directed at engaging students' prior knowledge or beliefs were relatively under-represented in this study. This may reflect the relatively low prior knowledge of our sample. At the same time, we were somewhat surprised by these results given that the topic of mass incarceration is a controversial one in the United States and, in this study, was described across texts presenting partially conflicting information. As such, we expected the controversial nature of this topic to potentially elicit students' strategy use directed toward their prior beliefs. Finally, students' metacognitive-regulatory strategy use was found to be comparatively well-represented in this study, whereas prior work has found students to engage in metacognition only to a limited extent (Du and List, 2020). Of course, this may be in large-part attributable to the task instructions that students in the metacognitive processing condition received, prior to reading. Nevertheless, results from this study seem to be an encouraging indicator that metacognitive monitoring during reading can be cued via the task instructions provided, as was previously done by Stadler and Bromme (2008).

The fourth and final conclusion is methodological in nature. In this study, we used students' highlights and associated explanations as indicators of strategy use during reading. We found doing so to be an effective method of assessing processing. Indeed, capturing the nature of students' strategy use during task completion has long proven to be a formidable challenge (Fryer and Dinsmore, 2020). On the one hand, think-aloud procedures have been effective at capturing students' online processing. On the other hand, think-alouds are data



and labor-intensive procedures that may be overly taxing for some learners (Van Gog et al., 2005; Muñoz et al., 2006). In this study, rather than using a think-aloud approach, we asked students to highlight particular information in texts, as an indicator of strategy use, and to explain these highlights. To the extent that the information students highlighted and their corresponding explanations were found to differ across task conditions in the ways expected, this approach to capturing strategic processing may be a promising one. Students' highlights and associated explanations, in this study, seemed to be effective at providing information regarding students' attendance to specific content in texts (i.e., according to the information highlighted) and associated cognitive processes (i.e., through the explanations provided). Moreover, our use of highlighting allowed us to capture students' strategic processing in an accessible and scalable way. As such, using the highlighting tool, with associated explanations, may constitute a useful method for capturing strategic processing during reading in future work.

## Predicting Task Performance Based on Strategy Use

A second finding to emerge from this study is that while students' overall task performance did not differ by strategy condition, differences in manifest strategic processing, and multiple text directed critical-analytic strategy use, in particular, were predictive of students' responses to open-ended questions tapping integration. This latter finding seems logical given that these strategies most reflected deep-level intertextual processing, while the open-ended questions used in this study assessed students' integrative understanding of complementary and conflicting issues discussed across texts.

Still, we were somewhat surprised to find that strategy use was not significant in predicting students' research report quality. In part, this lack of findings may be attributable to some limitations in sample size. Nevertheless, interpreted through the lens of the IF-MT, it may be that while the nature of students' strategic processing was associated with the cognitive outcomes that students generated, such strategy use did not carry forward to the external products (i.e., research reports) that students composed. This constitutes an explanation for why open-ended integration performance, capturing students' cognitive outcomes, was predicted by strategic processing during reading, while research report writing quality, considered to be an external outcome, was not. As suggested by List and Alexander (2019), in their description of the IF-MT, the external products that students compose based on multiple texts, in addition to reflecting a set of cognitive outcomes, also demand that students employ a variety of writing skills, not expressly specified in the IF-MT. Still, the link between students' cognitive outcomes and writing performance is clearly exhibited in this study via the strong association among these two outcome measures of interest (see Table 6).

As a more general theory of the case, our understanding of these results is that the mode of processing prompted in students in the preparation stage of the IF-MT, impacts their strategic

processing during execution. This strategic processing, in turn, then results in particular differences in task performance (i.e., the types of cognitive outcomes that students construct as a result of learning from multiple texts). Such an explanation is consistent with our not finding significant difference in task performance to manifest across strategy conditions (Research Question 2) but, nevertheless, our determining strategy use to differ by task condition (Research Question 1) and multiple text directed strategy use to predict task performance (Research Question 3). Validating such an understanding requires replicating this study and exploring mediation analyses, as we aim to do in future work. Implications for the IF-MT are, in part, that even if task assignments can be used to engender particular forms of strategic processing during execution, the degree to which such strategies are engaged and their quality are the ultimate determinants of students' production, or resultant task performance.

## Implications

There are at least four implications for theory and practice associated with this study. To start, this study is among the first to expressly use the IF-MT as a framework for understanding students' learning from multiple texts. In this study, we were able to link aspects of the task assignment, to students' processing during execution, to the quality of the external products that students developed after reading a set of multiple texts. Second, in this study we adopted an innovative and analytic approach to capturing students' processing when learning from multiple texts. That is, we were able to decompose the nature of students' strategic processing both in terms of its functions and referents. Indeed, and thirdly, we did this by adopting a novel methodological approach to capture the nature of students' strategic processing, namely the use of a highlighting tool with associated explanations. In doing so, we demonstrated that strategic processing, when captured in this manner, corresponded to the task assignment that students received prior to reading. Finally, and in line with much of the literature, we demonstrated that learners' engagement in multiple-text directed strategies, in particular, had benefits for students' integration-related task performance when learning from multiple texts.

## Limitations

Despite the strengths of this study, at least four limitations must be acknowledged.

First, in this study, we assigned students to engage in intratextual, intertextual, or metacognitive processing when learning from multiple texts. This was done to isolate the effects of each mode of strategy engagement on students' multiple text task performance. And, indeed, task assignment was found to be associated with differences in strategy engagement, as captured via the information that students highlighted and their associated explanations. Nevertheless, within the context of real-world multiple text tasks, students are likely to need to engage a variety of strategies, including all three of these types, for successful task completion. In other words, learning from multiple texts simultaneously requires students to identify relevant and important information, to connect information across texts, and to monitor text quality and their

own understanding. To the extent that cuing students' use of all of the strategies that they may need for successful task completion is unreasonable and that strategic processing should, by its very nature, be deliberately and dynamically engaged by learners, this study only demonstrates the association between particular types of strategy engagement (i.e., directed toward multiple texts) and task performance. More work is still needed to understand how such strategy engagement may be best fostered in learners.

Second, our coding of students' strategy use in this study was based on the information that students marked, using the highlighting tool, and their associated explanations for the information highlighted. While we considered this to be an effective and unobtrusive way to collect data on processing, this approach carried with it a number of limitations. For instance, there was not always a one-to-one mapping between the information that students highlighted and the associated explanations that they wrote. This requiring us to generalize that, for example, when students reported that they highlighted relevant information, this explanation pertained to all of the sentences that they indicated in-text. Further, we, as researchers, interpreted students' explanations of strategy use as serving particular functions and as directed toward specific referents. However, these interpretations, although supported by the information that students highlighted in text, should be validated with behavioral measures, like eye-tracking, in future work. Finally, asking students to type their explanations for information highlighted in association with each text rather than to report strategic processing continuously (i.e., as during a think-aloud) may have resulted in incomplete or overly-crystallized strategy reports. That is, students may have either under-reported all of the processing that they engaged during reading or refined their explanations, perhaps to better comport with task demands. Both of these possibilities are suggested by prior work (Van Gog et al., 2005).

Third, the two performance outcomes examined in this study were scored in such a way that prioritized students' multiple text integration. Although content integration, or connection-information across texts, has been identified as a central outcome in students' learning from multiple texts (Britt et al., 1999; Perfetti et al., 1999), additional factors (e.g., writing quality, organization) were not well-captured by our rubric, as aspects of external product composition. A broader range of multiple text learning outcomes, scored in a more comprehensive fashion, should potentially be considered in future work. As an added point, the emphasis on integration reflected in the rubrics used to score both students' research reports and open-ended responses may have unduly benefited students belonging to the intertextual processing strategy condition. Still, these effects were somewhat mitigated both by the truly essential role of integration in students' learning from multiple texts (i.e., we considered our prioritizing of integration to be appropriate) and by the lack of differences in task performance identified across conditions. Further, asking students to compose a research report is not a task assignment that has frequently been examined in prior work, with directing students to engage in argument composition being much more common (Wiley and Voss, 1999;

Anmarkrud et al., 2014). Nevertheless, research report writing was the task assignment used in the present study both because we wanted to encourage students' comprehensive discussion of the various key issues introduced across texts (List and Alexander, 2019; List et al., 2019a) and because the experimental texts used in this study did not have a clear, two-sided argument structure.

Finally, students completed this study online, at a time and location of their choosing and in the midst of the COVID-19 pandemic. These factors may have resulted in lower than desired recruitment and performance in this study. As such, replicating results, in both lab-based and classroom settings, constitute important areas for future work.

## CONCLUSION

In this study, we sought to contribute new insights into undergraduate students' ability to learn about a complex and controversial topic (i.e., mass incarceration) through their engagement in a multiple text task. The design of this investigation was theory-based, reflecting the phases of multiple text learning and the strategic processes articulated in the Integrated Framework of Learning from Multiple Text. In keeping with the goals of this special issue on information processing assessment and online thinking and reasoning in higher education, we presented the undergraduate students in our study with three varied task directives intended to orient their processing of information contained in four carefully orchestrated texts. Moreover, to externalize students' thinking and reasoning during task completion, without disrupting or distorting processing too much, we asked students to highlight information in texts corresponding to their particular task condition (i.e., intratextual, intertextual, and metacognitive). We then created a unique system for scoring these highlights based on the strategy functions and referents represented by each highlighted segment of text.

To extend what is known about college students' multiple text task performance, we also incorporated several measures of learning. Specifically, we assessed the quality of the research reports that students composed based on multiple texts, as well as students' responses to a series of open-ended questions specifically created to capture their integrated understanding of content introduced across the four study texts. In terms of the IF-MT, we expected that the varied processing directives that students had been given in the preparation stage, and the specific highlighter tools that students were asked to use during execution, would translate into differential research report quality and responses to open-ended integration questions in the production stage. As hypothesized, the three directives, indeed, resulted in changes in learners' processing and task performance.

All in all, what this investigation has contributed to the literature on information processing and online thinking and reasoning assessment is clear evidence that even mature readers can benefit from scaffolds that serve to orient their text processing in facilitative ways. In addition, the current study has offered alternative ways that students' thinking and reasoning can be

effectively captured during the course of task completion, along with innovative methods for scoring such thinking. Without question, there is much more to be learned about university students' information processing in online contexts and the thinking and reasoning that give rise to learning within those contexts. Nonetheless, we regard this study as a step in the right direction.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## REFERENCES

- Anmarkrud, Ø., Bråten, I., and Strømso, H. I. (2014). Multiple-documents literacy: strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learn. Individ. Differ.* 30, 64–76. doi: 10.1016/j.lindif.2013.01.007
- Barzilai, S., and Weinstock, M. (2015). Measuring epistemic thinking within and across topics: a scenario-based approach. *Contemp. Educ. Psychol.* 42, 141–158. doi: 10.1016/j.cedpsych.2015.06.006
- Bazelon, E. (2019). *If Prisons Don't Work, What Will?*. New York, NY: The New York Times.
- Brante, E. W., and Strømso, H. I. (2018). Sourcing in text comprehension: a review of interventions targeting sourcing skills. *Educ. Psychol. Rev.* 30, 773–799. doi: 10.1007/s10648-017-9421-7
- Bråten, I., Anmarkrud, Ø., Brandmo, C., and Strømso, H. I. (2014). Developing and testing a model of direct and indirect relationships between individual differences, processing, and multiple-text comprehension. *Learn. Instruct.* 30, 9–24. doi: 10.1007/s11455-018-9868-z
- Bråten, I., and Strømso, H. I. (2011). Measuring strategic processing when students read multiple texts. *Metacogn. Learn.* 6, 111–130. doi: 10.1007/s11409-011-9075-7
- Britt, M. A., Perfetti, C. A., Sandak, R. L., and Rouet, J. F. (1999). "Content integration and source separation in learning from multiple texts," in *Essays in Honor of Tom Trabasso*, ed. S. R. Goldman (Mahwah, NJ: Lawrence Erlbaum Associates, Inc), 209–233.
- Britt, M. A., and Sommer, J. (2004). Facilitating textual integration with macro-structure focusing tasks. *Read. Psychol.* 25, 313–339. doi: 10.1080/02702710490522658
- Cerdán, R., and Vidal-Abarca, E. (2008). The effects of tasks on integrating information from multiple documents. *J. Educ. Psychol.* 100, 209–222. doi: 10.1037/0022-0663.100.1.209
- Cho, B. Y., Afflerbach, P., and Han, H. (2018). "Strategic processing in accessing, comprehending, and using multiple sources online," in *Handbook of Multiple Source Use*, eds J. L. G. Braasch, I. Bråten, and M. T. McCrudden (New York, NY: Routledge), 133–150. doi: 10.4324/9781315627496-8
- Daher, T. A., and Kiewra, K. A. (2016). An investigation of SOAR study strategies for learning from multiple online resources. *Contemp. Educ. Psychol.* 46, 10–21. doi: 10.1016/j.cedpsych.2015.12.004
- Datig, I. (2016). Citation behavior of advanced undergraduate students in the social sciences: a mixed-method approach. *Behav. Soc. Sci. Librar.* 35, 64–80. doi: 10.1080/01639269.2016.1214559
- Dinsmore, D. L., and Alexander, P. A. (2012). A critical discussion of deep and surface processing: what it means, how it is measured, the role of context, and model specification. *Educ. Psychol. Rev.* 24, 499–567. doi: 10.1007/s10648-012-9198-7
- Du, H., and List, A. (2020). Researching and writing based on multiple texts. *Learn. Instruct.* 66:e0101297. doi: 10.1016/j.learninstruc.2019.101297

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Penn State IRB: STUDY00008166. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

- Fryer, L. K., and Dinsmore, D. L. (2020). The Promise and pitfalls of self-report. *Front. Learn. Res.* 8, 1–9. doi: 10.14786/flr.v8i3.623
- Gerjets, P., Kammerer, Y., and Werner, B. (2011). Measuring spontaneous and instructed evaluation processes during Web search: integrating concurrent thinking-aloud protocols and eye-tracking data. *Learn. Instruct.* 21, 220–231. doi: 10.1016/j.learninstruc.2010.02.005
- Gil, L., Bråten, I., Vidal-Abarca, E., and Strømso, H. I. (2010a). Summary versus argument tasks when working with multiple documents: which is better for whom? *Contemp. Educ. Psychol.* 35, 157–173. doi: 10.1016/j.cedpsych.2009.11.002
- Gil, L., Bråten, I., Vidal-Abarca, E., and Strømso, H. I. (2010b). Understanding and integrating multiple science texts: summary tasks are sometimes better than argument tasks. *Read. Psychol.* 31, 30–68. doi: 10.1080/02702710902733600
- Goldman, S. R., Braasch, J. L., Wiley, J., Graesser, A. C., and Brodowska, K. (2012). Comprehending and learning from internet sources: processing patterns of better and poorer learners. *Read. Res. Q.* 47, 356–381. doi: 10.1002/RRQ.027
- Hagen, Å.M., Braasch, J. L., and Bråten, I. (2014). Relationships between spontaneous note-taking, self-reported strategies and comprehension when reading multiple texts in different task conditions. *J. Res. Read.* 37, 141–157. doi: 10.1111/j.1467-9817.2012.01536.x
- Hendley, M. (2012). Citation behavior of undergraduate students: a study of history, political science, and sociology papers. *Behav. Soc. Sci. Librar.* 31, 96–111. doi: 10.1080/01639269.2012.679884
- Kobayashi, K. (2009). Comprehension of relations among controversial texts: effects of external strategy use. *Instruct. Sci.* 37, 311–324. doi: 10.1007/s11251-007-9041-6
- Le Bigot, L., and Rouet, J. F. (2007). The impact of presentation format, task assignment, and prior knowledge on students' comprehension of multiple online documents. *J. Literacy Res.* 39, 445–470. doi: 10.1080/10862960701675317
- List, A. (2020). "Six questions regarding strategy use when learning from multiple texts," in *Handbook of Strategies and Strategic Processing*, eds D. L. Dinsmore, L. K. Fryer, and M. M. Parkinson (New York, NY: Routledge), 119–140. doi: 10.4324/9780429423635-8
- List, A., and Alexander, P. A. (2017). Text navigation in multiple source use. *Comput. Human Behav.* 75, 364–375. doi: 10.1016/j.chb.2017.05.024
- List, A., and Alexander, P. A. (2019). Toward an integrated framework of multiple text use. *Educ. Psychol.* 54, 20–39. doi: 10.1080/00461520.2018.1505514
- List, A., Du, H., and Wang, Y. (2019a). Understanding students' conceptions of task assignments. *Contemp. Educ. Psychol.* 59:101801. doi: 10.1016/j.cedpsych.2019.101801
- List, A., Du, H., Wang, Y., and Lee, H. Y. (2019b). Toward a typology of integration: examining the documents model framework. *Contemp. Educ. Psychol.* 58, 228–242. doi: 10.1016/j.cedpsych.2019.03.003
- McCarthy, K. S., and Goldman, S. R. (2015). Comprehension of short stories: effects of task instructions on literary interpretation. *Discourse Process.* 52, 585–608. doi: 10.1080/0163853X.2014.967610

- McCrudden, M. T. (2018). "Text relevance and multiple-source use," in *Handbook of Multiple Source Use*, eds J. L. G. Braasch, I. Bråten, and M. T. McCrudden (New York, NY: Routledge), 168–183. doi: 10.4324/9781315627496-10
- McCrudden, M. T., Magliano, J. P., and Schraw, G. (2011). The effect of diagrams on online reading processes and memory. *Discourse Process.* 48, 69–92. doi: 10.1080/01638531003694561
- McCrudden, M. T., and Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educ. Psychol. Rev.* 19, 113–139. doi: 10.1007/s10648-006-9010-7
- McCrudden, M. T., and Sparks, P. C. (2014). Exploring the effect of task instructions on topic beliefs and topic belief justifications: a mixed methods study. *Contemp. Educ. Psychol.* 39, 1–11. doi: 10.1016/j.cedpsych.2013.10.001
- McNamara, D. S. (2004). SERT: self-explanation reading training. *Discourse Process.* 38, 1–30. doi: 10.1207/s15326950dp3801\_1
- Muñoz, B., Magliano, J. P., Sheridan, R., and McNamara, D. S. (2006). Typing versus thinking aloud when reading: implications for computer-based assessment and training tools. *Behav. Res. Methods* 38, 211–217. doi: 10.3758/BF03192771
- Parkinson, M. M., and Dinsmore, D. L. (2018). Multiple aspects of high school students' strategic processing on reading outcomes: the role of quantity, quality, and conjunctive strategy use. *Br. J. Educ. Psychol.* 88, 42–62. doi: 10.1111/bjep.12176
- Perfetti, C. A., Rouet, J.-F., and Britt, M. A. (1999). "Towards a theory of documents representation," in *The Construction of Mental Representations During Reading*, eds H. van Oostendorp and S. R. Goldman (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc), 99–122.
- Potocki, A., Ros, C., Vibert, N., and Rouet, J. F. (2017). Children's visual scanning of textual documents: effects of document organization, search goals, and metatextual knowledge. *Sci. Stud. Read.* 21, 480–497. doi: 10.1080/10888438.2017.1334060
- Stadtler, M., and Bromme, R. (2008). Effects of the metacognitive computer-tool met. a. ware on the web search of laypersons. *Comput. Hum. Behav.* 24, 716–737. doi: 10.1016/j.chb.2007.01.023
- Uhrmacher, K. (2020). *Where 2020 Democrats Stand on Criminal Justice*. Washington, DC: The Washington Post.
- Van Gog, T., Paas, F., Van Merriënboer, J. J., and Witte, P. (2005). Uncovering the problem-solving process: cued retrospective reporting versus concurrent and retrospective reporting. *J. Exper. Psychol. Appl.* 11, 237–244. doi: 10.1037/1076-898X.11.4.237
- Wang, Y., and List, A. (2019). Calibration in multiple text use. *Metacogn. Learn.* 14, 131–166. doi: 10.1007/s11409-019-09201-y
- Weston-Sementelli, J. L., Allen, L. K., and McNamara, D. S. (2018). Comprehension and writing strategy training improves performance on content-specific source-based writing tasks. *Intern. J. Artif. Intellig. Educ.* 28, 106–137. doi: 10.1007/s40593-016-0127-7
- Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., and Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. *Am. Educ. Res. J.* 46, 1060–1106. doi: 10.3102/0002831209333183
- Wiley, J., Jaeger, A. J., and Griffin, T. D. (2018). "Effects of instructional conditions on comprehension from multiple sources in history and science," in *Handbook of Multiple Source Use*, eds J. L. G. Braasch, I. Bråten, and M. T. McCrudden (New York, NY: Routledge), 341–361. doi: 10.4324/9781315627496-20
- Wiley, J., and Voss, J. F. (1999). Constructing arguments from multiple sources: tasks that promote understanding and not just memory for text. *J. Educ. Psychol.* 91, 301–311. doi: 10.1037/0022-0663.91.2.301
- Wineburg, S. S. (1991). Historical problem solving: a study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *J. Educ. Psychol.* 83, 73–87. doi: 10.1037/0022-0663.83.1.73
- Wolfe, M. B., and Goldman, S. R. (2005). Relations between adolescents' text processing and reasoning. *Cogn. Instruct.* 23, 467–502. doi: 10.1207/s1532690xc2304\_2
- Woloshyn, V. E., Paivio, A., and Pressley, M. (1994). Use of elaborative interrogation to help students acquire information consistent with prior knowledge and information inconsistent with prior knowledge. *J. Educ. Psychol.* 86, 79–89. doi: 10.1037/0022-0663.86.1.79

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 List and Alexander. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# More Than (Single) Text Comprehension? – On University Students' Understanding of Multiple Documents

Nina Mahlow<sup>1\*</sup>, Carolin Hahnel<sup>2,3</sup>, Ulf Kroehne<sup>2</sup>, Cordula Artelt<sup>1,4</sup>, Frank Goldhammer<sup>2,3</sup> and Cornelia Schoor<sup>5</sup>

<sup>1</sup> Leibniz Institute for Educational Trajectories (LIfBi), Bamberg, Germany, <sup>2</sup> DIPF Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany, <sup>3</sup> Centre for International Student Assessment (ZIB), Frankfurt am Main, Germany, <sup>4</sup> Department of Longitudinal Educational Research, University of Bamberg, Bamberg, Germany, <sup>5</sup> Department of Educational Research, University of Bamberg, Bamberg, Germany

## OPEN ACCESS

### Edited by:

Olga Zlatkin-Troitschanskaia,  
Johannes Gutenberg University  
Mainz, Germany

### Reviewed by:

Helge I. Strømsø,  
University of Oslo, Norway  
Alexandra List,  
Pennsylvania State University (PSU),  
United States

### \*Correspondence:

Nina Mahlow  
nina.mahlow@lifbi.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Psychology

**Received:** 15 May 2020

**Accepted:** 18 September 2020

**Published:** 23 October 2020

### Citation:

Mahlow N, Hahnel C, Kroehne U,  
Artelt C, Goldhammer F and  
Schoor C (2020) More Than (Single)  
Text Comprehension? – On University  
Students' Understanding of Multiple  
Documents.  
Front. Psychol. 11:562450.  
doi: 10.3389/fpsyg.2020.562450

The digital revolution has made a multitude of text documents from highly diverse perspectives on almost any topic easily available. Accordingly, the ability to integrate and evaluate information from different sources, known as multiple document comprehension, has become increasingly important. Because multiple document comprehension requires the integration of content and source information across texts, it is assumed to exceed the demands of single text comprehension due to the inclusion of two additional mental representations: the integrated situation model and the intertext model. To date, there is little empirical evidence on commonalities and differences between single text and multiple document comprehension. Although the relationships between single text and multiple document comprehension can be well distinguished conceptually, there is a lack of empirical studies supporting these assumptions. Therefore, we investigated the dimensional structure of single text and multiple document comprehension with similar test setups. We examined commonalities and differences between the two forms of text comprehension in terms of their relations to final school exam grades, level of university studies and university performance. Using a sample of  $n = 501$  students from two German universities, we jointly modeled single text and multiple document comprehension and applied a series of regression models. Concerning the relationship between single text and multiple document comprehension, confirmatory dimensionality analyses revealed the best fit for a model with two separate factors (latent correlation: 0.84) compared to a two-dimensional model with cross-loadings and fixed covariance between the latent factors and a model with a general factor. Accordingly, the results indicate that single text and multiple document comprehension are separable yet correlated constructs. Furthermore, we found that final school exam grades, level of university studies and prior university performance statistically significant predicted both single text and multiple document comprehension and that expected future university performance was predicted by multiple document

comprehension. There were also statistically significant relationships between multiple document comprehension and these variables when single text comprehension was taken into account. The results imply that multiple document comprehension is a construct that is closely related to single text comprehension yet empirically differs from it.

**Keywords:** multiple document comprehension, single text comprehension, university students, reading comprehension, assessment

## INTRODUCTION

Reading is a core competence for societal and other forms of participation (OECD, 2010). It is assumed to be necessary for knowledge acquisition and skills development. However, reading *per se* as well as the demands readers need to meet have changed a lot as a result of the digital revolution. The ubiquity of the internet allows people to retrieve information and generate knowledge anytime and everywhere. This has led not only to changes in the modality of reading sources from paper-based to computer-based (e.g., Singer and Alexander, 2017; Kroehne et al., 2019), but increasingly requires readers to be able to integrate and evaluate information from different sources (List and Alexander, 2017a) due to the accessibility and multitude of available information. This competence, known as multiple document comprehension (MDC; e.g., Bråten and Strømsø, 2010a), entails the successful understanding, representation and integration of information from texts on the same subject matter stemming from different sources (also referred to as multiple documents).

Integrating and evaluating text information is especially relevant for university students, who need to become familiar with different topics and must be able to autonomously find information in order to study for an exam, give a presentation or review available literature for a term paper. In the course of such tasks, they might encounter multiple documents that provide redundant, complementary or even conflicting information (Bråten et al., 2014). Students have to determine the similarities and differences between texts in order to establish a coherent representation of who said what. There is evidence that a large number of students have problems with the demands of processing more than a single text (for an overview see Britt and Rouet, 2012). However, there are studies indicating that MDC can be improved through interventions (Britt and Aglinskis, 2002) and that it increases over the course of students' university studies (Schoor et al., 2020b; von der Mühlen et al., 2016).

Until the mid-1990s, models of reading comprehension focused on single text comprehension (STC; e.g., Kintsch, 1988; Trabasso et al., 1989; Graesser et al., 1994; Zwaan et al., 1995), leading to the development of reading comprehension tests based on the extraction of meaning from single texts with a single source. This changed in the late 1990s, when Perfetti et al. (1999) published the documents model framework. This framework addressed the expanded demands of MDC compared to STC by adding additional mental representations. These additional representations are referred to as the integrated situation model (integration of the content of multiple documents) and the intertext model (integration of source information from multiple

documents, e.g., the author or publishing date), both of which are part of the documents model (Perfetti et al., 1999; Britt and Rouet, 2012). Since then, numerous researchers have drawn upon this theoretical foundation to build various models that shed light on the different conditions and mechanisms involved in building a documents model. These models concern the interaction between person, task and text, although they focus on different elements (D-ISC model by Braasch and Bråten, 2017; CAEM by List and Alexander, 2017b; two-step validation model by Richter and Maier, 2017; RESOLV model by Rouet et al., 2017; content-source integration model by Stadler and Bromme, 2014).

Despite these theoretical efforts, the question regarding particular requirements of MDC – and therefore the structure of the relationship between single text and multiple document comprehension – remains insufficiently clarified (Stadler, 2017; Strømsø, 2017). Since the dimensionality of STC and MDC has not yet been examined, the present study addresses this research gap. Specifically, this study applies a newly developed test measuring MDC (Schoor et al., 2020a,b). The test covers all facets of mental representations within MDC; it thus includes not only the integrated situation model component, as it has often been addressed in former studies using expressive and receptive tasks (for an overview see Primor and Katzir, 2018), but also the intertext model and documents model components. To measure STC, a standardized and approved instrument from the National Educational Panel Study that taps important cognitive requirements for reading was used (Gehrer et al., 2013). Both MDC and STC are abilities that students should learn in school before entering university, but it can be expected that they further develop during students' university studies (e.g., Schoor et al., 2020b; von der Mühlen et al., 2016, for MDC). Due to our focus on MDC, we also examined the relation between MDC test scores and students' level of university studies, final school exam grades and university performance. Furthermore, we examined these relations when including STC in the models in order to investigate whether this provides additional insights into the relationship between MDC and STC.

## THEORETICAL BACKGROUND

### Single Text Comprehension

Single text comprehension (also often referred to as text or reading comprehension) is the result of a process of extracting meaning from text and establishing a coherent mental representation of the text content. It comprises several cognitive

component skills at the word, sentence, and text level (e.g., Perfetti et al., 2005) which also differentiate skilled and poor readers. Reading comprehension is required for literally all higher-level cognitive activities, such as learning, logical thinking, problem solving and decision making (Kintsch, 1988).

Since there is a long research tradition in the field of reading comprehension, different comprehension models have been developed (for an overview of seven prominent models see McNamara and Magliano, 2009). Some have focused on basic and general comprehension processes and verbal efficiency (Perfetti, 1985; Kintsch, 1988; Gernsbacher, 1991; van den Broek et al., 1999), others focus primarily on inference processes and on retrieving prior knowledge (e.g., Trabasso et al., 1989; Graesser et al., 1994; Zwaan et al., 1995). In his seminal paper, Kintsch (1988), suggested that readers construct three layers of representations. (1) The *surface code* or surface level is created through decoding processes of the verbatim text in order to construct a representation of the text string (lexical and syntactical structure). (2) The *textbase level* is the first level of meaning, in which the explicit content of the text is represented by the reader. (3) Deep meaning is established through the construction of a *situation model*. This represents the interpretation level, since prior knowledge and inferences are used here to build an elaborate and coherent interpretation of the information provided in the text.

Taken together, STC involves several cognitive activities and strategies, such as establishing local and global coherence relations, drawing knowledge-based inferences (e.g., Graesser et al., 1994; Oakhill et al., 2003), monitoring the plausibility of the text (Isberner and Richter, 2014) and monitoring the comprehension process itself (Cain, 2009).

When assessing overall reading comprehension abilities, test developers often adopt a result-oriented perspective condensing internal structures and processes. At the same time care has to be taken to ensure that the demands implemented in a reading test match with the findings of cognitive research. For example, STC assessments in large-scale assessments often follow a functional perspective on reading. Reading literacy in this sense encompasses “an individual’s capacity to understand, use, reflect on and engage with written texts, in order to achieve one’s goals, to develop one’s knowledge and potential, and to participate in society” (OECD, 2010, p. 14). Such studies are the Programme for International Student Assessment (PISA; OECD, 1999), the International Adult Literacy Survey (IALS; OECD and Statistics Canada, 1995) or the National Educational Panel Study (NEPS; Blossfeld and Roßbach, 2019).

## Multiple Document Comprehension

Research on understanding multiple documents indicates the need to expand models of STC (e.g., Goldman, 2004; Rouet, 2006; Bråten et al., 2011; Britt and Rouet, 2012). Out of this thought, the documents model framework (Perfetti et al., 1999; Britt and Rouet, 2012) was developed. It suggests that, in addition to these demands of STC, the comprehension of multiple documents requires two additional mental representations: the *integrated situation model* and the *intertext model*, which are the components of the documents model. The former represents

the integration of the situation models for each single text, resulting in a global representation of the situation or phenomena described across the texts. This can be challenging when the reader encounters a conflict due to contradictory or incompatible information. In this case, students can either ignore the conflict, reconcile it or accept it as being due to different sources (Stadtler and Bromme, 2014). The intertext model represents information about single sources (e.g., information about the author, purpose or publication medium) as well as its integration across texts. The whole documents model encompasses the linkage between content and source information (e.g., who stated what). Beyond these cognitive representations, Wineburg (1991) found that the strategies of corroboration (comparing information across documents), contextualization (relating information about the documents’ context to prior knowledge), and sourcing (considering information about sources) are important for understanding multiple documents in history, since experts engaged more in such behaviors compared to novices. These strategies have also been identified as important in other domains like science (e.g., Bråten et al., 2011) even if they are occasionally viewed somewhat differently. Due to domain-specific characteristics, benevolence and expertise are important source attributes in science (Stadtler and Bromme, 2014), whereas in history sources are needed particularly in order to contextualize documents (Wineburg, 1991). Although variability has been identified between domains with respect to related practices, there are commonalities, such as engaging in close reading or constructing arguments that explain the logic of claims which apply to nearly every domain (Goldman et al., 2016), which justify considering MDC as a cross-disciplinary competence.

The enhanced demands of MDC are especially important for university students, who face multiple documents regularly when searching for literature in scientific databases or reading texts assigned by their course instructors. Nevertheless, even high school graduates should be able to extract the meaning of multiple documents, analyze and evaluate their content and employ the texts in their own learning process (Common Core State Standards, 2010; Kultusministerkonferenz, 2012). Studies suggest that upper elementary school children are already capable of processing multiple documents (Beker et al., 2019; Florit et al., 2019). Schoor et al. (2020b) found that the MDC of university students correlated statistically significant with their final school exam grades, indicating that high-performing students performed better in an MDC test than low-performing students. Britt and Aglinskias (2002) provided high school and college students with training in handling multiple documents and found that MDC can be modified. Even though disciplines differ in the extent to which they require students to handle multiple documents, there are indications that MDC develops positively during the course of university studies (von der Mühlen et al., 2016). This is in line with Schoor et al. (2020b), who found that master’s students outperform bachelor’s students on MDC-related tasks. Thus, MDC is a competence that seems to develop in school and is needed in order to successfully graduate from university, since university graduates should be able to gather, evaluate and interpret relevant information and derive scientifically sound judgments. Accordingly, relations to final

school exam grades and level of university studies can be found (Schoor et al., 2020a,b).

To date, most studies have assessed MDC by means of essays or intertextual inference verification tasks (for an overview see Primor and Katzir, 2018; for an exception see Schoor et al., 2020b). Essay tasks are defined as expressive tasks in which participants have to write a summary based on multiple documents they have read. Since essays are rated with regard to numerous aspects, their scoring is time-consuming. Additionally, they might measure writing skills in addition to MDC (Griffin et al., 2012). Intertextual inference verification tasks, in contrast, are receptive tasks in which participants have to evaluate the veracity of statements by combining information from different texts. Although this method is time-saving and can be objectively scored, it has the drawback of capturing only the integrated situation model and therefore not taking into account the intertext model or the whole documents model. On the other hand, there are studies which investigate sourcing during multiple text comprehension. Some of these focus on source memory (Maier and Richter, 2013; Braasch et al., 2016; Bråten et al., 2016), others on think aloud assessments (Anmarkrud et al., 2014; Barzilai et al., 2015; Strømsø and Bråten, 2014) which are more process-oriented measures of trustworthiness and refer less to sourcing as a retrospective mental model. To overcome the issue of focusing on subcomponents of MDC, Schoor et al. (2020a,b) developed an MDC test addressing all components of the documents model framework (Britt and Rouet, 2012, see section “Multiple Document Comprehension Measure”). However, the authors state that the relation between STC and MDC remains unclear.

## Relation of Multiple Document Comprehension to Single Text Comprehension

To summarize the abovementioned information, MDC exceeds the demands of STC in several ways. It frequently requires readers to compare and integrate information not only within but across documents, which becomes apparent in the integrated situation model and intertext model component of the documents model framework (Britt and Rouet, 2012).

Afflerbach and Cho (2009) showed that the differentiation between three categories of strategy use for traditional (single) texts can also be applied to (more extensive) reading strategies used with multiple documents. These categories are identifying and remembering important information, monitoring, and evaluating. Differences between single-text and multiple-document strategies within these categories are due to the different demands of MDC and STC.

According to Rouet (2006), the demands of MDC differ from the demands of STC in three ways: the relationship between documents, the distinction between texts and situation, and the role of source information. Imagine reading a newspaper article about a battle scene claiming that several people were wounded (Text A). First, multiple documents can complement each other in different ways. To illustrate this, a second text on this issue (Text B) might be complementary, and thus fill in gaps left

by the first text, or contradictory, thus representing different aspects of a situation. Secondly, multiple documents emphasize the distinction between a text and the situation described. Two or more texts can either describe different situations or – and this is what is typically referred to as “multiple documents” – describe one situation from different or similar perspectives. Imagine that the second text (Text B) states that no people were wounded. Since you do not have prior knowledge about this event, you consider that each scenario has a 50% chance of being true. Another text on the same issue (Text C) provides support for the point of view presented in Text A, thus claiming that several people were injured. Maybe you now think that this scenario is more likely to be true than the scenario claiming that no people were wounded. This example illustrates how the updating of prior knowledge and beliefs affects the comprehension of multiple documents (Richter and Maier, 2017). Thirdly, source information is especially important when reading multiple documents. The documents model framework assumes that readers experience texts as social entities which are embedded in a specific context (Britt et al., 2012). Readers can evaluate each text by devoting attention to the characteristics of the author(s), genre, publication date, intended audience and so on which in turn helps them build a representation of the situation described in the texts (especially whom to believe). Imagine that Text B is from a trustworthy source (e.g., an eyewitness or governmental organization), while Texts A and C were written by less objective authors, such as a protester or politicians who might benefit from the conflict. This will probably lead you to believe that the scenario with no casualties is the true one. Source information is especially important when the reader detects a conflict between the texts, as was shown by Braasch and Bråten (2017). This is more often the case when reading multiple documents than when reading a single text, since authors of multiple documents generally do not coordinate their work.

To date, studies examining the relation between MDC and STC are scarce. A study by Stadtler et al. (2013) found that reading multiple documents increased awareness and description of conflicts in comparison to reading a single text with the same content. This finding is in line with previous research stating that information arranged as multiple documents results in a better integrated mental representation than reading the same information within a single text (Wiley and Voss, 1996, 1999; Britt and Aglinskas, 2002; Bråten and Strømsø, 2006). For example, Britt and Aglinskas (2002) investigated whether the effectiveness of the Sourcer's Apprentice (a computer-based environment for teaching sourcing skills with multiple documents) was due to the nature of the environment or to the particular materials. They found that the students who were trained with the Sourcer's Apprentice showed better sourcing performance and tended to write better connected essays than the students who read a text-book version (single text) of the same training materials. These results are consistent with findings of Wiley and Voss (1999), who showed that writing an argumentative essay results in better information integration when reading multiple documents and not a single-text website. Furthermore, some studies measuring MDC via the intertextual verification task also measured participants' understanding of

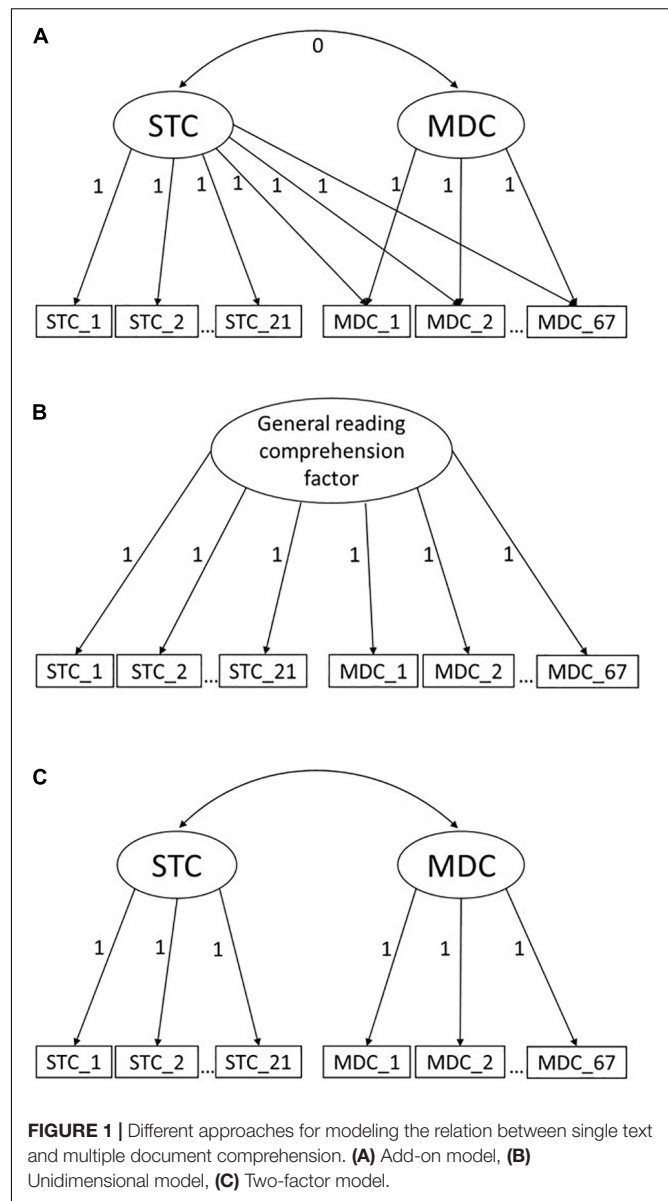


single texts via an intratextual inference verification task. In the latter task, participants had to evaluate the veracity of statements combining information from different parts of a single text. The results indicate that these STC measures correlate moderately, but statistically significantly with the MDC measures, with correlations between 0.41 and 0.58 (Strømsø et al., 2008, 2010; Bråten et al., 2009; Strømsø and Bråten, 2009; Bråten and Strømsø, 2010b, 2011; Gil et al., 2010; Hagen et al., 2014). However, there are currently no studies comparing the dimensionality of the constructs STC and MDC.

## The Present Study

Given the recent changes in reading-related demands and parallel changes in theoretical models and assessments of reading comprehension, we were primarily interested in the relation between MDC and STC from an individual differences perspective. Theoretical assumptions, such as the documents model framework (Britt and Rouet, 2012), suggest that the comprehension of multiple documents demands more from readers than STC, which raises the question of how the constructs STC and MDC are related to each other. In the documents model framework by Britt and Rouet (2012), STC is needed in order to complete a multiple document task. Indeed, the STC is inherent in the documents model framework since each document has to be read and understood in order to establish a documents model. Therefore, it can be expected that there is a common underlying factor for both STC and MDC, namely the ability to read and comprehend text. Considering the mentioned changes in reading-related requirements in addition, we hypothesize that not only the ability to read is necessary to solve MDC tasks, but also another ability that is independent of this, such as the ability to integrate multiple perspectives or keep multiple perspectives in mind (in the sense of a working memory-related ability). We call this the add-on hypothesis (for the underlying model see **Figure 1A**). It suggests that there may be students who are generally very skilled in reading and understanding texts, but who struggle with this additional requirement of MDC tasks. However, two alternative relations between MDC and STC are conceivable as well, as described below.

As a first alternative, STC and MDC might represent the same construct (unidimensionality hypothesis, see **Figure 1B** for the underlying model). Like our hypothesis, this alternative assumes a common underlying factor for both STC and MDC. In contrast, there is no additional source of inter-individual differences related to solving MDC tasks. Accordingly, students with a low overall ability are expected to score low in both STC and MDC tasks, and students with a high overall ability are expected to score high in both STC and MDC tasks. Differences in performance on STC and MDC tasks are reflected in the item difficulty. This is the approach used by PISA, which is a triennial international study that measures 15-year-olds' competences in reading, mathematics and science. The new PISA 2018 reading framework (OECD, 2010) expands the reading literacy concept through the use of a few multiple documents tasks. However, in the PISA framework, the multiple document items are not necessarily presented as more difficult than single text items, although single text and multiple document items capture a unidimensional reading construct.



As a second alternative, MDC and STC might be two separable constructs that have different characteristics and refer to distinct single and multiple text reading situations (two-factor hypothesis, see **Figure 1C** for the underlying model). Accordingly, there will be students who will be more able to comprehend texts in one situation than in the other. However, by expecting a high correlation between these factors, it is recognized that they are related and to some extent based on a number of common underlying abilities (e.g., decoding words and language comprehension). This is the case for competences assessed in different domains within international large-scale assessments. For example, in the PISA studies, reading literacy, mathematical literacy, and scientific literacy are separate constructs but still highly correlated (for PISA 2009 see OECD, 2012). A whole host of factors are probably involved in this covariation, intelligence being one of them (Baumert et al., 2009). Furthermore, STC

is also required for mathematical or scientific tasks when formulated in text form (e.g., problem solving tasks, OECD, 1999; Baumert et al., 2009). The two-factor hypothesis differs from the add-on hypothesis in the relation and operationalization of the constructs. In the add-on hypothesis, the additional competence to deal with multiple documents is independent of the competence of reading and comprehending texts. In the two-factor hypothesis the constructs MDC and STC are related since the factors are allowed to correlate. A high correlation between the two factors can be expected.

Based on these considerations, we postulate three competing and mutually exclusive hypotheses:

H1a: MDC is based on STC, but goes beyond reading-specific requirements, since additional cognitive processes are also required (add-on hypothesis).

H1b: MDC and STC represent the same construct (unidimensionality hypothesis).

H1c: MDC and STC represent separable constructs that are highly correlated (two-factor hypothesis).

Furthermore, we were interested in how MDC test scores are related to final school exam grades, the level of university studies and university performance. Since MDC is a competence that can already be observed in upper elementary school children (Beker et al., 2019; Florit et al., 2019) and develops further during the course of university studies (von der Mühlen et al., 2016; Schoor et al., 2020b), we expected that MDC test scores...

... are predicted positively by students' final school exam grades (H2),

... are predicted more strongly by the final school exam grades among bachelor's students than among master's students (H3),

... are higher among master's students than among bachelor's students (H4).

To the best of our knowledge, the relation between university performance (indicated by bachelor's and master's degree grade point averages) and MDC has not yet been investigated. Nevertheless, since there is evidence that MDC develops positively during the course of university studies (von der Mühlen et al., 2016), and MDC is a necessary component for the successful completion of university, we expected that MDC test scores...

... are predicted positively by prior university performance (H5),

... positively predict expected future university performance (H6).

In light of H1, we were also interested in exploring the relationships specified in H2–H6 conditional on STC. If MDC still relates to these variables even when the shared variance with STC is removed, this will deliver additional evidence that MDC

represents a separable construct providing relevant additional information about readers.

## MATERIALS AND METHODS

### Sample

The original sample consisted of 508 university students from two German universities enrolled in different programs within the humanities and social sciences. In order to prevent bias due to fluency in German, we excluded seven non-native speakers who had been learning German for less than 10 years. The resulting sample of 501 university students still included four non-native speakers who had spoken German for at least 17 years. The participants' age in the reduced sample spanned from 17 to 42 years ( $M = 22.76$ ,  $SD = 3.77$ , 78% female). The sample consisted of 53% ( $n = 264$ ) first-semester bachelor's students and 46% ( $n = 232$ ) master's students (who had been studying for 1–14 semesters). One percent of the sample ( $n = 5$ ) were teacher education students or students enrolled in old qualification formats like the university diploma (who had been studying for 8–18 semesters). The participants' final school exam grades (German Abiturnoten) ranged from 1.0 to 3.7 ( $M = 2.12$ ,  $SD = 0.66$ ,  $n = 493$ ). German Abitur grades and final university grades range from 1 ("very good") to 4 ("sufficient", pass mark). The bachelor's degree grade point averages of the master's students ranged from 1.1 to 2.8 ( $M = 1.83$ ,  $SD = 0.39$ ,  $n = 230$ ). The anticipated master's degree grade point average of the master's students ranged from 1.0 to 3.0 ( $M = 1.70$ ,  $SD = 0.34$ ,  $n = 229$ ).

### Design and Procedure

After the students provided their informed consent for participation, the study started with a questionnaire about demographic variables, such as the students' final school exam grades and level of university studies. Master's students were also asked for their bachelor's degree grade point average and anticipated master's degree grade point average. Afterward, the participants had to complete three blocks, which were presented in randomized order (booklet design). Between these blocks, participants had the chance to take a short break. The blocks consisted of either the MDC test, the STC test or a working memory test, which was not the focus of this study. Each participant completed two out of five units of the MDC test, which were administered in a balanced incomplete block design. They had the opportunity to take a break between units. The entire test session took about two hours. Both tests had a unit structure, as described in the following section.

### Measures

In order to ensure the comparability of the STC and MDC constructs, structurally similar tests were employed in the present study. This means that both tests had a unit structure [a unit is defined by text(s) plus items] and a similar navigation (e.g., participants could return to the texts at any time and texts and items were presented on different pages).

**TABLE 1** | Text characteristics of the MDC test.

Unit name	Number of texts	Unit content	Number of items	Number of words <sup>1</sup>	Readability (LIX) <sup>2</sup>	Readability (FRE) <sup>3</sup>	Claimed sources
Universe	3	Texts provide information about the end of the universe from a physics and cosmology perspective	15	455, 464, 448	41.5–45.5	50–55	Newspaper articles
Catalano	2	Biographies on the life of the fictitious mafia boss Catalano	11	644, 584	46.4–49.6	45–52	Online article from a criminological institute; economic newspaper article
2134	3	Texts describe an event in the year 2134: the arrival of aliens on earth	11	491, 434, 381	50.7–54.2	29–43	Internal laboratory report; internal government report; political speech
Nothing	2	Reviews of the fictitious novel ‘Nothing’	13	723, 562	47.1–51.8	43–51	Newspaper articles
Animals	3	Texts talk about different fictitious approaches to interpreting animals in novels	17	629, 1057, 451	51.1–55.0	32–40	Introductory textbook texts

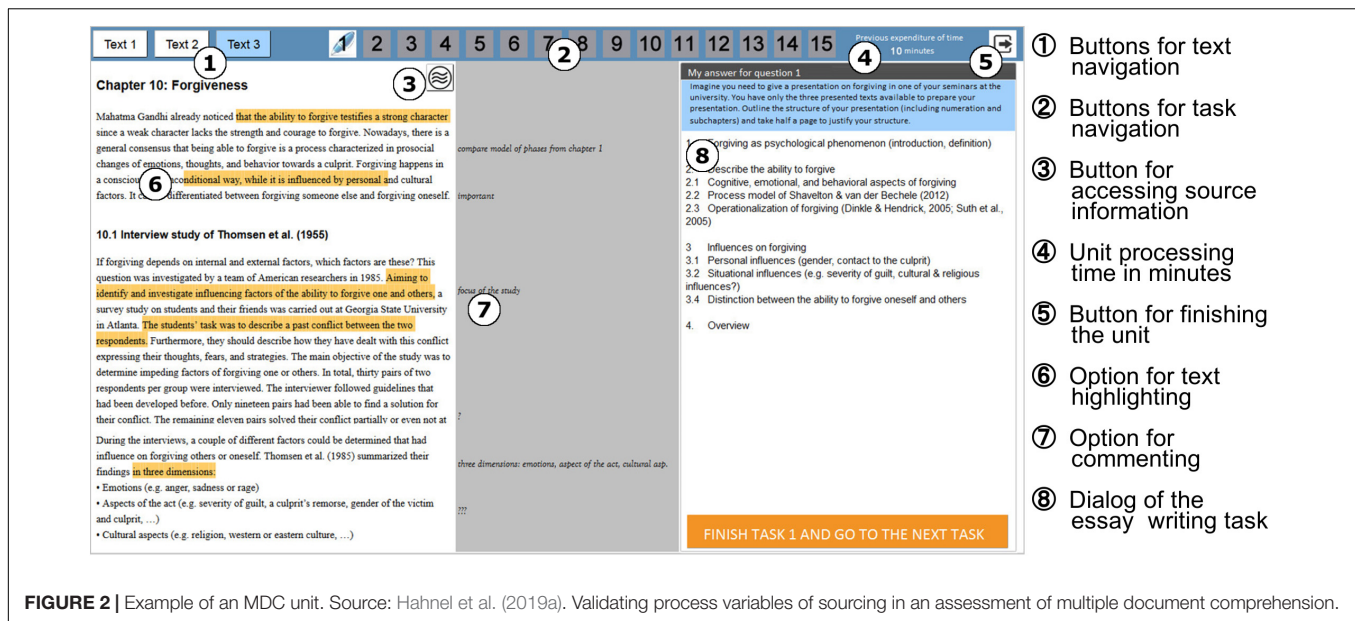
<sup>1</sup>Number of words per text. <sup>2</sup>Readability index (LIX) calculated by psychometrica.de (Lenhard and Lenhard, 2014). <sup>3</sup>Readability index (FRE) calculated by fleschindex.de. The readability indices are measures of text difficulty.

## Multiple Document Comprehension Measure

The MDC test began with a tutorial that explained the basic functions of the test, such as navigation, how to access source information, note-taking and highlighting. Actually, 74% of the sample paid attention to at least one source. Each MDC unit included two or three texts on the same issue with 11 to 17 items each. The five MDC units consisted of texts from four different domains (2× history, physics, literature, and literature studies) in order to assess MDC as a generic cross-disciplinary competence (for details see Schoor et al., 2020a,b). The texts and items were in German and developed by Schoor et al. (2020b). Most of the texts within a unit were redundant or complementary, only a few contained conflicting information. Conflicts were not fundamental and rather on a detail level, such as information on the age of a protagonist. In order to avoid prior knowledge effects, the text contents were fictitious, except for those in the physics domain. However, since this unit contained texts of a very specific nature, students were not expected to have much prior knowledge about the topic. Two units included an essay task, which had to be completed before the other items could be accessed. However, the essays were not included in the MDC test score, but instead used as a validation criterion. An overview of the text characteristics of the MDC test can be found in **Table 1**. We report two readability indices. The LIX (“Lesbarkeitsindex”) aims to determine the difficulty of a text by using a formula proposed by Björnsson (1968) and adapted to the German language by Bamberger and Vanecek (1984). It considers the average sentence length of a text and the percentage of words with more than six letters. These are calculated to a total value and compared with experience values of different text genres (high values indicate a more complex test). The Flesch-Reading-Ease (FRE) was originally established by Flesch (1948) and adapted to the German language by Amstad (1978). It results from the average number of syllables per word and the average sentence length. The FRE ranges between 0 and 100 where higher values indicate better comprehensibility (easiness).

Each unit started with an introductory page informing students about the number of texts and items, the time limit and setting a reading goal (e.g., “Please read the texts as if afterward you would have to describe how animals in novels can be interpreted”). The next page displayed the first text of the corresponding unit (for a screenshot of the text see **Figure 2**). The screen had a top bar with buttons for navigating between texts and items as well as information about the elapsed time and an exit button. During the test session, participants could navigate freely between the texts and items. Each text page included a button that produced a popup dialogue presenting source information about the text. Students could also highlight text passages, write comments in the margins, and receive feedback on their unit processing time and task progress (symbolized by green ticks on the item number buttons). The time restrictions were unit-specific and varied between 27 and 38 min. The units could be exited at any time before the time limit expired by clicking on the exit button. This evoked a popup window with a reminder of unsolved tasks (if any), asking students whether they wanted to exit the unit or return to the tasks. Ten minutes before time ran out, a popup window reminded participants of the already elapsed time and the unit-specific time limit. When the time limit expired, another popup window informed the students that time had run out. They then had to click a button in order to continue with the next task.

The 67 MDC items each measured one out of four cognitive requirements. Items requiring the corroboration of information (19 items) required participants to locate and compare information across different texts and were inspired by Wineburg (1991). The other three item types were constructed with reference to the documents model framework by Britt and Rouet (2012). In order to solve items requiring the integration of information, relevant information had to be identified from the texts and integrated with one another (integrated situation model, 18 items). For items requiring the comparison of sources, text characteristics and source information had to be assessed and compared (intertext model, 16 items). The most complex items



were those requiring the comparison of source-content links, since they combine the integrating information and comparing sources requirements (documents model, 14 items). In order to solve items of this type, readers had to build a mental model combining content and source information (who stated what). The MDC items required the consideration of at least two texts in order to identify the correct answer. The items were administered in a single-choice format (1 out of 4) or a verification format (yes/no or true/false). Example items of an excluded unit for each of the cognitive requirements can be found in the **Appendix**. Schoor et al. (2020a) showed that this MDC test is objective, reliable and valid, and represents a unidimensional construct (rather than a four-dimensional model representing the cognitive requirements or a five-dimensional model representing the unit structure).

Due to technical problems, the MDC data from 8 participants could not be used. For the remaining 493 participants, there were only a small number of missing values due to omitted or not reached items (0.57%). Because of this small amount, missing values were treated as if the respective item had not been administered (Pohl et al., 2014).

### Single Text Comprehension Measure

The administered STC measure is a computer-based reading comprehension test for university students based on the literacy concept from the National Educational Panel Study (NEPS; Gehrer et al., 2013). The same test was used as an online test for the university student cohort of the NEPS (Rohm et al., 2019). It consists of 21 items spread over five single texts on different topics with three to six items each. The texts represented a range of different text types (i.e., an information text, a commentary text, a literary text, an instructional text and an advertising text). No source information was presented for the texts since some items asked in particular for the source of the text. An exception was the short story unit, in which source information were presented

before the text content in order to contextualize the text. See **Table 2** for more information on the STC test.

The STC test started with a tutorial explaining to students the structure of the test (5 units with one text plus several items each), the total time limit of 28 min for the whole test, and the navigation and item response formats. There was no reading goal presented in the STC test. During the test, students could navigate freely between the text and items within a unit. Highlighting and commenting on text passages was not possible in the STC test. To exit the unit, an arrow button could be clicked any time. Clicking on the arrow button produced a popup window asking students whether they wanted to exit the unit or return to the tasks. When the time limit expired, a popup window informed the students that the time had run out. They had to click a button in order to continue with the next task.

The STC test consisted of items with different cognitive requirements, which had to be answered for each text. Items of Type 1 required students to find detailed information in the text (e.g., "What is xy?"; 2 items). In Type 2 items, text-related conclusions had to be drawn (e.g., "Which assumption about xy can be derived from the text?"; 9 items). The third item type required students to reflect on and assess statements made in the text (10 items). This included the ability to either comprehend the central message of the text, recognize its intention and judge its trustworthiness, or integrate prior knowledge in order to answer the items correctly. A situation model or mental model of the text was required in order to correctly answer Type 3 items, which were mostly items where headings had to be matched to certain paragraphs (matching items) or a new sentence had to be integrated into the text (text enrichment items). These formats are described in the following paragraph. **Table 3** shows how the requirements of the STC test correspond to the requirements of the MDC test.

The items were presented in one of the following four item formats. Most of the items were administered in a single-choice



**TABLE 2 |** Text characteristics of the STC test.

Unit name	Text characteristics	Number of items	Number of words	Readability (LIX) <sup>1</sup>	Readability (FRE) <sup>2</sup>	Text type
Handicraft	Text conveyed user guidance through work instructions; it is action-oriented and explains an activity step by step	4	238	45.4	51	Instruction text
Journalism	Text takes a particular stance; characterized by an argumentative text structure which is rather complex	5	258	51.2	51	Commenting text
False color photography	Sophisticated text for learning, advanced acquisition of knowledge, and finding detailed information	6	305	57	36	Information text
Law changes	Sophisticated call/claim with a persuasive function; the text language is purpose-oriented	3	250	64.3	22	Advertising text
Short story	Short story with many linguistic means; text with demanding interpretation because of its ambiguity, complexity, compression and openness	3	395	30.3	72	Literary text

<sup>1</sup>Readability index (LIX) calculated by *psychometrica.de* (Lenhard and Lenhard, 2014). <sup>2</sup>Readability index (FRE) calculated by *fleschindex.de*. The readability indices are measures of text difficulty.

**TABLE 3 |** Common requirements of the MDC and STC test.

Requirements MDC test	Corresponding requirement STC test	Common requirement
(1) Corroboration of information across texts: find information in text and compare it across texts.	(1) Finding information in text: find detailed information on sentence level.	Find information.
(2) Integration of information across texts: information has to be combined additively or by means of an inference.	(2) Drawing text-related conclusions: construct local or global coherence. (3a) Reflecting and assessing: comprehend the central idea, integration of background and world knowledge.	Integrate information.
(3) Comparison of sources and source evaluations across texts: judge each single source and compare.	(3b) Reflecting and assessing: recognize purpose and intention of a text, judge credibility.	Judge information with regard to source features.
(4) Comparison of source-content links across texts.	–	–

format, where one answer out of four is correct. Another item format comprised decision-making items where statements about the content of the text had to be judged as correct or incorrect (similar to the verification format in the MDC test). In the matching item format, headings had to be selected and assigned to a text section via drag and drop. Examples for these item formats can be found in Gehrer et al. (2012). The fourth item format comprised text enrichment task items. In these items, participants were asked to enrich a text meaningfully with three to four additional sentences. In order to do so, they had to drag a symbol marking a specific sentence to the correct gap within the text (the sentence could be dragged to any gap between two sentences). More information on the last item format can be found in Rohm et al. (2019). All item formats except for the single-choice items consisted of several subitems, which were summarized during data analysis in order to produce partial credit item solutions. Rohm et al. (2019) showed that the STC test represents a unidimensional construct (rather than a three-dimensional model representing the cognitive requirements or a five-dimensional model representing the unit structure).

Due to technical problems, we had to exclude the STC data from 2 participants. The remaining 499 participants omitted

0.36% of the items and did not reach 3.86% of the items. The missing responses were ignored and thus treated if not administered (same approach as for the MDC test). This is the approach used in the NEPS for scaling data from competence tests (Pohl and Carstensen, 2012). The MDC and STC tests were implemented using the CBA ItemBuilder (Roelke, 2012).

## Background Variables

In addition to the tests we asked the participants about their final school exam grades and their level of university studies (bachelor, master, and others). Master students were also asked about their bachelor's degree grade point average and their anticipated master's degree grade point average (assessed with the following question: "With what grade do you expect to complete your master's degree?"). All of these background variables were self-reports.

## Data Analysis

In order to investigate the three H1 alternatives, confirmatory (multi-dimensional) item response models specifying the dimensionality of MDC and STC were estimated and compared, using the software R version 3.6.0 (R Core Team, 2019) and the R package *TAM* (Robitzsch et al., 2019). MDC and STC were

modeled as latent variables, assuming a Rasch model for MDC responses and a partial credit model (Masters, 1982) for STC responses. This resulted in fixing the discriminations across all items in each model to one (see Schoor et al., 2020a,b).

In order to evaluate the hypothesis that MDC is an additional requirement to STC (H1a), we first specified a two-dimensional model with crossloadings and fixed covariance. That is, both STC and MDC items loaded onto one factor called “STC,” since we assumed that STC is required in order to solve MDC test items correctly. The MDC items additionally loaded onto a second factor which accounted for the “additional requirements” needed to solve the MDC items (add-on model, see **Figure 1A**). The covariance between the STC and MDC factors was fixed to be zero. In order to evaluate the hypothesis that MDC and STC reflect the same construct (H1b), we specified a unidimensional model where both STC items and MDC items loaded onto a joint factor, which we call “general reading comprehension factor” (unidimensional model, see **Figure 1B**). In order to examine the hypothesis that MDC and STC are two separable constructs (H1c), we specified a two-dimensional model where STC items loaded onto one factor and MDC items loaded onto another factor (two-factor model, see **Figure 1C**). The covariance between the two factors was freely estimated. Models were compared by using the Akaike information criterion (AIC), the Bayesian Information Criterion (BIC) and the  $\chi^2$ -difference test. We tested the  $\chi^2$ -difference for nested models, specifically for the unidimensional and the two-dimensional model as well as between the unidimensional and the add-on model.

Latent regression analyses for H2-H6 were conducted in Mplus 8.1 (Muthén and Muthén et al., 1998-2017). In order to account for the missing data structure of the MDC variables (missing by design), the MLR estimator was used, which allows maximum likelihood estimation with robust standard errors. MDC and STC were modeled as latent variables in a format that depended on the results of H1. Relevant predictors were final school exam grades, level of university studies (0 = bachelor, 1 = experienced students, i.e., master's or diploma program), bachelor's degree grade point average (only for master's students) and anticipated master's degree grade point average (only for master's students). Two regression models were tested for each predictor: the first estimated the effect of the predictor (e.g., final school exam grades) on MDC and the second added STC as a further predictor. In order to test whether bachelor's students differed from more experienced students in the relation of their final school exam grades to MDC (H3), a Wald test was performed.

## RESULTS

### Dimensionality of Multiple Document Comprehension and Single Text Comprehension (H1)

**Table 4** shows the results of the structural analyses of the STC and MDC test items. Both the AIC and the BIC showed higher values for the add-on model and the unidimensional model than for

the two-dimensional model, indicating that the two-dimensional model better fit with the data. Furthermore, the  $\chi^2$ -difference test showed that the two-dimensional model was statistically significantly different from the unidimensional model. Therefore, the results support H1c. The latent correlation between the latent factors MDC and STC was  $r = 0.84$ .

### Relations to Final School Exam Grades, Level of University Studies and University Performance (H2–H6)

Below, results are reported regarding the relations between the MDC test and final school exam grades, level of university studies and university performance. For comparative purposes, we also report the impact of the aforementioned variables on the STC test. MDC and STC were modeled as latent variables following the two-dimensional model.

#### Final School Exam Grades (H2)

Final school exam grades statistically significantly predicted STC ( $\beta = -0.39, p < 0.001$ ) as well as MDC ( $\beta = -0.43, p < 0.001$ ). This means that a better (lower) final school exam grade was associated with a better MDC test score, supporting H2. When STC was included as predictor in the regression model of MDC on final school exam grades, the impact of final school exam grades on MDC became smaller, but was still statistically significant ( $\beta = -0.24, p < 0.001$ ). As to be expected, STC also statistically significantly impacted MDC test scores (see **Table 5**).

#### Level of University Studies and Final School Exam Grades (H3)

The (negative) relation between the final school exam grades and MDC was higher for bachelor's than for experienced students ( $\beta_{BA} = -0.49, p_{BA} < 0.001$ ;  $\beta_{MA} = -0.40, p_{MA} < 0.001$ ), but the difference was not statistically significant [ $\chi^2(1) = 0.39, p = 0.534$ ]. The opposite was found for STC ( $\beta_{BA} = -0.38, p_{BA} < 0.001$ ;  $\beta_{MA} = -0.44, p_{MA} < 0.001$ ), but again the difference was not statistically significant [ $\chi^2(1) = 1.97, p = 0.161$ ]. When STC was included as predictor in the regression model, the impact of final school exam grades on MDC became smaller and was only still statistically significant for bachelor students ( $\beta_{BA} = -0.33, p_{BA} < 0.001$ ;  $\beta_{MA} = -0.13, p_{MA} = 0.106$ ). However, the difference was not statistically significant [ $\chi^2(1) = 2.28, p = 0.131$ ]. STC also statistically significantly impacted MDC test scores in both groups (see **Table 5**).

#### Level of University Studies (H4)

Level of university studies statistically significantly predicted MDC test scores ( $\beta = 0.24, p < 0.001$ ). The same was true for STC test scores ( $\beta = 0.23, p < 0.001$ ). The positive  $\beta$  coefficient indicates that more experienced students, such as master's students, performed better on the MDC test than bachelor's students. When STC was included as predictor, the impact of the level of university studies diminished, but was still statistically significant on the 5% level ( $\beta = 0.11, p = 0.030$ ). STC had a statistically significant impact on MDC (see **Table 5**).

**TABLE 4 |** Structural analysis of STC and MDC test items and model comparison with the unidimensional model.

Model	AIC	BIC	$n_{Par}$	$\Delta \chi^2$	$\Delta df$	$p$
Add-on model	30615.70	31050.01	103	4.90	1	0.03
Unidimensional model	30618.59	31048.69	102			
Two-factor model	<b>30574.04</b>	<b>31012.57</b>	104	48.55	2	0.00

$n_{Par}$ , number of parameters. Lowest values of the AIC and BIC indicating best fit are in bold.

**TABLE 5 |** Standardized effect sizes of the impact of the level of university studies, final school exam grades and prior university performance and STC on MDC per hypothesis.

		H2: Final school exam grades	H3: Level of university studies and final school exam grades		H4: Level of university studies	H5: Prior university performance
			Bachelor's students	Master's students		
Impact on STC	$\beta_{Predictor}^1 (SE)$	−0.39*** (0.05)	−0.38*** (0.07)	−0.44*** (0.07)	0.23*** (0.05)	−0.28*** (0.08)
Impact on MDC	$\beta_{Predictor}^1 (SE)$	−0.43*** (0.05)	−0.49*** (0.06)	−0.40*** (0.08)	0.24*** (0.06)	−0.31*** (0.07)
Impact on MDC when including STC as predictor	$\beta_{Predictor}^1 (SE)$	−0.24*** (0.05)	−0.33*** (0.07)	−0.13 (0.08)	0.11* (0.05)	−0.17* (0.08)
	$\beta_{STC} (SE)$	0.78*** (0.05)	0.73*** (0.06)	0.81*** (0.05)	0.82*** (0.04)	0.75*** (0.06)

Prior university performance (bachelor's degree grade point average, H5) could only be examined for the subsample of master's students. <sup>1</sup>Predictor varies depending on the hypothesis (H2 and H3: final school exam grades, H4: level of university studies, H5: bachelor's degree grade point average). \*\*\* $p < 0.001$ , \* $p < 0.05$ .

**TABLE 6 |** Standardized effect sizes of the impact of STC and MDC on expected future university performance.

	H6: Expected future university performance	
Impact of STC on anticipated master's degree grade point average	$\beta_{STC} (SE)$	−0.15 (0.08)
Impact of MDC on anticipated master's degree grade point average	$\beta_{MDC} (SE)$	−0.32*** (0.08)
Impact of MDC on anticipated master's degree grade point average when including STC as predictor	$\beta_{MDC} (SE)$	−0.49** (0.18)
	$\beta_{STC} (SE)$	0.23 (0.18)

Expected future university performance could only be examined for the subsample of master's students. \*\*\* $p < 0.001$ , \*\* $p < 0.01$ .

### Prior University Performance (H5)

University performance could only be examined for the subsample of master's students. The bachelor's degree grade point average of master's students statistically significantly predicted their MDC test scores ( $\beta = -0.31$ ,  $p < 0.001$ ) and STC test scores ( $\beta = -0.28$ ,  $p < 0.001$ ). When STC was included as predictor, the impact of bachelor's degree grade point average on MDC diminished, but was still statistically significant on the 5% level ( $\beta = -0.17$ ,  $p = 0.025$ ). STC had a statistically significant impact on MDC (see Table 5).

### Expected Future University Performance (H6)

Multiple document comprehension statistically significantly predicted the students' anticipated master's degree grade point average ( $\beta = -0.32$ ,  $p < 0.001$ , see Table 6). This was not true for STC ( $\beta = -0.15$ ,  $p = 0.075$ ). The impact of MDC on anticipated master's degree grade point average was even higher when STC was included as predictor as well ( $\beta = -0.49$ ,  $p < 0.001$ ). The shared variance of STC and MDC could not be responsible for this effect, since the effect of MDC on anticipated master's degree grade point average already existed before including STC. Therefore, the statistically significant relationship is not

traceable to STC, but only to MDC, indicating that MDC explains anticipated master's degree grade point average.

## DISCUSSION

The main purpose of this study was to examine the relation between single text and multiple document comprehension, since theoretical assumptions, such as the documents model framework (Britt and Rouet, 2012), suggest that the comprehension of multiple documents demands more from readers than the comprehension of single texts. We further investigated the relation between final school exam grades, the level of university studies and university performance with test scores for each form of text comprehension and explored these variables' relations to MDC while including STC as predictor in order to shed light on the relationship of MDC and STC.

### Discussion of the Results

With regard to the relation of STC and MDC, confirmatory dimensionality analyses revealed that a model with two separable but correlated factors (i.e., STC and MDC, latent correlation: 0.84) had a better fit compared to an add-on model and a

unidimensional model. Accordingly, there is evidence in favor of two highly correlated, but separable constructs (H1c) rather than MDC representing an add-on in terms of cognitive demands (H1a) or a single reading construct (H1b). The latent correlation found between STC and MDC resembles, for instance, the latent correlation between science literacy and reading literacy ( $r = 0.87$ ) or science literacy and mathematics ( $r = 0.89$ ) in PISA 2009 (OECD, 2012), indicating that STC and MDC are indeed separable constructs. This implies that single and multiple text reading situations are different in terms of the cognitive requirements they place on readers. There are students who perform better in single text reading situations than in multiple text reading situations and vice versa. We assume that the high correlation between these factors can be traced back to common underlying abilities, such as the decoding of words and sentences or intelligence. Thus, we cannot tell if MDC items are more difficult than STC items, but they require different – although related – abilities. This supports the view that additional (different) cognitive requirements are needed in order to represent multiple documents compared to single texts (Britt and Rouet, 2012) and is in line with Rouet (2006), who postulated that the demands of single text and multiple document comprehension differ. The results also suggest that the MDC test was successful in focusing on the nature of multiple document comprehension.

Furthermore, we could replicate the relations between the level of university studies (bachelor or master studies), final school exam grades and MDC (H2-H4) found by Schoor et al. (2020b), who used the same MDC test on a different sample. Our results show that final school exam grades statistically significantly predicted MDC (supporting H2), that MDC is not predicted more strongly by final school exam grades among bachelor's students than among master's students (rejecting H3), and that MDC is higher for master's students than for bachelor's students (supporting H4). Additionally, we added new findings to the existing literature since we found that prior university performance positively predicted MDC test scores (supporting H5) and that expected future university performance in terms of anticipated master's degree grade point average was predicted by MDC test scores (supporting H6). The same relations were found with STC. The only exception was that STC did not statistically significantly predict anticipated master's degree grade point average, while MDC did. This indicates that students' estimation of their expected master's degree grade point average to some point relied on their MDC and not on their STC, which is reasonable since MDC is particularly important during master's degree programs. However, a reciprocal relationship of grade point averages and both MDC and STC would also be conceivable.

In addition, we found smaller but still statistically significant relationships between MDC and the analyzed variables when STC was included in the models except for the impact of final school exam grades on MDC for master's students. This delivers additional evidence that MDC represents a construct that differs from STC, since it provides relevant additional information about readers.

The finding that STC and MDC are highly correlated, yet separable constructs is interesting. It suggests that theoretical

models explicitly addressing MDC – like those proposed by many researchers since the late 1990s – are reasonable and necessary. However, it has to be considered that we did not assess situation models for each single text in the MDC test. Although this is true for an explicit assessment, individual situation models were assessed implicitly since at least two texts had to be read and understood in order to answer the MDC items correctly. Our results suggest that MDC and STC are highly related (and not independent) constructs; therefore they support the assumption of Britt et al. (1999) and Perfetti et al. (1999) that in most circumstances situation models are not built for each text, but that the initial situation model is updated during the course of reading. Separate situation models are only created in special circumstances, such as when sources are distinct and elaborated (separate representation model) or when encountering conflicting information which necessitates the creation of an intertext model [tagging of information and corresponding sources, see Britt et al. (1999)]. This view is also consistent with Kintsch who postulates that a network is iteratively created, modified, and updated during the course of comprehension (Kintsch, 1998). Since the documents used in the MDC test are mostly redundant or complementary, it may not have been necessary to build separate situation models.

The strength and novelty of the present study lies in its operationalization of the MDC construct, since the employed test measures the concepts in its pure form. Measuring MDC with a test that covers all components of the documents model framework rather than only the integrated situation model or intertext component is quite novel. However, this implies that the constructs used in the present study differ from recent assessments of reading (literacy) implemented in studies like PISA or the Programme for the International Assessment of Adult Competencies (PIAAC). These large-scale assessments focus on the demands of authentic reading situations, which are conceptualized as a mixture of single text and multiple document comprehension. In the present study, we defined a multiple document task as a task that is not solvable with only one text. This is not necessarily the case in PISA 2018 (OECD, 2019), where multiple documents are viewed as a text characteristic and formats like online forums are also considered to be multiple documents. The approach taken here is thus different from the one taken by the PISA or PIAAC reading assessments.

However, measuring reading comprehension is actually a challenging task. There are different perspectives on reading which will influence how this competence is measured. In the present study we focused on an individual difference perspective by understanding reading as a product. However, there are also other perspectives, such as the cognitive psychological perspective which focusses on the process of reading such as decoding (word level and sentence level) or the educational-psychological perspective which focusses more on fostering reading comprehension. Even large-scale studies measure reading comprehension in different ways. Therefore, the results cannot be generalized to other STC or MDC tests than the ones used in the present study. However, it would be interesting to examine the relations between the MDC test



and other STC tests in order to see if the results can be replicated (especially with regard to the challenges associated with measuring reading competence).

Beyond the perspective of reading as a product, computer-based reading assessments can shed light on the behavioral process of reading by means of process (log file) data, that is, how readers proceed when reading single or multiple texts. It would be possible to compare successful with unsuccessful readers and if they differ in the strategies they used or in time they spent on the texts. For example, Heyne et al. (2020) found that instructed highlighting correlated significantly with reading competence. Therefore it might be one of the strategies used by successful readers. Another factor that plays a role in comprehending documents is the readers' working memory. Research relating working memory and MDC is still in the early stages (e.g., Hahnel et al., 2019b), but shows that MDC is cognitive demanding for university students. Future research should further investigate the revealed commonality between STC and MDC by identifying the common source of variance.

## Limitations

As a first limiting factor, it should be noted that the results of the present study are based on an *ad hoc* sample. Accordingly, the participating university students were not representative for the respective overall student population in the social sciences or humanities, but rather drawn from an easily accessible part of the population. Therefore, the results of the study are not generalizable to other student populations. In this regard, attention should also be paid to the anticipated master's degree grade point average. Students might tend to overrate future test results since the expected mean grade point average of the master's degree is descriptively slightly better than the mean of the bachelor's degree ( $M_{\text{Master}} = 1.70$ ;  $M_{\text{Bachelor}} = 1.83$ ). However, this could also indicate a potential selection bias due to self-selection (e.g., regarding the decision to continue studying) or external selection (e.g., numerus clausus or entrance tests). This would mean that especially students with better (lower) bachelor's degree point averages decide to study in a masters' program and that they are more likely to be admitted to these programs.

Secondly, it should be noted that students were provided with fictitious information in the MDC test (except for the unit "Universe"). This was done in order to minimize the impact of prior knowledge and prior beliefs, but goes along with the restriction that the MDC test can only rudimentarily capture what is often needed when dealing with multiple documents in everyday life: setting one's own preferences aside and processing information that is inconsistent with one's convictions as well as assessing the credibility of sources. Since the documents of the MDC test do not address critical or ambiguous topics, students were not explicitly informed about the fact of reading fictitious information. Nevertheless, debriefing should be considered in future studies.

A third limiting factor concerns the comparability of the MDC test and the STC test. First of all, an explicit reading

goal was only provided in the MDC test, but not in the STC test. Research has shown that reading goals can affect reading decisions, reading processes, and reading outcomes (e.g., Rouet, 2006; McCrudden and Schraw, 2007). Basically, it is difficult to attribute the differences between the tests to the presence of a reading goal since they also differ in other respects, such as the text length, readability, functionalities (notetaking and highlighting) and time feedback. For example, while the MDC test provided per-minute time feedback on the top of the screen and a time limit reminder 10 min before the time expired, the STC test did not provide such feedback. However, the percentage of not reached items in the STC test was rather low, indicating that most participants were able to complete the entire test in time. As a general limitation, there are several confounding variables related to the test specifications (e.g., reading goal, special functionalities, and time feedback) that may have had an effect on the results. Since we cannot exclude these effects, it would be highly desirable to replicate the present results with other tests.

## Conclusion

The present study closes a research gap by analyzing the dimensionality of STC and MDC assessed using tests which are structurally comparable and capture the measured concepts in their pure form. We found first evidence that STC and MDC are separable constructs, indicating that single and multiple text reading situations differ from each other in terms of the requirements they place on readers. However, the high correlation between the constructs indicates that fundamental abilities, such as decoding abilities or reasoning, are needed in both situations. This finding is not only important for the context of university studies, but for reading internet texts in general, where texts from multiple sources are prominent. Therefore, reading online can be seen as a special situation of reading multiple documents, since the use of search engines (Google, Bing, etc.) usually leads to information on a topic from different sources with different perspectives. Our work shows that a lack of the ability to understand and integrate information of such multiple texts cannot be compensated by reading skills (even if they are central), but that skills are necessary which are part of critical online reasoning. The present study also contributes to research on the assessment of MDC, since we could replicate the findings of Schoor et al. (2020a) regarding the relations between MDC and the level of university studies and final school exam grades. Furthermore, we could add results on the relation between university performance and MDC.

In summary, the present study enhances our understanding of the MDC construct and its relation to STC as well as to students' level of university studies, final school exam grades and university performance. We thereby add empirical evidence to the existing research regarding commonalities and differences between MDC and STC, which is currently mostly of a theoretical nature. The present study also shows that the MDC test developed by Schoor et al. (2020a) is an instrument that validly distinguishes MDC from STC and can therefore serve as a diagnostic instrument for university students.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher. The MDC test is currently available on request. It is planned to make the test and the data of the study available to researchers. The STC test is part of a long-term large-scale assessment in Germany (National Educational Panel Study) and is therefore subject to special protective regulations. Access to the test is possible by submitting an application if there is a well-founded interest in research. The use of the test is possible with a cooperation agreement.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

## REFERENCES

- Afflerbach, P., and Cho, B.-Y. (2009). "Identifying and describing constructively responsive comprehension strategies in new and traditional forms of reading," in *Handbook of Research on Reading Comprehension*, eds S. Israel and G. Duffy (Abingdon: Routledge), 69–90.
- Amstad, T. (1978). *Wie verständlich sind unsere Zeitungen? [How Understandable are Our Newspapers?]*. Ph.D. Dissertation, Universität Zürich, Zurich.
- Anmarkrud, Ø., Bråten, I., and Strømso, H. I. (2014). Multiple-documents literacy: strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learn. Individ. Dif.* 30, 64–76. doi: 10.1016/j.lindif.2013.01.007
- Bamberger, R., and Vanecek, E. (1984). *Lesen – Verstehen – Lernen – Schreiben [Reading – Understanding – Learning – Writing]*. Los Angeles, CA: Jugend und Volk.
- Barzilai, S., Tzadok, E., and Eshet-Alkalai, Y. (2015). Sourcing while reading divergent expert accounts: pathways from views of knowing to written argumentation. *Instruct. Sci.* 43, 737–766. doi: 10.1007/s11251-015-9359-4
- Baumert, J., Lüdtke, O., Trautwein, U., and Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: evidence in support of the distinction between intelligence and student achievement. *Educ. Res. Rev.* 4, 165–176. doi: 10.1016/j.edurev.2009.04.002
- Beker, K., van den Broek, P., and Jolles, D. (2019). Children's integration of information across texts: reading processes and knowledge representations. *Read. Writ.* 32, 663–687. doi: 10.1007/s11145-018-9879-9
- Björnsson, C. H. (1968). *Läsbarhet [Readability]*. The Hague: Liber.
- Blossfeld, H.-P., and Roßbach, H.-G. (eds.) (2019). *Education as a Lifelong Process*, Edition ZfE Edn, Vol. 3. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-23162-0
- Braasch, J. L. G., and Bråten, I. (2017). The discrepancy-induced source comprehension (D-ISC) model: basic assumptions and preliminary evidence. *Educ. Psychol.* 52, 167–181. doi: 10.1080/00461520.2017.1323219
- Braasch, J. L. G., McCabe, R. M., and Daniel, F. (2016). Content integration across multiple documents reduces memory for sources. *Read. Writ.* 29, 1571–1598. doi: 10.1007/s11145-015-9609-5
- Bråten, I., Anmarkrud, Ø., Brandmo, C., and Strømso, H. I. (2014). Developing and testing a model of direct and indirect relationships between individual differences, processing, and multiple-text comprehension. *Learn. Instruct.* 30, 9–24. doi: 10.1016/j.learninstruc.2013.11.002
- Bråten, I., Britt, A. M., Strømso, H. I., and Rouet, J.-F. (2011). The role of epistemic beliefs in the comprehension of multiple expository texts: toward an integrated model. *Educ. Psychol.* 46, 48–70. doi: 10.1080/00461520.2011.538647

## AUTHOR CONTRIBUTIONS

NM: data preparation and analysis, manuscript preparation, and final writing. CS and CH: study design, material development, data collection, and final writing. CA: supervision, review, and final approval. FG: review of the manuscript. UK: technical implementation of the study and review of the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This research was funded by the German Federal Ministry of Education and Research (BMBF; Funding Number: 01PK15008), for the collaborative research project "MultiTex – Process-based assessment of multiple documents comprehension" within the research program KoKoHs ("Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor – Validierungen und methodische Innovationen").

- Bråten, I., Salmerón, L., and Strømso, H. I. (2016). Who said that? Investigating the plausibility-induced source focusing assumption with norwegian undergraduates. *Contemp. Educ. Psychol.* 46, 253–262. doi: 10.1016/j.cedpsych.2016.07.004
- Bråten, I., and Strømso, H. I. (2006). Effects of personal epistemology on the understanding of multiple texts. *Read. Psychol.* 27, 457–484. doi: 10.1080/0270210600848031
- Bråten, I., and Strømso, H. I. (2010a). Effects of task instruction and personal epistemology on the understanding of multiple texts about climate change. *Discourse Process.* 47, 1–31. doi: 10.1080/01638530902959646
- Bråten, I., and Strømso, H. I. (2010b). When law students read multiple documents about global warming: examining the role of topic-specific beliefs about the nature of knowledge and knowing. *Instruct. Sci.* 38, 635–657. doi: 10.1007/s11251-008-9091-4
- Bråten, I., and Strømso, H. I. (2011). Measuring strategic processing when students read multiple texts. *Metacogn. Learn.* 6, 111–130. doi: 10.1007/s11409-011-9075-7
- Bråten, I., Strømso, H. I., and Britt, M. A. (2009). Trust matters: examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Read. Res. Q.* 44, 6–28. doi: 10.1598/rrq.44.1.1
- Britt, M. A., and Aglinskias, C. (2002). Improving students' ability to identify and use source information. *Cogn. Instruct.* 20, 485–522. doi: 10.1207/S1532690XC12004\_2
- Britt, M. A., Perfetti, C. A., Sandak, R., and Rouet, J.-F. (1999). "Content integration and source separation in learning from multiple texts," in *Narrative Comprehension, Causality, and Coherence. Essays in Honor of Tom Trabasso*, eds S. R. Goldman, A. C. Graesser, and P. van den Broek (Mahwah NJ: Lawrence Erlbaum Associates Publishers), 209–233.
- Britt, M. A., and Rouet, J.-F. (2012). "Learning with multiple documents: component skills and their acquisition," in *Enhancing the Quality of Learning: Dispositions, Instruction, and Learning Processes*, eds J. R. Kirby and M. J. Lawson (New York, NY: Cambridge University Press), 276–314.
- Britt, M. A., Rouet, J.-F., and Braasch, J. L. G. (2012). "Documents as entities: extending the situation model theory of comprehension," in *Reading – From Words to Multiple Texts*, eds M. A. Britt, S. R. Goldman, and J.-F. Rouet (New York, NY: Routledge), 160–179.
- Cain, K. (2009). Making sense of text: skills that support text comprehension and its development. *Perspect. Lang. Lit.* 35, 11–14.
- Common Core State Standards (2010). Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects. Available online at: [http://www.corestandards.org/assets/CCSSI\\_ELA%20Standards.pdf](http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf) (accessed September 28, 2020).

- Flesch, R. (1948). A new readability yardstick. *J. Appl. Psychol.* 32, 221–233. doi: 10.1037/h0057532
- Florit, E., Cain, K., and Mason, L. (2019). Going beyond children's single-text comprehension: the role of fundamental and higher-level skills in 4th graders' multiple-document comprehension. *Br. J. Educ. Psychol.* 90, 449–472. doi: 10.1111/bjep.12288
- Gehrer, K., Zimmermann, S., Artelt, C., and Weinert, S. (2012). *The Assessment of Reading Competence (Including Sample Items for Grade 5 and 9)*. Scientific Use File 2012, Version 1.0.0. Bamberg: University of Bamberg, National Educational Panel Study.
- Gehrer, K., Zimmermann, S., Artelt, C., and Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *J. Educ. Res. Online* 5, 50–79.
- Gernsbacher, M. A. (1991). "Cognitive processes and mechanisms in language comprehension: the structure building framework," in *The Psychology of Learning and Motivation*, ed. G. H. Bower (New York, NY: Academic Press), 217–263. doi: 10.1016/s0079-7421(08)60125-5
- Gil, L., Bråten, I., Vidal-Abarca, E., and Strømso, H. (2010). Understanding and integrating multiple science texts: summary tasks are sometimes better than argument tasks. *Read. Psychol.* 31, 30–68. doi: 10.1080/0270210902733600
- Goldman, S. R. (2004). "Cognitive aspects of constructing meaning through and across multiple texts," in *Uses of Intertextuality in Classroom and Educational Research*, eds N. Shuart-Faris and D. Bloome (Greenwich: Information Age), 313–347.
- Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., Greenleaf, C., et al. (2016). Disciplinary literacies and learning to read for understanding: a conceptual framework for disciplinary literacy. *Educ. Psychol.* 51, 219–246. doi: 10.1080/00461520.2016.1168741
- Graesser, A. C., Singer, M., and Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychol. Rev.* 101, 371–395. doi: 10.1037/0033-295x.101.3.371
- Griffin, T. D., Wiley, J., Britt, M. A., and Salas, C. R. (2012). The role of CLEAR thinking in learning science from multiple-document inquiry tasks. *Int. Electron. J. Element. Educ.* 5, 63–78. doi: 10.4324/9781315458618-5
- Hagen, Å.M., Braasch, J. L. G., and Bråten, I. (2014). Relationships between spontaneous note-taking, self-reported strategies and comprehension when reading multiple texts in different task conditions. *J. Res. Read.* 37, S141–S157. doi: 10.1111/j.1467-9817.2012.01536.x
- Hahnel, C., Kröhne, U., Goldhammer, F., Schoor, C., Mahlow, N., and Artelt, C. (2019a). Validating process variables of sourcing in an assessment of multiple document comprehension. *Br. J. Educ. Psychol.* 89, 524–537. doi: 10.1111/bjep.12278
- Hahnel, C., Schoor, C., Kröhne, U., Goldhammer, F., Mahlow, N., and Artelt, C. (2019b). The role of cognitive load for university students' comprehension of multiple documents. *Zeitschrift für Pädagogische Psychologie*. 33, 105–118. doi: 10.1024/1010-0652/a000238
- Heyne, N., Artelt, C., Gnams, T., Gehrer, K., and Schoor, C. (2020). Instructed highlighting of text passages – indicator of reading or strategic performance? *Lingua* 236:102803. doi: 10.1016/j.lingua.2020.102803
- Isberner, M.-B., and Richter, T. (2014). "Comprehension and validation: separable stages of information processing? A case for epistemic monitoring in language comprehension," in *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences*, eds D. N. Rapp and J. L. G. Braasch (Boston, MA: MIT Press), 245–276.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. Rev.* 95, 163–182. doi: 10.1037/0033-295x.95.2.163
- Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.
- Kroehne, U., Buerger, S., Hahnel, C., and Goldhammer, F. (2019). Construct equivalence of PISA reading comprehension measured with paper-based and computer-based assessments. *Educ. Meas. Issues Pract.* 38, 97–111. doi: 10.1111/emip.12280
- Kultusministerkonferenz (2012). *Bildungsstandards im Fach Deutsch für die Allgemeine Hochschulreife (Beschluss der Kultusministerkonferenz vom 18.10.2012) [Educational Standards in the Subject German for the Acquisition of the Higher Education Entrance Qualification (Resolution of the Conference of Ministers of Education of October 18, 2012)]*. Available online at: [www.kmk.org/fileadmin/veroeffentlichungen\\_beschluesse/2012/2012\\_10\\_18-Bildungsstandards-Deutsch-Abi.pdf](http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Deutsch-Abi.pdf) (accessed September 28, 2020).
- Lenhard, W., and Lenhard, A. (2014). *Berechnung des Lesbarkeitsindex LIX nach Björnson [Calculation of the Readability Index LIX by Björnson]*. Biberburg: Psychometrica.
- List, A., and Alexander, P. A. (2017a). Analyzing and integrating models of multiple text comprehension. *Educ. Psychol.* 52, 143–147. doi: 10.1080/00461520.2017.1328309
- List, A., and Alexander, P. A. (2017b). Cognitive affective engagement model of multiple source use. *Educ. Psychol.* 52, 182–199. doi: 10.1080/00461520.2017.1329014
- Maier, J., and Richter, T. (2013). Text belief consistency effects in the comprehension of multiple texts with conflicting information. *Cogn. Instruct.* 31, 151–175. doi: 10.1080/07370008.2013.769997
- Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika* 47, 149–174. doi: 10.1007/BF02296272
- McCrudden, M. T., and Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educ. Psychol. Rev.* 19, 113–139. doi: 10.1007/s10648-006-9010-7
- McNamara, D. S., and Magliano, J. P. (2009). Toward a comprehensive model of comprehension. *Psychol. Learn. Motiv.* 51, 297–384. doi: 10.1016/s0079-7421(09)51009-2
- Muthén, L. K., and Muthén, B. O. (1998–2017). *Mplus User's Guide*, 8th Edn. Los Angeles, CA: Muthén & Muthén. doi: 10.1016/s0079-7421(09)51009-2
- Oakhill, J., Cain, K., and Bryant, P. E. (2003). The dissociation of word reading and text comprehension: evidence from component skills. *Lang. Cogn. Process.* 18, 443–468.
- OECD (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Paris: OECD Publishing. doi: 10.1787/9789264173125-en
- OECD (2010). *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science*, PISA. Paris: OECD Publishing. doi: 10.1787/9789264062658-en
- OECD (2012). *PISA 2009 Technical Report*, PISA. Paris: OECD Publishing. doi: 10.1787/9789264167872-en
- OECD (2019). *PISA 2018 Assessment and Analytical Framework*, PISA. Paris: OECD Publishing. doi: 10.1787/b25efab8-en
- OECD, and Statistics Canada (1995). *Literacy, Economy and Society: Results of the 1st International Adult Literacy Survey*. Paris: OECD Publishing.
- Perfetti, C. A. (1985). *Reading Ability*. New York, NY: Oxford University Press.
- Perfetti, C. A., Landi, N., and Oakhill, J. (2005). "The acquisition of reading comprehension skill," in *The Science of Reading: A Handbook*, eds M. J. Snowling and C. Hulme (Oxford: Blackwell), 227–247. doi: 10.1002/9780470757642.ch13
- Perfetti, C. A., Rouet, J.-F., and Britt, M. A. (1999). "Toward a theory of documents representation," in *The Construction of Mental Representations During Reading*, eds H. van Oostendorp and S. R. Goldman (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 99–122.
- Pohl, S., and Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests* (NEPS Working Paper No. 14). Bamberg: National Educational Panel Study.
- Pohl, S., Gräfe, L., and Rose, N. (2014). Dealing with omitted and not-reached items in competence tests. *Educ. Psychol. Meas.* 74, 423–452. doi: 10.1177/0013164413504926
- Primor, L., and Katzir, T. (2018). Measuring multiple text integration: a review. *Front. Psychol.* 9:2294. doi: 10.3389/fpsyg.2018.02294
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Richter, T., and Maier, J. (2017). Comprehension of multiple documents with conflicting information: a two-step model of validation. *Educ. Psychol.* 52, 148–166. doi: 10.1080/00461520.2017.1322968
- Robitzsch, A., Kiefer, T., and Wu, M. (2019). *TAM: Test Analysis Modules*. R package version 3.3-10. Available online at: <https://CRAN.R-project.org/package=TAM> (accessed May 06, 2020).
- Roelke, H. (2012). "The ItemBuilder: a graphical authoring system for complex item development," in *Proceedings of E-Learn 2012–World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 1, Montréal, QC*, eds T. Bastiaens and G. Marks (Morgantown, WV: Association for the Advancement of Computing in Education (AACE)), 344–353.
- Rohm, T., Scharl, A., Ettner, J., and Gehrer, K. (2019). *NEPS Technical Report for Reading: Scaling Results of Starting Cohorts 4 (wave 10), 5 (wave 12), and 6*

- (wave 9) (NEPS Survey Paper No. 62). Bamberg: National Educational Panel Study.
- Rouet, J.-F. (2006). *The Skills of Document use: From Text Comprehension to Web-Based Learning*. Mahwah, NJ: Lawrence Erlbaum Associates, doi: 10.4324/9780203820094
- Rouet, J.-F., Britt, M. A., and Durik, A. M. (2017). RESOLV: readers' representation of reading contexts and tasks. *Educ. Psychol.* 52, 200–215. doi: 10.1080/00461520.2017.1329015
- Schoor, C., Hahnel, C., Artelt, C., Reimann, D., Kroehne, U., and Goldhammer, F. (2020a). Entwicklung und Skalierung eines Tests zur Erfassung des Verständnisses multipler Dokumente von Studierenden [Developing and scaling a test of multiple document comprehension in university students]. *Diagnostica* 66, 123–135. doi: 10.1026/0012-1924/a000231
- Schoor, C., Hahnel, C., Mahlow, N., Klagges, J., Kroehne, U., Goldhammer, F., et al. (2020b). "Multiple document comprehension of university students: test development and relations to person and process characteristics," in *Student Learning in German Higher Education – Innovative Measurement Approaches and Research Results*, eds O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, and C. Lautenbach (Wiesbaden: Springer VS), 221–240. doi: 10.1007/978-3-658-27886-1\_11
- Singer, L. M., and Alexander, P. A. (2017). Reading on paper and digitally: what the past decades of empirical research reveal. *Rev. Educ. Res.* 87, 1007–1041. doi: 10.3102/0034654317722961
- Stadtler, M. (2017). The art of reading in a knowledge society: commentary on the special issue on models of multiple text comprehension. *Educ. Psychol.* 52, 225–231. doi: 10.1080/00461520.2017.1322969
- Stadtler, M., and Bromme, R. (2014). "The content–source integration model: a taxonomic description of how readers comprehend conflicting scientific information," in *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences*, eds D. N. Rapp and J. Braasch (Cambridge, MA: MIT Press), 379–402.
- Stadtler, M., Scharrer, L., Brummernhenrich, B., and Bromme, R. (2013). Dealing with uncertainty: readers' memory for and use of conflicting information from science texts as function of presentation format and source expertise. *Cogn. Instruct.* 31, 130–150. doi: 10.1080/07370008.2013.769996
- Strømso, H. I. (2017). Multiple models of multiple-text comprehension: a commentary. *Educ. Psychol.* 52, 216–224. doi: 10.1080/00461520.2017.1320557
- Strømso, H. I., and Bråten, I. (2009). Beliefs about knowledge and knowing and multiple-text comprehension among upper secondary students. *Educ. Psychol.* 29, 425–445. doi: 10.1080/01443410903046864
- Strømso, H. I., and Bråten, I. (2014). Students' sourcing while reading and writing from multiple documents. *Nord. J. Digital Literacy* 9, 92–111.
- Strømso, H. I., Bråten, I., and Britt, M. A. (2010). Reading multiple texts about climate change: the relationship between memory for sources and text comprehension. *Learn. Instruct.* 20, 192–204. doi: 10.1016/j.learninstruc.2009.02.001
- Strømso, H. I., Bråten, I., and Samuelstuen, M. S. (2008). Dimensions of topic-specific epistemological beliefs as predictors of multiple text understanding. *Learn. Instruct.* 18, 513–527. doi: 10.1016/j.learninstruc.2007.11.001
- Trabasso, T., van den Broek, P., and Suh, S. Y. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse Process.* 12, 1–25. doi: 10.1080/01638538909544717
- van den Broek, P., Young, M., Tzeng, Y., and Linderholm, T. (1999). "The landscape model of reading," in *The Construction of Mental Representations During Reading*, eds H. van Oostendorp and S. R. Goldman (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 71–98.
- von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E. M., and Berthold, K. (2016). The use of source-related strategies in evaluating multiple psychology texts: a student–scientist comparison. *Read. Writ.* 29, 1677–1698. doi: 10.1007/s11145-015-9601-0
- Wiley, J., and Voss, J. F. (1996). The effects of 'playing historian' on learning in history. *Appl. Cogn. Psychol.* 10, 63–72. doi: 10.1002/(sici)1099-0720(199611)10:7<63::aid-acp438>3.0.co;2-5
- Wiley, J., and Voss, J. F. (1999). Constructing arguments from multiple sources: tasks that promote understanding and not just memory for text. *J. Educ. Psychol.* 91, 301–311. doi: 10.1037/0022-0663.91.2.301
- Wineburg, S. S. (1991). Historical problem solving: a study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *J. Educ. Psychol.* 83, 73–87. doi: 10.1037/0022-0663.83.1.73
- Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995). The construction of situation models in narrative comprehension: an event-indexing model. *Psychol. Sci.* 6, 292–297. doi: 10.1111/j.1467-9280.1995.tb00513.x

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Mahlow, Hahnel, Kroehne, Artelt, Goldhammer and Schoor. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



## APPENDIX

### Example-Items:

#### Requirement 1: Corroboration of information across texts

- Do the texts agree with regard to the following issues? (agree/disagree)
  - The question whether forgiving is dependent from culture or not.
  - The role of rage for forgiving.

#### Requirement 2: Integration of information across texts

- Based on the information of all 3 texts: who is the most probable to forgive another person? (1 out of 4).
  - A very religiously living male Asian with difficulties to decide.
  - A female Asian who is very decisive.
  - An atheist male American without contact to the wrongdoer.
  - A Greek nun with lots of contact to the wrongdoer.

#### Requirement 3: Comparison of sources and source evaluations

- Are the following statements correct? (yes/no).
  - The authors of the texts probably pursue very similar goals.
  - The texts could have been written independently from each other with the authors not knowing the other texts.

#### Requirement 4: Comparison of source-content links (several sources)

- Compare the three dimensions of factors influencing forgiveness according to Thompson et al. with the process model by Shavelton and van den Bechele. Which statement is correct? (1 out of 4).
  - All factors influencing forgiveness in the model by Shavelton and van den Bechele can be classified into the dimensions according to Thompson et al., but not all dimensions according to Thompson et al. can be assigned to one or more phases of the model by Shavelton and van den Bechele.
  - All dimensions according to Thompson et al. can be assigned to one or more phases of the model by Shavelton and van den Bechele but not all factors influencing forgiveness in the model by Shavelton and van den Bechele can be classified into the dimensions according to Thompson et al.
  - Both all dimensions according to Thompson et al. can be assigned to one or more phases of the model by Shavelton and van den Bechele, and all factors influencing forgiveness in the model by Shavelton and van den Bechele can be classified into the dimensions according to Thompson et al.
  - Neither can all dimensions according to Thompson et al. be assigned to one or more phases of the model by Shavelton and van den Bechele nor can all factors influencing forgiveness in the model by Shavelton and van den Bechele be classified into the dimensions according to Thompson et al.



# Multiple Texts as a Limiting Factor in Online Learning: Quantifying (Dis-)similarities of Knowledge Networks

Alexander Mehler<sup>1\*</sup>, Wahed Hemati<sup>1</sup>, Pascal Welke<sup>2</sup>, Maxim Konca<sup>1</sup> and Tolga Uslu<sup>1</sup>

<sup>1</sup> Text-Technology Lab, Institute of Computer Science, Faculty of Computer Science and Mathematics, Goethe University Frankfurt, Frankfurt, Germany, <sup>2</sup> Machine Learning and Artificial Intelligence Lab, Institute for Computer Science, University of Bonn, Bonn, Germany

## OPEN ACCESS

### Edited by:

Olga Zlatkin-Troitschanskaia,  
Johannes Gutenberg University  
Mainz, Germany

### Reviewed by:

Guillermo Solano-Flores,  
Stanford University, United States  
Mary Frances Rice,  
University of New Mexico,  
United States

### \*Correspondence:

Alexander Mehler  
mehler@em.uni-frankfurt.de

### Specialty section:

This article was submitted to  
Digital Education,  
a section of the journal  
Frontiers in Education

Received: 15 May 2020

Accepted: 23 September 2020

Published: 03 November 2020

### Citation:

Mehler A, Hemati W, Welke P,  
Konca M and Uslu T (2020) Multiple  
Texts as a Limiting Factor in Online  
Learning: Quantifying (Dis-)similarities  
of Knowledge Networks.  
Front. Educ. 5:562670.  
doi: 10.3389/feduc.2020.562670

We test the hypothesis that the extent to which one obtains information on a given topic through Wikipedia depends on the language in which it is consulted. Controlling the size factor, we investigate this hypothesis for a number of 25 subject areas. Since Wikipedia is a central part of the web-based information landscape, this indicates a language-related, linguistic bias. The article therefore deals with the question of whether Wikipedia exhibits this kind of linguistic relativity or not. From the perspective of educational science, the article develops a computational model of the information landscape from which multiple texts are drawn as typical input of web-based reading. For this purpose, it develops a hybrid model of intra- and intertextual similarity of different parts of the information landscape and tests this model on the example of 35 languages and corresponding Wikipedias. In the way it measures the similarities of hypertexts, the article goes beyond existing approaches by examining their structural and semantic aspects intra- and intertextually. In this way it builds a bridge between reading research, educational science, Wikipedia research and computational linguistics.

**Keywords:** multiple texts, information landscape, knowledge graphs, intratextual similarity, intertextual similarity, three-level topic model, network similarity measurement, linguistic relativity

## 1. INTRODUCTION

Reading is increasingly carried out by means of online multiple texts, which can simultaneously consist of (segments of) texts of diverse genres, registers, authorships, credibilities etc. (Barzilai and Zohar, 2012; Goldman et al., 2012; Britt et al., 2018). That is, learning takes place, so to speak, on the basis of “document collages” whose components are gathered from a constantly growing, nowadays mostly web-based information landscape (Zlatkin-Troitschanskaia et al., 2018) or space (Hartman et al., 2018)<sup>1</sup>. The multiplicity of the texts involved and the diversity of their genres and register

<sup>1</sup>The term information landscape (Zlatkin-Troitschanskaia et al., 2019a,b) includes but is not limited to the web as both a huge and highly diverse text data repository or source of online reading (Cho and Afflerbach, 2015; Wolf, 2018), which is substructured along countless web genres (Mehler et al., 2010), registers and thematic domains. Online reading resembles a traversal of this landscape, each of which involves numerous decisions about what to read, in what sequence and in what depth. The term “landscape” manifests this dual character: on the one hand, the information landscape is a repository that offers innumerable decision possibilities (intertext-as-product perspective), which on the other hand are to be decided by the reader (Cho and Afflerbach, 2015; Britt et al., 2018) in such a way that for each reading process a (usually different) multiple text is delimited in this landscape (intertext-as-process perspective).

(Halliday and Hasan, 1989) are text-linguistic characteristics of online reading (Britt et al., 2018). A third, so to speak macroscopic aspect of this process is the starting point of this article. It is about the *Information Landscape* (IL) from which innumerable readers in countless reading processes delineate ever new multiple texts and thus manifest a distributed process through which this landscape is opened up. To introduce our research agenda regarding this IL, we start from the *Documents Model* (DM) of Perfetti et al. (1999) and Britt et al. (2012). While multiple texts are studied by a wide range of approaches<sup>2</sup>, the reason for choosing the DM as a starting point is due to its text-linguistic heritage—based on the *Construction-Integration Model* (CIM) of Kintsch (1998)—and its context model, which facilitates modular extensions. As far as the text-linguistic orientation of the DM is concerned, its notion of the so-called intertext model is of particular interest for our study of the IL.

Generally speaking, the DM distinguishes two outcomes of multiple text comprehension: the *Intertext Model* (IM), which comprises representations of the constituents of multiple texts and their links, and the *Mental Model* (MM)<sup>3</sup> as a result of comprehension processes that operate within and beyond the boundaries of these constituents. This includes the process of integration, a term borrowed from the CIM, which in the DM also concerns information from different texts. In contrast to text linguistics, which predicts that cohesion and coherence relations should be resolvable within the boundaries of a text to facilitate its understanding (Kintsch, 1998), this condition does not usually apply to multiple texts: they induce additional intrinsic cognitive loads (Sweller, 1994) as a result of interacting elements of separate texts (such as conflicting, contradicting, or otherwise incoherent statements about the same event; Barzilai and Zohar, 2012) and increased efforts in decision making as a result of hyperlinkage (DeStefano and LeFevre, 2007)<sup>4</sup>. Consequently, Goldman et al. (2012, p. 357) speak of (as we may add: online) reading as an *intertextual process*.

Britt et al. (2012) assume that intertext models represent selected constituents of multiple texts as “document entities” together with entity-related information (e.g., on authorship). This is supplemented by three types of links: IM-related source-to-source (e.g., *x supports or contradicts y*), MM-related content-to-content and source-to-content links (see **Figure 1**). A prediction of the DM, which is crucial for our work, is that the probability of generating an intertext model as a result of reading a multiple text is a function of the number of the texts involved, their authors, the perspectives they provide on the corresponding described situation (Britt et al., 2012, p. 171), the tasks to be accomplished and other contextual factors (Britt

et al., 2018). This suggests to speak of the intertext model as a kind of *cognitive map* (Downs and Stea, 1977) of the underlying multiple text, where the MM abstracts from this textbase (e.g., by applying macro operations; van Dijk, 1980; van Dijk and Kintsch, 1983): that is, readers produce intertext models as cognitive maps of multiple texts as parts of the underlying IL, while groups or communities of readers produce distributed cognitive maps (Mehler et al., 2019) of larger sections of the IL or the IL as a whole. This duality of small- and large-scale reading processes leads to the object of this article. That is, we ask how the IL looks like from the perspective of these distributed cognitive maps or vice versa, how it presents itself to different reader communities.

The latter question will be in the focus of this article, which is organized as follows: section 2 outlines a model of distributed reading of multiple texts by multiple readers that overcomes the single-reader perspective of the DM. Section 3 explains the relevance of Wikipedia for the educational science pursued here and gives an overview of related research. Section 4 explains our research questions and describes in detail the methods we have developed to answer them. In section 5, we describe our experiments and discuss their results. More specifically, in section 5.2 we explain how our approach relates to research on linguistic relativity and in section 5.3 we discuss two implications of our findings for research on information processing and online reasoning in the context of higher education, which is the focus of the special issue to which this article contributes. This includes two implications: The first one concerns a reformulation of the closed world assumption which describes a problematic attitudinal basis for reading learning resources like Wikipedia; the second concerns the problem of blurred domain boundaries regarding a microstructural, learner-related and a macrostructural perspective in terms of distributed reading processes. Finally, in section 6 we give a conclusion and an outlook on future work.

## 2. TOWARD A MODEL OF DISTRIBUTED READING OF MULTIPLE TEXTS

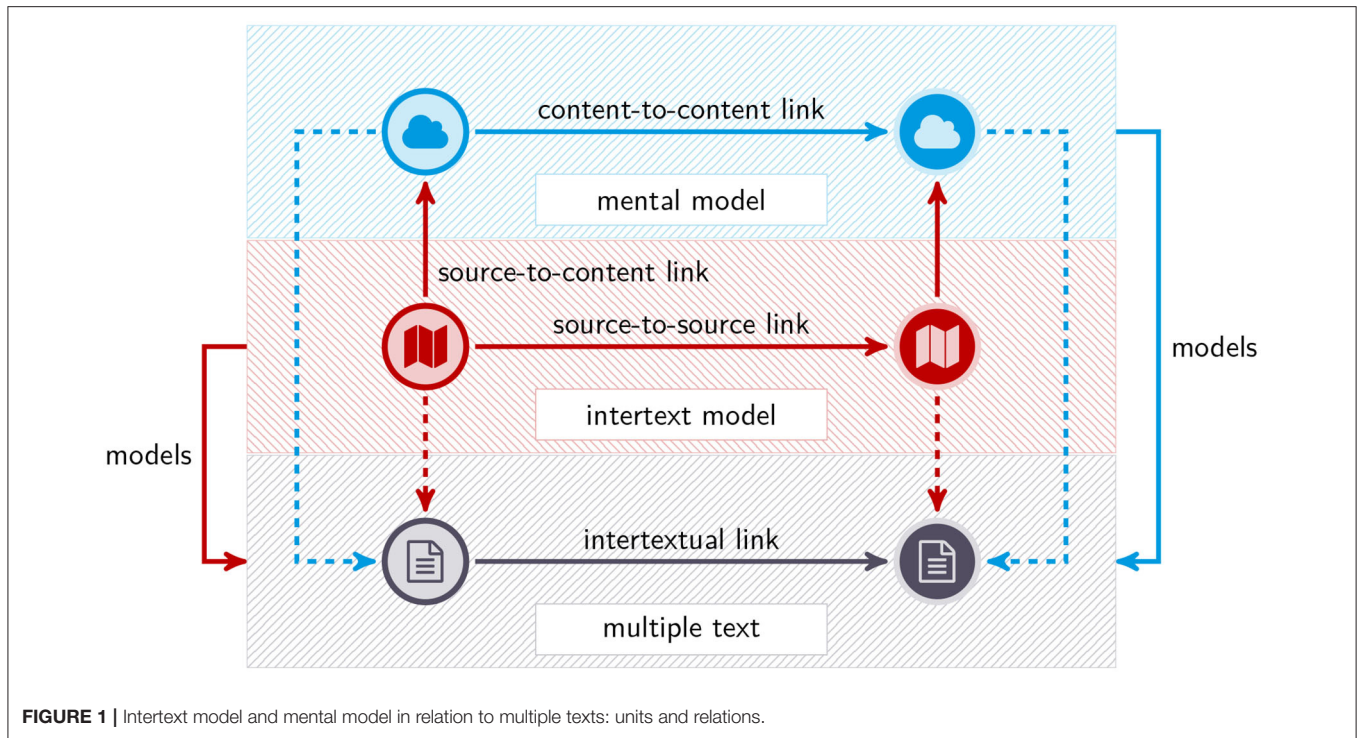
Although the DM takes the necessary step of generalizing the CIM toward modeling multiple texts, it is largely single reader-oriented. To broaden this focus, we generalize the DM conceptually in two steps:

1. Step 1: *Microscopic alignment*: In a first step, which is still within the boundaries of the DM, we consider intertext models as reader-centered approximations of parts of the IL with varying degrees of explicitness (reflecting the probability mentioned above). This predicts that different readers can approximate different parts (i.e., multiple texts) of the IL just as they can align (Pickering and Garrod, 2004) their intertext models of overlapping parts depending on their interaction, which according to Salmerón et al. (2018, p. 286) is a characteristic of non-academic online reading and also regards online collaborative learning (Coiro et al., 2018, p. 487). Starting from the context model of Britt et al. (2018, p. 53), **Figure 2** illustrates this alignment scenario in terms of a situation semantic adaptation (Mehler and

<sup>2</sup>See Primor and Katzir (2018) for a current overview and List and Alexander (2019) for an integrated view of this model landscape.

<sup>3</sup>Instead of situation models, Britt et al. (2018) speak of integrated models as mental models of the semantic content of multiple texts, to emphasize that this content is not limited to descriptions of situations (as provided, e.g., by narrative texts). In this way, depending on the task, texts of different types get into the focus of their reading model. We prefer to speak of mental models to emphasize the openness of our approach to the DM.

<sup>4</sup>For a review of aspects of cognitive load of Internet-based online reading see Loh and Kanai (2016).



**FIGURE 1** | Intertext model and mental model in relation to multiple texts: units and relations.

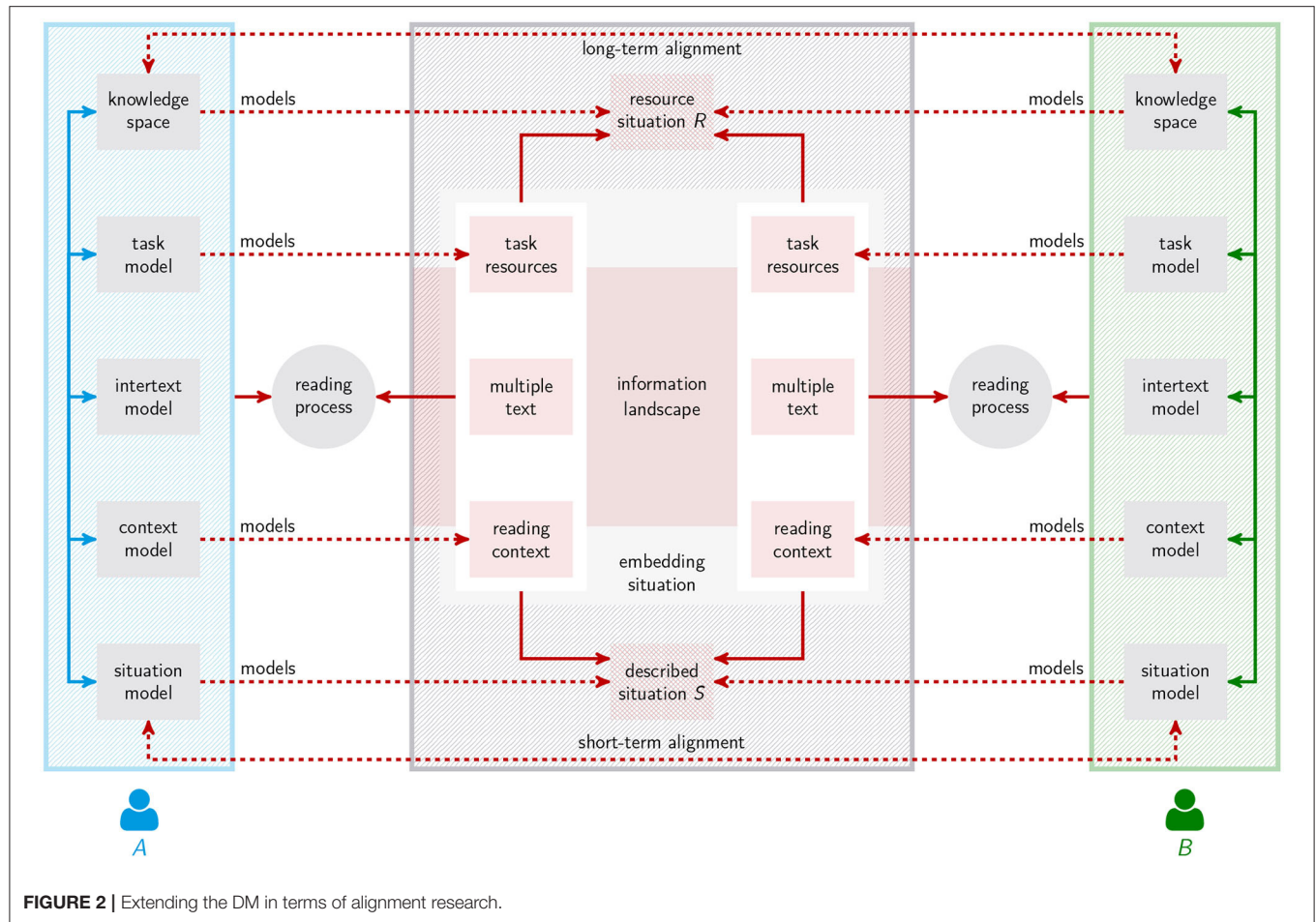
Ramesh, 2019): two readers *A* and *B* read not necessarily different multiple texts in related contexts (related to space, time, etc.) to solve the same or related tasks (manifested by task instructions etc.) in order to achieve the same or related goals. We assume that the texts originate from an IL in whose context they describe a situation *S* and that *A* and *B* refer to the same resource situation *R* to understand which situation their multiple texts actually describe, where all references to contextual units are indirect: they are mediated through mental representations (cf. Britt et al., 2018, p. 45) of multiple texts (IM), described situations (MM), reading contexts (context model), task contexts (task model) and resource situations [long-term memory (LTM) or knowledge space]. As a result of collaborative, cooperative or simply parallel reading processes, the representations of the readers may align with each other, in the short-term (concerning, e.g., their MMs) or long-term (regarding their LTMs). That is, as illustrated in **Figure 2**, *A* and *B* have the possibility to align their mental representations so that they understand the same or similar multiple texts as descriptions of the same or similar situations<sup>5</sup>. Evidently, such an alignment requires many things, but at least the chance that both readers have access to the same or semantically sufficiently similar texts from which they can extract the same or sufficiently similar

multiple texts. *But do they?* This question brings us to the second step of generalizing the DM:

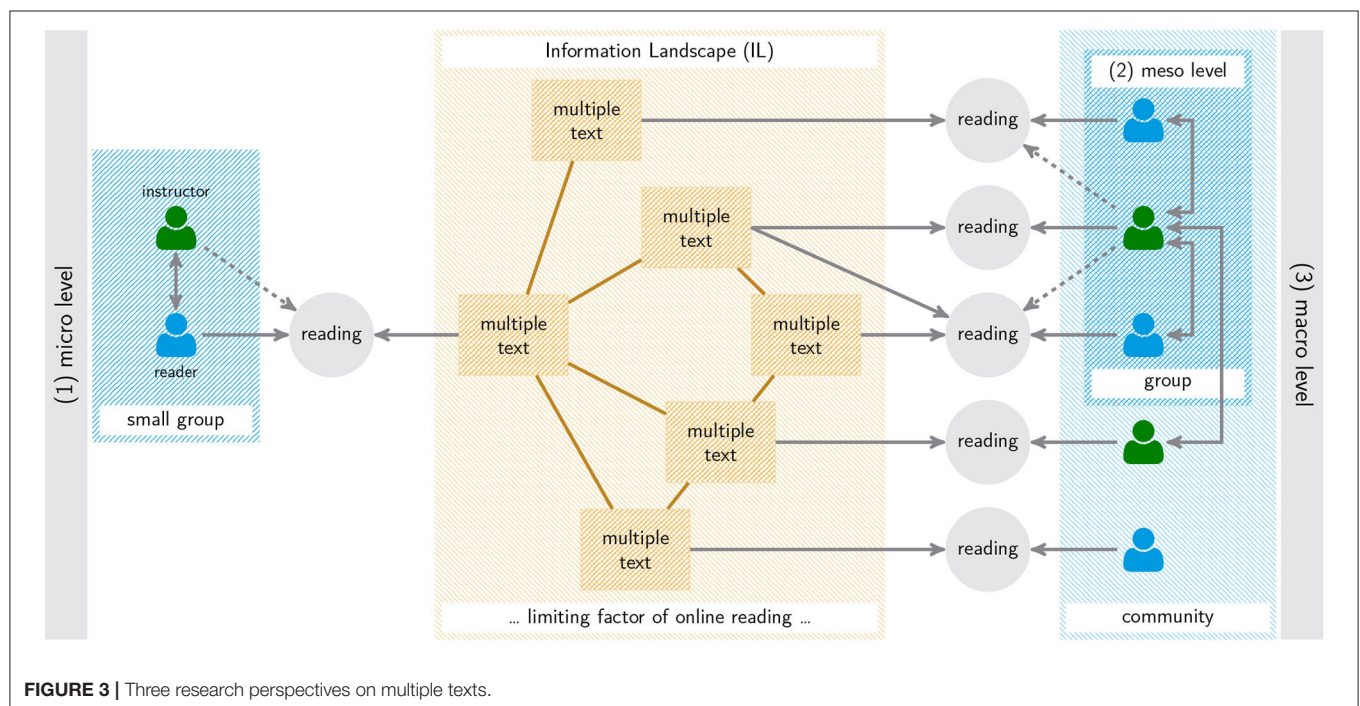
2. Step 2: *Macroscopic alignment*: From the perspective of reader communities, reading is a distributed process that approximates a multifaceted IL, so that both the IL and its distributed representation by innumerable intertext models jointly develop. Obviously, parallel to the diversity of the IL, communities of readers are also diverse as a result of a wide range of factors (Hsieh, 2012; Hargittai and Dobransky, 2017; Braasch et al., 2018b): for example, membership in different language communities (as focused by this article), ethnicities, cultures, age groups, social groups (e.g., families), residency in different places (or geographies; Graham et al., 2015) or practice of different social roles. In any event, in analogy to Step 1, we may expect that different communities dealing with the same or semantically similar parts of the IL should be able to align their corresponding intertext models among each other. In shorter terms: different groups should be able to represent similar parts of the IL in a similar way. *But do they actually have access to the same or at least sufficiently similar parts of the IL—especially under the condition that they deal with the same topic?*
3. Step 1 concerns the microscopic alignment of intertext models as a result of reading situations that are paired together via shared, overlapping, or otherwise related multiple texts, tasks, etc. Step 2 refers to the macroscopic alignment of reader communities as a result of countless such pairings, whereby these communities and consequently their alignments are subdivided according to their social structure. In the range of these extremes there are meso-level alignment processes manifested by smaller groups of agents (such as learning

<sup>5</sup>Adding the notion of a described situation to the context model of Britt et al. (2018) allows for distinguishing communication scenarios, such as misunderstandings (*A* and *B* represent the same situation *S* in mutually incompatible ways), disinformation (*A* pretends to describe a situation he or she knows to be unreal, whereas *B* considers it to be factual) or misinformation (*A* and *B* represent the same unreal situation *S*, which both consider real) (Kendeou et al., 2019) and related phenomena.





**FIGURE 2 |** Extending the DM in terms of alignment research.



**FIGURE 3 |** Three research perspectives on multiple texts.

groups) (see **Figure 3**). The three levels have in common that the underlying IL is temporally on a different scale: it is subject to a lower dynamic than the multiple texts that are extracted from it as a result of reading processes—even in the case of algorithmically (generated and) linked documents (where the underlying algorithms may reflect user profiles). But how uniformly does the IL present itself to its (communities or groups of) readers? Obviously this question is currently outside the scope of the framework of the DM and its relatives.

Step 2 concerns precisely the viewpoint of this article. That is, we are concerned with a central prerequisite for alignable intertext models among readers as members of large communities. This refers to the intertextual shape of the IL from the perspective of different communities who may have different accesses to it or “see” different landscapes, even in situations where the opposite would be assumed. The DM and related approaches do not model what the multiple texts are extracted from and what countless intertext models in their distributed totality ultimately represent, that is, an underlying multifaceted, highly dynamic IL, its numerous document nodes and their relational, intertextual embeddings.

According to Hartman et al. (2018, p. 56), reading research mostly considers small amounts of offline texts pre-selected by the experimenter rather than open ILs in which users decide what to read. But if reading is a kind of problem solving that involves multiple search and decision processes (Britt et al., 2018, p. 43) (e.g., about what to search for and where to find it), then the question arises as to the limits of these processes as imposed by the IL and how they differ for which reader communities. Apparently, approaches to multiple texts focus on micro-models that leave the corresponding macro-models, which inform about the shape of the IL and its organizational laws, under-specified. The present paper takes a step in the direction of filling this gap: it develops a macroscopic model of the IL and examines how its shape appears from the perspective of certain large-scale reader communities. Our aim is, so to speak, to impart knowledge about the “wild” in which the sort of reading takes place which according to Braasch et al. (2018b, p. 535) is to become the subject of reading research. Thus, our approach is complementary to current research on the intertext model: we study the IL underlying the construction of intertext models from a macroscopic perspective, in contrast to reading research, which starts from a microscopic perspective of small groups or individual readers (see **Figure 3**). In terms of the integrated framework of multiple texts (List and Alexander, 2019) we are concerned with the intertextuality of those information units to which the cognitive strategies and behavioral skills of readers are related. That is, in modification of the fourth goal of future research on the use of multiple sources according to Braasch et al. (2018b), we deal with the phenomenon that different communities are offered different information, especially in the context of the same topic. The extent to which this phenomenon applies to different language communities will be examined using the example of the most frequently used knowledge resource on the Web, that is Wikipedia (RRID:SCR\_004897).

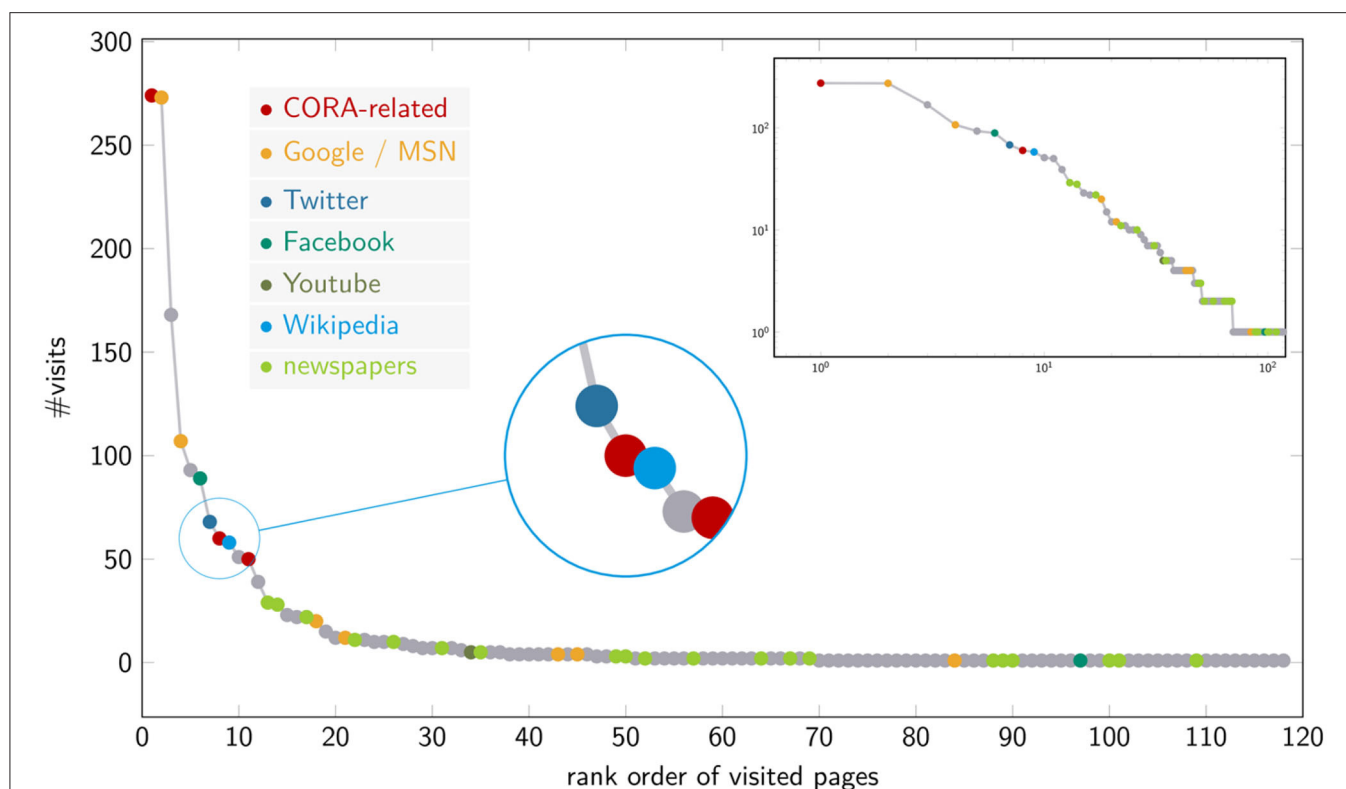
### 3. WIKIPEDIA: EDUCATIONAL RELEVANCE AND BIAS

Wikipedia is a primary multilingual source for web-based knowledge acquisition and online learning (Lucassen and Schraagen, 2010; Okoli et al., 2012, 2014; Head, 2013; Mesgari et al., 2015; Konieczny, 2016; Singer et al., 2017; Lemmerich et al., 2019). It not only offers releases for numerous languages, but is changing and growing to a degree that makes it a first reference for an open topic universe, a resource which—besides the web as a whole—reflects the breadth, depth, and dynamics of human knowledge generation to an outstanding degree: whenever a new topic comes up, little time passes before Wikipedia provides information about how it is classified and related to already established topics<sup>6</sup>. Its dynamics and multilingualism even make it a source for the automatic extraction of learning content (Pentzold et al., 2017; Conde et al., 2020). The educational relevance refers to Wikipedia as a whole as well as to specific domains, such as law or health (Okoli et al., 2014; Smith, 2020). Singer et al. (2017) show that 16% of English Wikipedia uses are work/school-related tending to involve long page views in thematically coherent sessions. Lemmerich et al. (2019) additionally show that these uses vary considerably across languages, especially in the case of work/school-related uses. Independent of academic reservations (cf. Mesgari et al., 2015; Konieczny, 2016), Wikipedia establishes itself as a source for reading by students (Okoli et al., 2014), as an additional learning resource (Konieczny, 2016), partly with advantages over textbooks (Scaffidi et al., 2017), a resource to which students themselves contribute (Okoli et al., 2012). **Figure 4** illustrates the relevance of Wikipedia for online academic reading: it shows the Zipfian distribution of online sources frequented by students in the CORA study (Nagel et al., accepted). In this context, Wikipedia appears as an outstanding reference. Even more so: taking into account the “mutual beneficial relationship” of Google and Wikipedia (McMahon et al., 2017) (according to which traffic on Wikipedia originates mainly from Google), both resources largely dominate this usage scenario.

Given the exceptional position of Wikipedia, the research community has investigated numerous biases regarding its content and use. Denning et al. (2005) speak of accuracy, motive, expertise, volatility, coverage, and source as risks of Wikipedia, which are potential reference points for biases. In terms of coverage, this concerns, for example, the “self-focus bias” (Hecht and Gergle, 2009) or the tendency that topic selections reflect author preferences (cf. Holloway et al., 2007; Halavais and Lackaff, 2008). In this context, Hecht and Gergle (2010a) show for 25 language editions that they differ enormously in the coverage of topics, that this diversity is not explained by their size and that English Wikipedia does not cover its sister editions. A similar approach is taken by Warncke-Wang et al. (2012), who investigate a variant of Tobler’s first law (Tobler, 1970) according to which geographically nearby Wikipedias are more similar

<sup>6</sup>Given its relevance, it is astonishing how rarely Wikipedia is mentioned as a source for reading in Braasch et al. (2018a).





**FIGURE 4 |** Rank-frequency distribution of accesses to web resources in the CORA project. The embedded figure shows the frequency distribution as a log-log plot.

in terms of *Inter-Language Links* (ILL). What distinguishes these studies from ours is that, with few exceptions, conceptual alignments of Wikipedias concern only paired articles identified by ILLs, so that hypertext structure, which is crucial for online learning in terms of the DM, is ignored. Furthermore, similarities of articles are quantified by degree statistics, so that one can hardly speak of a content-based comparison. Comparable to the latter studies, however, we also refer to aligned articles to map shared topics (cf. Bao et al., 2012), but do so via Wikidata (RRID:SCR\_018492) to identify commonly referred concepts and retain their network structure as manifested by Wikipedia's article graph.

Biased topic coverage is related to what Massa and Scrinzi (2012) call the linguistic point of view (contradicting Wikipedia's NPOV), which predicts that different (e.g., cultural) communities tend to present the same topics differently. In line with this view, thematic biases reflect cultural differences. This perspective is further developed by Miquel-Ribé and Laniado (2016), who speak of cultural identities, according to which editors tend to write about topics related to their culture. Using geo-referenced data and keyword-related heuristics, they identify cultural identity related articles (on average 25% of articles in 40 languages) to diagnose language-specific thematic preferences (regarding 15 languages) and translation-related associations of editions that are dominated by certain topics (e.g., *geography*). Thematic preferences are modeled using 18 topics derived with the help of Wikipedia's category system. Based on this analysis,

they distinguish types of culture-related articles: language-specific articles shared by a few editors and articles appearing in many languages (in terms of ILLs). Another example is Miz et al. (2020), who examine English, French and Russian Wikipedia by exploring clusters of trending articles using topic modeling based on eight topics. In contrast to these studies, we not only consider a larger number of languages and many more topics, but especially semantically coherent article subnetworks, which are examined for their differences along intra- and intertextual dimensions.

Biases of Wikipedia were also analyzed regarding selected areas: Lorini et al. (2020) observe a variant of Tobler's Law according to which authors tend to write about geographically close events. Similarly, Samoilenko et al. (2017) describe a preference for recency. Oeberst et al. (2019) investigate a bias of groups who present their views more positively (cf. Álvarez et al., 2020). A related example regarding biographical articles that includes linguistic analyses is given by Callahan and Herring (2011). Wagner et al. (2016) also present a multidimensional content analysis, now of a gender bias. Given the importance of Wikipedia as a knowledge repository and taking into account its various biases, the question of its influence on knowledge formation on the part of readers comes up (Oeberst et al., 2018).

The research considered so far shares the observation of a biased topic coverage, which relativizes Wikipedia's domain independence (Jiang et al., 2017), since certain topics (Kittur et al., 2009) or views dominate, be it due to cultural preferences

(Massa and Scrinzi, 2012; Laufer et al., 2015; Miquel-Ribé and Laniado, 2016), language differences (Hecht and Gergle, 2010a; Massa and Scrinzi, 2012; Warncke-Wang et al., 2012; Samoilenko et al., 2016), geographical factors (Hecht and Gergle, 2010b; Karimi et al., 2015; Laufer et al., 2015; Samoilenko et al., 2016; Lorini et al., 2020), or the fact that group membership influences POV (Oeberst et al., 2019). However, though these observations should be based on content-related analyses of large amounts of data, they often concern rather non-content related features (e.g., degree statistics) taking into account a maximum of 20 topics, so that topic resolution is kept low while hypertext structure is underrepresented. On the other hand, concentrating on selected areas allows accurate linguistic analyses to be carried out, but these are difficult to automate and thus difficult to apply across languages. What is needed, therefore, is a procedure that allows for more precise thematic and article network-related analyses and which can be automatically applied to many languages. Exactly such a procedure is presented here: it uses Wikidata to identify subjects of articles of whatever subject areas, and about 100 topic categories to model their diversity. This will allow us to investigate biased topic coverage intra- and intertextually.

## 4. RATIONALE AND METHOD

To investigate the topic coverage of Wikipedia in educationally relevant areas, we investigate how the descriptions of the same entities or knowledge objects (from the fields of economics, physics, chemistry, biology, etc.) are distributed across its language editions. That is, we investigate how Wikipedia presents itself to its readers as part of the IL in the area of education-related reading. We test the hypothesis that the extent to which one obtains information on a given topic depends on the language in which Wikipedia is consulted. Given the skewed size distribution of Wikipedia's releases for different languages, this may sound obvious at first. But we will control the size factor and examine this hypothesis for individual subject areas and topics. Since Wikipedia is a highly frequented part of the IL, this would indicate a *language-related bias*, that is, a sort of *linguistic relativity*. Thus, our article is ultimately concerned with the question of whether Wikipedia exhibits this kind of relativity or not. We test this hypothesis on the example of 35 Wikipedias. To this end, we focus on three research questions:

**Q1** How do languages resemble each other in terms of the knowledge networks that manifest themselves in the associated Wikipedias?

Obviously, high dissimilarities in the latter sense mean that students who consult the associated Wikipedias are informed very differently about the same field of knowledge. Differences in knowledge between learners of different languages may then be consolidated or even expanded as a result of such a bias. An example of such a scenario is shown in **Figure 5**, which contrasts networks of articles about paintings from German and Dutch Wikipedia. The extracted networks are obviously very different; they show very different parts of the information landscape, although on the same subject area.

**Q2** How do the similarity ratings after Q1 differ depending on the underlying knowledge domain?

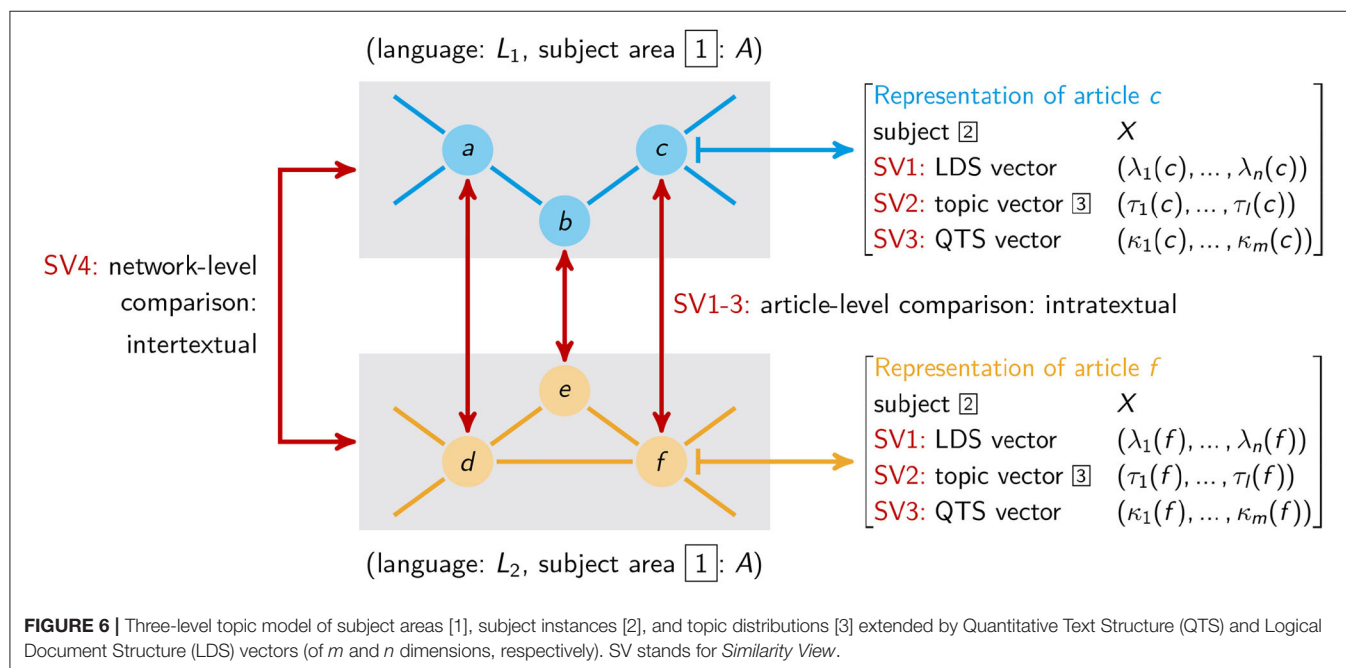
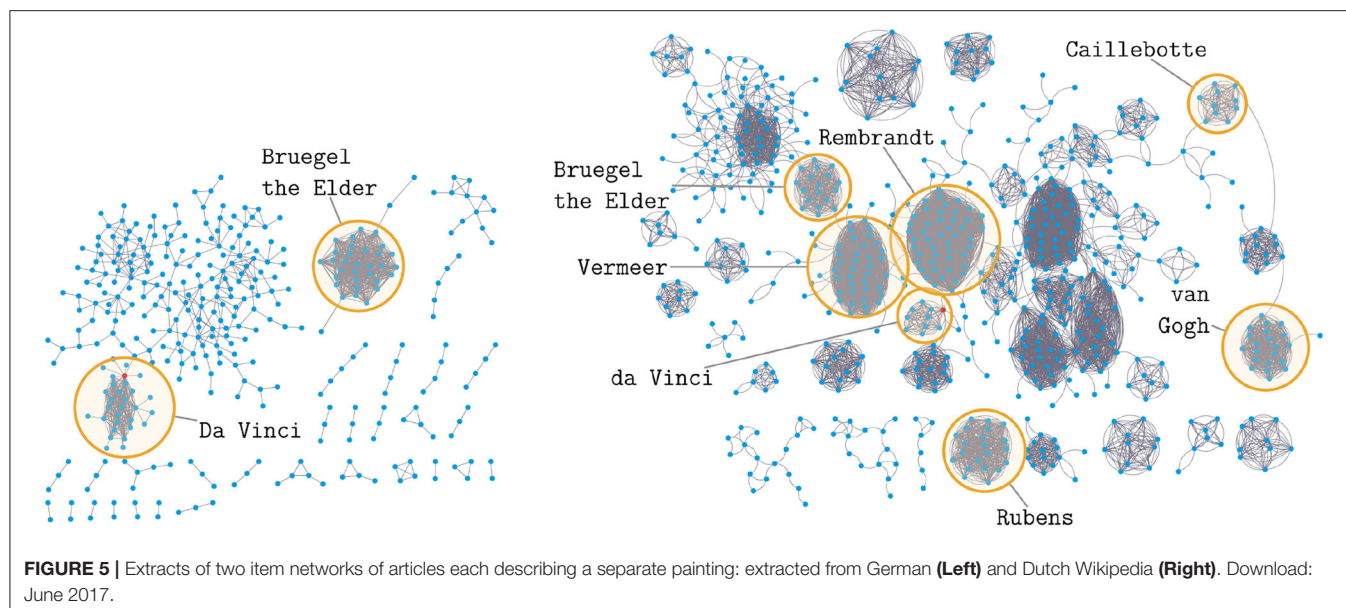
The various fields of knowledge and scientific disciplines that contribute to their development have been developed in different ways. Domain-specific learning can therefore benefit considerably from discipline-specific strategies of knowledge acquisition and processing (List and Alexander, 2019). If this is true, then we might expect a shadowing of these differences in Wikipedia: it is then likely that different fields of knowledge manifest themselves differently in Wikipedia, while the similarities of languages as manifested in Wikipedia are strongly conditioned by the reference to these fields. By answering Q2, we inform educational research about the manifestation density of certain knowledge domains in certain language editions. This research can help to avoid wrong conclusions from generalizations for knowledge domains or languages: different outcomes of learners of different languages, for example, may be the result of differences in such resources and not necessarily the result of different linguistic structures of the underlying task descriptions (Mehler et al., 2018).

**Q3** Regardless of such differences, is there a knowledge-related “lingua franca” which, by its Wikipedia, makes the dissemination of knowledge in other languages predictable and thus serves as a reference for knowledge dissemination?

English Wikipedia could play the role of such a reference due to its size and status as the primary source of translation between Wikipedias (cf. Warncke-Wang et al., 2012). However, several studies question this role (Hecht and Gergle, 2010a; Samoilenko et al., 2016). Thus, the question arises whether these results also apply to our combined intra- and intertextual model.

To answer Q1–3, we develop a method to extract and compare Wikipedia article graphs on the same topic, henceforth called *item networks*. Since we have to apply this method across languages, it must be both easy to implement and systematically reproducible, in such a way as to ensure that we are dealing with the same subject(s) regardless of the languages under consideration. To realize this measurement operation with the help of Wikidata we compare item networks intra- and intertextually by means of a three-level topic model as depicted in **Figure 6**: starting from any subject area (e.g., *painting*) (1st level), we identify all its instances in Wikidata, which we use to extract all the articles within Wikipedia's language editions that address these instances. In this way, we identify the 2nd level of our topic model, whose elements we refer to here as subjects or subject instances (see **Figure 6**). That is, a Wikipedia article is assigned a unique subject (e.g., *Mona Lisa*) based on its Wikidata mapping, which specifies the entity (i.e., a *Wikidata item* in the role of the *definiendum*) that the article describes. At this stage we get two topic assignments for each article: the corresponding subject area (e.g., painting) and its subject instance (e.g., *Mona Lisa*). Using Wikipedia's article graph, where connections between articles are given by hyperlinks, we then obtain one article network per subject area and language, with the semantic coherence of these item networks resulting from the reference to the underlying subject area common to their articles. In the third

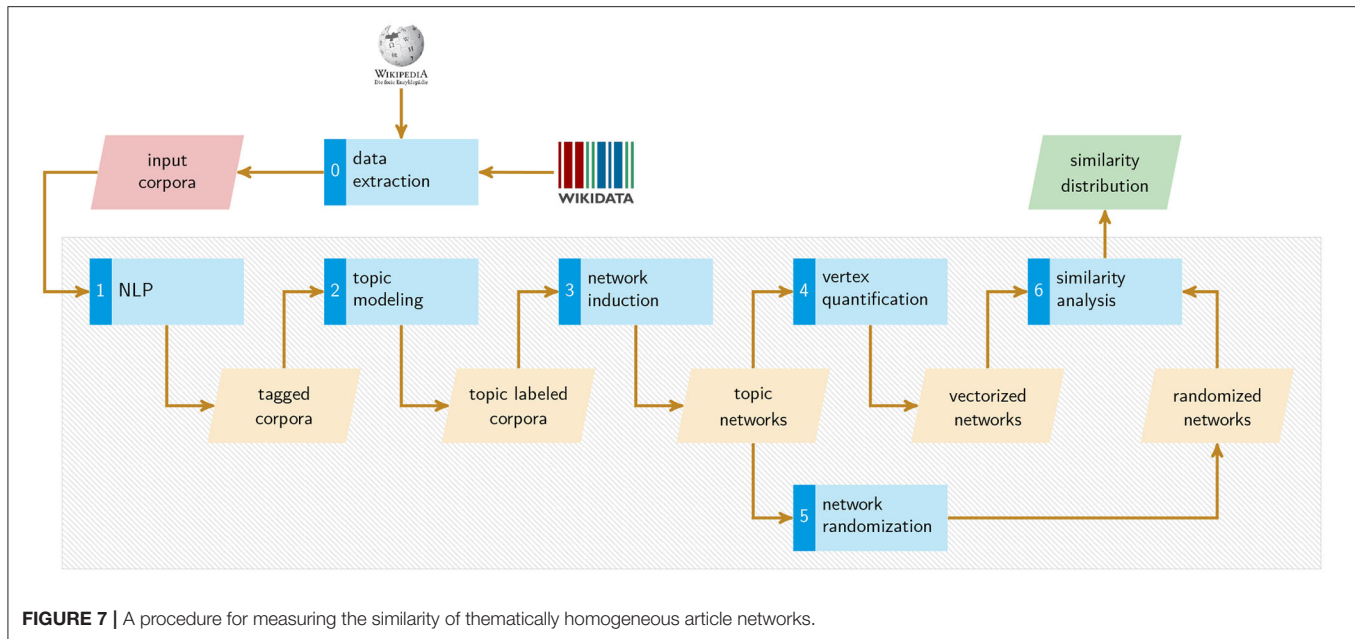




step we characterize each node of these item networks by three vectors (see Figure 6). This applies in particular to the thematic perspectives under which subjects are treated. In this way we reach the 3rd level of our topic model: articles as manifestations of the same subject area are characterized according to the thematic perspectives (contributing to the *definiens*) under which they describe their subject where these perspectives are modeled as topic vectors.

Suppose the subject of a given article in a certain language is the painting *Bal du moulin de la Galette* (by Pierre-Auguste Renoir) which instantiates the subject area *painting*. Further,

assume that our algorithm detects that this article deals with the topic economy (since it reports that this painting is one of the most expensive ever auctioned). Then we get three topic-related assignments (see Figure 6): regarding the article's (abstract) subject area, its (concrete) subject and the thematic perspective of their descriptions. Having done this for whole article networks of different languages, we can then compare these networks at the subject level by asking whether articles about the same subject thematize it in similar ways—this concerns what we call *intratextual similarity*—and at the subject area level by asking whether the article networks are structurally similar in terms of



**TABLE 1** | Variants of using Wikidata items to identify Wikipedia articles as vertices of INs.

		<i>B</i> -hierarchy	
		Unexpanded	Expanded
<b>A-hierarchy</b>	Unexpanded ( $A_{\hat{x}}$ )	$U_{x \in \mathbb{O}} \left\{ \hat{B}_{\hat{x}} \right\}, U_{x \in \mathbb{O}} \left\{ \hat{\mathbb{B}}_{\hat{x}} \right\}$ 1,333,421	$U_{x \in \mathbb{O}} \left\{ \hat{B}_{\hat{x}} \right\}, U_{x \in \mathbb{O}} \left\{ \hat{\mathbb{B}}_{\hat{x}} \right\}$ 24410338
	Expanded ( $A_{\tilde{x}}$ )	$U_{x \in \mathbb{O}} \left\{ \tilde{B}_{\tilde{x}} \right\}, U_{x \in \mathbb{O}} \left\{ \tilde{\mathbb{B}}_{\tilde{x}} \right\}$ 3,190,562	$U_{x \in \mathbb{O}} \left\{ \tilde{B}_{\tilde{x}} \right\}, U_{x \in \mathbb{O}} \left\{ \tilde{\mathbb{B}}_{\tilde{x}} \right\}$ >32,949,992

what we call *intertextual similarity*. In this way, we implement the procedure to answer Q1–3, as shown in **Figure 7**. Under the null hypothesis, different language article networks for the same subject area are very similar, both in terms of the articles' descriptions of the same subjects (intratextual similarity) and in terms of their hypertext structure (intertextual similarity). We will essentially falsify this hypothesis and answer Q1–3 accordingly. We now turn to explaining Step 0–6 of **Figure 7** in detail.

#### 4.1. Data Extraction

In order to extract article networks on the same subject area relevant to educational science, that is, in order to perform Step 0 and 3 of **Figure 7**, we use the classification of the fields of science and technology of OECD (2007) and its correspondents in Wikidata. The resulting networks, induced separately for each Wikipedia, are called *Item Networks* (IN) to emphasize the way they are induced by means of Wikidata items. In this way we guarantee three things: (i) the reference to subject areas, such as those addressed by the PISA studies, (ii) a thematic breadth of the selected topics, and (iii) their transferability between Wikipedias. INs are generated as follows: for each of the OECD categories that

can be assigned to Wikidata, we consult its *studies*-statements to determine all Wikidata classes that are related to the respective OECD category in this way. This is necessary because we need to move from OECD categories (e.g., *art*) (which we refer to for ensuring thematic diversity of general topics) to subject areas (e.g., *painting*) as likely destinations of searches in the context of the former: while the OECD categories and their Wikidata subclasses induce *subclass of*-hierarchies of fields of science and technology (henceforth called *A-hierarchies*), the latter induce *subclass of*-hierarchies of subject areas (cf. **Figure 6**) that are studied in these fields (called *B-hierarchies*). At this point we need an extension to ensure a higher coverage rate of OECD categories. The reason for this is that some of them do not have *studies*-statements in Wikidata, so that they would fall out of the selection process to the detriment of the targeted thematic breadth. Therefore, we additionally examine the descendants of OECD categories in *A-hierarchies* to determine additional classes from *B-hierarchies* by means of these descendants' *studies* statements. This leads to a number of alternatives for sampling INs (see **Table 1**), of which only a subset is feasible, as is now formally explained: let  $\mathcal{D} = (V_{\mathcal{D}}, A_{\mathcal{D}}, \lambda)$  be a representation of Wikidata as a directed graph with the set  $V_{\mathcal{D}}$  of vertices (so-called Wikidata items), the set  $A_{\mathcal{D}}$  of arcs (links) between these items and the arc labeling function  $\lambda$ . Further, let  $\mathbb{O} = \{x, y, \dots\}$  be the set of OECD categories. For a given OECD category  $x \in \mathbb{O}$ , we generate the set  $A_{\hat{x}}$  of all items belonging to the *A-hierarchy* dominated by  $x \in V_{\mathcal{D}}$ :

$$A_{\hat{x}} = \{v \in V_{\mathcal{D}} \mid 0 < \text{ged}_{\text{scot}}(x, v) < |V_{\mathcal{D}}| \} \cup \{x\} \quad (1)$$

where  $\text{ged}_{\text{scot}}(x, v)$  is the length of the shortest directed path from  $x$  to  $v$  in  $\mathcal{D}$  crossing only *subclass of*-links. Alternatively, we dispense with this expansion and get the set  $A_{\tilde{x}} = \{x\}$ . In the case of  $A_{\tilde{x}}$ , we then explore *studies*-links of  $x$  and of its subclasses,

while in the case of  $A_{\tilde{x}}$ , only those of  $x$  are explored. The resulting sets of  $B$ -level items that are “studied” in this sense are denoted by

$$B_{\tilde{x}} = \{w \in V_{\mathcal{D}} \mid \exists v \in A_{\tilde{x}}: (v, w) \in A_{\mathcal{D}} \wedge \lambda((v, w)) = \text{studies}\} \quad (2)$$

$$B_{\tilde{x}} = \{w \in V_{\mathcal{D}} \mid (x, w) \in A_{\mathcal{D}} \wedge \lambda((x, w)) = \text{studies}\} \quad (3)$$

Obviously,  $A_{\tilde{x}} \subseteq A_{\hat{x}}$  and  $B_{\tilde{x}} \subseteq B_{\hat{x}}$ . Now we have two alternatives again: either, we recursively explore *subclass of*-links to get more subject area-related Wikidata items for comparing Wikipedias (this leads to  $\hat{B}_{\tilde{x}}$  and  $\check{B}_{\tilde{x}}$ ) or not (generating  $\check{B}_{\tilde{x}} = B_{\tilde{x}}$  and  $\hat{B}_{\tilde{x}} = B_{\tilde{x}}$ ). A test shows that recursive expansions on the side of  $B$ -hierarchies (generating  $\hat{B}_{\tilde{x}}, \check{B}_{\tilde{x}}$ ) leads to overly large subject area representations that induce computationally hardly processable article networks. Take the example of the OECD category *Mathematics*: if we expand this category on the side of  $B$ -hierarchies, we get candidate items like *set*, which are dominated by *mathematical concept* as the target of a *study*-statement starting from *Mathematics*. But *set* has many instances in Wikidata, many of which are not mathematical concepts. Such examples, which occur frequently in the variants  $\hat{B}_{\tilde{x}}$  and  $\check{B}_{\tilde{x}}$ , realize unwanted changes of subject area, so that  $B$ -sided expansions are mentioned here only as a theoretical alternative. To get processable networks, we alternatively represent the elements of item sets  $C_x \in \{\hat{B}_{\tilde{x}}, \check{B}_{\tilde{x}}, \check{B}_{\tilde{x}}, \check{B}_{\tilde{x}}\}$  as singletons, each containing a single Wikidata item that is finally used to extract Wikipedia article graphs:

$$\mathbb{B}_x = \{\{y\} \mid y \in C_x\} \quad (4)$$

Take the example of  $\check{\mathbb{B}}_{\tilde{x}}$ , that is, the set of singletons, each containing a topic studied either directly or indirectly under OECD category  $x$  –  $\cup_{x \in \mathbb{O}} \check{\mathbb{B}}_{\tilde{x}}$  currently contains 172 labels of corresponding subject areas (see **Table 1**). Obviously,  $\check{\mathbb{B}}_{\tilde{x}} \subseteq \hat{\mathbb{B}}_{\tilde{x}}, \hat{\mathbb{B}}_{\tilde{x}} \subseteq \hat{\mathbb{B}}_{\tilde{x}}, \check{\mathbb{B}}_{\tilde{x}} \subseteq \hat{\mathbb{B}}_{\tilde{x}}$ . Having identified all Wikidata *item sets* for each OECD category  $x$ , we get an expression for the set of candidate subject areas:

$$\mathbb{B} = \bigcup_{x \in \mathbb{O}} \{\hat{B}_{\tilde{x}}, \hat{B}_{\tilde{x}}, \check{B}_{\tilde{x}}, \check{B}_{\tilde{x}}\} \cup \hat{\mathbb{B}}_{\tilde{x}} \cup \hat{\mathbb{B}}_{\tilde{x}} \cup \check{\mathbb{B}}_{\tilde{x}} \cup \check{\mathbb{B}}_{\tilde{x}} \quad (5)$$

An element  $B_x \in \mathbb{B}$  is a set of Wikidata items that are (in-)directly accessible from OECD category  $x$  via *studies*-links. Item sets like  $B_x$  serve to identify Wikipedia articles on the same subject across languages. This is achieved as follows: each item in  $B_x$  allows for identifying a corresponding set of instances (e.g., the painting *Bal du moulin de la Galette* by Pierre-Auguste Renoir) by exploring *instance of*-links in Wikidata. These instances are linked from Wikipedia articles and correspond to subjects in **Figure 6** as instances of subject area  $B_x$ , which in turn is derived from OECD category  $x$  to ensure that we address educationally relevant topics. At this point, an important difference to section 3 becomes clear: the approaches mentioned there operate on small, closed lists of abstract topic categories (similar to the OECD categories used here), whereas we use OECD categories only to address the level of subject areas and their subject instances (cf. **Figure 6**). The

reference to ILL does not solve the problem of these approaches, since ILLs merely define groups of articles on the same subject, to put it in our terminology. In contrast to this, we use Wikidata as a whole to extract arbitrary alignable article networks from different Wikipedias, differentiating between subject areas, subject instances and article-wise topic distributions—in this sense, our approach is both thematically stratified and open. In any event, we limit sampling to  $\check{\mathbb{B}}_{\tilde{x}}$  for extracting article networks. The reason is that  $\hat{\mathbb{B}}_{\tilde{x}}$  and  $\check{\mathbb{B}}_{\tilde{x}}$  explore to few *studies*-links, while variants that expand  $B$ -level items induce unwanted topic changes (see above). From the 172 candidate subject areas belonging to  $\check{\mathbb{B}}_{\tilde{x}}$  (see **Table 1**), we select the 19 largest ones supplemented by six areas. **Figure 8** shows the boxplots of the article networks’ orders (i.e., number of their nodes) for each of these subject areas, which we derived from 35 Wikipedia language editions (see **Table 2**). For these 35 languages we trained topic models to detect the topic distributions of their articles (see section 4.2, Level 3 of **Figure 6** and Step 2 of **Figure 7**).

Now, let  $\mathbb{W} = \{\mathcal{W}_1 \dots, \mathcal{W}_l\}$  be the set of all Wikipedias  $\mathcal{W}_i = (V_i, A_i) \in \mathbb{W}$  each represented as a directed graph with the set of vertices (i.e., articles)  $V_i$  and arc set  $A_i$ . Further, let  $a \in \mathbb{B}$  be a subject area, then we induce for each Wikipedia  $\mathcal{W}_i$  the *Item Network* (IN)  $I_a(\mathcal{W}_i)$  of all articles on subjects that are directly or indirectly studied under the OECD category corresponding to  $a$ : it is defined as the subgraph of  $\mathcal{W}_i$ ’s article graph that consists only of articles which by their Wikidata links are mapped onto instances of elements of  $a$ . To generate this graph, we explore *instance of*-links from elements of  $a$  to Wikidata items addressed by Wikipedia articles. Let  $\iota: \cup_{i=1}^l V_i \rightarrow V_{\mathcal{D}}$  denote the function that links articles as instances to Wikidata items in the latter sense, then we get the following expression for  $I_i^a$  (see Step 3 of **Figure 7**):

$$I_a(\mathcal{W}_i) = I_i^a = (V_i^a, A_i^a) = (\{v \in V_i \mid \iota(v) \in a\}, \{(v, w) \in A_i \mid v, w \in V_i^a\}) \quad (6)$$

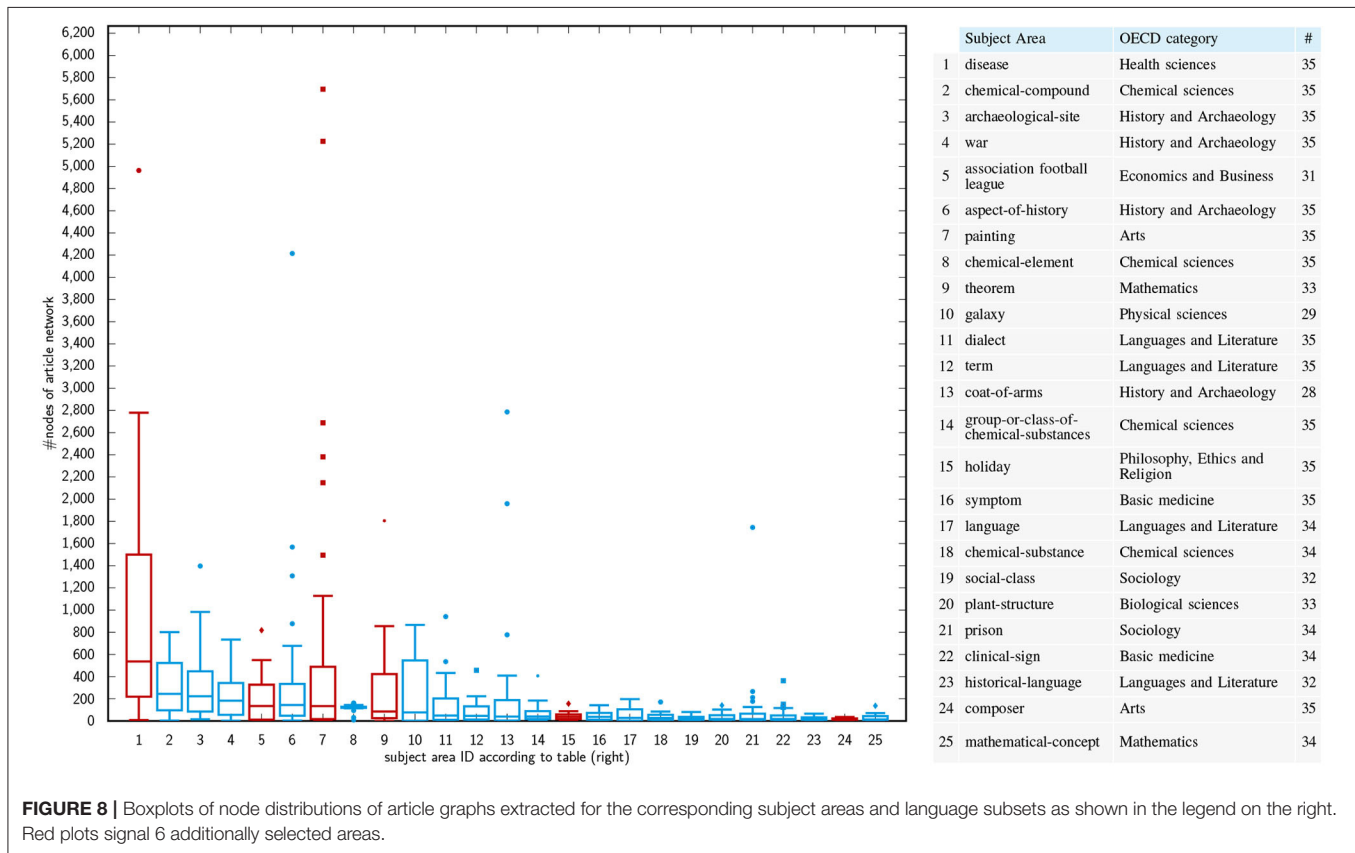
This allows us to finally define the so-called *alignment set* of any pair of INs  $I_i^x = (V_i^x, A_i^x)$ ,  $I_j^y = (V_j^y, A_j^y)$  which contains all pairs of articles of these INs that are alignable because of treating the same subject:

$$\mathcal{W}_{ij}^{xy} = \{(v, w) \mid v \in V_i^x \wedge w \in V_j^y \wedge \iota(v) = \iota(w)\} = (\mathcal{W}_{ji}^{yx})^{-1} \quad (7)$$

## 4.2. A Hybrid Approach to Measuring the Similarity of INs

So far, we described the object of measuring thematic dissimilarities of Wikipedia’s language editions. Now we define the aspects under which we measure these dissimilarities. To this end, we consider (1) *syntactic* (text-structural), (2) *semantic* (topic-related), and (3) text statistical aspects of the similarity of INs. While these *intratextual* aspects are all vertex content-related, a fourth *Similarity View* (SV) focuses on the network structure of INs and thus on *intertextual* aspects of similarity. Our starting point is the observation that each Wikidata item can be addressed by several Wikipedia articles, and vice versa,





the same article can describe several items, even if it is injectively mapped to Wikidata. The calculation of similarities of INs then requires the mapping of network similarities on the level of node content (intratextual) and network structure (intertextual). For this purpose, we vectorize the textual nodes of INs according to the similarity views SV1–4. That is, given a SV  $s$ , we assume that each node  $v \in V_i^x$  of each IN  $I_i^x = (V_i^x, A_i^x)$  is mapped onto a corresponding vector  $\vec{v} \in \mathbb{R}^{k_s}$ . This is expressed by the function

$$h_s: V_i^x \rightarrow \mathbb{R}^{k_s} \quad (8)$$

where  $k_s$  is the dimensionality associated with SV  $s$ . For INs  $I_i^x = (V_i^x, A_i^x)$ ,  $I_j^y = (V_j^y, A_j^y)$  we calculate their similarity node- and network-wise using the alignment set  $\mathcal{W}_{ij}^{xy}$  (with a few exceptions, elements of vectors  $\vec{v}$  are non-negative):  $\forall (v, w) \in V_i^x \times V_j^y$ :

$$\sigma_{s_1}(I_i^x, I_j^y) = \begin{cases} \frac{\sum_{(v,w) \in \mathcal{W}_{ij}^{xy}} \sum_{n=1}^{k_s} h_s(v)[n] h_s(w)[n]}{\sqrt{\sum_{v \in V_i^x} \sum_{n=1}^{k_s} h_s(v)[n]^2} \sqrt{\sum_{w \in V_j^y} \sum_{n=1}^{k_s} h_s(w)[n]^2}} & V_i^x \neq \emptyset \neq V_j^y \\ 0 & \text{else} \end{cases} \in [-1, 1] \quad (9)$$

Computing Formula 9 is part of performing Step 6 of Figure 7. It requires defining vectorization functions  $h_s$  for SV  $s = 1..3$  (cf. Step 4 of Figure 7) (SV4 is introduced below):

1. **SV1: Logical Document Structure (LDS):** according to this SV, pairs of aligned articles are the more similar the more their LDS (Power et al., 2003) resemble each other, and the more such pairs of aligned articles of two INs resemble each other in this sense, the more similar these INs are. For measuring similarities of the LDS of aligned articles we vectorize articles along  $k_1 = 11$  dimensions (cf. Callahan and Herring, 2011): *number of characters, number of sections, breadth of the table of content tree, depth of this tree, number of outgoing links to pages inside Wikipedia, number of outgoing links to pages outside Wikipedia, number of pictures, number of tables, number of links to the Integrated Authority File and related norm data, number of references, and number of categories*. In this way, we identify text pairs that address the same Wikidata item by similar document structures, e.g., with regard to the length of the presentation, the use of images, tables, or hyperlinks. Due to its orientation on surface structural features, this method can easily be calculated across languages.
2. **SV2: Thematic structure:** according to this SV, pairs of aligned articles are the more similar, the more similar the distributions of topics they address when describing their subjects (see Figure 6), and the more such pairs of aligned articles of two INs resemble each other in this sense, the more similar they are. SV2 is implemented with text2ddc (Uslu et al., 2019) (Step 2 in Figure 7), a neural network based on fastText (Joulin et al., 2017) which uses TextImager (Hemati et al., 2016) to preprocess texts (Step 1 in Figure 7). That is, topics are



**TABLE 2 |** Wikipedia language editions, which were analyzed thematically: “Topics” is the number of DDC-based topic classes trained for the corresponding language, “Train” is the number of training examples and “F-score” the harmonic mean of precision and recall of the corresponding test.

	Code	Language	Topics	Train	F-score
1	ar	Arabic	96	14,991	0.80
2	bs	Bosnian	87	5,599	0.83
3	ceb	Cebuano	68	2,069	0.87
4	ckb	Sorani	82	3,090	0.75
5	da	Danish	94	12,761	0.78
6	de	German	98	21,933	0.87
7	el	Greek	92	8,395	0.76
8	en	English	97	19,772	0.85
9	es	Spanish	95	16,951	0.85
10	fr	French	94	17,313	0.86
11	he	Hebrew	94	12,262	0.83
12	hi	Hindi	85	7,271	0.74
13	hu	Hungarian	91	10,854	0.85
14	id	Indonesian	93	11,265	0.81
15	it	Italian	94	15,894	0.85
16	ja	Japanese	93	16,390	0.84
17	ko	Korean	92	13,557	0.76
18	lv	Latvian	89	7,572	0.83
19	mk	Macedonian	88	5,750	0.76
20	ml	Malayalam	84	5,465	0.85
21	mr	Marathi	83	3,061	0.85
22	nl	Dutch	97	15,507	0.85
23	pl	Polish	96	16,356	0.84
24	pt	Portuguese	93	15,663	0.84
25	ro	Romanian	93	10,690	0.77
26	ru	Russian	97	17,302	0.85
27	sh	Serbo-cro.	94	9,536	0.82
28	si	Sinhala	81	2,521	0.83
29	simple	Simple English	93	10,882	0.83
30	sr	Serbian	91	10,607	0.82
31	sv	Swedish	95	16,458	0.80
32	te	Telugu	80	3,916	0.84
33	vi	Vietnamese	88	10,279	0.83
34	war	Waray	85	4,282	0.75
35	zh	Chinese	92	15,595	0.84

identified as elements of the 2nd level of the *Dewey Decimal Classification* (DDC), a topic model widely used in the field of libraries. To this end, each article is mapped onto a 98 dimensional topic vector, whose membership values encode the degree to which text2ddc estimates that the article deals with the topic corresponding to the respective dimension (cf. Mehler et al., 2019; Uslu et al., 2019). In this way, we identify text pairs that tend to describe the same subject of the same area in different languages under the perspective of similar topic distributions.

Since we do not have *Part of Speech* (POS) taggers for all target languages, we basically pursued a word-form-related

approach to train text2ddc: we generated training and test corpora by retrieving information from Wikidata, Wikipedia and the Integrated Authority File of the German National Library. Since Wikipedia is offered for a variety of languages, such corpora can be created for many languages. We optimized text2ddc with regard to selected linguistic features as a result of various pre-processing steps, such as lemmatization and disambiguation. In the last column of **Table 2** we show the *F*-values obtained for the corresponding tests. The highest *F*-value (87%) is achieved for German (where we also explored POS data), the lowest for Hindi (74%). Although this is a wider range of values, it is currently the only way to compare the content of texts in different languages in terms of the way their subjects are treated. And although text2ddc was trained for a larger number of languages<sup>7</sup>, we concentrated on those for which it achieves an *F*-value of at least 74%.

3. *SV3: Quantitative Text Structure (QTS)*: according to this SV, pairs of aligned articles are the more similar the more their QTSs resemble each other, and again, the more such pairs of aligned articles of two INs resemble each other in this sense, the more similar they are. To get comparable vector representations of the QTS of articles, we use a subset of 17 dimensions of quantitative linguistics as evaluated by Konca et al. (2020): *adjusted modulus, alpha, Gini coefficient, h-point, entropy, hapax legomena percentage, curve length, lambda, vocabulary richness, repeat rate, relative repeat rate, thematic concentration, secondary thematic concentration, type-token-ratio, unique trigrams, average sentence length, and number of difficult words*. Two text characteristics from Konca et al. (2020), which require POS tagging, are excluded from SV3, since POS-tagging tools were not available for all languages considered here. We additionally compute autocorrelations (lag 1–10) of consecutive sentence-related association probabilities with BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019). BERT is a language model based on a bidirectional transformer (Vaswani et al., 2017), which uses encoder-decoder attention (Bahdanau et al., 2015) and self-attention (Cheng et al., 2016) mechanisms to calculate token representations conditioned on both left and right context. In total, we compute  $k_3 = 27$  characteristics to vectorize the QTS of any article of the 35 languages considered here.

While SV1-3 are article content-related, graph similarity measures consider intertextual structures (SV4). We can calculate SV4 independently of SV1-3, or so that the contributions of SV1-3 are embedded at the node level of SV4. To elaborate the first variant, we consider four measures (thereby implementing Step 6 of **Figure 7**):

1. As a baseline, we compute the *Graph Edit distance-based Similarity* (GES) using the graph edit distance for labeled directed graphs (cf. Mehler et al., 2019). In this way we arrive

<sup>7</sup>See <https://textimager.hucompute.org/DDC/>

at a measure that calculates graph similarities as a function of overlaps of node and arc sets, respectively.

2. *Edge-Jaccard-Similarity (EJS)*: As a second baseline, we compute the arc set-based Jaccard similarity:

$$\text{EJS}(I_i^x, I_j^x) = \frac{|E_i^x \cap E_j^x|}{|E_i^x \cup E_j^x|} \in [0, 1]$$

$\text{EJS}(I_i^x, I_j^x) = 1$  if and only if  $I_i^x$  and  $I_j^x$  have identical non-empty arc sets (where the Jaccard similarity of empty sets is defined to be zero). EJS decreases with increasing symmetric difference in the arc sets.

3. *DeltaCon with Personalized PageRank*: Koutra et al. (2016) propose DeltaCon, a family of similarity measures for node-aligned graphs. It is based on the distance function

$$d(I_i^x, I_j^x) = \sqrt{\sum_{v \in V_i^x \cup V_j^x} \sum_{w \in V_i^x \cup V_j^x} \left( \sqrt{s_i(v, w)} - \sqrt{s_j(v, w)} \right)^2}$$

where the columns of  $s_i, s_j \in \mathbb{R}^{|V_i^x \cup V_j^x| \times |V_i^x \cup V_j^x|}$  store *affinity* values between vertices  $v, w \in V_i^x \cup V_j^x$  in the union graphs  $I_i^x = (V_i^x \cup V_j^x, E_i^x)$  and  $I_j^x = (V_i^x \cup V_j^x, E_j^x)$ . As  $d(I_i^x, I_j^x)$  is a metric, Koutra et al. (2016) propose to define the DeltaCon similarity as  $\text{sim}(I_i^x, I_j^x) = \frac{1}{1+d(I_i^x, I_j^x)}$ . DeltaCon can be parameterized by various node similarity measures, of which we consider a variant that is semantically meaningful in our context: Personalized PageRank (for vertices  $v, w \in V(I_i^x)$ ) measures the probability of being in vertex  $w$  in the stable distribution of a random walk on  $I_i^x$  that has a probability of  $1 - \alpha$  to be reset to  $v$  in each step. This models the probability of landing at an article  $w$  when starting at article  $v$ , randomly following links and going back to article  $v$  when one is “lost.” It thus models, considering all articles in turn as starting articles, a crowd of Wikipedia users following hyperlinks to navigate the information landscape. We proceed similarly to compute  $s_j$  and set  $\alpha = 0.85$ .

4. As both EJS and DeltaCon are highly sensitive to non-overlapping node sets, we restrict the input INs  $I_i^x = (V_i^x, A_i^x)$  and  $I_j^x = (V_j^x, A_j^x)$  for these two similarity measures to the subgraphs induced by the intersection of their vertex sets, that is,  $I_i^x = I_i^x[V_i^x \cap V_j^x]$  and  $I_j^x = I_j^x[V_i^x \cap V_j^x]$ . This approach yields high similarities if the subgraphs on the aligned nodes are similar and disregards dissimilarity induced by unaligned nodes in both graphs. Hence, the latter two measures can be seen as “optimistic” and may give high similarities even if the overlap of two vertex sets is rather small. To provide an alternative to this view, we compute the arc set-based cosine graph similarity of Mehler et al. (2019). For this purpose we start from the following “axioms” concerning the similarity of INs:

A1 The higher the number of shared subject instances, the more similar the INs.

- A2 The lower the proportion of shared subject instances in the total number of instances of the underlying subject area, the less similar the INs.
- A3 The higher the number of shared paths, the more similar the INs.

A1 also concerns isolated vertices: two INs can be identical even if all their vertices are isolated. A2 damps this depending on the total number of shareable subjects: the smaller the orders of the INs in relation to this number, the less similar they are. A3 prefers pairs of INs that share many edges, but only in cases where equal paths start from aligned subjects. A3 essentially states that two networks are the more similar, the more similar they look like from the perspective of the more aligned nodes. A1 and A3 are measured with the apparatus of Mehler et al. (2019); to satisfy A2, we damp the resulting cosine similarity by the quotient of shared nodes and the total number of shareable items; we can do the same regarding the size of the networks, but refrain from this “pessimistic” variant in the present paper. We will refer to this variant by *Cosine Graph Similarity* (CGS): it rates pairs of networks as similar that link large proportions of candidate Wikidata items in a similar way.

We calculate four graph similarity measures ranging from set-based to spherical measures, where EJS and DeltaCon weight similarities more optimistically and CGS more pessimistically.

#### 4.2.1. Assessing Observed Similarities

Any similarity found between INs on the same subject area has to be evaluated according to how far it is higher than what is randomly expected and lower than what is ideally expected (see Step 5 of **Figure 7**). For this purpose, we consider the following bounds:

1. To get a *lower bound* we compute the similarities of random counterparts of INs: we consider Erdős-Rényi (Erdős and Rényi, 1959) graphs  $R(I_i^x)$  chosen uniformly at random from all graphs of the same order and size as  $I_i^x$ . This randomization concerns SV4 without node-related similarities, assuming a bijection between the nodes of  $R(I_i^x)$  and  $I_i^x$  based on their Wikidata items. Randomizations are performed 100 times per IN; similarity values are averaged accordingly.
2. As an *upper bound* we consider the maximum similarity of different language INs observed over all subject areas in our corpus of INs. Since we display these values as a function of the minimum  $\min(|V_i^x|, |V_j^x|)$ , we get an estimate of the maximum similarity observed for this minimum among all similarities observed in our experiment. Though the theoretical maximum is always 1 for identical graphs, this maximum is unlikely to be observable in practice even under the condition of comparatively similar INs. Therefore, by our re-estimation we achieve a more realistic upper bound that is actually observable.
3. To obtain the lower bound for SV1-3, the similarity of the non-aligned, randomly chosen articles from each pair of INs were calculated (using Formula 9). The articles in each pair were drawn from two independent random permutations,

whereupon the number of pairs (of articles) was kept the same as in the aligned case. To reduce the impact of possible outliers, the results of each calculation were averaged over 100 independent runs.

## 5. EXPERIMENT

Applying the methods of section 4 along the procedure of **Figure 7** to the Wikipedia editions of **Table 2** to generate language-specific INs for the subject areas of **Figure 8** produces the results of **Table 3**: based on the number of INs per subject area listed in **Figure 8** (whereby languages whose INs according to Formula 6 correspond to the empty graph for the given subject area are not listed), we arrive at 103,299 graph comparisons using our 7 similarity measures, three of which are vertex (SV1-3) and four hypertext structure-oriented<sup>8</sup>. Since we randomize each IN 100 times to perform the same procedure for each random setting, the final number of graph comparisons performed equals 10322900.

We start our experiment with three subject areas, whose analyses span the similarity spectrum observed by us: *chemical element* (OECD class *Chemical sciences*), *disease* (class *Health sciences*), and *language* (*Languages and Literature*). Let us first consider the set-based similarity measure GES. In the first column of Row 4 of **Table 3** we see the corresponding heatmap of the 35 languages' INs from **Table 2**: the greener, the more similar the INs of the respective language pair for this subject area. Apparently, we find that *chemical element* is treated very uniformly across languages in terms of hypertext structure. This subject area is the maximum of what we observed regarding this uniformity: it best approximates the ideal under the hypothesis that different language Wikipedias report uniformly on the same topic—note that the rows and columns of the heatmaps are ordered according to the orders  $|V_i^x|$  of the INs  $I_i^x$ . This is confirmed by the curve displayed below the heatmap: it shows the similarity values listed in the heatmap as a function of  $\min(|V_i^x|, |V_j^x|)$  (minimum of the orders of the input graphs  $I_i^x, I_j^x$ ). In this way we see the influence of graph order: similarity values of comparisons with smaller graphs move to the left, those where both graphs are larger move to the right. In fact, we see for this subject area that most comparisons concern (equally) large graphs achieving high GES-values. This is contrasted by the last column of Row 4 which depicts the heatmap of subject area *language*: now, small graphs dominate the distribution with high similarity values, while pairs of significantly larger INs tend to have much lower values. This example demonstrates a specialization of a few language editions on a broad representation of this subject area, while the majority of editions tend to underrepresent it. All in all, we arrive at a zigzag curve of similarity values, which does not indicate a clear trend. In the mid of the range of these two examples in Row 4, we find the subject area *disease*: the heatmap now suggests that the language pairs are rather dissimilar from the perspective of

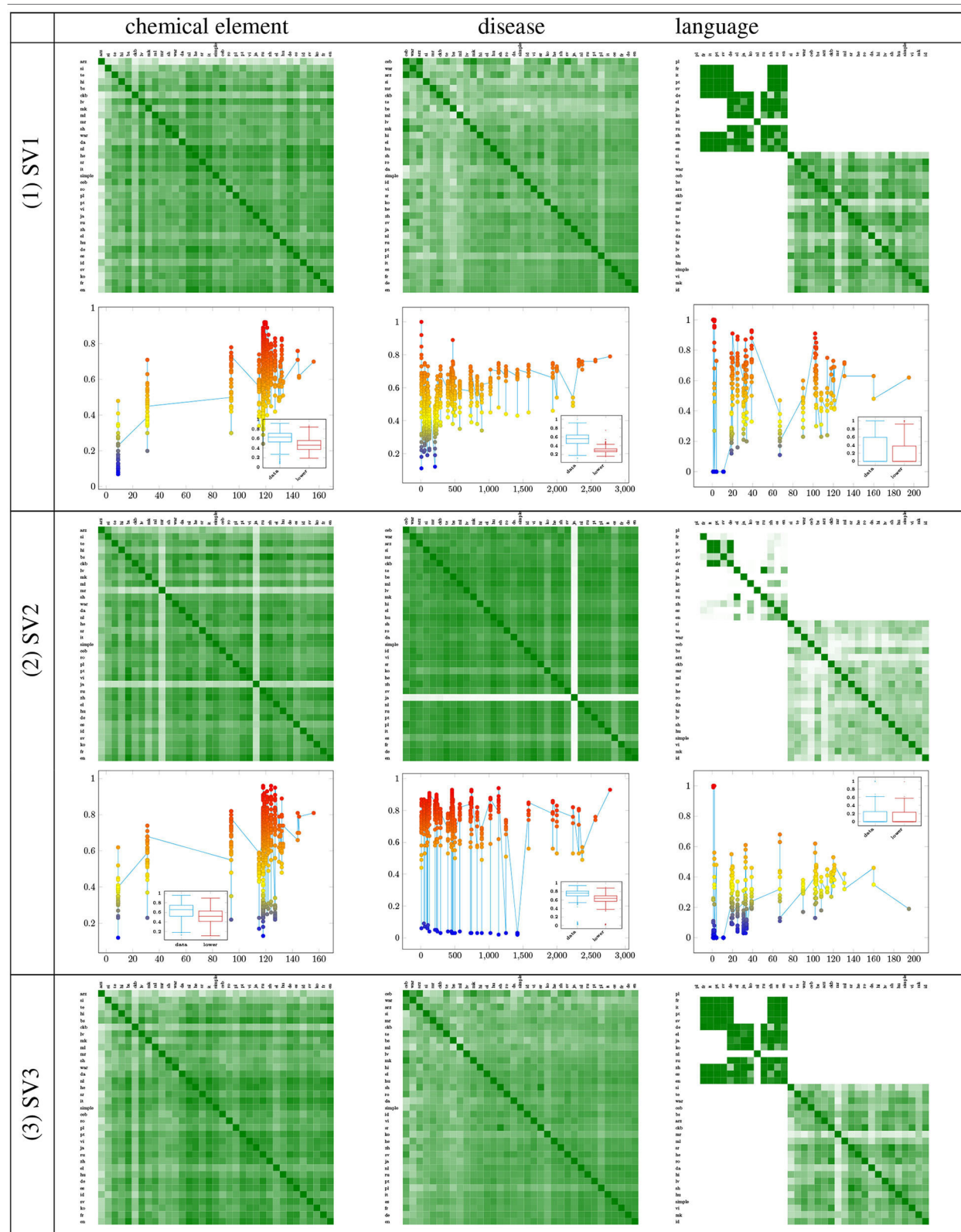
this subject area. This is particularly evident from the similarity curve, which shows an almost constant trend of small values with a small increase toward larger graphs. Let us now look at the same examples from the perspective of DeltaCon (Row 5); the situation remains essentially the same: *chemical element* induces a more homogeneous mass of simultaneously larger and similar graphs, *language* exhibits again a zigzag pattern, and *disease* reveals slightest similarities regardless of the INs' orders. Obviously, the latter subject area seems to contradict the ideal of homogeneous, equally informative Wikipedias very strongly. This is also confirmed by the randomizations according to section 4.2.1, which are presented as boxplots embedded in the similarity curves: blue is the boxplot of the similarity values shown in the curve itself and red is the boxplot of the similarity values of the randomizations. Interestingly, in the case of *language* randomized similarities are even higher than their empirical counterparts. This can be explained by the topology of ER graphs, which tend to have short diameters making shorter paths more probable. Anyhow, the situation is almost the same as in the case of Row 4: observed similarities seem to deviate only slightly from their randomized counterparts; similarities between these INs resemble those of corresponding random graphs. Now we look at CGS (Row 6), which evaluates graph similarities more “pessimistically” than DeltaCon. This is confirmed by the heatmaps: in the case of *chemical element*, the high similarity values change in favor of a “chessboard” view, while the binary regime disappears in the case of *language*, whose similarity progression now resembles that for *disease*. In the latter two cases, low similarities dominate, with the tendency of higher similarities only for pairs of larger graphs: based on CGS, pairs of small graphs are highly dissimilar (damping effect). Interestingly, the graph similarities as a function of the minimum of the orders of the input graphs no longer predominantly show a zigzag pattern as in the case of GES and DeltaCon: the impression of gradual transitions now prevails (2nd line of Row 6).

The latter assessments are confirmed by the set of 25 subject areas. **Table 4** shows the DeltaCon-based boxplots displayed as a curve: medians are represented by straight lines while the value ranges between the 25th and 75th percentile are colored accordingly; the blue curve represents observed similarities, the red one their random counterparts (“lower bound”) and the orange one the corresponding upper bounds (see section 4.2.1). **Table 4**, in which the subject areas are arranged according to decreasing median similarities of their INs, shows a clear trend: in almost all cases, observed similarities are below both the similarities of randomized INs and the upper bounds. That is, observed similarities are far away from ideally equally informative Wikipedias, which would cover the corresponding subject areas uniformly for all languages. Even more: as far as DeltaCon considers transitive dependencies of nodes along the same paths, it turns out that the INs' random counterparts even tend to have a greater hypertext-structural similarity—as explained above, this is partly due to their small diameters. Obviously, randomness makes networks seem more similar than if one follows existing walks in the real networks simultaneously. This even applies to the maximum values measured (upper bound). At the same time, we observe a broad spectrum of

<sup>8</sup>If all INs were non-empty, this number would equal  $\#_{\text{similarities}} \cdot \#_{\text{topics}} \cdot \#_{\text{comparisons/language}} = 7 \cdot 25 \cdot \left(\binom{35}{2} + 35\right) = 110,250$ .



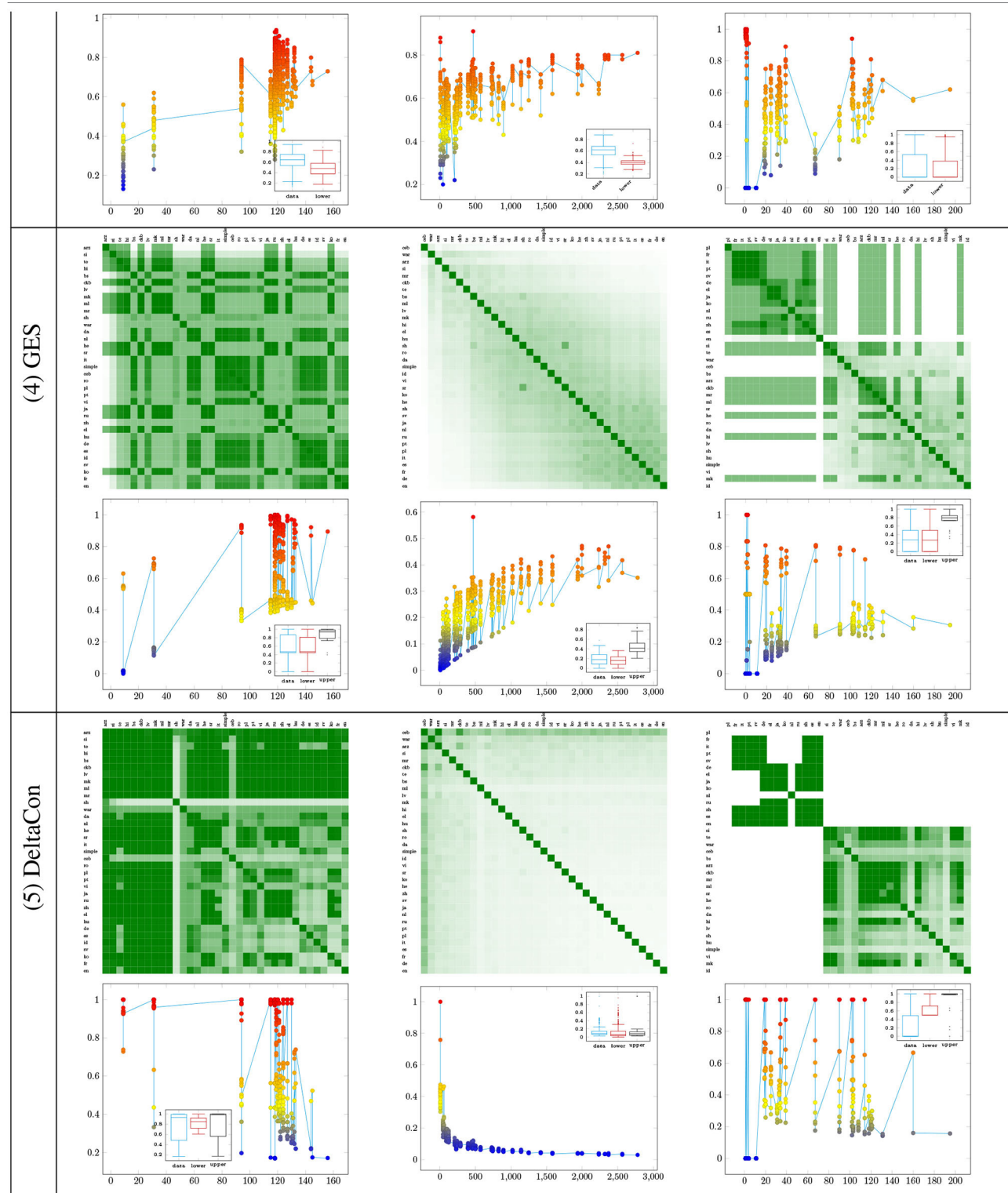
**TABLE 3 |** Six measures (row) for computing the similarities of INs by example of three subject areas (column): rows and columns of the heatmaps correspond to languages (INs); curves below the heatmaps display similarity values as a function of the minimum order of the input graphs.



(Continued)

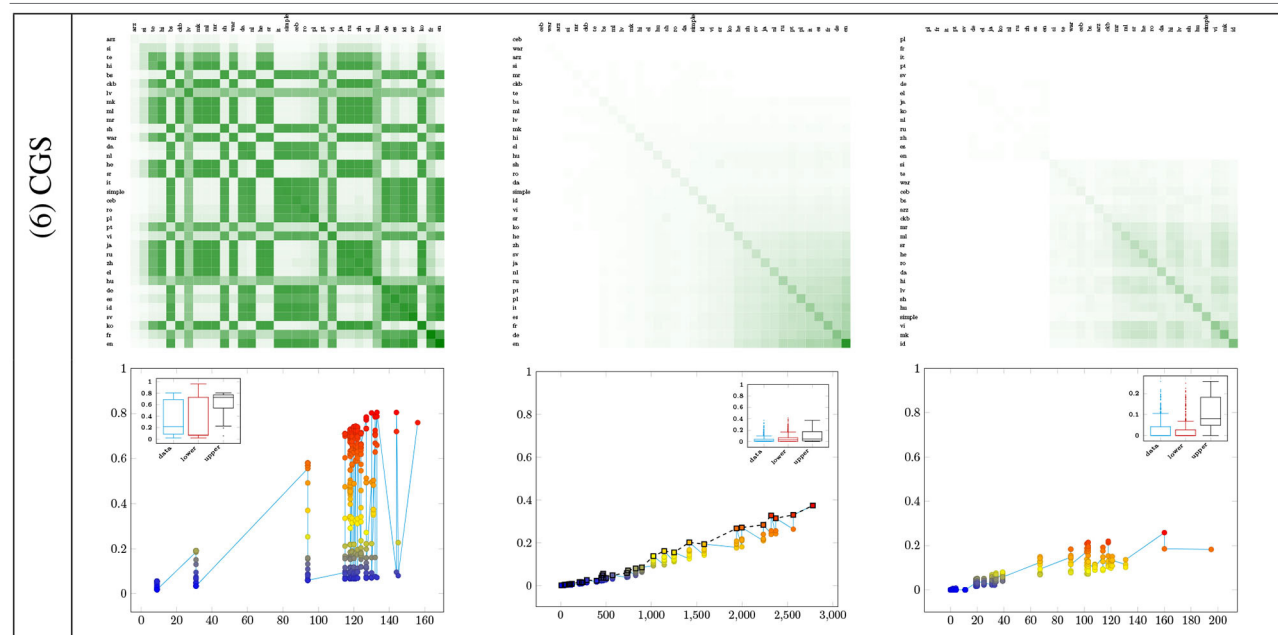


TABLE 3 | Continued



(Continued)

TABLE 3 | Continued



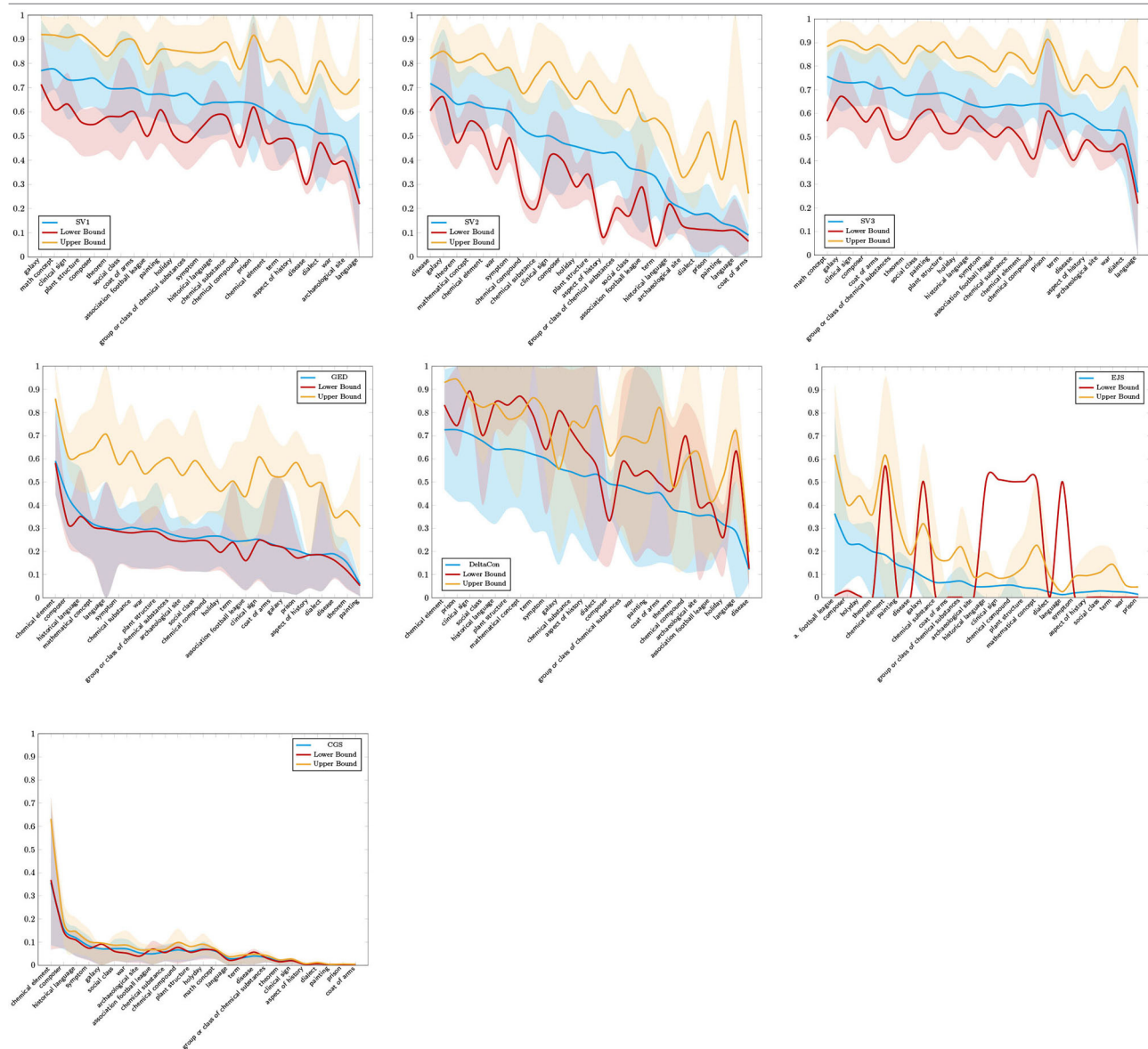
similarity values ranging from a minimum of about 20% (*disease*) to a maximum of about 75% (*chemical element*). Therefore it matters very much in which subject area one reads Wikipedia—our approach shows this in a fine-grained way for educationally relevant topics using a three-level topic model. The picture becomes even clearer when we look at the distribution of GES-based values in **Table 4**: observed similarities are now hardly distinguished from their random counterparts and far away from their upper bounds. A more chaotic picture, somewhat reminiscent of DeltaCon, results from the Jaccard-based variant EJS: observed similarities tend to be even smaller than their random counterparts, but not always. In **Table 4**, the least chaotic pattern is produced by CGS: we observe a monotonically decreasing function with small differences for minimum and maximum values, where the similarity values fall below those of the other measures.

From this lesson we learn that subject areas are rather unevenly distributed across Wikipedia's language editions, whether one measures their similarities using simpler (GES, EJS) or more complex measures (DeltaCon, CGS). But what picture do we get from looking at the content-related similarities of articles? In **Table 3** we start with SV1 (Row 1, LDS). Again, the picture is tripartite: *chemical element* marks the upper limit of observed similarities (the larger the INs, the higher their similarity), *language* the lower limit, and *disease* a middle case. *language* also shows that LDS-based similarities break down into two groups: for larger and for smaller graphs, while members of both groups show no or little similarities between each other. In any event, observed similarities tend to exceed their random counterparts (embedded blue (observed) and red (randomized) boxplots). This tripartition is basically confirmed by SV2 (thematic similarities measured by text2ddc)

(Row 2 in **Table 3**). But now *disease* turns out as a subject area whose article graphs (INs) consist of thematically more homogeneous articles—more or less irrespective of the number of articles considered. In any event, thematic similarity is again concentrated more or less on large graphs in the case of *chemical element*. The third case (Row 3 in **Table 3**) concerns quantitative text structures. Here too, the general picture is confirmed: concentration on high values for large graphs in the case of *chemical element*, more evenly distributed, but high values in the case of *disease* and a bipolar picture in the case of *language*. As before, randomized counterparts are exceeded.

The rather exceptional case of SV2 (thematic similarities) is confirmed by the overall view for the 25 subject areas in **Table 3**: observed similarities along SV2 are below the values for LDS (SV1) and QTS (SV3), while the latter are more evenly distributed across the subject areas, possibly reflecting a law-like behavior as described by quantitative linguistics (Köhler et al., 2005). In any case, with few exceptions, the observed values are again within the interval spanned by their randomized variants and upper bounds. This points to similarity distributions far away from ideally homogeneously structured Wikipedia articles, which in addition would manifest almost the same topic distributions: actually, they do not. Apparently hypertextual dissimilarity is parallel to textual dissimilarity: *what is rather dissimilar in terms of intertextual structure, tends to be dissimilar in terms of intratextual structure as well*. However, we also observe the case of examples, such as *chemical element*, which has both high similarity values in terms of DeltaCon and SV2 (thematic similarity): this is, so to speak, the maximum of simultaneous inter- and intratextual similarity observed here. In any event, our study combines intra- and intertextual measurements where the former are based on three views, regarding LDS, QTS and

**TABLE 4 |** Distributions of similarity values of INs calculated by the similarity measures of section 4.2 and displayed per subject area. From left to right, top-down: SV1-3, GES, DeltaCon, EJS, CGS.



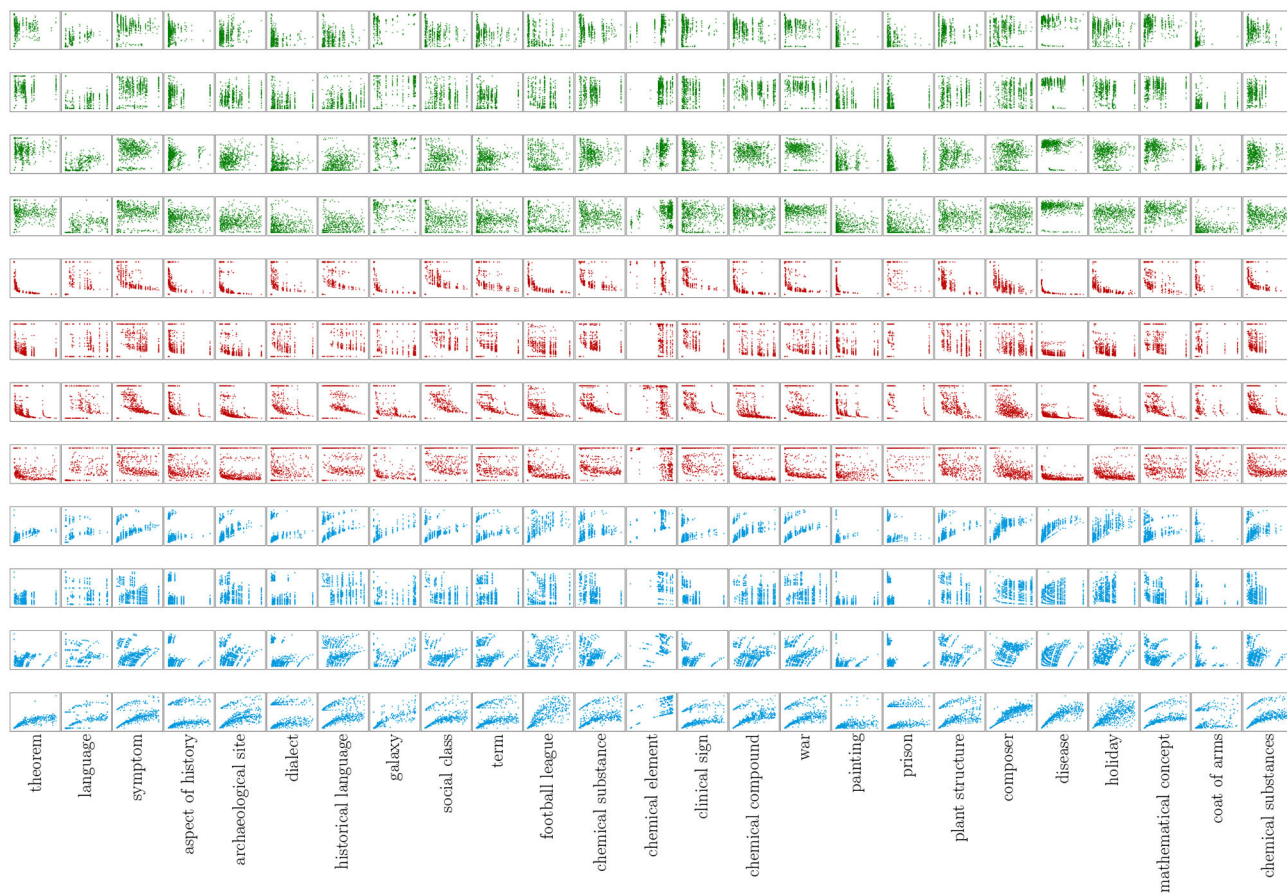
thematic text structures. In this way, we obtain a more precise, broader view of article content than has been possible with the methods of related research.

We now turn to correlation analysis and ask about the dependence of our similarity analysis on the size of the graphs involved, based on the hypothesis that the size (and thus indirectly the degree of development or activity) of a subject area explains our results. For this purpose, we calculate Spearman’s rank correlation with respect to four data series, each of which is generated for two similarity views or three similarity measures (SV2, DeltaCon and GES).

The data series contrast measured similarities with the orders  $|V_i^x|, |V_j^x|$  of the networks  $I_i^x, I_j^x$  involved.

That is, we ask whether the rank of a pair of networks [the higher their similarity  $\sigma(I_i^x, I_j^x)$ , the higher the rank] correlates with its rank according to size  $s(I_i^x, I_j^x)$  (the larger the networks, the higher the rank), distinguishing four alternatives:  $s(I_i^x, I_j^x) = \min(|V_i^x|, |V_j^x|)$ ,  $s(I_i^x, I_j^x) = \max(|V_i^x|, |V_j^x|)$ ,  $s(I_i^x, I_j^x) = |V_i^x| + |V_j^x|$ , and  $s(I_i^x, I_j^x) = \min(|V_i^x|, |V_j^x|) / \max(|V_i^x|, |V_j^x|)$ . The distributions are shown in **Figure 9**, the corresponding correlations in **Figure 10** (which additionally plots the values for CGS). For SV2, we observe a very strong effect regarding *language* (“visually” confirmed by **Table 3**) (note that the curves are ordered according to variant  $\min(|V_i^x|, |V_j^x|)$ , sorted in descending order); in most other cases correlations are





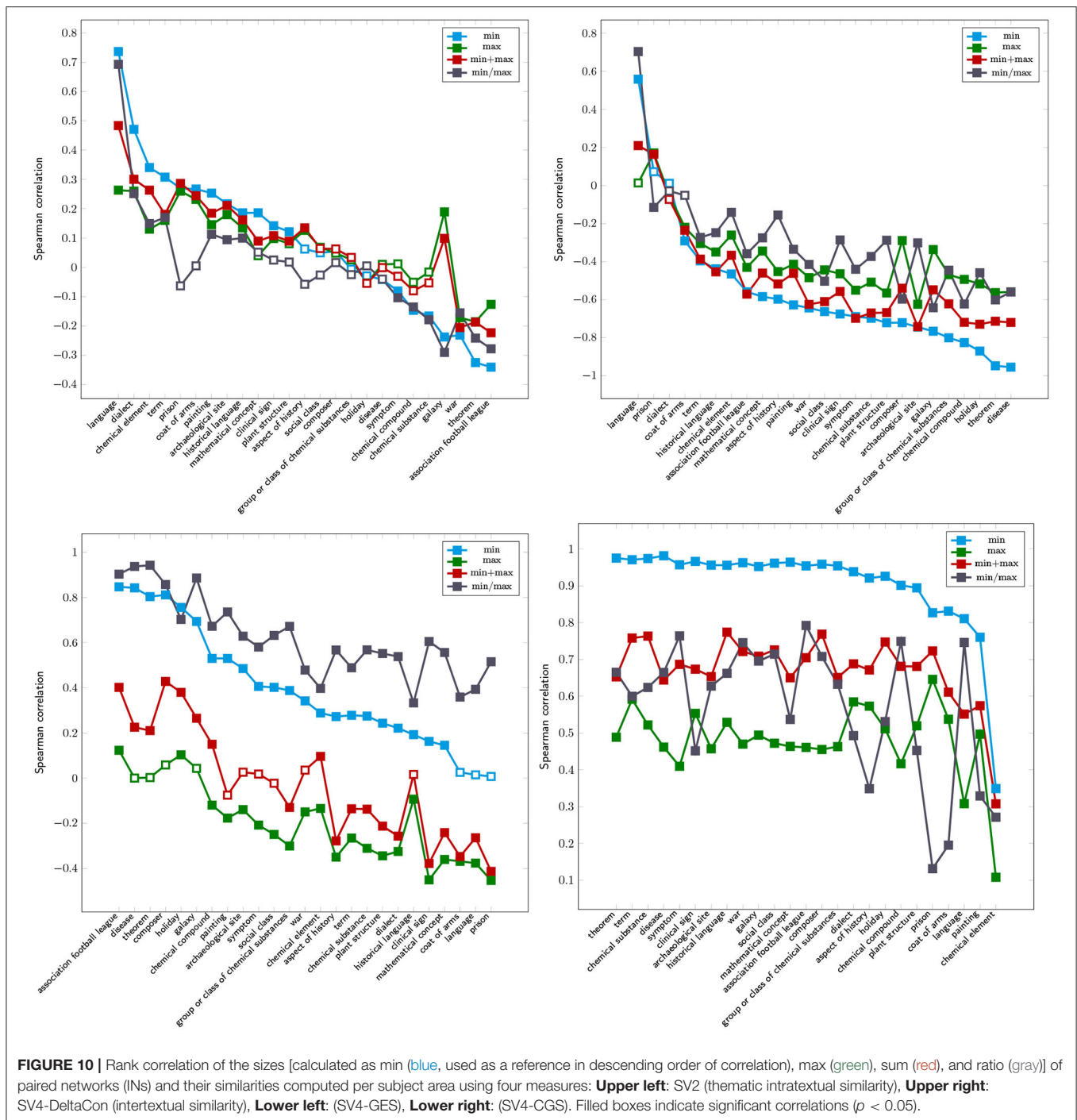
**FIGURE 9 |** Similarities of languages as a function of the size (number of vertices) of the networks (INs) involved. Four alternatives of calculating the size  $s(I_i^x, I_j^x)$  of pairs  $(I_i^x, I_j^x)$  of network: 1st row:  $s(I_i^x, I_j^x) = \min(|V_i^x|, |V_j^x|)$ , 2nd row:  $s(I_i^x, I_j^x) = \max(|V_i^x|, |V_j^x|)$ , 3rd row:  $s(I_i^x, I_j^x) = |V_i^x| + |V_j^x|$  and 4th row:  $s(I_i^x, I_j^x) = \min(|V_i^x|, |V_j^x|) / \max(|V_i^x|, |V_j^x|)$ . Green: similarity view SV2 (thematic intratextual similarity), red: similarity view SV4-DeltaCon (intertextual similarity), blue: SV4-GES.

rather low (whether positive or negative). From this picture we conclude that observed thematic similarities of articles in different languages on the same subject cannot be attributed to the sizes of the networks involved. Remarkably, for SV2 we observe that the data series mostly coincide with variant  $\min(|V_i^x|, |V_j^x|)$  regarding the correlations' order. This more or less also applies to DeltaCon in **Figure 10**. However, we now mostly observe higher negative correlations. Apparently, ranks in terms of structural similarity correlate negatively with size-related ranks. This means that if the INs are small, their similarities are likely to be large and vice versa. In the case of GES (**Figure 10**) we find this assessment more or less reversed: the correlations (blue) are almost all significantly higher than 0.2; that is, the larger the networks involved, the higher the similarity. From this perspective we can conclude that either both measures (GES and DeltaCon) contradict each other or (the more likely interpretation) they measure orthogonal aspects of graph similarity (the one set intersection-, the other walk-based). An exception is again CGS, which shows stable, high rank correlations not only for small graphs, *for which it computes very*

*high values of dissimilarity*, but also depends more on size than SV2 and DeltaCon—as motivated by the definition of CGS, size is a better predictor of it: the smaller the INs, the less their similarity and vice versa. In any case, the picture we get from this analysis is ambiguous, so that we hesitate to conclude that hypertextual similarity is reliably correlated with the orders of the networks involved: *intratextually, the similarity of INs does not depend on size and intertextually it does not show a clear trend.*

Next, we ask about the status of subject areas as a function of their analyses along SV1-4. Since each similarity measure induces a ranking of areas based on the average of the similarities observed for the language pairs (see **Table 4**), we can ask whether the rankings induced by different measures correlate or not. Lower rank correlations would then indicate unsystematic similarity relations in the sense that intra- and intertextual similarities do not point in the same direction. Lower rank correlations for either intra- or intertextual measures, in turn, would point to contradictory results. All in all, such findings would indicate that the INs under consideration exhibit incoherent similarities—contrary to the assumption of their





uniform similarity along intra- and intertextual dimensions. This is essentially what we find in **Table 5**: rank correlations are mostly low and not significant. One exception is the negative correlation of EJS and DeltaCon, which, apparently, measure different things. Another exception are the few examples of high positive correlations, such as those of GES and CGS. This correlation analysis shows that, with few exceptions, the similarity-based ordering of subject areas along one similarity

dimension (whether intra- or intertextual) does not allow us to infer their order along another dimension.

Next, we consider language networks whose edges correspond to the similarity values of the underlying language pairs (related network analyses have been conducted by Miquel-Ribé and Laniado, 2016; Samoilenko et al., 2016). We want to know in which subject areas which language clusters arise and whether different languages are center-forming in different subject areas

**TABLE 5 |** Rank correlations of the orderings of subject areas according to the average similarities obtained for the corresponding INs of language pairs where the similarities are computed by means of seven different similarity measures.

	SV1	SV2	SV3	GES	DeltaCON	EJS	CGS
SV1	1.00	0.27	0.95	-0.02	0.30	0.36	0.07
SV2	0.27	1.00	0.26	0.13	0.07	0.23	0.40
SV3	0.95	0.26	1.00	0.05	0.28	0.40	0.06
GES	-0.02	0.13	0.05	1.00	0.28	-0.06	0.72
DeltaCON	0.30	0.07	0.28	0.28	1.00	-0.40	0.18
EJS	0.36	0.23	0.40	-0.06	-0.40	1.00	0.17
CGS	0.07	0.40	0.06	0.72	0.18	0.17	1.00

	SV1	SV2	SV3	GES	DeltaCON	EJS	CGS
SV1	0.00	0.19	0.00	0.93	0.15	0.07	0.74
SV2	0.19	0.00	0.22	0.55	0.74	0.28	0.05
SV3	0.00	0.22	0.00	0.82	0.17	0.05	0.78
GES	0.93	0.55	0.82	0.00	0.18	0.78	0.00
DeltaCON	0.15	0.74	0.17	0.18	0.00	0.05	0.38
EJS	0.07	0.28	0.05	0.78	0.05	0.00	0.41
CGS	0.74	0.05	0.78	0.00	0.38	0.41	0.00

(Left) Heatmap of correlation values, (Right) p-Values.

(see question Q3). Since the underlying similarity matrices (see **Table 3**) induce complete graphs, we filter out all edges whose similarity values are below the average similarity measured across all subject areas for the respective measure. **Figure 11** illustrates the result by the example of three subject areas (*composer*, *painting*, and *language*) and two similarity views (SV2 and SV4-DeltaCon). The impression that we get already by this selection is rather confusing: the graphs look very different, from the perspective of the similarity measure and from the perspective of the subject area. A clear trend is not discernible (although English Wikipedia is usually prominently positioned). From this brief network analysis, we conclude that the different languages may have different degrees of salience depending on the subject area, while there is no single language that dominates in all these cases.

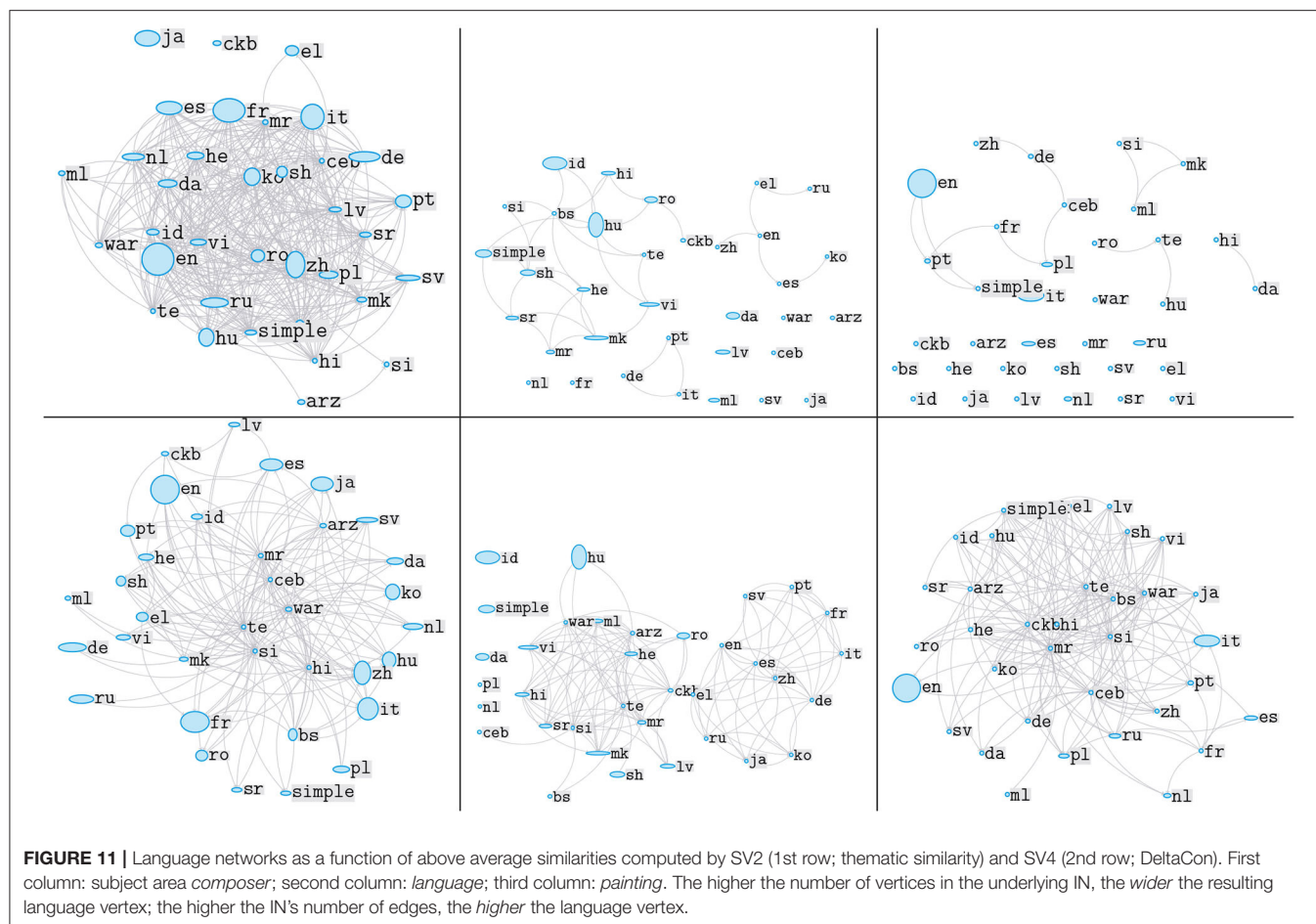
To conclude this analysis, we ask about the similarities of the positions of all 35 languages in the networks derived from the heatmaps of **Table 3** and weight them by the ratio of these networks' strength centralities to the respective maximum. That is, we take the completely connected language similarity network for each of the 25 topics, in which each edge is weighted according to the similarity of the associated languages in the sense of the underlying similarity measure. Regardless of the actual structure of the INs, in the ideal state of equal INs for the different languages about the same topic, it is to be expected that the edges in the language networks are weighted by 1. Under this regime, the distance correlations of the strength centralities of languages should be maximal when being compared for each pair of topic-related language networks<sup>9</sup>: the languages then occupy identical network positions—independent of the topics. However, if all INs are maximally dissimilar, we likewise get maximum distance correlations of the languages according to their network positions. Thus, to differentiate between these two cases, we weight observed distance correlations by the quotient of observed network strength and maximum possible strength. For pairs of the fully connected language networks considered here,

this maximum is  $35 * 34$ , assuming a maximum edge weight of 1. To avoid unrealistically low weights, we multiply this maximum by the actually observed maximum value per similarity measure. The resulting heatmaps are shown in **Figure 12**: If INs on the same topic tend to be similar, both heatmaps are expected to be saturated, the weighted and the unweighted. Conversely, if these INs tend to be dissimilar, only the unweighted map is likely to be saturated. In **Figure 12** we show the corresponding results for the optimistic variant of similarity measurement (DeltaCon) and its pessimistic counterpart (CGS). Obviously, with CGS we obtain higher distance correlations of the strength centralities of the languages in the 25 language networks than with DeltaCon, but very small weighted variants of these correlations, which indicate that the INs are extremely far from ideal similarities. The only exception is *chemical element*: this topic is described similarly in the different languages and more close to the ideal scope than any other topic. In the case of DeltaCon the weighted distance correlations are much higher, but still far from ideal. In **Figure 13** we add the distance correlations generated by GES and SV2: for GES the unweighted correlations are similar to those of DeltaCon, for SV2 those of CGS. The situation is the opposite in the case of the weighted correlations. The general tendency is that the languages tend to have similar network positions, but at the price of lower similarities of the INs.

## 5.1. Discussion

Given the scope and growth dynamics of Wikipedia, it is not surprising that a reader who reads it in a particular language expects to be sufficiently broadly and deeply informed about the subject of his or her information search, provided it is one of the larger Wikipedias. We have shown that such an assumption about Wikipedia as a central building block of the IL does not apply—at least in the context of the subject areas examined here. Their different language versions differ so much in their treatment of the same subject area that it is necessary to know which area in which language someone is consulting if one wants to know how much the part of the IL he or she is traversing is biased. It may be the case that a reader's consultation of Wikipedia is accompanied by the assumption, that it is an open, dynamically

<sup>9</sup>For the notion of vertex strength (i.e., weighted degree values) see Barrat et al. (2004), for computing network centralities based on vertex indices see Feldman and Sanger (2007).



growing resource that largely covers relevant fields of knowledge, where one likely finds what one is looking for, and vice versa, that what is irrelevant is excluded. One might even assume that if a language version of Wikipedia is only sufficiently large, it will probably show this pattern of coverage—regardless of the underlying language. We have shown that at the current stage of development, such assumptions do not apply.

Our analysis has shown that subject areas covered by Wikipedia differ from language to language in a way that is hardly predictable by the size of the networks involved. Consequently, with regard to question Q1, we have to state that the similarities between the languages vary from subject area to subject area. It is therefore necessary to define the thematic reference (subject area) in order to say something about comparable languages. This may seem obvious, but it shows that Wikipedia's language editions are designed differently for different research fields and subject areas. With few exceptions (e.g., *chemical element*), we find that inequalities prevail within the same field. This makes it difficult to say which field is the more evenly distributed, language-independent one, which ensures that students can expect nearly equal information coverage regardless of language. Thus, in relation to question Q2, we conclude that the choice of the subject area has a major influence on the similarity assessment. And because of this influence, we do not discover a lingua franca

which, due to its size and coverage would serve as a kind of reference for the similarity relations of the different languages, so that although they may be dissimilar among themselves, they would be predominantly similar with respect to this central language. As things stand, the data do not support the attribution of this central role to English Wikipedia—at least from the point of view of the fields of knowledge and subject areas considered here (question Q3). In other words, knowing that Wikipedia is small in your first language and therefore probably does not cover the subject area of your task, it is hardly a way out of this dilemma to recommend reading English Wikipedia instead: it would be better to read at least both. However, we also see that depending on the subject area, other languages could play the role of primary sources (e.g., Hungarian in the case of subject area *language*—see Figure 11—or French in the case of subject area *composer*)—of course, this presupposes that one has the language skills for them.

Samoilenko et al. (2016) point out the influence of multiple dimensions on the commonalities, similarities and differences of Wikipedia's language editions, including language, culture, and geographical proximity. Apparently, we shed light on this view from the perspective of our four-part similarity analysis, which distinguishes between intratextual (text structural, quantitative, and thematic similarity) and intertextual (hypertextual) aspects,



where the latter are simultaneously examined by means of four measures. The important aspect is that these reference points may be influenced to varying degrees by linguistic, cultural or even geographical factors. In the case of SV3 (quantitative text profiles), for example, we observed rather evenly distributed similarities of the INs of the different subject areas, at a higher similarity level. The differences observed in this context may indicate the influence of the underlying languages, which their authors are less or barely able to “escape” through their own control. The extent to which SV3 models law-like behavior of

texts would then make INs of different or related languages look more similar—and differences could indicate a strong influence of the respective languages. SV1 (LDS) shows a similar pattern, although this may be due to the underlying webgenre, its editing rules and the way Wikipedia monitors compliance with them. However, such a finding could negate the role of this SV. In any case, one should not underestimate the source of information on which SV1 is based, as it concerns elements of documents that are likely to be the focus of reading: tables and especially figures and pictures. In this respect, a further development is





needed that integrates text analysis with image analysis and related approaches.

SV2 (thematic structure) differs from SV1 and SV3: It decreases rapidly and reaches a very low level of similarity (in case of subject area *disease* (see **Table 4**). Part of this dynamic is likely caused by the diverging *F*-values of our topic model (see **Table 2**). But let us assume its effectiveness. What could be the cause then, if we do not look for factors, such as size or age of a Wikipedia? Take the example of the subject area *disease*: could it be that it is cultural differences that determine which diseases

are described in which language edition by means of which intertextual structures (see the low similarity values measured with DeltaCon in **Table 3**)? Is there, so to speak, a cultural or any comparable disposition for the arrangement of intertextual contexts, while the corresponding intratextual similarities (SV2 in **Table 3**) tend to be much higher? In the case of subject area *holiday* (which according to SV2 and GES occupies a position in the middle, according to DeltaCon in the lower range and according to EJS in the upper range of the similarity spectrum), cultural references are rather likely. Whether true or

not, this gives rise to the question of which intertextual structures, which link-based factors for the production of multiple texts are culturally determined or have a cultural imprint. Is there, so to speak, a linguistic, cultural or knowledge area-related fingerprint that could be read from intertextual structures, from different parts of the IL, a fingerprint which could help to explain its dynamics beyond what is done by lexical-semantic analyses of link anchors? Even though our research does not answer this question, it does raise it and thus builds a bridge between the kind of data science we pursue and text-linguistic questions from the field of reading research.

To answer these questions, we need a multiple text model that includes the underlying IL as a limiting factor of what results from reading processes as multiple text; a model that considers linguistic, cultural, genre- and register-related (Halliday and Hasan, 1989) as well as social factors when asking about the function or meaning of a given or missing link. It is obvious that such a model-theoretical extension of reading research benefits considerably from browsing models, such as those developed by more recent hypertext research (Dimitrov et al., 2017; Lamprecht et al., 2017): it may not be surprising that readers are more likely to select links in the initial sections of Wikipedia articles according to text-structural criteria. However, this research has also highlighted the importance of semantic criteria for link selection. It is now necessary to further explore this system of motives and to extend it to the generation of multiple texts as a whole. In this context, our approach to DeltaCon is of interest, which evaluates random walks as a source of information to assess vertex affinities. The reason for this is that it can easily be linked to empirical research on the reading or navigation behavior of users in Wikipedia and comparable resources. In this way, it would integrate intertextual structural analysis with the pragmatics of real hypertext use. By additionally integrating the article-content-related similarity views SV1-3 as developed here, it would open up a very broad spectrum of information sources for the analysis and comparison of multiple texts, namely syntactic, semantic and pragmatic sources. This could ultimately pave the way to go beyond the detection of limiting factors, as the IL in the form of Wikipedia imposes on reading, to gain models of how this IL is actually represented by its readers and entire reader communities through distributed reading processes, that is, as a distributed cognitive map of the IL.

Irrespective of these findings, conclusions and prospects for future work, a number of boundary conditions must be considered with regard to our research, which at the same time affect its limitations:

1. *Coverage*: Although we have implemented an extensible procedure for extracting INs as intertextual manifestations of subject areas, which in principle can take into account any of the millions of Wikidata items to conduct cross-linguistic studies, we have only analyzed a subset of 25 such areas. This approach could be extended by asking about the convergence of similarities/dissimilarities between languages, as a result of studying a much wider range of subject areas. In this sense, we provided the starting point for a more detailed examination of the thematic biases of Wikipedia, compared to what has been

studied so far. This level of detail originates from our three-level topic model (**Figure 6**), which should be expanded into a model of thematic-rhematic intertextuality (cf. Mehler et al., 2019).

2. *Similarity analysis*: We implemented a hybrid approach to measure the similarities of INs, the structure-oriented part of which includes four approaches to graph similarity measurement. However, the spectrum of relevant measures is much wider (Emmert-Streib et al., 2016), so that their expressiveness and significance for intertextuality research should be examined more thoroughly. One may even think of a multiple source similarity measure that simultaneously maps various structural and other informational sources to assess the similarity of multiple texts and hypertexts.
3. *Network analysis*: The modeling of reading requires the modeling of cognitive processes, which in the case of distributed reading means the modeling of processes in social networks. As explained in our introduction, such an endeavor requires modeling the “fluent alignment” of two processes: on the one hand, the multi-authorial writing of Wikipedia and its embedding in the larger IL and, on the other hand, the diversity of reading situations, their task contexts and contextual resources by which they are conditioned. In this way, our approach could be further developed by integrating models of web-based writing research<sup>10</sup> and social network analysis.
4. *Content mining*: Although we integrate a semantic model regarding the thematic structure of single texts, we do not yet consider their content beyond this level. This means, for example, that although our approach can find strong similarities between different Wikipedias on the same topic, the hypertexts in question can inform about this topic in very different ways, for example by making different statements or by embedding them in different argumentative contexts. By extending the semantic part of our text representation model (in particular by including knowledge graphs and representations based on semantic role labeling (Palmer et al., 2010) to detect, e.g., assertions or claims) we would be able to more reliably detect semantic (dis-)similarities between texts and overcome the corresponding limitation of our model.

A more comprehensive model of distributed reading and writing that meets these extensions to overcome the related limitations is certainly a challenge for future research, especially if it is to be based on thorough linguistic analysis. Currently, we see no way out of such a research direction for education science, that is, for studying learning in the age of information.

## 5.2. How Does Our Approach Relate to Linguistic Relativity?

Starting from a selected set of topics, we have shown that different language Wikipedias produce quite different networks for informing about the same topics. That is, we detected a

<sup>10</sup>For a recent network theoretical sentiment analysis of online writing see, for example, Stella et al. (2020). For a review of network theoretical approaches to knowledge networks in education science see Siew (2020).

bias: extent and organization of a topic's representation depend on the underlying Wikipedia—the former are biased by the latter. Since we related this bias to the languages in which the Wikipedias are written, we spoke of a *linguistic bias*. This raises the question what our approach contributes to research on what is known as *Linguistic Relativity* (LR) (Lakoff, 1987; Lucy, 1997) or *Cultural Relativity* (CR) (Gumperz and Levinson, 1991, 1996) (cf. also Sharifian, 2017). Are the differences we observe caused by differences in the underlying languages (LR) or even by cultural differences (CR) between the communities of writers producing these Wikipedias? Our approach does not allow for a direct answer to this question, as it is not based on the linguistic or social data required for such an undertaking. Nevertheless, it is worth explaining how it relates to this research.

To clarify this we utilize the distinction of *structure-*, *domain-*, and *behavior-centered* approaches according to Lucy (1997): the first start from observed semantic differences between languages to examine their influence on thought, the second from “domain[s] of experienced reality” (Lucy, 1997, p. 298) to ask how languages represent them, the third from language-specific practices of use. According to Lucy (1997, p. 298), the distinguishing feature of domain-specific approaches is that they characterize domains independently of the target languages. This is what we intend to do when deriving topic representations from Wikidata items and their relations (cf. Mehler et al., 2011): we explore this data to gain access to conceptual representations of parts of experienced reality, ask how they are described in different Wikipedias and whether the networked descriptions based thereon are commensurable or not. Our first assessment is that most approaches to Wikipedia's LR or CR (cf. Massa and Scrinzi, 2012; Laufer et al., 2015; Miquel-Ribé and Laniado, 2016; Miz et al., 2020) are domain-centered in such a way. Before assessing what they can say about LR, we go one step further in characterizing our approach, this time with the help of Lakoff (1987, p. 322) who discusses conceptual organization as a reference point for assessing the commensurability of conceptual systems, with which our approach to graph similarity is apparently compatible. In particular, Lakoff (1987, p. 334) concludes that organizational differences point to different conceptual systems and thus to LR. Let us assume that INs *manifest* such conceptual systems hypertextually. The differences we find in these INs could then reflect conceptual differences of the underlying languages and, if these differences are culturally determined, CR: conceptualizations of the subject areas investigated here would then be linguistically relative and ultimately culture-specific. If languages encode world views (Gumperz and Levinson, 1996, 2), Wikipedias can then be seen as manifestations of parts of such views, and since languages differ in encoding them, their Wikipedias are consequently non-trivially different. Moreover, due to the distributed authorship of Wikipedia, there is a direct link to the concept of distributed cognition (Hollan et al., 2000), to which approaches to CR are connected (Gumperz and Levinson, 1991; Sharifian, 2017). From this point of view, it seems plausible to assume effects of different communities, each of which produces representations of conceptual systems

or world views in the form of hypertexts, that are more or less incommensurable. From this it follows that by examining such differences we should get access to the differences of the underlying worldviews and their encodings. To sum up: we use Wikidata to identify cross-linguistic conceptual units (as a bridge to experienced reality), examine their language-specific lexicalizations (article names) and interconnections (hyperlinks) using corresponding excerpts from Wikipedia, interpret their differences as evidence of LR and speculate on CR as its cause (see above).

Apparently this consideration already brings us to the end of the analogy. The reason for this assessment concerns the research objects of the areas compared here. More precisely, the question is to which entities or systems the observed differences are finally ascribed. While research on LR aims to make statements about *language systems*—beyond the level of lexis—or *conceptual systems* (by asking whether the same parts of reality (e.g., the color spectrum) are conceptualized and coded in a language-specific way), we and our relatives in Wikipedia make statements about *textual instances* of such systems. The former deal with differences in language systems, while we study hypertextual differences without directly drawing conclusions about the underlying systems—beyond the lexical level. That is, we do not directly contribute to research on LR at the level of *language as system* but rather at the level of *language as text* (Hjelmslev, 1969). Moreover, beyond the question of whether or not a concept is manifested and networked in a language, we do not consider cases of conceptual splitting, fusion, etc. From this point of view, it is certainly no overstatement to claim that we found evidence of the type of linguistic bias described above. But it would be an overstatement to claim that we thereby measure LR on the level of language systems and the underlying cultures. Though we can speculate that cultural differences are responsible for the differences we measure, we cannot yet prove this with our apparatus. To consider language as a system, linguistic analyses are required, such as those provided by comparative linguistics (Bisang and Czerwinski, 2019). For example, we can ask about differences in the linguistic manifestation of topics (e.g., information structure, density and uncertainty, relevance, salience, etc.) and what effects this has on the organization and linking of articles. Based on this, we could ask for language-specific text patterns and attribute them to author communities and thus to cultural differences, but we would need independent data to support such conclusions. At least the linguistic part of this task now lies within the interdisciplinary reach of comparative and computational linguistics, so that we can put research on *reading/writing multiple texts* on a broader methodological basis. This is work for the future.

### 5.3. Information Processing and Online Reasoning

With regard to information processing and online reasoning in higher education, our approach has at least two implications:

1. As reviewed in section 3, Wikipedia is one of the most important resources in education. In accordance with the

Matthew principle, it is the primary target of first (and often final) information searches, either directly or via Google searches. This prominent position is contrasted by (i) the skewness of the thematic similarity relations we found, (ii) the unevenly distributed depths and widths of thematizations of the same topics across different Wikipedias, and (iii) the contextual dependence of the thematic similarities of these Wikipedias. This tripartite skewness and dependence, which is associated with a certain conception of linguistic relativity (see above), is now in turn in conflict with a way of using Wikipedia according to a reformulation of the *Closed World Assumption* (CWA) (Reiter, 1978); this reformulation assumes that Wikipedia is (almost) complete in the sense that it describes (almost) everything that is “relevant” to a particular subject area, while leaving out everything that is “irrelevant” in that sense<sup>11</sup>: *what is relevant is in Wikipedia, or it is not relevant*. Such an assumption, which is not seriously supported by anyone, is obviously wrong, and we have given three more reasons for its rejection. The question is, however, whether even learners in the field of higher education search for and process information in *implicit* agreement with such an assumption or a similar view (possibly only for reasons of effort reduction). From the perspective of reading scenarios corresponding to this view, our research reveals a considerable potential of “false negatives,” that is, of knowledge units or components that have not been described but could have been described in Wikipedia. Our similarity analyses show that the same topics are described in different Wikipedias in different densities, and thus what is described in one Wikipedia is likely omitted in others (leading to “false negatives” in the latter sense); otherwise we would have observed much higher similarities of these Wikipedias. Thematic relevance (and consequently accessibility for the online reading process) is then either language-specific—which, given the topics selected here as examples, is unlikely to be the case—or one must acknowledge that resources, such as Wikipedia strictly contradict the latter variant of the CWA. Reading habits in the sense of this variant and actual information offer stand then in fundamental contradiction to each other. This assessment should be seen from the perspective that online accessibility attracts readers and that the status of Wikipedia as a widely accepted, unrivaled learning resource increases its likelihood of being read in higher education in whatever scenario (thus confirming a variant of the Matthew principle).

2. With the amount of information available online on numerous fields of knowledge and its accessibility through a medium with a unique selling point, such as Wikipedia, we can observe another phenomenon that is relevant to higher education but closed to the first glance. Starting from the concept of domain-specific learning and reading skills (List and Alexander, 2019), the question arises as to the online transparency and visibility of domain boundaries, which allow a learner to notice, in the context of his domain-specific task completion, at which

points he or she leaves the content area of his or her domain while switching to a domain that is possibly unfamiliar to him or her and for which he or she lacks the knowledge background and corresponding learning skills. This question is related to problems on two scales: on the micro-level, it concerns the cognitive load on the part of the learner due to an increased need for the integration of domain-external knowledge; on the macro-level, it corresponds to the blurring of domain boundaries, of established disciplinary differences, which are characterized by the development of the aforementioned domain-specific learning and reading skills. In speculative terms, a distributed reading process takes place as a result of numerous such events, in which each topic can be contextualized by any other topic of any domain (small-world effect). As a consequence of such a process, the imparting of domain differences, peculiarities and boundaries can be complicated even in higher education. At first glance, our research does not contribute to studying such processes, since we focused on specific topics for our similarity analyses, for which we have developed an extraction procedure, since Wikipedia does not make the relevant topic-related subnetworks visible. At second glance, however, our research also provides insights in this respect, as readers will hardly always follow the paths on which our similarity analyses were based. But if dissimilarity is already the predominant diagnosis for the latter paths, then this applies all the more to the free, unbound search for information (*what must already be considered dissimilar in our sense can only appear as more dissimilar under the latter regime*), so that the assessment made under the previous bullet point is reinforced. This implies the simplified consequence that with the increasing importance of learning resources like Wikipedia, the transfer of domain-specific knowledge can become more difficult and thus gains in importance.

We may add a third point concerning the divergence of time scales, since while the importance of Wikipedia is likely to grow steadily, this does not necessarily apply to the closing of the knowledge gaps diagnosed here: the idea of a convergence of all Wikipedias with regard to the increasingly similar presentation of the same subject areas is likely to be as daring as it is unrealistic. In any case, our approach shows that when it comes to investigating web-based resources with regard to their effects on (higher) education, the computational linguistic approach developed here appears indispensable if concrete models of the affected sections of the corresponding IL are required. That aims certainly beyond Wikipedia, but remains methodically in the context of what has been elaborated here.

## 6. CONCLUSION

We introduced a three-level topic model in combination with a graph-theoretical model for measuring the intra- and intertextual similarities of article networks from different language editions of Wikipedia. In this way we built a bridge between reading

<sup>11</sup>This learning resource-related reformulation of the CWA replaces the concept of *truth* with the concept of *relevance with respect to specific information needs*.



research, educational science and computational linguistics. To this end, we described a new perspective for reading research that focuses more on the information landscape as a limiting factor of online reading. We have continued research showing that Wikipedia exhibits a topical coverage bias. However, we have done this using a much more elaborate topic and text structure model in conjunction with a quantitative model of hypertext structure, a hybrid model that is more realistic from the perspective of hypertext linguistics (Storrer, 2002). Finally, we have derived two implications from our findings; the first concerns a variant of the closed world assumption and the related prediction that learning processes based on shallow reading and quick search processes are likely to be affected by the thematic dissimilarities and contextual dependencies of online information resources as we have observed in Wikipedia, and that this effect may also have negative effects on the acquisition of domain-specific knowledge and corresponding learning skills. In future work we will continue elaborating our computational hypertext model. This will be done with special attention to hypertext usage to obtain graph models that integrate browsing behavior into graph similarity analysis. In this way, we will try to model students' online learning by combining information about *how they read* with information about *what they read* and how this is networked. This will possibly give an outlook on two things: the construction and integration of knowledge on the part of single students, learners, or readers and its distributed counterpart regarding larger communities of students, learners or readers and their distributed reading processes.

## REFERENCES

- Álvarez, G., Oeberst, A., Cress, U., and Ferrari, L. (2020). Linguistic evidence of in-group bias in english and spanish Wikipedia articles about international conflicts. *Discourse Context Media* 35:100391. doi: 10.1016/j.dcm.2020.100391
- Bahdanau, D., Cho, K., and Bengio, Y. (2015). "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations* (San Diego, CA), 1–15.
- Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., and Gergle, D. (2012). "Omnipedia: bridging the wikipedia language gap," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, TX), 1075–1084. doi: 10.1145/2207676.2208553
- Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proc. Nat. Acad. Sci. U.S.A.* 101, 3747–3752. doi: 10.1073/pnas.0400087101
- Barzilai, S., and Zohar, A. (2012). Epistemic thinking in action: evaluating and integrating online sources. *Cogn. Instruct.* 30, 39–85. doi: 10.1080/07370008.2011.636495
- Bisang, W., and Czerwinski, P. (2019). "Performance in knowledge assessment tests from the perspective of linguistic typology," in *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)* (Cham: Springer), 207–235. doi: 10.1007/978-3-030-26578-6\_16
- Braasch, J. L. G., Bråten, I., and McCrudden, M. T. (Eds.). (2018a). *Handbook of Multiple Source Use*. New York, NY; London: Routledge.
- Braasch, J. L. G., McCrudden, M. T., and Bråten, I. (2018b). "Reflections and future directions," in *Handbook of Multiple Source Use, Chapter 29*, eds J. L. Braasch, I. Bråten, and M. T. McCrudden (New York, NY; London: Routledge), 527–537. doi: 10.4324/9781315627496-29

## DATA AVAILABILITY STATEMENT

Code and datasets used for this study can be found on github: <https://github.com/texttechnologylab/WikiSim>.

## AUTHOR CONTRIBUTIONS

AM has written sections 1, 2, 3, and 5 (together with PW), generated **Figures 1–4, 6, 7, 11–13** and **Table 1**. AM, MK, PW, TU, and WH have written section 4. WH extracted INs from Wikidata and Wikipedia and generated **Figures 8–10** as well as **Tables 3–5**, together with AM, he generated **Figure 5**. TU generated **Table 2**, trained and tested text2ddc. AM (CGS), MK (SV1, SV3), PW (EJS, DeltaCon), TU (SV2), and WH (GES) implemented, computed, and described the similarity measures. AM, MK, PW, and WH proofread and revised the final manuscript. All authors contributed to the conception of the paper, its algorithmization, and the interpretation of the results.

## FUNDING

WH was partly financed by the State of Hessen through the LOEWE research focus on *minority studies: language and identity*. PW was partially funded by the Federal Ministry of Education and Research of Germany as part of the competence center for machine learning ML2R (01|S18038C) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2070-390732324.

- Britt, M. A., Rouet, J.-F., and Braasch, J. L. (2012). "Documents as entities: extending the situation model theory of comprehension," in *Reading - From Words to Multiple Texts*, eds M. A. Britt, S. R. Goldmann, and J. F. Rouet (New York, NY: Routledge), 161–179. doi: 10.4324/9780203131268
- Britt, M. A., Rouet, J.-F., and Durik, A. M. (2018). *Literacy Beyond Text Comprehension: A Theory of Purposeful Reading*. New York, NY: Routledge.
- Callahan, E. S., and Herring, S. C. (2011). Cultural bias in Wikipedia content on famous persons. *J. Am. Soc. Inform. Sci. Technol.* 62, 1899–1915. doi: 10.1002/asi.21577
- Cheng, J., Dong, L., and Lapata, M. (2016). "Long short-term memory-networks for machine reading," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, TX: Association for Computational Linguistics), 551–561. doi: 10.18653/v1/D16-1053
- Cho, B.-Y., and Afflerbach, P. (2015). Reading on the internet: realizing and constructing potential texts. *J. Adolesc. Adult Liter.* 58, 504–517. doi: 10.1002/jaal.387
- Coiro, J., Sparks, J. R., and Kulikowich, J. M. (2018). "Assessing online collaborative inquiry and social deliberation skills as learners navigate multiple sources and perspectives," in *Handbook of Multiple Source Use*, eds J. L. Braasch, I. Bråten, and M. T. McCrudden (New York, NY; London: Routledge), 485–501. doi: 10.4324/9781315627496-27
- Conde, A., Arruarte, A., Larrañaga, M., and Elorriaga, J. A. (2020). How can Wikipedia be used to support the process of automatically building multilingual domain modules? A case study. *Inform. Process. Manage.* 57:102232. doi: 10.1016/j.ipm.2020.102232
- Denning, P., Horning, J., Parnas, D., and Weinstein, L. (2005). Wikipedia risks. *Commun. ACM* 48, 152–152. doi: 10.1145/1101779.1101804

- DeStefano, D., and LeFevre, J.-A. (2007). Cognitive load in hypertext reading: a review. *Comput. Hum. Behav.* 23, 1616–1641. doi: 10.1016/j.chb.2005.08.012
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Minneapolis, MN), 4171–4186.
- Dimitrov, D., Singer, P., Lemmerich, F., and Strohmaier, M. (2017). “What makes a link successful on wikipedia?” in *Proceedings of the 26th International Conference on World Wide Web, WWW 17* (Geneva: International World Wide Web Conferences Steering Committee), 917–926. doi: 10.1145/3038912.3052613
- Downs, R. M., and Stea, D. (1977). *Maps in Minds: Reflections on Cognitive Mapping*. New York, NY: Harper & Row.
- Emmert-Streib, F., Dehmer, M., and Shi, Y. (2016). Fifty years of graph matching, network alignment and network comparison. *Inform. Sci.* 346–347, 180–197. doi: 10.1016/j.ins.2016.01.074
- Erdős, P., and Rényi, A. (1959). On random graphs. *Publ. Math.* 6, 290–297.
- Feldman, R., and Sanger, J. (2007). *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press.
- Goldman, S. R., Braasch, J. L., Wiley, J., Graesser, A. C., and Brodowinska, K. (2012). Comprehending and learning from internet sources: processing patterns of better and poorer learners. *Read. Res. Q.* 47, 356–381. doi: 10.1002/RRQ.027
- Graham, M., Straumann, R. K., and Hogan, B. (2015). Digital divisions of labor and informational magnetism: mapping participation in Wikipedia. *Ann. Assoc. Am. Geograph.* 105, 1158–1178. doi: 10.1080/00045608.2015.1072791
- Gumperz, J. J., and Levinson, S. C. (1991). Rethinking linguistic relativity. *Curr. Anthropol.* 32, 613–623. doi: 10.1086/204009
- Gumperz, J. J., and Levinson, S. C. (1996). “Introduction: Linguistic relativity re-examined,” in *Rethinking Linguistic Relativity* (Cambridge: Cambridge University Press), 1–18.
- Halavais, A., and Lackaff, D. (2008). An analysis of topical coverage of Wikipedia. *J. Comput. Mediat. Commun.* 13, 429–440. doi: 10.1111/j.1083-6101.2008.00403.x
- Halliday, M. A. K., and Hasan, R. (1989). *Language, Context, and Text: Aspects of Language in a Sociosemiotic Perspective*. Oxford: Oxford University Press.
- Hargittai, E., and Dobransky, K. (2017). Old dogs, new clicks: digital inequality in skills and uses among older adults. *Can. J. Commun.* 42, 195–212. doi: 10.22230/cjc.2017v42n2a3176
- Hartman, D. K., Hagerman, M. S., and Leu, D. J. (2018). “Toward a new literacies perspective of synthesis: multiple source meaning construction,” in *Handbook of Multiple Source Use, Chapter 4*, eds J. L. Braasch, I. Bråten, and M. T. McCrudden (New York, NY; London: Routledge), 55–78. doi: 10.4324/9781315627496-4
- Head, A. (2013). Project information literacy: what can be learned about the information-seeking behavior of today's college students? *SSRN Electron. J.* doi: 10.2139/ssrn.2281511. [Epub ahead of print].
- Hecht, B., and Gergle, D. (2009). “Measuring self-focus bias in community-maintained knowledge repositories,” in *Proceedings of the Fourth International Conference on Communities and Technologies* (University Park, PA), 11–20. doi: 10.1145/1556460.1556463
- Hecht, B., and Gergle, D. (2010a). “The Tower of Babel meets Web 2.0: user-generated content and its applications in a multilingual context,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10* (Atlanta, GA), 291–300. doi: 10.1145/1753326.1753370
- Hecht, B. J., and Gergle, D. (2010b). “On the “localness” of user-generated content,” in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW '10* (New York, NY: ACM), 229–232. doi: 10.1145/1718918.1718962
- Hemati, W., Uslu, T., and Mehler, A. (2016). “TextImager: a distributed UIMA-based system for NLP,” in *Proceedings of COLING 2016: System Demonstrations* (Osaka), 59–63.
- Hjelmslev, L. (1969). *Prolegomena to a Theory of Language*. Madison, WI: University of Wisconsin Press.
- Hollan, J., Hutchins, E., and Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput. Hum. Interact.* 7, 174–196. doi: 10.1145/353485.353487
- Holloway, T., Bozicevic, M., and Börner, K. (2007). Analyzing and visualizing the semantic coverage of Wikipedia and its authors: research articles. *Complexity* 12, 30–40. doi: 10.1002/cplx.20164
- Hsieh, Y. P. (2012). Online social networking skills: the social affordances approach to digital inequality. *First Monday* 17. doi: 10.5210/fm.v17i4.3893
- Jiang, Y., Bai, W., Zhang, X., and Hu, J. (2017). Wikipedia-based information content and semantic similarity computation. *Inform. Process. Manage.* 53, 248–265. doi: 10.1016/j.ipm.2016.09.001
- Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). “Bag of tricks for efficient text classification,” in *Proceedings of the 15th Conference of the EACL: Volume 2, Short Papers* (Valencia: Association for Computational Linguistics), 427–431. doi: 10.18653/v1/E17-2068
- Karimi, F., Bohlin, L., Samoilenko, A., Rosvall, M., and Lancichinetti, A. (2015). Mapping bilateral information interests using the activity of Wikipedia editors. *Palgrave Commun.* 1, 1–7. doi: 10.1057/palcomms.2015.41
- Kendeou, P., Robinson, D. H., and McCrudden, M. T. (2019). “Modeling the dissemination of misinformation through discourse dynamics,” in *Misinformation and Disinformation in Education: An Introduction* (Charlotte: Information Age Publishing), 1–4.
- Kintsch, W. (1998). *Comprehension. A Paradigm for Cognition*. Cambridge: Cambridge University Press.
- Kittur, A., Chi, E. H., and Suh, B. (2009). “What's in Wikipedia?: mapping topics and conflict using socially annotated category structure,” in *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI '09* (New York, NY: ACM), 1509–1512. doi: 10.1145/1518701.1518930
- Köhler, R., Altmann, G., and Piotrowski, R. G. (Eds.). (2005). *Quantitative Linguistics. An International Handbook*. Berlin; New York, NY: Mouton de Gruyter.
- Konca, M., Mehler, A., Baumartz, D., and Hemati, W. (2020). *From Distinguishability to Informativity: A Quantitative Text Model for Detecting Random Texts*.
- Konieczny, P. (2016). Teaching with Wikipedia in a 21st-century classroom: perceptions of Wikipedia and its educational benefits. *J. Assoc. Inform. Sci. Technol.* 67, 1523–1534. doi: 10.1002/asi.23616
- Koutra, D., Shah, N., Vogelstein, J. T., Gallagher, B., and Faloutsos, C. (2016). DeltaCon: principled massive-graph similarity function with attribution. *ACM Trans. Knowl. Discov. Data* 10, 28:1–28:43. doi: 10.1145/2824443
- Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago, IL: University of Chicago Press.
- Lamprecht, D., Lerman, K., Helic, D., and Strohmaier, M. (2017). How the structure of wikipedia articles influences user navigation. *New Rev. Hypermed. Multimed.* 23, 29–50. doi: 10.1080/13614568.2016.1179798
- Laufer, P., Wagner, C., Flöck, F., and Strohmaier, M. (2015). “Mining cross-cultural relations from Wikipedia: a study of 31 European food cultures,” in *Proceedings of the ACM Web Science Conference* (Oxford), 1–10. doi: 10.1145/2786451.2786452
- Lemmerich, F., Sáez-Trumper, D., West, R., and Zia, L. (2019). “Why the world reads Wikipedia: beyond english speakers,” in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne, VIC), 618–626. doi: 10.1145/3289600.3291021
- List, A., and Alexander, P. A. (2019). Toward an integrated framework of multiple text use. *Educ. Psychol.* 54, 20–39. doi: 10.1080/00461520.2018.1505514
- Loh, K. K., and Kanai, R. (2016). How has the internet reshaped human cognition? *Neuroscientist* 22, 506–520. doi: 10.1177/1073858415595005
- Lorini, V., Rando, J., Saez-Trumper, D., and Castillo, C. (2020). Uneven coverage of natural disasters in Wikipedia: the case of flood. *arXiv* 2001.08810.
- Lucassen, T., and Schraagen, J. M. (2010). “Trust in Wikipedia: how users trust information from an unknown source,” in *Proceedings of the 4th Workshop on Information Credibility, WICOW '10* (Raleigh, NC), 19–26. doi: 10.1145/1772938.1772944
- Lucy, J. A. (1997). Linguistic relativity. *Annu. Rev. Anthropol.* 26, 291–312. doi: 10.1146/annurev.anthro.26.1.291
- Massa, P., and Scrini, F. (2012). “Manypedia: comparing language points of view of wikipedia communities,” in *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration, WikiSym 12* (New York, NY: Association for Computing Machinery), 1–9. doi: 10.1145/2462932.2462960

- McMahon, C., Johnson, I., and Hecht, B. (2017). "The substantial interdependence of Wikipedia and Google: a case study on the relationship between peer production communities and information technologies," in *Eleventh International AAAI Conference on Web and Social Media* (Montréal, QC), 142–151.
- Mehler, A., Gleim, R., Gaitsch, R., Uslu, T., and Hemati, W. (2019). From topic networks to distributed cognitive maps: Zipfian topic universes in the area of volunteered geographic information. *Complexity*. 4, 1–47. doi: 10.1155/2020/4607025
- Mehler, A., Hemati, W., Uslu, T., and Lücking, A. (2018). "A multidimensional model of syntactic dependency trees for authorship attribution," in *Quantitative Analysis of Dependency Structures*, eds J. Jiang and H. Liu (Berlin; New York, NY: De Gruyter), 315–348. doi: 10.1515/9783110573565-016
- Mehler, A., Pustynnikov, O., and Diewald, N. (2011). Geography of social ontologies: testing a variant of the Sapir-Whorf hypothesis in the context of Wikipedia. *Comput. Speech Lang.* 25, 716–740. doi: 10.1016/j.csl.2010.05.006
- Mehler, A., and Ramesh, V. (2019). "TextInContext: on the way to a framework for measuring the context-sensitive complexity of educationally relevant texts—a combined cognitive and computational linguistic approach," in *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*, ed O. Zlatkin-Troitschanskaia (Cham: Springer International Publishing), 167–195. doi: 10.1007/978-3-030-26578-6\_14
- Mehler, A., Sharoff, S., and Santini, M. (Eds.). (2010). *Genres on the Web: Computational Models and Empirical Studies*. Dordrecht: Springer.
- Mesgari, M., Okoli, C., Mehdi, M., Nielsen, Årup Nielsen F., and Lanamäki, A. (2015). The sum of all human knowledge: a systematic review of scholarly research on the content of Wikipedia. *J. Assoc. Inform. Sci. Technol.* 66, 219–245. doi: 10.1002/asi.23172
- Miquel-Ribé, M., and Laniado, D. (2016). "Cultural identities in wikipeidias," in *Proceedings of the 7th 2016 International Conference on Social Media & Society* (London), 1–10. doi: 10.1145/2930971.2930996
- Miz, V., Hanna, J., Aspert, N., Ricaud, B., and Vanderghelynst, P. (2020). "What is trending on Wikipedia? Capturing trends and language biases across Wikipedia editions," in *Companion Proceedings of the Web Conference 2020* (Taipei), 794–801. doi: 10.1145/3366424.3383567
- Nagel, M.-T., Schäfer, S., Zlatkin-Troitschanskaia, O., Schemer, C., Maurer, M., Molerov, D., et al. (accepted). How do university students' web search behavior, website characteristics, and the interaction of both influence students' critical online reasoning? *Front. Educ.*
- Oeberst, A., von der Beck, I., Back, M. D., Cress, U., and Nestler, S. (2018). Biases in the production and reception of collective knowledge: the case of hindsight bias in Wikipedia. *Psychol. Res.* 82, 1010–1026. doi: 10.1007/s00426-017-0865-7
- Oeberst, A., von der Beck, I., Matschke, C., Ihme, T. A., and Cress, U. (2019). Collectively biased representations of the past: Ingroup bias in Wikipedia articles about intergroup conflicts. *Br. J. Soc. Psychol.* 59, 791–818. doi: 10.1111/bjso.12356
- OECD (2007). *Revised Field of Science and Technology (FOS)*. Available online at: [www.oecd.org/science/inno/38235147.pdf](http://www.oecd.org/science/inno/38235147.pdf)
- Okoli, C., Mehdi, M., Mesgari, M., Årup Nielsen, F., and Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: a systematic review of scholarly research on wikipedia readers and readership. *J. Assoc. Inform. Sci. Technol.* 65, 2381–2403. doi: 10.1002/asi.23162
- Okoli, C., Mehdi, M., Mesgari, M., Årup Nielsen, F., and Lanamäki, A. (2012). The people's encyclopedia under the gaze of the sages: a systematic review of scholarly research on Wikipedia. *SSRN Electron. J.* 1–138. doi: 10.2139/ssrn.2021326
- Palmer, M., Gildea, D., and Xue, N. (2010). *Semantic Role Labeling*. Morgan & Claypool Publishers.
- Pentzold, C., Weltevrede, E., Mauri, M., Laniado, D., Kaltenbrunner, A., and Borra, E. (2017). Digging wikipedia: the online encyclopedia as a digital cultural heritage gateway and site. *J. Comput. Cult. Herit.* 10, 1–19. doi: 10.1145/3012285
- Perfetti, C. A., Rouet, J.-F., and Britt, M. A. (1999). "Toward a theory of documents representation," in *The Construction of Mental Representations During Reading*, eds H. van Oostendorp and S. R. Goldman (Mahwah, NJ: Erlbaum), 99–122.
- Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–226. doi: 10.1017/S0140525X04000056
- Power, R., Scott, D., and Bouayad-Agha, N. (2003). Document structure. *Comput. Linguist.* 29, 211–260. doi: 10.1162/08912010322145315
- Primor, L., and Katzir, T. (2018). Measuring multiple text integration: a review. *Front. Psychol.* 9:2294. doi: 10.3389/fpsyg.2018.02294
- Reiter, R. (1978). "On closed world data bases," in *Logic and Data Bases*, eds H. Gallaire and J. Minker (Boston, MA: Springer US), 55–76. doi: 10.1007/978-1-4684-3384-5\_3
- Salmerón, L., Kammerer, Y., and Delgado, P. (2018). "Non-academic multiple source use on the internet," in *Handbook of Multiple Source Use*, eds J. L. Braasch, I. Bråten, and M. T. McCrudden (New York, NY; London: Routledge), 285–302. doi: 10.4324/9781315627496-17
- Samoilenko, A., Karimi, F., Edler, D., Kunegis, J., and Strohmaier, M. (2016). Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity. *EPJ Data Sci.* 5:9. doi: 10.1140/epjds/s13688-016-0070-8
- Samoilenko, A., Lemmerich, F., Weller, K., Zens, M., and Strohmaier, M. (2017). "Analysing timelines of national histories across Wikipedia editions: a comparative computational approach," in *Eleventh International AAAI Conference on Web and Social Media*, 210–219.
- Scaffidi, M. A., Khan, R., Wang, C., Keren, D., Tsui, C., Garg, A., et al. (2017). Comparison of the impact of wikipedia, UpToDate, and a digital textbook on short-term knowledge acquisition among medical students. *JMIR Med. Educ.* 3:e20. doi: 10.2196/mededu.8188
- Sharifian, F. (2017). Cultural linguistics and linguistic relativity. *Lang. Sci.* 59, 83–92. doi: 10.1016/j.langsci.2016.06.002
- Siew, C. S. Q. (2020). Applications of network science to education research: quantifying knowledge and the development of expertise through network analysis. *Educ. Sci.* 10:101. doi: 10.3390/educsci10040101
- Singer, P., Lemmerich, F., West, R., Zia, L., Wulczyn, E., Strohmaier, M., et al. (2017). "Why we read Wikipedia," in *Proceedings of the 26th International Conference on World Wide Web* (Perth, WA), 1591–1600. doi: 10.1145/3038912.3052716
- Smith, D. A. (2020). Situating Wikipedia as a health information resource in various contexts: a scoping review. *PLoS ONE* 15: e0228786. doi: 10.1371/journal.pone.0228786
- Stella, M., Restocchi, V., and Deyne, S. D. (2020). #lockdown: Network-enhanced emotional profiling in the time of COVID-19. *Big Data Cogn. Comput.* 4:14. doi: 10.3390/bdcc4020014
- Storrier, A. (2002). Coherence in text and hypertext. *Document Des.* 3, 156–168. doi: 10.1075/dd.3.2.06sto
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learn. Instruct.* 4, 295–312. doi: 10.1016/0959-4752(94)90003-5
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* 46, 234–240. doi: 10.2307/143141
- Uslu, T., Mehler, A., and Baumartz, D. (2019). "Computing classifier-based embeddings with the help of text2ddc," in *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)* (La Rochelle).
- van Dijk, T. A. (1980). *Macrostructures. An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Hillsdale, NJ: Erlbaum.
- van Dijk, T. A., and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York, NY: Academic Press.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.
- Wagner, C., Graells-Garrido, E., Garcia, D., and Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Sci.* 5:5. doi: 10.1140/epjds/s13688-016-0066-4
- Warncke-Wang, M., Uduwage, A., Dong, Z., and Riedl, J. (2012). "In search of the ur-wikipedia: universality, similarity, and translation in the Wikipedia inter-language link network," in *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (Linz), 1–10. doi: 10.1145/2462932.2462959
- Wolf, M. (2018). *Reader, Come Home: The Reading Brain in a Digital World*. New York, NY: Harper.

- Zlatkin-Troitschanskaia, O., Bisang, W., Mehler, A., Banerjee, M., and Roeper, J. (2019a). "Positive learning in the internet age: developments and perspectives in the plato program," in *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)*, ed O. Zlatkin-Troitschanskaia (Cham: Springer International Publishing), 1–5. doi: 10.1007/978-3-030-26578-6\_1
- Zlatkin-Troitschanskaia, O., Brückner, S., Molerov, D., and Bisang, W. (2019b). "What can we learn from theoretical considerations and empirical evidence on learning in higher education? Implications for an interdisciplinary research framework," in *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)* (Cham: Springer), 287–309. doi: 10.1007/978-3-030-26578-6\_21
- Zlatkin-Troitschanskaia, O., Schmidt, S., Molerov, D., Shavelson, R. J., and Berliner, D. (2018). "Conceptual fundamentals for a theoretical and empirical framework of positive learning," in *Positive Learning in the Age of Information*

*(PLATO)-A Blessing or a Curse?* eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Springer), 29–50. doi: 10.1007/978-3-658-19567-0\_4

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Mehler, Hemati, Welke, Konca and Uslu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# How Do University Students' Web Search Behavior, Website Characteristics, and the Interaction of Both Influence Students' Critical Online Reasoning?

Marie-Theres Nagel<sup>1\*</sup>, Svenja Schäfer<sup>2</sup>, Olga Zlatkin-Troitschanskaia<sup>1</sup>, Christian Schemer<sup>3</sup>, Marcus Maurer<sup>3</sup>, Dimitri Molerov<sup>1</sup>, Susanne Schmidt<sup>1</sup> and Sebastian Brückner<sup>1</sup>

<sup>1</sup> Department of Business and Economics Education, Johannes Gutenberg University Mainz, Mainz, Germany, <sup>2</sup> Department of Communication, University of Vienna, Vienna, Austria, <sup>3</sup> Department of Communication, Johannes Gutenberg University Mainz, Mainz, Germany

## OPEN ACCESS

### Edited by:

Raquel Gilar-Corbi,  
University of Alicante, Spain

### Reviewed by:

Jacqueline P. Leighton,  
University of Alberta, Canada  
Gonzalo Lorenzo Lledo,  
University of Alicante, Spain

### \*Correspondence:

Marie-Theres Nagel  
marie.nagel@uni-mainz.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Education

**Received:** 21 June 2020

**Accepted:** 19 October 2020

**Published:** 19 November 2020

### Citation:

Nagel M-T, Schäfer S, Zlatkin-Troitschanskaia O, Schemer C, Maurer M, Molerov D, Schmidt S and Brückner S (2020) How Do University Students' Web Search Behavior, Website Characteristics, and the Interaction of Both Influence Students' Critical Online Reasoning? *Front. Educ.* 5:565062. doi: 10.3389/feduc.2020.565062

The Internet has become one of the main sources of information for university students' learning. Since anyone can disseminate content online, however, the Internet is full of irrelevant, biased, or even false information. Thus, students' ability to use online information in a critical-reflective manner is of crucial importance. In our study, we used a framework for the assessment of students' *critical online reasoning (COR)* to measure university students' ability to critically use information from online sources and to reason on contentious issues based on online information. In addition to analyzing students' COR by evaluating their open-ended short answers, we also investigated the students' web search behavior and the quality of the websites they visited and used during this assessment. We analyzed both the number and type of websites as well as the quality of the information these websites provide. Finally, we investigated to what extent students' web search behavior as well as the quality of the used website contents are related to higher task performance. To investigate this question, we used five computer-based performance tasks and asked 160 students from two German universities to perform a time-restricted open web search to respond to the open-ended questions presented in the tasks. The written responses were evaluated by two independent human raters. To analyze the students' browsing history, we developed a coding manual and conducted a quantitative content analysis for a subsample of 50 students. The number of visited webpages per participant per task ranged from 1 to 9. Concerning the type of website, the participants relied especially on established news sites and Wikipedia. For instance, we found that the number of visited websites and the critical discussion of sources provided on the websites correlated positively with students' scores. The identified relationships between students' web search behavior, their performance in the CORA tasks, and the qualitative website characteristics are presented and critically discussed in terms of limitations of this study and implications for further research.

**Keywords:** critical online reasoning assessment, online information, web search, log file analysis, content analysis, quality of online information, higher education

## INTRODUCTION

In the context of digitalization, society's overall media behavior has changed fundamentally. Digital technologies are opening up new opportunities for accessing and distributing information (Mason et al., 2010; Kruse, 2017; Tribukait et al., 2017). The Internet has become one of the main sources of information for university students' learning (Brooks, 2016; Newman and Beetham, 2017). Prior research indicates that the way students process and generally handle online information can be strongly influenced not only by personal characteristics but also by the quality of the accessed websites and their content (Tribukait et al., 2017; Braasch et al., 2018). Possible relationships between qualitative website characteristics, students' web search behavior and their judging of online information, however, have hardly been studied to date. In particular, there are hardly any studies that examine the connection between different quality criteria of websites and students' evaluation of website quality. In addition, most of the existing studies are based on students' self-reports and/or were conducted in a simulated test environment, so that their generalizability regarding students' actual web search behavior in the real online environment (Internet) remains questionable.

To bridge this gap, the study presented here aims to provide empirical insights into the complex relationship between (1) students' search behavior, (2) students' evaluation of websites, and (3) the qualitative characteristics of the websites students evaluated and used in their written responses in a free and unrestricted web search, and (4) the real online environment (Internet) where they find their sources. Therefore, the study focuses on the research question: *To what extent are students' web search behavior – regarding the number and type of accessed websites and webpages – as well as the quality of the used website contents related to students' critically reflective use of online information?*

While the multitude of online information and sources may positively affect learning processes, for instance by providing access to a wide variety of learning resources at low effort and cost (Beaudoin, 2002; Helms-Park et al., 2007; Yadav et al., 2017), online information might also have multiple negative impacts on learning (Maurer et al., 2018, 2020). First, information available on the Internet is not sufficiently structured (Kruse, 2017), so that students may, for example, feel overwhelmed by the amount of information ("information overload," Eppler and Mengis, 2004). Second, since anyone can publish content online, the Internet is full of irrelevant, biased, or even false information. As a consequence, mass media rarely offer complete information and sometimes even provide inaccurate information as they are designed to exploit mental weak points that may present judgmental traps or promote weak reasoning (Ciampaglia, 2018; Carbonell et al., 2018). This holds particularly true for social media (European Commission, 2018; Maurer et al., 2018), whereby social networks and messengers, together with online newspapers and news magazines, online videos and podcasts, are considered the least trustworthy sources of news or information. As a result, when learning with and from the Internet, students face the heightened challenge of judging the

quality of the information they find online. The extent to which the advantages of the Internet prevail or the disadvantages result in students being overwhelmed or manipulated depends on their abilities concerning search behavior and critical evaluation of online information.

Especially considering students' intensive use of the Internet during their higher education studies, it is important to assess and foster their critical online reasoning skills with suitable measures. One promising approach consists of performance assessments in an open-ended format using tasks drawn from real-world judgment situations that students and graduates face in academic and professional domains as well as in their private lives (McGrew et al., 2017; Shavelson et al., 2019). We therefore used a corresponding framework for assessing students' ability to deal critically with online information, the Critical Online Reasoning Assessment (CORA), which was adapted and further developed from the American Civic Online Reasoning Assessment (Wineburg et al., 2016; Molerov et al., 2019). With this framework, we assessed university students' ability to critically evaluate information from online sources and to use this information to reason on contentious issues (Wineburg et al., 2016). To investigate the research question of this study, we used students' response data, including (i) their web browser history with log files from the CORA task processing, (ii) their written responses to the tasks, and (iii) the websites they visited and the content they used during this assessment, and performed a quantitative analysis of website characteristics.

## THEORETICAL FOUNDATIONS

### Students as Digital Natives?

The evaluation of information sources is crucial for successfully handling online information and learning from Internet-based inquiry (Wiley et al., 2009; Mason et al., 2010), and using online information in a critical-reflective manner is a necessary skill. Critically analyzing and evaluating digitally represented information is necessary to cope with the oversupply of unstructured information and to analyze make judgments about the information found online (Gilster, 1997; Hague and Payton, 2010; Ferrari, 2013; Kruse, 2017). In higher education, it has long been assumed that students, as the generation of digital natives, are skilled in computer use and information retrieval and thus use digital media competently (Prensky, 2001; Murray and Pérez, 2014; Blossfeld et al., 2018; Kopp et al., 2019). However, recent studies have shown that students perform poorly when it comes to correctly judging the reliability of web content (e.g., McGrew et al., 2018). Although being familiar with a variety of digital media (e.g., social networking sites, video websites; Nagler and Ebner, 2009; Jones and Healing, 2010; Thompson, 2013), students use them primarily for private entertainment or social exchange, and are not capable of applying their digital skills in higher education and critically transferring information-related skills to the learning context (Gikas and Grant, 2013; Persike and Friedrich, 2016; Blossfeld et al., 2018). Students often base their judgment of websites on irrelevant criteria such as the order of search results and authority of a search engine, the website design,

or previous experience with the websites and the information provided there, while they neglect the background of a website or the credibility of the author (McGrew et al., 2017). For instance, Wikipedia and Google were the most frequently used despite students rating them as rather unreliable and students' overall use of all web search tools was rather unsophisticated (Judd and Kennedy, 2011; Maurer et al., 2020).

The core aspect of a successful search strategy is the correct evaluation and, in particular, selection of reliable websites and content therein, as students form their opinions and make judgments on this basis. If students refer to websites with biased or misrepresented information, this inevitably leads to a lower-quality or even completely incorrect judgment. It is therefore particularly problematic that, according to first studies, students struggle to evaluate the trustworthiness of the information they encounter online and to distinguish reliable from unreliable websites (McGrew et al., 2019). The expectation that today's students generally have a digital affinity is therefore not tenable (Kennedy et al., 2008; Bullen et al., 2011). To be able to deal successfully with online information, it is urgently necessary that today's students first learn to critically question, examine and evaluate it (Mason et al., 2010; Blossfeld et al., 2018).

## Critical Online Reasoning

The ability to successfully deal with online information and distinguish, for instance, reliable and trustworthy sources of information from biased and manipulative ones (Wineburg et al., 2016) regarded *Critical Online Reasoning* (COR), which we define as a key facet of critical and analytic thinking while using online media (Zlatkin-Troitschanskaia et al., 2020). In contrast to other concepts related to critical thinking, COR is explicitly limited to use in the online information environment. Besides *critical thinking* (Facione, 1990), COR refers to some aspects of *digital literacy*, i.e., the ability to deal with digital information and the technology required for it in a self-determined and critical manner (Gilster, 1997; Hague and Payton, 2010; JISC, 2014), which can be placed in the broader field of media competence and communication (Gilster, 1997; Hague and Payton, 2010; McGrew et al., 2017). Since the search for information using suitable strategies is an important aspect of COR, there is a conceptual overlap with 'information problem solving' (Brand-Gruwel et al., 2009). It refers in particular to metacognition, which regulates the entire COR process, including the development of an appropriate (search) strategy to achieve the objectives, reflecting on the status of information procurement and the search process.

In this respect, COR refers in particular to three superordinate dimensions: *searching and source evaluation*, *critical reasoning* and *decision making* (Zlatkin-Troitschanskaia et al., 2020). *Searching and source evaluation* describes the evaluation of the information and sources found online and includes the ability to select, understand and evaluate relevant texts on a website and to judge whether a source is credible, using additional resources available online and by cross-checking with other search results. *Critical Reasoning* means to recognize and evaluate arguments and their components used in the sources found online with regard to evidentially, objectivity, validity, and consistency. *Decision-making* refers to the process of making a correct,

evaluative judgment and reaching a conclusion based on reliable sources, which also includes explaining the decision in a well-structured and logically cohesive way (for a more comprehensive description of the COR construct, see Zlatkin-Troitschanskaia et al., 2020).

## Students' Internet Search Behavior

When navigating the web, users can either directly access websites that might provide the information they are looking for. However, the most common way of finding a way through the overwhelming amount of available information is to use a search engine (Beisch et al., 2019). Students have to narrow down their search and to select an appropriate website out of thousands of search results presented by the search engine. In a next step, the available information on a selected website has to be assessed in view of the task or the general informational need. The process of accessing websites, either directly or through search engines, and browsing websites to examine the available information is repeated until the user is satisfied with the findings (Hölscher and Strube, 2000) and is able to construct a mental model that meets their need for information.

When it comes to evaluating this process, there are two levels need to be addressed: the websites themselves (search results) and the information (content) provided by these websites. At the website-level, both the depth and the quality of the search behavior have to be considered (Roscoe et al., 2016). Depth means that it is important that the search for information online is extensive, in terms of both the number of search inquiries and the number of visited websites. The extensive use of search engines and various sources is a typical behavior applied by experienced Internet users and has been shown to improve the solving of web-related tasks (Hölscher and Strube, 2000; White et al., 2009). In this context, Wineburg and McGrew (2016) emphasize the importance of lateral reading: They found that if professional fact checkers have to evaluate websites, they quickly open various other tabs to verify the information with other sources. Other groups like students, who focused more on single websites and their features without cross-checking the content on other websites, performed worse in the given tasks.

Moreover, it is not only important that a variety of different sources is considered, but also that the information comes from sources that can be trusted (Bråten et al., 2011). In today's digital information environment, where everyone can publish and spread information online (e.g., via Blogs, Wikis, or Social Network Sites), the Internet is also used to disseminate misinformation or other manipulative content (Zimmermann and Kohring, 2018). Thus, users need to be able to effectively evaluate the sources they use when searching for online information (Brand-Gruwel et al., 2017). Several studies confirmed that experts from various domains such as history, finance or health pay much more attention to the authors and sources of online information than novices (Stanford et al., 2002; Bråten et al., 2011), which ultimately contributes to a higher task score (Brand-Gruwel et al., 2017).

In sum, it can be assumed that web search behavior is decisive for successfully solving online information problems and tasks. More precisely, the number of search queries, the number of

visited websites as well as the type of the sources used may contribute to a higher task score. To gain a deeper insight into students' web search behavior, we investigated how it can be described regarding the number and type of visited websites.

As studies have shown, a person's information seeking behavior is influenced by various factors such as their information needs (Tombros et al., 2005). One important variable is task complexity, which can influence the search process in that, as complexity increases, searchers make more search queries and use more sources of information (Kim, 2008). In studies, the search behavior of the users was also affected by whether a task had a clear answer or was rather open-ended (Kim, 2008). Beyond the task, there are inter-individual preferences for search strategies among users, as various attempts to assign users to different search behavior profiles show (Heinström, 2002). Similarly, differences between users and between tasks can be assumed with regard to the preferred types of websites. Therefore, it can be assumed that:

*H1: There are differences related to both student characteristics and CORA task characteristics in students' web search behavior in terms of the number and type of websites and webpages students used to solve these CORA tasks.*

## Website Characteristics

Web search behavior refers both to judging the different websites as well as to evaluating the information the websites provide (Roscoe et al., 2016). In this respect, content quality can be evaluated on different levels of website content. On a more formal level, the topic of interest should be exhaustively covered, in terms of the amount/scope of information provided as well as the variety sources referenced, for the learner to gain knowledge and a broad understanding of a topic (Gadiraju et al., 2018). For example, in the field of digital news media and political knowledge, findings show that exposure to established news sites, which provide full-length articles and usually disclose the sources they use, has positive outcomes for the gain of information (Dalrymple and Scheufele, 2007; Andersen et al., 2016) while the use of social network sites as a source for information has no (Dimitrova et al., 2011; Feezell and Ortiz, 2019) or even negative (Wolfsfeld et al., 2016) effects. Unlike established news sites, social network sites only provide news teasers and thus cover only selected aspects of a topic that are not necessarily verified (Wannemacher and Schulenberg, 2010; Guess et al., 2019), which might explain the different effects in terms of information gain.

Additional criteria that are brought up frequently when judging the quality of journalistic content usually focus on specific content features. From a normative point of view, news "should provide citizens with the basic information necessary to form and update opinions on all of the major issues of the day, including the performance of top public officials" (Zaller, 2003, p. 110). To fulfill this purpose, media content has to be objective. That means media content should be neutral, meaning without any kind of bias that is manipulative, for example by favoring political actors or taking a certain side on a controversial issue (Kelly, 2018). Closely related to neutrality is *balance*, i.e., media should cover a topic by mentioning different points of view

(Steiner et al., 2018). This is especially important for solving social conflicts in a democracy (McQuail, 1992) but also for increasing the knowledgeability of citizens (Scheufele et al., 2006). Another facet of objectivity concerns *factuality*, i.e., media content should be based on relevant and true facts that can be verified.

In sum, a neutral and balanced coverage based on facts is a prerequisite for an informed citizenry that possesses the necessary knowledge to form own opinions. Therefore, these are important indicators for the quality of news and online media (McQuail, 1992; Gladney et al., 2007; Urban and Schweiger, 2014). Considering hypothesis *H1*, that students' website preferences depend on the task, the quality of the used websites likely varies by task as well. Therefore we assume:

*H2: The quality of the used websites varies substantially between the different CORA tasks.*

## Students' Internet Search Behavior, Website Characteristics, and COR

With regard to the number and type of websites visited, previous findings have shown that relying on several trustworthy websites and the effective use of search engines is a web search behavior that is typically applied by experts and improves scores in online information seeking tasks (Hölscher and Strube, 2000; White et al., 2009). Wineburg et al. (2016), for instance, also emphasize the importance of using a multitude of different websites while searching for information. In their study, professional fact checkers, who verified information from a website with a variety of other sources, performed much better than other groups like students, who focused more on single websites and their features without cross-checking the content on other websites. Regarding COR, it can be assumed that there is a comparable relationship with search behavior. Consequently, we assume:

*H3: During students' web search, a larger number and variety of websites used by a student is positively correlated with a higher COR score, compared to using fewer websites.*

Since COR includes not only the correct evaluation of websites using other sources but also the critical handling of website content and the integration of the information found into a final judgment, it also relates to the content that websites provide. As websites differ with regard to content quality, they provide different baselines for COR. Regarding political information, for instance, if media coverage was too short, not exhaustive enough and from unreliable sources like social network sites, there was no (Dalrymple and Scheufele, 2007; Andersen et al., 2016) or even a negative effect on information gain (Wolfsfeld et al., 2016). The same applies to normative quality criteria such as neutrality, balance and facticity, which were positively related to information gain (Scheufele et al., 2006). Therefore, it can be assumed that:

*H4: There is a positive correlation between the quality of the media content students used to solve the CORA tasks and their COR score, i.e., higher quality corresponds to a better COR score.*



## MATERIALS AND METHODS

### Assessment of Critical Online Reasoning

To measure critical online reasoning (COR) we used five newly developed computer-based performance tasks (hereafter referred to as CORA) which were adapted from the US-American Civic Online Reasoning assessment (Wineburg and McGrew, 2016; for details on the adaptation, development and validation of CORA, see Molerov et al., 2019). Each task requires the participants to judge (a) whether a given website or tweet is a reliable source of information on a certain topic or (b) whether a given claim is true or untrue by performing a time-restricted open web search to respond to the task questions. In CORA, which comprises tasks on four different topics (Task 1: Vegan protein sources, Task 2: Euthanasia, Task 3: Child development, Task 4: Electric mobility, Task 5: Government revenue; for an example, see **Supplementary Appendix 1**), the participants had to evaluate the strengths and weaknesses of given claims, evaluate the credibility and reliability of different sources using any resources available online, and explain their judgments.

CORA aims to measure students' *generic* COR. Thus, the five tasks were designed in a way that, although they addressed certain social or political issues, students do not need prior content knowledge to answer the CORA tasks. Rather, each task prompt asks students to use the Internet to solve these tasks and formulate their responses as written statements. In particular, the prompts for tasks 1 and 3–5 provided a link to an initial website, which the test takers were asked to evaluate. The written responses in an open-ended format (short statements) to each task were scored according to a newly developed and validated rating scheme by two to three independent (trained) raters (for details, see section "Procedure").

### Procedure

To explore our hypotheses, we conducted a laboratory experiment. Prior to the survey, the students were informed that their web history would be recorded and that their participation in the following experiment was voluntary; all participants signed a declaration of consent to the use of their data for scientific purposes. Subsequently, the participants' socio-demographic data and media use behavior were measured with a standardized questionnaire (approx. 10 min). Afterward, participants were randomly assigned three out of the five CORA tasks to answer. For each task, the participants had a total time of 10 min to conduct the web search and write a short response (30 min in total).

Students were asked to use the preinstalled Firefox browser and when they closed the browser, their browser history was automatically saved using the "Browsing History View" feature for Windows. Every change to the URL, caused either by clicking on a link, entering a URL in the address line or searching with a search engine, was logged. By giving participants one-time guest access to the computers, we maintained their anonymity and ensured that their Internet search results were not affected by previous browser usage.

For the assessment of the students' performance, their written responses to the open-ended questions were scored by two independent human raters using a three-step rating scheme that was specifically developed based on the COR construct definition (McGrew et al., 2017). Using defined criteria, the raters judged whether the participants had noticed existing biases in the websites linked in the tasks and had made a well-founded judgment with regard to the question. This resulted in a score of 0, 0.5, 1, 1.5, or 2 points per answer (with 2 as the highest possible score). We then calculated the interrater reliability and averaged the scores of both raters for each participant and for each task, whereby a sufficient interrater reliability was determined, with Cohen's kappa = 0.80 ( $p = 0.000$ ) for the overall COR score. To analyze the log file data, we conducted a quantitative content analysis as described in the next section.

To test the hypotheses, we first analyzed the data descriptively. The correlations expected in the hypotheses were tested subsequently by means of correlation analyses, chi<sup>2</sup>-tests, one-way ANOVAs and *t*-tests. All analyses were conducted using Stata Version 15 (StataCorp, 2017) and SPSS 25 (IBM Corp, 2017).

### Content Analysis of Log Files

To analyze the log files from the CORA tasks with a content analysis (Früh, 2017), we developed a corresponding coding manual (see **Supplementary Appendix 2**). The basic idea of this methodological approach is to aggregate textual or visual data into defined categories. Thereby, the coding process needs to be conducted in a systematic and replicable way (Riffe et al., 2019). The essential and characteristic instrument for this process is the coding manual, which contains detailed information about the categories that are part of the analysis. Moreover, the coding manual also provides basic information about the purpose of the study and the units of analysis that are used to code the text material in the coding process (Früh, 2017).

In our study, we introduced two units of analysis: The first was the browsing history with a log file for each participant and task. First, formal information about the *test taker ID* and the *number of the task* were coded. This coding was followed by the *number of URLs (websites)* and the *number of subpages of URLs (webpages)* that the participant visited to solve the CORA tasks. The second unit of analysis were the individual URLs the participants visited to solve the tasks. Thus, the raters followed the links provided in the log file to access the information required in the different categories of the coding manual. The coding process for this unit of analysis also started with the coding of the *test taker ID* and the *number of the task* to enable data matching. Then, after following the link and inspecting the website (and webpages), raters had to code the *type of source*. To determine the characteristics for these categories, we selected sources that varied in their degree of reliability. As reliable sources, we considered journalistic outlets of both public service broadcasters<sup>1</sup> as well as established private news organizations<sup>2</sup>. Studies confirm that these outlets provide information with a high qualitative standard (Wellbrock, 2011; Steiner et al., 2018). Moreover, other reliable sources are those

<sup>1</sup>tagesschau.de

<sup>2</sup>spon.de

provided by governmental institutions (e.g., Federal Agency for Civic Education). Furthermore, we also considered scientific publications as reliable sources, which are especially important for finding online information in the context of higher education (Strømsø et al., 2013).

Additionally, we also coded the responses according to whether students relied on social media for solving the tasks during the study, as they are frequently used as learning tools in higher education. Here, especially *Wikipedia* plays an important role as an information provider (Brox, 2012; Selwyn and Gorard, 2016). The non-profit online encyclopedia invites everyone to participate as a contributor by writing or editing new or existing entries. Thus, although it can be considered the largest contemporary reference resource that is freely available to everyone, these entries are not made by experts and are published without review (Knight and Pryke, 2012). Another important social media channel for higher education is Facebook (Tess, 2013), which also offers some opportunities for learning since students can connect with each other and share information (Barczyk and Duncan, 2013). However, information spread over Facebook is often unreliable or even false (Guess et al., 2019). Moreover, Facebook is a tool that is predominately used to keep in touch with friends and thus rather increases distraction when it comes to learning (Roblyer et al., 2010). Thus, both *Wikipedia* and Facebook but also blogs, Twitter, YouTube, Instagram, and forums were considered in the coding manual. Sources which have commercial interests like online shops of shops of organizations were considered even less reliable than social media are. Their purpose is not to provide neutral information but to convince users to buy their products and to increase financial profits.

In sum, the list of source types consisted of websites that were part of the CORA task, social media sites, research institutes, websites of governmental institutions, news sites, sites of specialist magazines, scientific publications, book uploads, online shops or sites of organizations. If none of these categories matched the source used by a student, “other” was coded. For each category, further subcategories were provided to differentiate, for example, whether social media means that the participant visited either Facebook or *Wikipedia*. **Figure 1** shows what types of websites students accessed according to this coding. **Figure 2** shows one example where a participant accessed the following websites to find information for one of the CORA tasks.

Each of these websites is assigned a numerical code depending on the type of source. The first entry is *Wikipedia*, which has the value 22. The next URL goes back to a university website, which is coded with 31. The same is true for the website of the PThU, which is also a university and thus received the value 31. The next two links are both established news websites and are coded with 52. Finally, youtube.com is a social media site and has the value 25 (for the coding manual, see **Supplementary Appendix 2**).

In a next step, we analyzed the quality of the content of the online sources used by students when solving a CORA task. To evaluate content quality, certain indicators were identified. In this part of the coding process, the raters followed the links of the log file and applied the coding scheme to the text on each webpage the participant visited. Since the *amount of information (text) on a webpage* and the *use of external sources* are particularly important

for learning and understanding concepts, these indicators were considered as quality dimensions in our study. With regard to the *amount of information*, the raters examined which sections of the individual webpages dealt with the topic covered in the task and counted the number of words in those sections. For coding the number of external sources (scientific and non-scientific), links in the texts as well as sources mentioned at the end of the texts were also counted. If sources were mentioned in the text, it was additionally coded if the text addressed the credibility of these sources (0 – no/1 – yes). Other quality indicators that are important for learning and understanding, and have thus been considered in various studies on media quality, are *balance* and *facticity* (McQuail, 1992; Gladney et al., 2007; Urban and Schweiger, 2014). They too were part of the coding manual. For balance, if the text has a clear stance that takes the side of, for example, a certain actor or on an issue, 1 was coded. If the text was rather balanced or did not take any side, 2 was coded. Facticity addressed the relation of opinion and facts. Three different codes indicated if the text contained (almost) exclusively opinions (1), was balanced in this regard (2) or was (exclusively) based on facts (3).

The coding process was conducted by two human raters. One of these raters was responsible for the first recording unit (log file), the other rater for the second (content of the websites). Before starting the coding process, the raters were intensively trained by the researchers of this study. To provide a sufficient level of reliability, the researchers and the raters coded the same material and compared the results of the coding until the level of agreement between researchers and coders reached at least 80% for each category.

## Sample

The sub-sample used in this study consisted of 45 economics students from one German university and is part of the overall sample ( $N = 123$ ; see below) used in the overarching CORA study (see Molerov et al., 2019; Zlatkin-Troitschanskaia et al., 2020). Participation in the CORA study, which was voluntary, was requested in obligatory introductory lectures at the beginning of the winter semester 2018/2019, the summer semester 2019, and the winter semester 2019/2020. To ensure more intrinsic test motivation, for their participation in the study, the students received credits for a study module.

For this article, a selected sub-sample of the participant data was used, since the coding and analysis of all websites the CORA participants used to solve these tasks was hardly feasible for practical research limitations. When selecting this subsample, we included students from all study semesters represented in the overall sample. Another important criterion for the sampling was the students' central descriptive characteristics such as gender, age, migration background and prior education, which may influence students' web search behavior and COR task performance.

The majority of the 45 participants were at the beginning of their studies ( $m = 1.76$ ,  $SD = 1.45$ , with an average age of 21.5 years ( $SD = 2.82$ ) and an average school-leaving grade of 2.44 ( $SD = 0.62$ ); 60% of the participants were women.

The subsample used in this study, which is relatively large with a view to the comprehensive analysis conducted in this

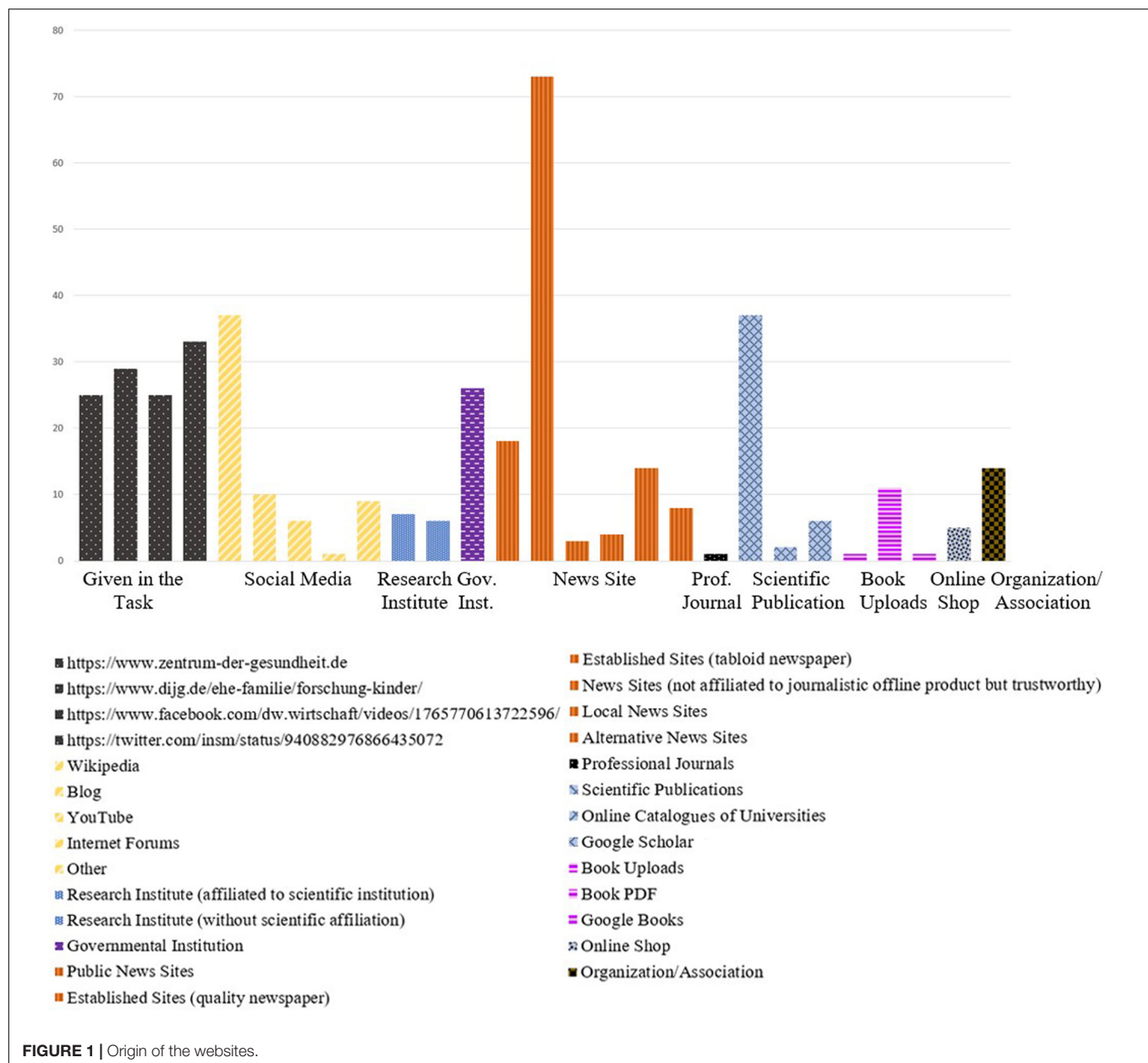


FIGURE 1 | Origin of the websites.

[https://de.wikipedia.org/wiki/Volker\\_Gerhardt](https://de.wikipedia.org/wiki/Volker_Gerhardt)  
<https://www.philosophie.hu-berlin.de/de/lehrbereiche/gerhardt/mitarbeiter/gerhardt>  
[https://www.pthu.nl/Over\\_PThU/Organisatie/Medewerkers/boer/Boer-Leben\\_bis\\_zuletzt-SDZ.pdf](https://www.pthu.nl/Over_PThU/Organisatie/Medewerkers/boer/Boer-Leben_bis_zuletzt-SDZ.pdf)  
<https://www.tagesspiegel.de/meinung/leserbriefe/sollte-sterbehilfe-erlaubt-werden/1272988.html>  
<https://www.welt.de/politik/deutschland/article106222145/Drei-Jahre-Haft-fuer-gewerbsmaessige-Suizid-Beihilfe.html>  
[https://www.youtube.com/watch?v=O9oghcz\\_Q\\_o](https://www.youtube.com/watch?v=O9oghcz_Q_o)

FIGURE 2 | Excerpt from a log file.

study, can be considered representative for both the total sample of the CORA study and for German economics students at the beginning of their studies in general: For instance, the overall

CORA sample consists of 123 students with 61% women, an average age of 22 years ( $SD = 2.82$ ), and an average study semester of 2.02 ( $SD = 1.82$ ). There are also hardly any differences



**TABLE 1** | Average number of visited websites and webpages per task.

Task	Sample	Websites (SD)	Webpages (SD)	Score (SD)
1	24	2.54 (SD = 1.77)	2.67 (SD = 2.35)	0.88 (SD = 0.82)
2	29	4.86 (SD = 1.87)	2.07 (SD = 1.79)	1.39 (SD = 0.62)
3	29	2.07 (SD = 1.13)	5 (SD = 4.18)	0.66 (SD = 0.64)
4	25	2.88 (SD = 1.67)	4.76 (SD = 4.91)	0.84 (SD = 0.53)
5	28	2.93 (SD = 1.80)	4.21 (SD = 4.28)	0.81 (SD = 0.40)

regarding the school-leaving grade, with an average of 2.41 (SD = 0.52). In comparison with a representative Germany-wide sample of 7111 students (Zlatkin-Troitschanskaia et al., 2019) in their first semester, there are also hardly any differences in terms of age ( $m = 20.41$ , SD = 2.69) and school-leaving grade ( $m = 2.37$ , SD = 0.57). Only the proportion of women in the German-wide representative sample is slightly lower (54%). However, since previous studies did not find any gender-specific effects on COR performance (Breakstone et al., 2019), this study assumes that there is no discrepancy of this kind that should be taken into consideration.

## RESULTS

### Students' Internet Search Behavior (H1)

To examine the online search behavior of students, both the number of websites and webpages as well as the type of websites students visited while solving the CORA tasks were examined. The findings are displayed in **Table 1**. On average, the number of visited websites was highest for task 2 and lowest for task 3. With regard to the visited webpages of a website, however, the pattern was exactly the opposite: Task 3, for example, had on average the most webpages visited per participant, whereas task 2, with  $m = 2.07$  (SD = 1.79), had the fewest. This pattern shows that for task 2, participants rather relied on several websites, while for the other tasks they browsed the webpages of websites more intensively.

With regard to the type of source, a descriptive analysis showed substantial differences between the tasks. In tasks 1, 3, 4, and 5, for instance, one of the most frequently visited pages was the one given in the task prompt ( $n = 24$  in task 1 –  $n = 33$  in task 5), whereas no website was given as a starting point for task 2 (see **Table 2** and **Figure 1**).

All students complied with the task and accessed the pages from the task prompts at least once. By contrast, for task 2 which did not have a given start page, pages in the categories social media ( $n = 21$ ), news ( $n = 65$ ) and scientific publications ( $n = 34$ ) were used more frequently than for the other tasks. In the other tasks, significantly fewer pages of these categories were visited: Social media pages were used four (task 3) to 17 times (task 5), news pages seven (task 5) to 31 times (task 4), and scientific publications or online catalogs were used not at all (task 4) to 5 five times (task 3).

Moreover, significantly more news pages were used for task 4 ( $n = 31$ ) than for tasks 1, 3 and 5 ( $n = 7 - n = 9$ ), and more governmental web pages for task 5 ( $n = 13$ ) than for the other

**TABLE 2** | Number of accessed websites by website type and per task<sup>1</sup>.

Type of website	Task					Total
	1	2	3	4	5	
1 Given in the task	24	0	29	25	33	112
2 Social media	12	21	4	9	17	63
3 Research Institute	2	4	4	0	3	13
4 Gov. Institution	1	7	4	1	13	26
5 News site	9	65	8	31	7	120
6 Professional journal	0	1	0	0	0	1
7 Scientific publication	3	34	5	0	3	45
8 Book uploads	4	2	3	2	2	13
9 Online shop	3	2	0	0	0	5
10 Other	3	2	3	3	3	14
	61	139	60	71	81	<b>N = 412</b>

<sup>1</sup>Referring to the number of individual websites that all participants visited across all tasks.

tasks ( $n = 1$  in tasks 1/4 –  $n = 7$  in task 2). Overall, the participants most frequently used news pages ( $n = 120$ ), pages from the tasks ( $n = 112$ ), social media ( $n = 63$ ), and scientific publications ( $n = 45$ ). An examination using the  $\chi^2$ -test confirmed that students' use of website categories significantly differed between the tasks ( $\chi^2 = 180.81$ ,  $df = 36$ ,  $p = 0.00$ ).

Since the differentiation observed so far was still rather rough, we conducted a more precise analysis based on the types of online sources most frequently used by the participants. In the case of news sites, the sites used by far most frequently (60.8%) were established news sites that can be assigned to a high-quality (national) newspaper or magazine<sup>3,4</sup> (see **Table 3**). Public broadcasting news sites<sup>3</sup> (e.g., 15%) and local news sites<sup>4</sup> were accessed occasionally (11.7%). In contrast, other sites that cannot be assigned to any journalistic offline product and/or that disseminate rather unreliable news were used little or not at all. When selecting news sites for research, students apparently mainly relied on well-known and established national and local news sites, while avoiding possibly less well-known sites without offline equivalents and well-known unreliable sites.

In the next most frequently used category, social media, the students' preference for Wikipedia was evident at 58.5% (see **Table 4**), followed by blogs (15.9%) and various sources that cannot be specifically assigned (14.3%). YouTube was also used a few times (9.5%), while Facebook, Twitter, Instagram, and forums were hardly or not at all used.

The third most frequently used type of website, scientific publications or online catalogs, was almost exclusively used in the form of pages of scientific journals (82.2%); online catalogs of universities (4.44%) or Google Scholar (13.33%) were hardly used at all. Overall, the analyses pertaining to H1 show clear differences in students' search behavior. Hypothesis H1 was therefore not rejected.

<sup>3</sup>faz.net

<sup>4</sup>zeit.de



**TABLE 3 |** Number of accessed news websites by subtype and per task<sup>1</sup>.

Type of Source	N = 120	
	Freq.	Percent
Public News Sites	18	15.00
Established Sites of a Quality Newspaper	73	60.83
Established Sites of a Tabloid Newspaper	3	2.5
News Sites not Affiliated to any Journalistic Offline-Product (Trustworthy News)	4	3.33
Local News Sites	14	11.67
Alternative News Sites	8	6.67
	120	100.00

<sup>1</sup>Referring only to the sub-group of news websites that the participants visited.

**TABLE 4 |** Number of accessed social media websites by subtype and per task<sup>1</sup>.

Type of Source	N = 63	
	Freq.	Percent
Wikipedia	37	58.73
Blog	10	15.87
YouTube	6	9.52
Online Forums	1	1.59
Other	9	14.29
	63	100.00

<sup>1</sup>Referring only to the sub-group of social media websites that the participants visited.

## Quality of the Visited Websites (H2)

The second analyzed aspect of students' web search behavior concerns the quality of the webpage content that the respondents visited to solve the CORA tasks. In terms of length, participants visited websites that provided on average  $m = 2,448.81$  (SD = 4,237.06) words that were relevant for the CORA tasks. An examination of the average number of words on the webpages used for each task showed that task 3 stands out. Here, the webpages provided an average number of 4,865.31 (SD = 6,631.01) words, followed by webpages visited to solve task 2 with  $m = 2,797.67$  (SD = 5,230.25) words, task 1 with  $m = 1,748.39$  (SD = 2,126.10) words, task 4 with  $m = 1,559.41$  (SD = 1,565.11) words and task 5 with  $m = 1,458.41$  (SD = 1,549.62) words. A one-way ANOVA of the extent of the webpages grouped by task revealed that there are significant differences between the tasks [ $F(4) = 7.70$ ;  $p < 0.001$ ]. A comparison between the groups with a *post hoc* test (Tamhane T2) showed that task 3 significantly differs from all other tasks except for task 2 ( $p < 0.05$ ). All other group comparisons were not significant.

Another quality dimension concerned external sources (see Table 5). Here, we investigated whether scientific sources or non-scientific sources were mentioned on the accessed webpages and whether the response text discussed the quality of the used external sources. Concerning the first two categories, on average, the websites used to respond to the tasks provided  $m = 6.92$

**TABLE 5 |** Average number of scientific and non-scientific sources per task<sup>1</sup>.

Task	Average number of scientific sources (SD)	Average number of non-scientific sources (SD)
1	7.48 (SD = 21.14)	3.09 (SD = 5.55)
2	6.38 (SD = 12.64)	2.88 (SD = 6.52)
3	22.16 (SD = 42.80)	3.90 (SD = 11.32)
4	0.13 (SD = 0.34)	4.28 (SD = 12.86)
5	2.19 (SD = 5.34)	8.41 (SD = 13.51)

<sup>1</sup>Referring only to the sub-group of scientific/non-scientific websites that the participants visited.

(SD = 21.12) scientific sources and  $m = 4.48$  (SD = 10.45) non-scientific sources.

The analysis of the average number of sources per task showed that for scientific sources, task 3 stands out. There, participants relied on websites that provided many scientific sources. For task 4, participants rather relied on sources with almost no scientific sources. According to a one-way ANOVA with the number of scientific sources as dependent and the respective tasks as grouping variable, the tasks had a significant effect for the number of scientific sources used [ $F(4) = 11.57$ ;  $p < 0.001$ ]. The *post hoc* tests (Tamhane T2) showed that except for task 1, all other tasks differed at least marginally from each other ( $p < 0.10$ ). Concerning non-scientific sources, participants preferred websites with a higher number of this kind of sources for task 4 and 5. All other mean values were rather similar. Here, the effect of the task was also significant [ $F(4) = 3.87$ ;  $p < 0.01$ ]. However, according to the *post hoc* tests (Tamhane T2), only the differences between task 5 and task 2 as well as task 5 and task 1 were significant ( $p < 0.05$ ).

Finally, we analyzed whether the visited webpages referenced additional sources that could allow for conclusions about the credibility of the websites to be drawn. On average, this was the case for 52 of the 379<sup>5</sup> analyzed webpages. If the tasks are also considered, for task 2, no webpage was used that addressed the reliability of the sources. For all other tasks, the share of webpages discussing the external sources was on a comparable level that varied between 21.2 and 28.8%. An examination using the chi<sup>2</sup>-test confirmed that there are highly significant differences ( $\chi^2 = 28.67$ , df = 4,  $p = 0.00$ ).

Further, we also investigated whether an article was balanced and based on facts. Concerning balance, 138 of the 379 (36.4%) webpages were rather unbalanced while the rest was classified as balanced. If the tasks were taken into account, for task 1, the share of webpages with rather unbalanced information was high with 70.7%, while for task 2 this share was rather low (10.6%). For task 3, about a third (32.8%) of the webpages was rather unbalanced while the share for task 4 was 46.5 and 41.8% for task 5. A chi<sup>2</sup>-test confirmed that the differences were highly significant ( $\chi^2 = 66.32$ , df = 4,  $p = 0.00$ ).

<sup>5</sup>This number differed from the number of 412 analyzed websites mentioned before because 33 of the links were expired when the content analysis was conducted. Even if, for example, the type of source could still be identified, the content of the websites could not be used for the content analysis.

**TABLE 6** | Level of facticity of the visited websites per task<sup>1</sup>.

Task	Mere opinion	Balanced	Mere facts	Total
1	33 (56.9%)	13 (22.4%)	12 (20.7%)	58
2	3 (2.7%)	58 (51.3%)	52 (46.0%)	113
3	14 (24.1%)	24 (41.4%)	20 (34.5%)	58
4	31 (43.7%)	32 (45.1%)	8 (11.3%)	71
5	31 (39.2%)	26 (32.9%)	22 (27.8%)	79
Total	112 (29.6%)	153 (40.4%)	114 (30.1%)	<b>N = 379</b>

<sup>1</sup>Referring to the number of individual websites that all participants visited across all tasks.

The last category was facticity (see **Table 6**). If the tasks were not considered, the participants chose roughly equal shares of websites that were more opinion-based, balanced, and fact-based. However, this depended again on the tasks, and every task had a different pattern. Participants chose almost no opinion-based websites for task 2 but rather preferred balanced or fact-based content. The highest values for opinion-based content were found for task 1, followed by tasks 4 and 5. In task 3, participants preferred especially balanced websites. A  $\chi^2$ -test confirmed that the differences were highly significant ( $\chi^2 = 79.70$ ,  $df = 8$ ,  $p = 0.00$ ). Based on the analysis results, H2 was not rejected.

### Correlation of Search Behavior With the COR Score (H3)

The descriptive examination of the average scores per task showed substantial differences between the tasks, with participants scoring best on average for task 2 with 1.39 points and worst for task 3 with 0.66 points (see **Table 2**). There were significant positive correlations between the number of websites visited and the task scores, for task 1 ( $n = 48$ ,  $r = 0.59$ ,  $p = 0.000$ ), task 3 ( $n = 52$ ,  $r = 0.33$ ,  $p = 0.02$ ) and task 5 ( $n = 53$ ,  $r = 0.32$ ,  $p = 0.02$ ). Even if no significant effects were found for items 2 and 5, a correlation of the total number of websites used by the participants with their summed up overall scores confirmed the overarching tendency that the use of a larger number of websites in the processing of CORA tasks was associated with a higher CORA score ( $n = 87$ ,  $r = 0.49$ ,  $p = 0.000$ ). Overall, the results indicated at least the tendency that visiting more websites during the search was associated with a better CORA test performance. Thus, H3 was not rejected.

### Relationship Between the Quality Characteristics of Visited Websites and COR Score (H4)

When analyzing the relationship between the types of websites used by students and their CORA task score, a corresponding single-factor analysis of variance with the ten groups of websites as independent variable and the total score of the participants as dependent variable was just barely not significant ( $p = 0.06$ ).

For a comprehensive analysis of the correlation between different aspects of search behavior, we considered whether the score was different between students who only visited the websites specified in the task and students who visited additional

websites. To obtain the overall score, each participant's scores on all three completed tasks were added up, resulting in ranging between 0 and 6 points. A corresponding  $t$ -test for the overall score of both groups showed that the students who visited additional websites, on average, achieved a significantly higher total score of 3.20 points than the students who only stayed on websites linked in the CORA tasks (2.72 points,  $p = 0.002$ ).

To analyze the effects of the quality characteristics of the website content on the score, the categories defined in the coding manual (length, use of scientific sources, use of non-scientific sources, discussion of the sources, balance and facticity) were examined using correlation analyses. Here, we only found a significant correlation for the discussion of external sources and the overall score  $r = 0.22$ , ( $p = 0.000$ ). All in all, with regard to the characteristics of the visited websites, two characteristics in particular had a significant correlation with the COR results: While participants who only remained on the websites specified in the tasks performed worse, the use of websites that critically report sources had a positive effect. In sum, H4 had to be partly rejected.

## DISCUSSION

### Interpretation of the Results

For this study, students' critical online reasoning (COR) was assessed using open-ended performance tasks, and their web search behavior was analyzed using log files that recorded their actions while solving the tasks. Concerning the CORA task performance, the students show a low level of skill in judging the reliability of websites, which confirms previous findings (Wineburg et al., 2016; Breakstone et al., 2019). The students' web search behavior differs between tasks, and the type and wording of the task appears to have a noticeable correlation with the students' search behavior. The identified differences may be explained by the specific characteristics of the respective tasks, as some of the tasks, for instance, referred to everyday topics frequently addressed on the news (e.g., task 4). When solving tasks that included a link to a website, the participants tended to spend more time on these websites and look at a larger number of subpages, visiting fewer or no additional websites in the free web search. In contrast, the students visited significantly more websites while solving task 2, which had not included any links to websites.

Furthermore, the students were found to have preferences for certain types of websites, especially news sites, social media (Wikipedia), and scientific journals, both across all CORA tasks and with regard to the individual tasks, whereas other website types such as blogs and online shops were neglected. Taking into account the limited processing time, it can be assumed that after reading the task and visiting the corresponding website, the participants tried to gain an overview by visiting well-known news sites, scientific journals, and online encyclopedias. The deviations in task 2 also show that the content of the task prompt and, in particular, whether a link to a website was included therein appears to correlate with the resulting search behavior.

The findings confirm that students have a strong preference for Wikipedia as a source of information (Maurer et al., 2020). This indicates that even though the students could certainly pay more attention to scientific sources and should not rely on Wikipedia as much as they do, they at least refrain from using completely unreliable sources such as alternative news sites, online shops, or Facebook. Wikipedia has a special status in this regard. It has been repeatedly proven to be a reliable source of information, which studies have attributed to the collaboration between Wikipedia users. However, the fact that Wikipedia articles are often written as a collaborative effort between numerous users is also the reason why the credibility of the Wikipedia articles cannot be guaranteed (Lucassen et al., 2013).

Overall, students tend to rely on sources they also typically use to gain information in their everyday life (Beisch et al., 2019) and might know from a university context (Maurer et al., 2020). This is consistent with earlier findings that indicated that people prefer to search for information on websites they are familiar with through their own experience or that are generally well known, and that students in particular are more likely to use a limited range of media or sources, depending on the nature of the task and immediacy considerations (Oblinger and Oblinger, 2005; Walraven et al., 2009; List and Alexander, 2017). In our study, the students may have tried to avoid wasting time by visiting uninformative websites or unknown websites.

The fact that the students had a tendency to spend a large amount of time on the websites linked in the CORA tasks also confirms prior findings (e.g., Flanagan and Metzger, 2007; Kao et al., 2008; Hargittai et al., 2010; Wineburg et al., 2016), where the study participants (with the exception of professional fact checkers) showed a tendency to focus on individual websites and their features. This tendency may have been more pronounced in the context of the CORA due to the time limit of 10 min. This time limit may have also caused a tendency to neglect online catalogs of universities and Google Scholar, as the students did not have the time to read in-depth scientific articles. In this respect, our study is in line with prior findings that students also consult non-scientific sources when they need information in regular university life (e.g., when preparing for exams) and have more time (Maurer et al., 2020). With regard to scientific sources, the question arises how elaborate findings can be presented in a more comprehensible way and how it can be ensured that they are easier to understand for a wider audience by paying more attention to the needs of the readership.

Concerning the quality of the websites' content, the participants tended to rely on news based on both scientific and non-scientific sources. This especially holds true for task 3, where the participants used a larger number of scientific sources and wrote significantly more words than they did when solving the other tasks. A possible explanation for this finding might be that the topic of the task (Child development) has a more scientific background than the other tasks, which makes it necessary to rely on more comprehensive and more scientific websites. This is an important finding, as sources with these kinds of characteristics also provide a suitable basis for learning (Dalrymple and Scheufele, 2007; Dimitrova et al., 2011). For the other two quality dimensions, balance and facticity, we find

a more mixed picture. In general, the students tended to also take into consideration unbalanced and opinion-based sources. However, this depends on the specific task. In this context, it has to be considered that some of the tasks contained links to websites that were categorized as unbalanced and opinion-based. Taking this into account, it can be concluded that students appear to know how to find websites with reliable and fact-based content, even though there is still some room for improvement.

Moreover, as assumed, we found a relationship between the number of visited websites (as an indicator of the students' web search behavior) and a higher CORA task performance, with the exception being task 2, where no significant correlation was found. This could be due to the fact that no links to websites were included in this task. In the other four CORA tasks that did include links to specific websites, these websites were usually either biased or only of limited reliability. Thus, spending a large amount of time on these websites alone may have a significantly negative effect on the final responses of the participants, whereas visiting other websites could cause them to detect the bias. In task 2, however, this situation did not apply, as no link was included in the task. Since the students with their heuristic approach of referring to well-known (news) sites and encyclopedias largely avoided particularly unreliable or biased websites when working on the CORA tasks, there may have been less of a correlation between the total number of websites visited and the quality of the students' task responses in task 2. Studies from the field of political education in particular indicate that established news sites that offer full-length articles can have a positive effect on knowledge acquisition (Dalrymple and Scheufele, 2007; Andersen et al., 2016).

Another finding is also indicative of a relationship between the task definition, search behavior, and score: Students who visited additional websites on average achieved a significantly higher COR score than those who only looked at the websites (and webpages) mentioned in the tasks. In addition, with regard to the quality dimension of facticity, the amount of "purely opinion-based argumentation" expressed in task 2 was particularly low and the amount of "purely facts-based argumentation" particularly high (see **Table 6**), while the average test score was significantly higher in task 2 than in the other tasks. These findings are consistent with prior research (e.g., Anmarkrud et al., 2014; Wineburg et al., 2016; List and Alexander, 2017), indicating that it is of great importance to at least check the reliability of a website and its contents and to cross-check the information stated on a website with that stated on other websites.

Contrary to our second hypothesis (H2), however, we hardly found any correlations between other website characteristics such as facticity or scientific/non-scientific sources and the students' test score. One reason for this finding could be that the participants used only a limited variety of different websites and had only a limited amount of time to perform their online searches, so that certain website characteristics such as the extent of task-related content included therein were not fully considered. In addition, it is not clearly evident from the log data how much time the participants spent on the individual webpages and which sections of these webpages they actually read, whereas the ratings based on the coding manual always refer to entire

webpages. Thus, higher correlations may have been determined for certain characteristics if only the sections of the webpages that the students actually read had been taken into account. The aspect of “number of scientific/non-scientific sources,” however, may be of limited use as a quality feature of a website, as biased websites can also use external sources to convey the impression that they are credible sources of information.

In conclusion, the students showed a great heterogeneity both in terms of their Internet use and their performance on the CORA tasks. Taking into account task-specific characteristics (in particular the wording of the task), the most frequently used websites were the ones that had already been included in the tasks as well as news sites, social media sites, and scientific websites. In particular, preferences for established news sites, websites of scientific journals, and Wikipedia were found. The quality of the visited websites also varied and depended on the task that the test-takers were working on. On average, the websites provided a large amount of relevant information and used both scientific and non-scientific sources. However, only a rather small number of websites critically reflected on their sources, and the participants showed a preference for websites that were based on opinions. When it comes to the relationship between these content features and the scoring, we found a positive relationship between the score and the number of visited websites, and the use of additional websites beyond the ones already included in the tasks. Although no significant correlation between the type of website used and the students' CORA test performance could be determined, using websites that critically reflect on their sources also increased the students' test performance.

## Limitations

There are also some limitations to our study that should be taken into account when interpreting the results. For instance, this paper used a subsample of the CORA study consisting only of students in the first phase of their economics studies at one German university. As Maurer et al. (2020) indicate, students' web search behavior may differ between different study domains and universities or change over the course of study. For example, Breakstone et al. (2019) found that students with more advanced education performed better on tasks on civic online reasoning. Therefore, both search behavior and CORA performance will be analyzed in a larger and more heterogeneous sample in follow-up studies to confirm the representativeness of the results found in this study.

The possible influences of personal characteristics on students' Internet search behavior and CORA performance were not considered in this paper. Previous studies found correlations between, for example, the influence of ethnicity and socio-economic background on civic online reasoning (Breakstone et al., 2019). The extent to which these correlations between the participants' personal characteristics and their CORA performance can be replicated and whether they also have an effect on search behavior is being clarified in further studies. With regard to the implications for university teaching, for example, it would be useful to learn more about the students' prior knowledge of strategies for searching information on the Internet

(acquired through, e.g., previously attended research courses at the university) and how they deal with misinformation.

To better understand the correlation between website characteristics and the students' performance on CORA tasks, their search processes should be analyzed in more detail, for instance, based on data from an eye-tracking study. In particular, the duration and frequency of the individual webpage visits should be included in further analyses and it should be examined in more detail which sections of the visited webpages the students focused on and what exactly they did there (e.g., reading certain sections). A useful empirical extension of our study might be to include a more qualitative approach to investigate the students' web search behavior. For example, using the think-aloud method (Leighton, 2017) would reveal in more detail which strategies the students apply when searching for information online, how they choose their sources, and how they judge the content of websites. For this purpose, it would be helpful to use experimental study designs that focus on examining specific aspects of the web search, for example, how trainings focused on using different web browsers and different search interfaces affect the search process. This might be considered in future research designs.

## Implications

In addition to the implications for future studies resulting from the limitations described above, further implications can be derived from the findings of this study, especially for (university) teaching. In particular, the overall rather poor CORA performance of the students confirms that there is a clear need for support when it comes to dealing with online information in an appropriate way (e.g., Allen, 2008). This is of particular importance, as the Internet is the main source of information for students enrolled in higher education (Maurer et al., 2020). Thus, COR should be promoted, for example, by offering courses on web search strategies at the university library and by fostering COR skills in lectures or seminars in a targeted manner. Students should not only be taught suitable strategies for searching the Internet but also criteria and techniques for judging the credibility of (online) sources and information (e.g., Konieczny, 2014).

Seeing as web searches are firmly ingrained in teaching and task requirements in the university context nowadays, it is important to consider the influence the prompt of a task can have on students' search behavior and their COR. The identified effects of whether or not a link to a specific website was included in the task prompt on the students' search behavior should also be taken into account, both in future studies and with regard to designing new exercises in teaching. Although first efforts and successes have been reported when it comes to promoting these skills through targeted intervention measures aimed at students (McGrew et al., 2019), COR is still not an integral part of teaching at many universities (Persike and Friedrich, 2016). There is, therefore, an urgent need for instructional action in this context, especially as studies indicate that the specific teaching methods of individual universities have an influence on the students' use of online media and on which sources they tend to use (Persike and Friedrich, 2016).



## CONCLUSION

This study provides an insight into the so far under researched relationship between students' web search behavior when evaluating online sources, the characteristics of the visited websites, and the information the students used. Our findings provide insights not only into students' preferences for certain types of websites in online searches and their quality, but also into the relationship between the characteristics of these websites and students' performance in the CORA tasks. In this respect, this study contributes to previous research, which had been mainly focused on students' website preferences for learning or for private purposes and, moreover, often collected this data through self-reports or in a simulated test environment. In particular, while previous studies on students' abilities and skills related to COR (e.g., searching and evaluation strategies) had been primarily focused on the test results (i.e., the score) and to what extent it is influenced by personal characteristics, this contribution analyzes the connection between students' search behavior during task processing and characteristics of the particular website they used in more detail.

Based on the unique analyses and results, this study highlights that there is a clear need for students to receive targeted support in higher education, which should be urgently addressed by implementing appropriate measures, as the ability to use online resources and critical online reasoning in a competent manner constitute not only an important basis for academic success but also for lifelong learning and for participation in society as an informed citizen.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation

## REFERENCES

- Allen, M. (2008). Promoting critical thinking skills in online information literacy instruction using a constructivist approach. *College Undergr. Libr.* 15, 1–2. doi: 10.1080/10691310802176780
- Andersen, K., Bjarnø, C., Albæk, E., and de Vreese, C. H. (2016). How news type matters. Indirect effects of media use on political participation through knowledge and efficacy. *J. Media Psychol.* 28, 111–122. doi: 10.1027/1864-1105/a000201
- Anmarkrud, Ø., Bråten, I., and Strømsø, H. I. (2014). Multiple-documents literacy: strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learn. Individ. Diff.* 30, 64–76. doi: 10.1016/j.lindif.2013.01.007
- Barczyk, B. C., and Duncan, D. G. (2013). Facebook in higher education courses: an analysis of students' attitudes, community of practice, and classroom community. *CSCCanada* 6, 1–11. doi: 10.3968/j.ibm.1923842820130601.1165
- Beaudoin, M. F. (2002). Learning or lurking? *Internet High. Educ.* 5, 147–155. doi: 10.1016/s1096-7516(02)00086-6

and institutional requirements. The participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

M-TN and SvS wrote the article and conducted the analyses. OZ-T developed the idea for the study, developed the assessment, supervised the analyses, and wrote the article. CS and MM developed the idea for the study, supported the development of the coding manual, and supervised the analyses. DM and SuS developed the assessment and the rating scheme. SB wrote the article and supervised the analyses. All authors contributed to the article and approved the submitted version.

## FUNDING

This study is part of the PLATO project, which was funded by the German Federal State of Rhineland-Palatinate.

## ACKNOWLEDGMENTS

We would like to thank the two reviewers and the editor who provided constructive feedback and helpful guidance in the revision of this manuscript. We would also like to thank all students from the Faculty of Law and Economics at Johannes Gutenberg University Mainz who participated in this study as well as the raters who evaluated the written responses.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2020.565062/full#supplementary-material>

- Beisch, N., Koch, W., and Schäfer, C. (2019). ARD/ZDF-Onlinestudie 2019. Mediale internetnutzung und video-on-demand gewinnen weiter an bedeutung. *Media Perspekt.* 9, 374–388.
- Blossfeld, H. P., Bos, W., Daniel, H. D., Hannover, B., Köller, O., Lenzen, D., et al. (2018). *Digitale Souveränität und Bildung*. Gutachten: Waxmann.
- Braasch, L. G., Bråten, I., and McCrudden, M. T. (2018). *Handbook of Multiple Source Use*. London: Routledge.
- Brand-Gruwel, S., Kammerer, Y., van Meeuwen, L., and van Gog, T. (2017). Source evaluation of domain experts and novices during Web search. *J. Comput. Assist. Learn.* 33, 234–251. doi: 10.1111/jcal.12162
- Brand-Gruwel, S., Wopereis, I., and Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Comput. Educ.* 53, 1207–1217. doi: 10.1016/j.compedu.2009.06.004
- Bråten, I., Strømsø, H. I., and Salmeron, L. (2011). Trust and mistrust when students read multiple information sources about climate change. *Learn. Instr.* 21, 180–192. doi: 10.1016/j.learninstruc.2010.02.002
- Breakstone, J., Smith, M., and Wineburg, S. (2019). *Students' Civic Online Reasoning. A National Portrait*. Available online at: <https://stacks.stanford.edu/>

- file/gf151tb4868/Civic%20Online%20Reasoning%20National%20Portrait.pdf (accessed May 16, 2020).
- Brooks, C. (2016). *ECAR study of students And Information Technology*. Louisville: ECAR.
- Brox, H. (2012). The elephant in the room. A place for wikipedia in higher education? *Nordlit* 16:143. doi: 10.7557/13.2377
- Bullen, M., Morgan, T., and Qayyum, T. (2011). Digital learners in higher education. Generation is not the issue. *Can. J. Learn. Technol.* 37, 1–24.
- Carbonell, X., Chamorro, A., Oberst, U., Rodrigo, B., and Prades, M. (2018). Problematic use of the internet and smartphones in University Students: 2006–2017. *Int. J. Environ. Res. Public Health* 15:475. doi: 10.3390/ijerph15030475
- Ciampaglia, G. L. (2018). “The digital misinformation pipeline,” in *Positive Learning in the Age of Information*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Wiesbaden: Springer), 413–421. doi: 10.1007/978-3-658-19567-0\_25
- Dalrymple, K. E., and Scheufele, D. A. (2007). Finally informing the electorate? How the Internet got people thinking about presidential politics in 2004. *Harvard Int. J. Press Polit.* 12, 96–111. doi: 10.1177/1081180X07302881
- Dimitrova, D., Shehata, A., Strömbäck, J., and Nord, L. W. (2011). The effects of digital media on political knowledge and participation in election campaigns. *Commun. Res.* 41:1. doi: 10.1177/0093650211426004
- Eppler, M. J., and Mengis, J. (2004). The concept of information overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines. *Inform. Soc.* 20, 325–344. doi: 10.1080/01972240490507974
- European Commission (2018). *Fake News and Disinformation Online*. Brussels: European Commission.
- Facione, P. A. (1990). *Critical Thinking: A Statement of Expert Consensus on Educational Assessment and Instruction. Research Findings and Recommendations*. Newark, NJ: American Philosophical Association (ERIC).
- Feezell, J., and Ortiz, B. (2019). ‘I saw it on Facebook’. An experimental analysis of political learning through social media. *Inform. Commun. Soc.* 92:3. doi: 10.1080/1369118X.2019.1697340
- Ferrari, A. (2013). *DIGCOMP: A Framework for Developing and Understanding Digital Competence in Europe*. Brussels: European Commission.
- Flanagin, A. J., and Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media Soc.* 9, 319–342. doi: 10.1177/1461444807075015
- Früh, W. (2017). *Inhaltsanalyse*. Stuttgart: UTB.
- Gadiraju, U., Yu, R., Dietze, S., and Holtz, P. (2018). “Analyzing knowledge gain of users in informational search sessions on the web,” in *CHIIR '18: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, New York, NY.
- Gikas, J., and Grant, M. (2013). Mobile computing devices in higher education. Student perspectives on learning with cellphones, smartphones & social media. *Internet High. Educ.* 19, 18–26. doi: 10.1016/j.iheduc.2013.06.002
- Gilster, P. (1997). *Digital Literacy*. New York: Wiley Computer Publications.
- Gladney, G. A., Shapiro, I., and Castaldo, J. (2007). Online editors rate Web news quality criteria. *Newspaper Res. J.* 28, 1–55. doi: 10.1177/073953290702800105
- Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think. Prevalence and predictors of fake news dissemination on Facebook. *Sci. Adv.* 5:eau4586. doi: 10.1126/sciadv.aau4586
- Hague, C., and Payton, S. (2010). *Digital Literacy Across the Curriculum*. Bristol: Futurelab.
- Hargittai, E., Fullerton, L., Menchen-Trevino, E., and Yates Thomas, K. (2010). Trust online: young adults’ evaluation of web content. *Int. J. Commun.* 4, 468–494.
- Heinström, J. (2002). *Fast Surfers, Broad Scanners and Deep Divers. Personality and Information-Seeking Behaviour*. Dissertation, Akademi University Press, Åbo.
- Helms-Park, R., Radia, P., and Stapleton, P. (2007). A preliminary assessment of Google Scholar as a source of EAP students’ research materials. *Internet High. Educ.* 10, 65–76. doi: 10.1016/j.iheduc.2006.10.002
- Hölscher, C., and Strube, G. (2000). Web search behaviour of Internet experts and newbies. *Comput. Netw.* 33, 1–6. doi: 10.1016/S1389-1286(00)00031-1
- IBM Corp (2017). *IBM SPSS Statistics for Windows, Version 25.0*. Armonk, NY: IBM Corp.
- JISC (2014). *Developing Digital Literacies*. Available online at: <https://www.jisc.ac.uk/guides/developing-digital-literacies> (accessed October 22, 2020).
- Jones, C., and Healing, G. (2010). Net generation students. Agency and choice and the new technologies. *J. Comput. Assist. Learn.* 26, 344–356. doi: 10.1111/j.1365-2729.2010.00370.x
- Judd, T., and Kennedy, G. (2011). Expediency-based practice? Medical students’ reliance on Google and Wikipedia for biomedical inquiries. *Br. J. Educ. Technol.* 42, 351–360. doi: 10.1111/j.1467-8535.2009.01019.x
- Kao, G. Y., Lei, P.-L., and Sun, C.-T. (2008). Thinking style impacts on Web search strategies. *Comput. Hum. Behav.* 24, 1330–1341. doi: 10.1016/j.chb.2007.07.009
- Kelly, D. (2018). Evaluating the news: (Mis)perceptions of objectivity and credibility. *Polit. Behav.* 41, 445–471. doi: 10.1007/s11109-018-9458-4
- Kennedy, G. E., Judd, T. S., Churchward, A., Gray, K., and Krause, K. (2008). First year students’ experiences with technology. Are they really digital natives? *AJET* 24, 108–122. doi: 10.14742/ajet.1233
- Kim, J. (2008). Task as a context of information seeking: an investigation of daily life tasks on the web. *Libri* 58:172. doi: 10.1515/libr.2008.018
- Knight, C., and Pryke, S. (2012). Wikipedia and the University, a case study. *Teach. High. Educ.* 17, 649–659. doi: 10.1080/13562517.2012.666734
- Konieczny, P. (2014). Rethinking Wikipedia for the classroom. *Contexts* 13, 80–83. doi: 10.1177/1536504214522017
- Kopp, M., Gröbinger, O., and Adams, S. (2019). Five common assumptions that prevent digital transformation at higher education Institutions. *INTED2019 Proc.* 160, 1448–1457. doi: 10.21125/inted.2019.0445
- Kruse, O. (2017). *Kritisches Denken und Argumentieren*. Konstanz: UVK.
- Leighton, J. (2017). *Using Think-Aloud Interviews and Cognitive Labs in Educational Research*. Oxford: Oxford University Press.
- List, A., and Alexander, P. A. (2017). Text navigation in multiple source use. *Comput. Hum. Behav.* 75, 364–375. doi: 10.1016/j.chb.2017.05.024
- Lucassen, T., Muilwijk, R., Noordzij, M. L., and Schraagen, J. M. (2013). Topic familiarity and information skills in online credibility evaluation. *J. Am. Soc. Inform. Sci. Technol.* 64, 254–264. doi: 10.1002/asi.22743
- Mason, L., Boldrin, A., and Ariasi, N. (2010). Epistemic metacognition in context: evaluating and learning online information. *Metacogn. Learn.* 5, 67–90. doi: 10.1007/s11409-009-9048-2
- Maurer, M., Quiring, O., and Schemer, C. (2018). “Media effects on positive and negative learning,” in *Positive Learning in the Age of Information*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Wiesbaden: Springer), 197–208. doi: 10.1007/978-3-658-19567-0\_11
- Maurer, M., Schemer, C., Zlatkin-Troitschanskaia, O., and Jitomirski, J. (2020). *Positive and Negative Media Effects on University Students’ Learning: Preliminary Findings and a Research Program*. New York, NY: Springer.
- McGrew, S., Breakstone, J., Ortega, T., Smith, M., and Wineburg, S. (2018). Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory Res. Soc. Educ.* 46, 165–193. doi: 10.1080/00933104.2017.1416320
- McGrew, S., Ortega, T., Breakstone, J., and Wineburg, S. (2017). The challenge that’s bigger than fake news. Civic reasoning in a social media environment. *Am. Educ.* 41, 4–9.
- McGrew, S., Smith, M., Breakstone, J., Ortega, T., and Wineburg, S. (2019). Improving university students’ web savvy: an intervention study. *Br. J. Educ. Psychol.* 89, 485–500. doi: 10.1111/bjep.12279
- McQuail, D. (1992). *Media Performance. Mass Communication and the Public Interest*. London: Sage.
- Molero, D., Zlatkin-Troitschanskaia, O., and Schmidt, S. (2019). Adapting the civic online reasoning assessment for cross-national use. *Paper presented at Annual Meeting of the American Educational Research Association*, (Toronto: APA).
- Murray, M. C., and Pérez, J. (2014). Unraveling the digital literacy paradox: how higher education fails at the fourth literacy. *Issues Inform. Sci. Inform. Technol.* 11, 85–100. doi: 10.28945/1982
- Nagler, W., and Ebner, M. (2009). “Is Your University ready for the ne(x)t-generation?,” in *Proceedings of 21st ED-Media Conference*, New York, NY, 4344–4351.
- Newman, T., and Beetham, H. (2017). *Student Digital Experience Tracker 2017: The Voice of 22,000 UK Learners*. Bristol: JISC.

- Oblinger, D. G., and Oblinger, J. L. (2005). *Educating the Net Generation*. n.p. Washington, DC: Educause.
- Persike, M., and Friedrich, J. D. (2016). *Lernen mit digitalen Medien aus Studierendenperspektive. Arbeitspapier*. Berlin: Hochschulforum Digitalisierung.
- Prensky, M. (2001). Digital natives, digital immigrants. *On Horizon* 9, 1–6. doi: 10.1002/9781118784235.eelt0909
- Riffe, D., Lacy, S., Fico, F., and Watson, B. (2019). *Analyzing Media Messages. Using Quantitative Content Analysis in Research*. New York: Routledge.
- Roblyer, M. D., McDaniel, M., Webb, M., Herman, J., and Witty, J. V. (2010). Findings on Facebook in higher education. A comparison of college faculty and student uses and perceptions of social networking sites. *Internet High. Educ.* 13, 134–140. doi: 10.1016/j.iheduc.2010.03.002
- Roscoe, R. D., Grebitus, C., O'Brian, J., Johnson, A. C., and Kula, I. (2016). Online information search and decision making. Effects of web search stance. *Comput. Hum. Behav.* 56, 103–118. doi: 10.1016/j.chb.2015.11.028
- Scheufele, D., Hardy, B., and Brossard, D. (2006). Democracy based on difference: examining the links between structural heterogeneity, heterogeneity of discussion networks, and democratic citizenship. *J. Commun.* 56, 728–753. doi: 10.1111/j.1460-2466.2006.00317.x
- Selwyn, N., and Gorard, S. (2016). Students' use of Wikipedia as an academic resource — Patterns of use and perceptions of usefulness. *Internet High. Educ.* 28, 28–34. doi: 10.1016/j.iheduc.2015.08.004
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. P. (2019). Assessment of University students' critical thinking: next generation performance assessment. *Int. J. Test.* 19, 337–362. doi: 10.1080/15305058.2018.1543309
- Stanford, J., Tauber, E. R., Fogg, B. J., and Marable, L. (2002). *Experts vs. Online Consumers: A Comparative Credibility Study of Health and Finance Web Sites*. Cham: Springer.
- StataCorp (2017). *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.
- Steiner, M., Magin, M., and Stark, B. (2018). Uneasy bedfellows. *Digital J.* 7, 100–123. doi: 10.1080/21670811.2017.1412800
- Strømso, H. I., Bråten, I., Britt, M. A., and Ferguson, L. (2013). Spontaneous sourcing among students reading multiple documents. *Cogn. Instr.* 31, 176–203. doi: 10.1080/07370008.2013.769994
- Tess, P. A. (2013). The role of social media in higher education classes (real and virtual) – A literature review. *Comput. Hum. Behav.* 29, A60–A68. doi: 10.1016/j.chb.2012.12.032
- Thompson, P. (2013). The digital natives as learners. Technology use patterns and approaches to learning. *Comput. Educ.* 65, 12–33. doi: 10.1016/j.compedu.2012.12.022
- Tombros, A., Ruthven, I., and Jose, J. M. (2005). How users assess web pages for information seeking. *J. Am. Soc. Inform. Sci. Technol.* 56, 327–344. doi: 10.1002/asi.20106
- Tribukait, M., Baier, K., Grzempa, H., Loukovitou, A., Sijakovic, R., Tettschlag, N., et al. (2017). Digital learning in European education policies and history curricula. *Eckert. Dossiers* 13:417.
- Urban, J., and Schweiger, W. (2014). News quality from the recipients' perspective. *J. Stud.* 15, 821–840. doi: 10.1080/1461670X.2013.856670
- Walraven, A., Brand-Gruwel, S., and Boshuizen, H. P. A. (2009). How students evaluate information and sources when searching the World Wide Web for information. *Comput. Educ.* 52, 234–246. doi: 10.1016/j.compedu.2008.08.003
- Wannemacher, K., and Schulenberg, F. (2010). “Wikipedia in academic studies: corrupting or improving the quality of teaching and learning?,” in *Looking Toward the Future of Technology-Enhanced Education. Ubiquitous Learning and The Digital Native*, eds M. Ebner and M. Schiefner (Hershey, PA: IGI Global), 296–312.
- Wellbrock, C. (2011). Die journalistische Qualität deutscher Tageszeitungen – Ein Ranking. *MedienWirtschaft* 8:2. doi: 10.15358/1613-0669-2011-2-22
- White, R. W., Dumais, S. T., and Teevan, J. (2009). “Characterizing the influence of domain expertise on web search behavior,” in *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*, New York, NY.
- Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., and Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. *Am. Educ. Res. J.* 46, 1060–1106. doi: 10.3102/0002831209333183
- Wineburg, S., Breakstone, J., McGrew, S., and Ortega, T. (2016). *Evaluating Information: The Cornerstone of Civic Online Reasoning*. Stanford History Education Group. Available online at: <https://stacks.stanford.edu/file/druid:fv751yt5934/SHEG%20Evaluating%20Information%20Online.pdf> (accessed May 10, 2020).
- Wineburg, S., and McGrew, S. (2016). *Why Students Can't Google Their Way to the Truth: Fact-checkers and students approach websites differently*. *Educational Week*. Available online at: <https://www.edweek.org/ew/articles/2016/11/02/why-students-cant-google-their-way-to.html> (accessed May 10, 2020).
- Wolfsfeld, G., Yarchi, M., and Samuel-Azran, T. (2016). Political information repertoires and political participation. *New Med. Soc.* 18:9. doi: 10.1177/1461444815580413
- Yadav, R., Tiruwa, A., and Suri, P. K. (2017). Internet based learning (IBL) in higher education: a literature review. *J. Int. Educ. Bus.* 10:2. doi: 10.1108/JIEB-10-2016-0035
- Zaller, J. (2003). A new standard of news quality: burglar alarms for the monitorial citizen. *Polit. Commun.* 20, 109–130. doi: 10.1080/10584600390211136
- Zimmermann, F., and Kohring, M. (2018). “Fake News” als aktuelle desinformation. Systematische bestimmung eines heterogenen Begriffs. *Med. Kommunikationswissenschaft* 66, 526–541. doi: 10.5771/1615-634x-2018-4-526
- Zlatkin-Troitschanskaia, O., Beck, K., Fischer, J., Braunheim, D., Schmidt, S., and Shavelson, R. J. (2020). The role of students' beliefs when critically reasoning from multiple contradictory sources of information in performance assessments. *Front. Educ.* 11:2192. doi: 10.3389/fpsyg.2020.02192
- Zlatkin-Troitschanskaia, O., Jitomirski, J., Happ, R., Molerov, D., Schlax, J., Kühling-Thees, C., et al. (2019). Validating a test for measuring knowledge and understanding of economics among university students. *Z. Pädagogische Psychol.* 33, 119–133. doi: 10.1024/1010-0652/a000239

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, JL, declared a past co-authorship with several of the authors, OZ-T and SB, to the handling editor.

Copyright © 2020 Nagel, Schäfer, Zlatkin-Troitschanskaia, Schemer, Maurer, Molerov, Schmidt and Brückner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Narratives and Their Impact on Students' Information Seeking and Critical Online Reasoning in Higher Education Economics and Medicine

Mita Banerjee<sup>1\*</sup>, Olga Zlatkin-Troitschanskaia<sup>2</sup> and Jochen Roeper<sup>3</sup>

<sup>1</sup> Department of English and Linguistics, Obama Institute for Transnational American Studies, Johannes Gutenberg University, Mainz, Germany, <sup>2</sup> Department of Business and Economics Education, Johannes Gutenberg University, Mainz, Germany, <sup>3</sup> Department of Neurophysiology, University Hospital of the Goethe University, Frankfurt (Main), Germany

## OPEN ACCESS

### Edited by:

Huei-Tse Hou,  
National Taiwan University of Science  
and Technology, Taiwan

### Reviewed by:

Nam Ju Kim,  
University of Miami, United States  
Niwat Srisawasdi,  
Khon Kaen University, Thailand

### \*Correspondence:

Mita Banerjee  
mita.banerjee@uni-mainz.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Education

**Received:** 08 June 2020

**Accepted:** 12 October 2020

**Published:** 19 November 2020

### Citation:

Banerjee M,  
Zlatkin-Troitschanskaia O and  
Roeper J (2020) Narratives and Their  
Impact on Students' Information  
Seeking and Critical Online Reasoning  
in Higher Education Economics  
and Medicine. *Front. Educ.* 5:570625.  
doi: 10.3389/feduc.2020.570625

The digital and information age has fundamentally transformed the way in which students learn and the study material they have at their disposal, especially in higher education. Students need to possess a number of higher-order cognitive and metacognitive skills, including effective information processing and critical reasoning to be able to navigate the Internet and use online sources, even those found outside of academically curated domains and in the depths of the Internet, and to solve (domain-specific) problems. Linking qualitative and quantitative research and connecting the humanities to empirical educational science studies, this article investigates the *role of narratives and their impact on university students' information seeking and their critical online reasoning (COR)*. This study focuses on the link between students' online navigation skills, information seeking behavior and critical reasoning with regard to the specific domains: economics and medicine. For the *empirical analysis* in this article, we draw on a study that assesses the COR skills of undergraduate students of economics and medicine at two German universities. To measure COR skills, we used five tasks from the computer-based assessment "Critical Online Reasoning Assessment" (CORA), which assesses students' skills in critically evaluating online sources and reasoning using evidence on contentious issues. The *conceptual framework* of this study is based on an existing methodology – *narrative economics and medicine* – and discusses its instructional potential and how it can be used to develop a new tool of "wise interventions" to enhance students' COR in higher education. Based on qualitative content analyses of the students' written responses, i.e., short essays, three distinct patterns of information seeking behavior among students have been identified. These three patterns – "Unambiguous Fact-Checking," "Perspective-Taking Without Fact-Checking," and "Web Credibility-Evaluating" – differ substantially in their potential connection to underlying narratives of information used by students to solve the CORA tasks. This analysis suggests that training university students in narrative analysis can strongly contribute to enhancing their critical online reasoning.

**Keywords:** online reasoning patterns, narrative medicine, narrative economics, narrative content analysis, instructional interventions, higher education, performance assessment



## INTRODUCTION

### Research Background

The digital and information age has fundamentally transformed the way in which students learn and the study material they have at their disposal, especially in higher education. To navigate the Internet and to successfully use online sources, even those found outside of academically curated domains and in the depths of the Internet, as well as to solve (domain-specific) problems, they need to possess a number of higher-order cognitive and metacognitive skills, including effective information processing and critical reasoning (e.g., Zhou and Ren, 2016; Shavelson et al., 2019). Learners who use the Internet must be able to assess the credibility and trustworthiness of sources and information (McGrew et al., 2018; Wineburg et al., 2018), they have to balance new information against their prior knowledge and any beliefs they may hold (van Strien et al., 2014; List and Alexander, 2017, 2018), and they must recognize how a given text or media format can affect not only their rational decision-making processes (Stanovich, 2018) but also their emotional judgment, which may lead to judgment errors, for instance, due to fast thinking and other biases such as motivated reasoning (Stanovich et al., 2013; Kahne and Bowyer, 2017).

This ever-changing information and learning environment has profound consequences for the teaching of domain-specific knowledge in higher education (e.g., Harrison and Luckett, 2019). A number of obstacles appear to make online learning challenging: first, students may acquire misconceptions by uncritically selecting sources that provide, for instance, misleading or even false information. Second, students may stop searching once they have arrived at a simple, unambiguous answer (Johnson et al., 2016). Third, their online search behavior may be limited by their previous knowledge and beliefs (van Strien et al., 2014; List and Alexander, 2018), which may cause them to stop short of taking full advantage of the wealth and diversity of information that Internet sources can provide. The dichotomy between knowledge and beliefs can hence be a particular obstacle to learning with the help of the Internet (Chiu et al., 2013; Hsu et al., 2014; van Strien et al., 2016).

In this context, researchers and educators attempt to create what Walton (2014) has called “wise interventions”: new instructional methodologies are required that will “vaccinate” students against biased information they find on the Internet, and that will guide their (self-directed) searches, information seeking, reasoning, and learning paths. These interventions, in turn, need to be closely related to what Pellegrino and Hilton (2012) have called “transferable knowledge” and “deeper learning,” i.e., developing students’ skills both to navigate the Internet successfully to gain domain-specific knowledge (and avoid the acquisition of erroneous knowledge and misconceptions) as well as to learn in a way that enables them to master concepts not merely superficially, but rather to apply their knowledge and skills to new contexts.

With the aim of developing effective teaching methodologies, recent empirical research has studied how students navigate the Internet, for instance, when solving (domain-specific) tasks

(Collins-Thompson et al., 2016; Brand-Gruwel et al., 2017). Prior research indicates that students may lack the skills to understand how the content they find on the Internet is also generally shaped by (covert) *narrative framing of information* (e.g., metaphors and analogies) (de los Santos and Nabi, 2019; Luong et al., 2020), which may lead to a *framing effect*, and *cognitive heuristics* (e.g., confirmation bias, Brand-Gruwel et al., 2009; Powell et al., 2019; Zollo, 2019), and thus a biased selection of information, where students’ information seeking and reasoning are influenced by, for instance, positive or negative connotations of the presented information. However, little is known about how and to what extent narratives may affect students’ information seeking (incl. selection, interpretation, and use of information) and their critical reasoning when solving (domain-specific) problems (Ulyshen et al., 2015; Hoppe et al., 2018; Yu et al., 2018).

### Study Framework and Research Objectives

Linking qualitative and quantitative research and connecting the humanities to empirical educational science studies, this article investigates the *role of narratives and their impact on university students’ information seeking and their critical online reasoning (COR)*. This study focuses on the link between students’ online navigation skills, information seeking behavior and critical reasoning with regard to two specific domains: economics and medicine.

The *conceptual framework* of this study is based on existing methodology – *narrative economics and medicine* – and discusses its potential and how it can be used to develop a new tool of “wise interventions” to enhance students’ COR. The key to narrative economics and medicine is a combination of domain-specific knowledge and its narrative framing: it applies methodology from literary studies and narrative analysis to fundamentally rethink the use of narratives in economics and medicine (see section “Narrative Medicine and Narrative Economics”).

Our conceptual framework of students’ COR has some overlaps with related concepts such as “information literacy” (Armstrong and Brunskill, 2018) or “digital literacy” (Hartley, 2017). However, we expand these conceptualizations through the specific, additional focus on COR, which is related to various well-established traditions on critical thinking (Oser and Biedermann, 2020; see section “The Project ‘Critical Online Reasoning 192 Assessment’ (CORA)”; Zlatkin-Troitschanskaia et al., 2020; Nagel et al., 2020). Moreover, we enhance the existing conceptual framework by adding the concept of narrative competence (see section “Narrative Medicine and Narrative Economics”).

In this article, we demonstrate that the information that is available on the Internet is never “neutral”: the content of this information cannot be disentangled from the narratives that this information is embedded in (Kay, 2000). Narrative carries knowledge: it has an effect on the reader, for instance through the metaphors and analogies it uses (Hallin, 2000) and the perspectives it takes (Trzebinski, 1995). We argue, therefore, that learners need a specific skillset to recognize and to understand

the (covert) narrative structure and framing of the information they use, and how this framing can influence their information processing and reasoning.

For the *empirical analysis* in this article, we draw on a study that assesses the COR skills of undergraduate students of economics and medicine at two German universities. To measure COR skills, we used five tasks from the computer-based assessment “Critical Online Reasoning Assessment” (CORA), which assesses the students’ skills in critically evaluating online sources and reasoning from evidence on contentious issues (Molero et al., 2019; Nagel et al., 2020). In addition to conducting an open web search and evaluating online information, the students were also prompted to write an open-ended, argumentative response to each task. The difficulty lies in making a judgment in a short period of time while recognizing (covert) bias in information sources. The participants’ browser history and online behavior data were recorded, and their written responses (short essays) were evaluated by independent human raters using a newly developed and validated rating scheme (for details, see Nagel et al., 2020). On the basis of these data, we investigate the narrative framing of the online information processed by the students while working on the CORA tasks, and how an impact of this framing on students’ information processing and reasoning may manifest in their written responses.

We propose here that investigating students’ COR using multiple sources of information can be approached from two directions, which ultimately need to converge: first, through a qualitative narrative analysis of sources that the students used in their reasoning process, we can identify the information processing approaches students used based on both the sources’ content and their underlying narrative frames. The second approach is to analyze the students’ written responses in terms of whether they recognized the way in which a given information source, its context and content, its (covert) motives, and its potential conflicts with other evidence and/or the students’ prior knowledge and beliefs guided their information processing and reasoning approach. These analyses can provide us with an empirical insight into the relationship between students’ information seeking, the (covert) narrative framing, and the affective influence that sources on the Internet may have on students’ information processing and critical reasoning.

Based on this analysis, this article discusses how the methodologies of narrative medicine and narrative economics can be used to teach students how to critically and competently use online information for learning, to enhance students’ COR and help them overcome the aforementioned obstacles in online learning in higher education. By fostering COR rather than rather than superficially covering a wide range of learning content, narrative fields of study, in this case economics and medicine, may be a promising approach to help students deal with the current explosion of information in the classroom (e.g., McQuiggan et al., 2008). Drawing on this methodology as a potential teaching tool, this study also discusses the role of emotion in learning economics and medicine in higher education. Moreover, we argue that, given their attention

to narrative and to the relationship between narrative and knowledge building, narrative medicine and narrative economics have much in common and can be transferred to other academic domains.

Consequently, we investigate narrative economics and narrative medicine as a way for students to identify the narrative framing of learning materials and texts, and hence foster their skills in recognizing that Internet sources are never completely “neutral” and may influence their information seeking behavior and reasoning through both argument and affect. By combining domain-specific knowledge with a narrative framing approach, students’ preconceptions and beliefs may be uncovered. In this way, narrative economics and medicine may enhance domain-specific learning by fostering students’ skills in understanding the difference between knowledge and beliefs as well as between informed reasoning and motivated reasoning (Kunda, 1990).

By using the methodology of literary studies, which stipulates that answers are seldom one-sided and stresses the role of *ambiguity*, narrative economics and narrative medicine might enable students to better deal with ambiguity – a crucial yet often neglected faculty in the classroom (Craig and Charon, 2017). Finally, by promoting skills required for critical reasoning and making decisions in the face of ambiguity, initially when dealing with diverging sources and contradictory information in the context of university learning and later in real-life practical situations, narrative economics and narrative medicine can encourage students to search for information in such a way that they do not stop once they have found an answer, but continue searching and eventually devise a complex, multi-layered, and potentially ambiguous answer and a more elaborate critical reasoning and problem-solving approach.

To explore narrative economics and narrative medicine as tools to enhance the online learning of higher education students in the Internet age, this article combines education and learning research with the humanities. As learning today consists of a combination of in-class teaching and (self-directed) learning using the Internet, this article proposes that narrative economics and narrative medicine may train and equip students with skills that help them use the Internet in a manner that enhances their COR and their domain-specific knowledge acquisition.

## PRIOR RESEARCH

### The Project “Critical Online Reasoning Assessment” (CORA)

To successfully learn and study in higher education in the Internet age, knowledge and skills for critically processing information including online reasoning are crucial. Since various studies reveal significant deficits among both graduate and undergraduate students (e.g., McGrew et al., 2018; Wineburg et al., 2018; Hahnel et al., 2019; Zlatkin-Troitschanskaia et al., 2019), more research on students’ COR and its determinants is required. As some studies show in particular, information processing is significantly determined by students’ individual beliefs and preconceptions (e.g., Alexander et al., 2018; Zlatkin-Troitschanskaia et al., 2020) indicating the impact of affective

factors. In this context, the question arises to what extent students themselves may recognize affective influences, i.e., to what extent do they understand that the presentation of a given topic in an Internet source and its underlying (covert) narrative may shape their perception, attention, their emotions and their own decision-making? We consider this skillset an important facet of COR, and therefore used the data from the CORA study to investigate this question.

In CORA, undergraduate students from different study domains (including economics and medicine) were presented with tasks that describe real-life judgment situations (for an example, see section “Analyses of Student Responses”) and require the students to form an opinion on a given topic using an unrestricted web search (Nagel et al., 2020). To solve the CORA tasks, students were required to navigate the Internet and find suitable information sources on their own. One of the aims of the CORA study was to analyze how students select, evaluate and use Internet sources while working on a given CORA task and writing their response.

In the CORA tasks, students were given one website as a main source (see an example in section “Analyses of Student Responses”). Based on this website, they were asked to use the Internet and find other sources to form an opinion about a given topic and to evaluate the trustworthiness and the quality of the information presented on this website. CORA thus combined a website chosen by the task developer with online sources the students selected to make the CORA tasks as authentic as possible. Moreover, as Alexander et al. (2018) have shown, being able to form their own “search path” by doing their own research and autonomously selecting sources may enhance students’ (test) motivation.

This part of the CORA study is specifically related to research on the use of multiple sources in student learning (for an overview, see Braasch et al., 2018). According to Britt and Rouet (2012, p. 276), “studying multiple documents to learn about a topic can lead to a deeper, more complex understanding of the topic.” Moreover, since the CORA study requires students to autonomously find source material online to evaluate the source (website) given in the CORA task and to verify this information, this study is also related to research on self-regulated learning based on metacognitive skills (e.g., Neuenhaus et al., 2013; for “search as learning,” see Hoppe et al., 2018) as well as on information problem-solving (Brand-Gruwel et al., 2009).

At the same time, the students in the CORA study are required not only to evaluate the credibility and trustworthiness of the given website and its information but also, ultimately, to form their own opinion and justify this opinion in a brief essay referencing online information they used. In this respect, the CORA study is related to research on “web credibility” (Metzger and Flanagin, 2015; for “credibility evaluation,” see Metzger et al., 2010; for “information trust,” Lucassen and Schraagen, 2011), as well as on “trustworthiness.” For instance, Hendriks et al. (2015) designed an “inventory measuring laypeople’s ascriptions of an expert’s trustworthiness” to measure this skill.

Moreover, the CORA tasks also incorporate challenging issues that sometimes include moral or ethical aspects, for instance framed in terms of conflicting interests. The resolution

thereof requires students to apply multiple aspects of ethical critical reasoning and decision-making. For instance, in one of the CORA tasks (illustrated in section “Analyses of Student Responses”), students were given a link to a website and encouraged to conduct additional research online, then asked to state and justify their decisions. This facet of COR can be linked to *critical reflection*, which was defined by Oser and Biedermann (2020, p. 90) as “a basic attitude that must be taken into consideration if (new) information is questioned to be true or false, reliable or not reliable, moral or immoral etc.” Therefore, critical reflection involves recognizing potential motives or (covert) interests and analyzing consequences of making a decision.

## Narrative Medicine and Narrative Economics

Narrative medicine is an approach that emerged in the 1980s. Its founder, Rita Charon, a medical doctor who also holds a Ph.D. in literature, argued that with the rise of biotechnology in medicine, doctors had stopped paying attention to narratives. Introducing the model of narrative medicine, Charon suggested using the tools of literary analysis while listening to patients’ stories. She stresses the necessity of “honoring the stories of illness” (Charon, 2008).

Charon proposes narrative medicine not just as a tool to change doctor-patient communication but also as a way to make physicians recognize that narratives are a key component of medical knowledge. We tested this assumption by focusing on one case in particular, the history of the so-termed “broken heart syndrome” (Efferth et al., 2017). For decades, physicians had been approached by patients who, after the traumatic loss of a loved one, complained that their heart had been broken. The *metaphor* of the broken heart, moreover, has been one of the most powerful images to convey the extent of trauma, grief, or loss. Researchers then began to wonder whether the metaphorical quality of the image which was used to convey a physical condition might in fact prevent physicians from taking this condition seriously as a somatic condition. The recognition of the broken heart syndrome as a medical condition was hence hampered by the metaphorical, literary quality of the narrative in which it was conveyed. This obstacle may in part account for the fact that it took until 1990 for the condition to be recognized as a medical condition (Goldman, 2014; Efferth et al., 2017).

Narrative medicine sets out to demonstrate that narrative is central to the practice of medicine, both for knowledge acquisition and for doctor-patient communication. As a methodology, narrative medicine can serve to enhance physicians’ narrative competence. In this framework, narrative competence can be defined as the ability to “listen closely” (Charon et al., 2017) and detect hidden meanings and sudden turns in the narrative. Charon suggests that the act of doctors listening to patients’ narratives is akin to the careful reading of literary texts (Charon et al., 2017). By enabling them to pay attention even to minor details in patients’ narratives, Charon proposes, physicians will be able to arrive at more valid diagnoses. Moreover, the narrative competence will also serve to

improve doctor-patient communication (Charon et al., 2017). Narrative medicine has substantial overlaps with narrative ethics (Craig and Charon, 2017), and it is also closely related to medical humanities (Banerjee, 2018; Spencer, 2020).

Narrative medicine is increasingly becoming an established methodology for the teaching of medicine (McAllister, 2015). One of the aims of narrative medicine is to enhance students' self-reflection about their role as medical practitioners and about the kind of knowledge and skills required for a successful professional development in this domain. Students are trained in narrative competence, and they are taught to recognize that knowledge in medicine is constructed not only through data and biotechnological diagnostics, but also through narrative. At the core of narrative medicine as a teaching methodology lies the idea of estrangement (Spiegel and Spencer, 2017). For instance, medical students are asked to read literary texts such as Michael Ondaatje's *The English Patient*. These texts often do not feature specific medical settings but rather deal with concepts of care: how friends or relatives care about one another, and the protagonist's need for care. In narrative medicine courses, students are asked to relate to the texts in an affective manner: they relate the text to their own understanding of care. Through the "detour" of literature, medical students hence come to reflect on their own practice as physicians. After this *intervention* by narrative medicine, they may approach the clinical setting in a new way, and they may listen differently to patients' narratives. While narrative medicine is increasingly becoming an established tool in the didactics of medicine, its effectiveness for teaching in higher education still needs to be explored empirically (see section "Narrative Analysis in Educational and Learning Research").

It may be indicative of a paradigm shift across academic disciplines in a particular period of time that after Charon et al. (2017) had developed the *narrative medicine* approach – which understood narrative to be essential to the practice of medicine – the "narrative economics" approach was developed by Yale economist and Nobel laureate Shiller (2017).

Shiller integrates the fields of economics, anthropology, psychology, and literary studies to create this approach. Narrative economics is conceived by Shiller as a methodology for redefining knowledge in economics: so far, knowledge in economics has been conceived mainly in terms of theories and data; the role of narrative has been underestimated. By contrast, Shiller suggests relating to economic events such as economic downturns or fluctuations in the stock market through the narratives that are created around these events. Consequently, he proposes that economists need to be equipped with narrative competence (Shiller, 2019).

Shiller evokes the work of biologist Gould (1980) and his image of the "*homo narrator*." Following Gould, Shiller (2017) suggests that humans are a "storytelling animal": "...the human brain is built around narratives." Shiller (2017) goes on to look at important events in the history of economics like the stock market crash of 1929 to focus on narratives and "narrative history" as a potential reason for why we remember and forget certain events. He argues that economists need to team up with narrative scholars, such as literary researchers,

to unpack the power of narratives in conveying economic meaning: "*Not everyone is equally proficient at understanding narratives, and economists are among the worst at appreciating them*" (Shiller, 2017). Through narrative analysis, we may be able to understand the role narratives plays in what we might call economic memory. For instance, the images and metaphors we connect with the crash of 1929 are those of people losing their life savings overnight, of the stock market crash sparking off the *Great Depression*, and "...we've been worried about it happening again all this time, because the narrative isn't forgotten" (Shiller, 2017).

Overall, according to Charon et al. (2017) and Shiller (2019), narratives are at the core of knowledge acquisition in medicine and economics. In this article, we focus on the overlaps between these two methodologies. Based on prior research (Charon et al., 2017; Shiller, 2019), we argue that it is fruitful to link the methodologies of narrative medicine and narrative economics. Moreover, we argue that these two methodologies can be used for both narrative analysis (see section "Qualitative Narrative Research") and teaching intervention (see section "Narrative Analysis in Educational and Learning Research"): first, in the following section, we show how narrative analysis – which lies at the core of both narrative medicine and narrative economics – can be used in qualitative narrative research. Second, based on the narrative analyses in this article, we propose that narrative medicine and narrative economics can be employed to change students' online information-seeking behavior and foster COR in the domains of medicine and economics, and that this approach can be transferred to other domains (see section "Limitations and Future Perspectives").

## METHODS AND ANALYSIS

### Qualitative Narrative Research

To investigate our research question and provide insights into the potential influence of narratives on the extent to which students critically evaluated the information they were confronted with on the Internet, and which led them to come to certain conclusions, we connected qualitative narrative analyses of both students' written responses and the online information they used. Narrative analysis, as proposed here, shows overlaps with reconstructive hermeneutics, which are widely established in educational research (Malpas and Zabala, 2010). Like reconstructive hermeneutics, narrative analysis aims to identify and to "reconstruct" implicit patterns through text analyses. However, we expand this existing research by particular focusing on framing, affect, and metaphoricity.

Narrative framing, for instance, the use of metaphors (e.g., "broken heart" and "economic crisis"), analogies (virus as an "invisible enemy"), change of perspective ("life-value" vs "money-value") is covert; i.e., students/learners often do not recognize narratives and their role in information processing and decision-making. Prior research has shown that linguistic framing influences reasoning (e.g., Gibbs, 1994): narratives have a powerful influence on reasoning, as students select and use information that is consistent with a certain narrative frame and



that confirms their initial knowledge and beliefs (“vaccination is poison”), while neglecting any contradictory information (e.g., Thibodeau and Boroditsky, 2011). Thus, narratives may cause a so-termed framing effect due to a cognitive bias (e.g., confirmation bias) and lead to a biased selection of information, and students’ reasoning may be influenced by, for instance, positive or negative connotations of the information (Rumelhart, 1979; Pinker, 2007).

Building on this research, we analyzed the influence of narrative framing on COR by assessing economics and medical students. In the CORA study (see section “Sample and Procedures”), we retraced the sources students used on the Internet as well as their simultaneous use of multiple documents from various sources in their responses to the CORA tasks (for details, see Nagel et al., 2020). According to Hahnel et al. (2019, p. 524), “however, to use variables generated from process data (e.g., mouse clicks with timestamps) sensibly for educational purposes, their interpretation needs to be validated with regard to their intended meaning.” Thus, these quantitative data from the CORA study are linked to qualitative narrative analysis as proposed in this article. Based on a methodology from narratology often used in literary studies, we analyze online information used by students in terms of its underlying narrative features: this narrative analysis explores the ways in which (domain-specific) content was put into a “story” format. The qualitative narrative analyses based on different categories, for example, the structure of the text, the main topic of the “story,” narrative perspective (first-, second-, or third-person perspectives), mode of speech (direct/indirect speech), choice of metaphors, as well as the affective dimension involved in these textual features.

Crucially, narrative analysis as a tool for the assessment of the linguistic framing of information is based on the fact that rhetorical strategies are not always intentional. The speaker might, for instance, use metaphors or formulations that may lead the readers to become predisposed in certain ways, without actually being conscious of the effects of their rhetoric. Narrative analysis, along with discourse analysis, has thus tended to focus less on the speaker or producer of an utterance, than on the effects produced by the utterance itself. In this context, the evaluation of the expert’s expertise in a given domain may be equally based on the students’ ability to unpack not just the argument, but its underlying narrative and to recognize the *narrative affect* which a source might convey.

Consequently, we analyzed students’ written responses to web search tasks to see whether they recognized the narrative framing of the information they used. Based on the data from the CORA study, we focus on the research question *did the narrative influence how the students perceived and processed the information and how they reasoned based on the online sources they used?*

When analyzing the students’ written responses, we therefore focused on identifying clues as to whether the narratives of the online information they used influences students’ information seeking behavior and their COR. We suggested that if affective influence is key to narratives, this notion can also be applied to the interpretation of students’ responses, for instance, how students assess the trustworthiness of expert opinions on topics

described in the CORA tasks (for an example, see section “Sample and Procedures”). We proposed that a narrative analysis may also contribute to an understanding of students’ opinion-forming processes and their decision-making when learning with the help of the Internet.

## Analyses of Student Responses

### Sample and Procedures

In the first step, we took an initial look at the students’ written responses, i.e., the short essays, in which they described their decision or conclusion related to the evaluation of the credibility of the information presented on the website linked in the CORA task. While we cannot discuss all student responses collected in the study, we instead focus particularly on one CORA task “assisted suicide,” which dealt with aspects of moral reasoning (see **Figure 1**). We focus on this task, since we had the most written answers from students in both domains for this task, and they were on average longer than for the other CORA tasks, which could be due to the special moral and ethical aspect of this task. For the narrative analyses, the length of the written student responses was an important qualitative factor of the data.

Thus, the subsample used in this narrative analysis consists of 19 medical and 47 economics students from two German universities (the data of 11 students from other domains were excluded in the following analyses). The data were collected in the winter semester of 2019/2020. The assessments took place in a research laboratory under controlled conditions. To ensure test motivation, for their participation in the study, the students received credits for a study module. The majority of the 66 participants were in their first study year; about two thirds of the participants were women.

The subsample can be considered very large with a view to the comprehensive qualitative analysis conducted in this study. Moreover, this sample can be considered representative for the total sample of the CORA study in terms of the gender ratio and the study semester. However, the medical students are underrepresented in this sample (for limitations, see section “Limitations and Future Perspectives”).

In this CORA task that deals with the topic of assisted suicide, students were asked to discuss, to evaluate and to justify an expert’s opinion on assisted suicide presented on a website linked in the task. Here, students were first not explicitly asked what they thought of assisted suicide or even whether the expert in the article was credible or trustworthy. Rather, the first question in the task was formulated on a much more pragmatic level: “Do you think that Volker Gerhardt [the expert cited on the website linked in the task] supports assisted suicide?” Remarkably, this question, which seems to be only content-based at first sight, actually elicited student responses that precisely addressed these more focused questions of trustworthiness and the credibility of sources. In the subsequent question, students were asked to explain why they think this source is credible. In particular, they were asked to find additional information on the Internet and justify their responses with evidence from the Internet sources they used. In the next section, we present the key results of our qualitative narrative analysis.

Some people claim that Volker Gerhardt, professor of philosophy, supports assisted suicide. During the next 10 minutes, browse the web to find additional information to decide whether you think this claim is true.

Do you think Volker Gerhardt supports assisted suicide?

You can use any information on this website, and you can freely search the Internet.

**Justify your answer with evidence from the Internet sources you used and include the corresponding URLs. Explain why the sources you used are credible.**

**FIGURE 1** | Example CORA task “assisted suicide.”

## Results

When analyzing student responses as described in section “Qualitative Narrative Research,” we identified three distinct patterns: “Unambiguous Fact-Checking,” “Perspective-Taking without Fact-Checking,” and “Web Credibility-Evaluating,” which can be linked to existing research: for instance, the role of fact-checking was established in recent studies about university students’ online search behavior and its role in their learning (e.g., McGrew et al., 2019). Moreover, the ability to evaluate the credibility and trustworthiness of a given source has been at the core of recent research on online learning (e.g., Gierth and Bromme, 2020).

The majority of the student answers correspond to the pattern of “web credibility-evaluating” (56 percent); this was closely followed by the pattern of “perspective-taking without fact-checking” (41 percent). Only very few answers fell into the pattern of “unambiguous fact-checking” (3 percent). Regarding the pattern of “web credibility-evaluating,” many of the students seemed to be at a loss for criteria they could use to evaluate the credibility of a source. While many students referred to the trustworthiness of the source, for instance, of websites hosted by national newspapers (Die Welt, Süddeutsche Zeitung), others simply remarked that the website “looked” trustworthy because “it cited experts.”

In the following, we describe students’ task responses with regard to these three patterns. While some students’ responses showed elements of two or more of these patterns, here, we will elaborate on the answers that fell squarely into one of the profiles. At the same time, however, we demonstrate that this pattern-based analysis can only be the first step toward a more complex investigation of students’ online search behavior and their reasoning. Therefore, we conclude by indicating perspectives for further, more fine-grained research (see sections “Analysis of the Narrative Framing of the Most Commonly Used Online Source” and “The Impact of the Narrative Framing on the Student Responses to the CORA Task”).

### Unambiguous fact-checking

One pattern of student responses distinguished among the short essays can be defined as “Unambiguous Fact-Checking.” Prior research has outlined the relevance of “fact-checking” for students’ critical evaluation of online sources (e.g., McGrew et al., 2019). In our analysis of student responses, we investigated in

particular whether students’ search behavior indicates that they verify the “facts” stated in the original source. For instance, the original article may cite the opinion of a specific expert. Students can then “check” if this person is really an expert on the topic at hand, or they can neglect to do so. This pattern was termed “unambiguous” fact-checking, since this initial evaluation of the facts presented and/or experts cited in the source was only one of the steps of a more critical evaluation of the source that students were asked to make in the CORA tasks (see **Figure 1**). The next steps might include a more critical reflection on the quality of the facts presented in the online information used by students when solving the CORA tasks. For instance, the expert may be from a discipline that is not central to the topic at hand: a professor of physics for example is an expert in his field, but his expertise may not be pertinent to the specific task topic (assisted suicide). In this context, our research relates to the well-established approaches in ‘web credibility’ research (Bromme and Thomm, 2016).

In the following, we describe the responses of the student group of unambiguous fact-checkers in more detail. This group of students thinks that Volker Gerhardt supports assisted suicide: *“Based on my research, I would agree with the statement. Especially on the basis of his answer to a reader’s question in the Tagesspiegel. He is also an expert on Nietzsche’s philosophy as well as on theological philosophy. Nietzsche, who, as is well known, advocated ‘atheistic’ theories and declared ‘God is dead,’ freed thinking from the obstinacy of a God and the interpretation of the Church. Thus, it is not morally wrong to kill a human being if it so desired.”*

A number of things are remarkable about this pattern of responses. The first aspect relates to the question whether the students accept the knowledge and trustworthiness of the expert himself. In the article, Volker Gerhardt is introduced as a professor, a philosopher, and a person deeply concerned with the question of assisted suicide. In this response, the student first follows the “lead” that the article has established, namely that Volker Gerhardt’s expertise as a philosopher is key to the debate on assisted suicide. Second, however, the student goes on to double-check what this philosophical expertise is based on. Crucially, she does not refer to the fact that Gerhardt is a professor at the Humboldt University Berlin nor that he is a member of the Berlin Brandenburg Academy of Sciences, but she focuses on his expertise regarding Nietzsche.

The student follows two leads in particular: first, she “checks” the website given in this task by googling a YouTube video: an interview with Volker Gerhardt. This suggests that the YouTube videos were used to double-check the impression that the student gained through the website. Moreover, she enhances the presence of Volker Gerhardt by adding the visual impression (YouTube video) to the sense that we get of him from the website. Secondly, the student sets out to verify Professor Gerhardt’s expertise by following up on Nietzsche, one of the philosophers that he specializes on in his work. The student thus not only googles an interview with Volker Gerhardt, but also consults a Wikipedia source on Friedrich Nietzsche. Finally, in the answer, she states that “everyone knows that Nietzsche declared that God is dead.” The URL she provides below her short response, however, suggests that she just looked Nietzsche up on Wikipedia, and may not have known or remembered all these aspects in detail before consulting Wikipedia.

With regard to an underlying pattern, this student’s search behavior and reasoning approach indicate that she is by no means uncritical in her use of sources: she verifies the trustworthiness of the source as well as the credibility of the expert witness (Volker Gerhardt). However, her search behavior and reasoning also indicate that she might not recognize the narrative patterns which underlie the framing of the expert by the sources she consults.

Theoretically, the Internet could be an ideal source for learning, as there is an almost infinite number of sources available. However, for COR, the students need to recognize and understand alternative perspectives and arguments in a given source and hence alternative forms of search behavior and online information processing. This student’s search behavior hence corroborates one of the findings from prior research (see section “Research Background”), namely that students may stop searching once they have arrived at a simple, unambiguous answer (Johnson et al., 2016).

At the same time, it can be debated whether the student’s response is only a form of simple fact-checking or whether it exceeds this process. The student quoted here goes on to investigate not only the expert himself, but the expert’s own “expert,” namely the philosophy of Nietzsche. Yet, the student may not understand that Nietzsche’s philosophy is a highly complex philosophical theory with its own political and ethical implications and cannot be reduced to atheism alone. Students may therefore not understand the political tendencies and ethical associations that come with the Nietzsche reference. For follow-up empirical research (see section “Limitations and Future Perspectives”), this opens up an important question: where does the student’s fact-checking end? Which facts do they assume require verification through further sources?

### **Perspective-taking without fact-checking**

Another group of students whose pattern we described as “Perspective-Taking without Fact-Checking” do not focus on Volker Gerhardt at all in their responses, but rather on the question of assisted suicide more generally. This can be demonstrated with the following statement: “In his interview, Volker Gerhardt states that it is an incredible imposition to demand that other people hold him (doctor) responsible for the death of

another person. Because the doctor would not know in this state what it means for him and his conscience. The website [www.bpb.de](http://www.bpb.de) is a credible internet source, as many concrete topics are worked out very specifically and there is a lot of input.”

In contrast to the first group of students, the unambiguous fact-checkers, this group of students does not investigate the trustworthiness of the cited expert. Given the key relevance of fact-checking for critical reasoning and the evaluation of online sources, these students may hence be easily misled by the original source. Student answers from this group show that they immediately form an opinion of their own about the topic, without recognizing that this opinion-forming may be guided by the underlying narrative of the source.

Remarkably, this pattern of responses focuses on an aspect which was only marginally mentioned in the original source, namely the ethical dilemma of those who are called upon to assist another person’s suicide. To this extent, this pattern does not follow the “lead” laid in the article, namely the framing of this debate through the person and professional expertise of Volker Gerhardt. Notably, however, the student refers to the trustworthiness of one source he consults. The source is the Federal Agency for Civic Education, and hence a federal, non-partisan institution. While the student rates this source as trustworthy, however, he does not refer to the owner of the site and hence the institution itself – the Federal Agency for Civic Education – but rather focuses on the content provided on this site: “The website is a credible internet source, as many concrete topics are worked out very specifically and there is a lot of input.” This response corroborates research proposed by Wineburg et al. (2018), who state that students often cannot be seen as “fact checkers” and that they do not check the origin of a given website to find out where they have “landed.”

### **Web credibility-evaluating**

Another pattern of student responses, which we defined as “Web Credibility-Evaluating,” focused not on the first question – whether or not Volker Gerhardt is in favor of assisted suicide – but directly responded to the second question and discussed the trustworthiness of the source (Scharer et al., 2019). Student answers in this group indicate recognition of the fact that they must first check the trustworthiness of the source (e.g., newspaper, journal, and blog). The answers show that for this evaluation, students rely on their own prior knowledge of the (German) media landscape. However, the responses also indicate that once students had established the trustworthiness of the source, they did not go on to question the narrative framing of the article given in this source.

Since students were asked in this CORA task to search for websites relating to Volker Gerhardt’s opinion on assisted suicide, some referred to the article in the *Süddeutsche Zeitung* (analyzed in section “Analyses of Student Responses” below), while others used other sources. One student in this group thus refers to an article about Gerhardt in the *Tagesspiegel*: “Volker Gerhardt is quoted on this page. This quote contains statements by Gerhardt which clarify his attitude toward assisted suicide. Basically I judge the *Tagesspiegel* as a rather serious site, but with journalistic ones it can never be ruled out that false information may creep in. This

can be seen in the Spiegel scandal, where false information was subsequently uncovered in articles (Relotius). In this case, however, I think it is unlikely that Gerhardt's statements in the article were falsified. However, one cannot assume this to be 100% true."

Remarkably, the student was familiar with a "scandal" in which the Spiegel news magazine was involved, and hence goes on to question the trustworthiness of even established newspapers in general. From this observation, however, she goes on to question whether Gerhardt's opinion, which was quoted in the *Tagesspiegel*, was "fake" as well. At the same time, the student neglected another relevant perspective – compared to the "Perspective-taking" pattern, and did not double-check the given information – compared to the "Fact-checking" pattern.

In terms of a typology of student online information seeking and their COR that emerges from these preliminary qualitative analyses, it became evident that students of this group paid much more attention to the source than to the narrative itself. They hence understood credibility mainly as pertaining to the source in which a given report was provided, for instance, weighing the *Tagesspiegel* against the *Süddeutsche Zeitung*. The students who had thus established credibility, might have followed the narrative framing of the source itself. This suggests that training students in narrative analysis can strongly contribute to enhancing their COR (see section "Narrative Analysis in Educational and Learning Research").

Overall, other students' responses assume some of these reasoning approaches and arguments as well, which can only be referred to in an exemplary fashion here. Some refer to Gerhardt as "the professor," suggesting that it is his status and expertise that makes him a trustworthy source (indicating the "authority bias," Metzger et al., 2015). Other test participants also consult the website of the Federal Agency for Civic Education but unlike the student quoted above, they understood the institutional, non-partisan character of this source.

The three patterns of student CORA task-solving behavior that we have identified in their written responses – "Unambiguous Fact-Checking, Perspective-Taking without Fact-Checking, and Web Credibility-Evaluating" – differ substantially in their potential connection to underlying narratives. Of these types, it could be argued that the third response pattern – Web Credibility-Evaluating – seems to be the least impacted by narrative. This type of response does not take into account the expert's narrative at all but is rather concerned with the source it is cited in (*Tagesspiegel*). This may be especially problematic in that students may not recognize how the narrative form in which the information was given impacted their own reasoning strategies. Since this group of students shows the least understanding of how narrative framing can guide or even manipulate their own reasoning, this group may be most susceptible to acquiring misleading information or even erroneous (domain-specific) knowledge through Internet searches. Compared to this type of response, the first pattern, the "Fact-Checking," shows the highest impact of narratives on their own reasoning. Remarkably, this group of students appears to at least implicitly recognize a number of related facts, which they proceeded to cross-check with the given information (the reference to a specific expert-philosopher, and the expert's

reference to another expert). The "Perspective-Taking and Non-Fact-Checking" group of students show the least impact of narratives on their own reasoning.

To establish the relevance of narrative knowledge (Carroll, 2001) and narrative competence for teaching economics and medicine in higher education, however, we must go beyond defining these initial patterns of students' search behavior. To engage in COR, students need to understand how sources can influence or even manipulate their opinion-making. They need to be able to detect narrative framings in all their complexity. To elucidate this complexity, we will now analyze the source (online article) that most students based their answers on. We will then reconstruct the narrative framings which may have influenced the students' responses.

## Analysis of the Narrative Framing of the Most Commonly Used Online Source

In the next step, we qualitatively analyzed the websites which were most commonly used by students in their written responses. The aim of this analysis step is to reconstruct the leads given by the source which students may follow in their responses without recognizing they were being "guided" by these leads (see section "The Impact of the Narrative Framing on the Student Responses to the CORA Task").

First, a content-based narrative analysis might start off by noting that the information source most commonly used by students when solving the CORA task "assisted suicide" was an article from one of the largest daily newspapers in Germany *Süddeutsche Zeitung*. In terms of credibility, it can thus be argued that this is a reliable and multi-perspective source of information. Upon closer analysis, however, we might delve into the question of perspectives: (i) Who are the experts that the article cites, and how exactly are they being cited? (ii) What are their credentials, how does the article frame their narrative authority and their authority on the subject? (iii) What metaphors are being used, what discursive or narrative frameworks are evoked?

Seen in these terms, the narrative framing of the (task) topic of assisted suicide may in fact be quite surprising. First, it should be noted that the narrative is woven around one expert in particular, a professor of philosophy at the renowned Humboldt University in Berlin, Volker Gerhardt. This framing has a number of implications for the way the debate on assisted suicide is being framed:

First, the topic at hand is looked at from an academic perspective. Moreover, it is framed less as a political or societal issue, but more as an ethical one. Philosophy is hence implicitly reframed as being integral to ethics. It may be notable in this context that the question of ethics is itself a highly complex one. In the field of bioethics, for instance, experts might be situated in the domain of theology (as in the case of the former head of the German National Ethics Committee, Peter Dabrock), or medicine. At the same time, however, the fields to which the expert refers in his own opinion on assisted suicide by far exceeds philosophy and contains references to legal parameters as well as social and cultural ones. The



point which might be made here, in particular, is that legal parameters are reported through the philosopher's perspective. The article does not cite or feature another interview with a legal expert. Even as on the surface, the fields that the article refers to as relevant for assessing the topic of assisted suicide range from philosophy to ethics and law, all of these fields are represented by just one particular expert, who is a professor of philosophy.

Second and perhaps even more importantly, while the article uses direct speech to convey this expert's opinion on the subject, all other experts or potential discussants on the subject are present in the article merely through reported speech. Thus, the article notes, in reported speech and as if in passing, that representatives of the church and palliative care physicians have also referred to palliative care as a relevant factor in the context of the debate on assisted suicide. Narrative analysis here needs to be complemented by linguistic research to provide an insight into the differences in using direct or reported speech in a given text, and the different effects this will have on the reader.

Third and just as importantly, there is one metaphor used in the expert's direct speech which evokes a very particular historical context and a very particular emotional register. At what can be said to be an argumentative pivot of the article, the expert evokes the question of human rights. What happens, then, once the paradigm of human rights and its historical and ethical significance is evoked? Once the question of assisted suicide is framed in terms of human rights, the emotional subtext may have shifted imperceptibly. The absence of human rights, both historically and geographically, is implicitly framed as a context in which authorities can arbitrarily exert their power; where members of marginal communities – communities of color, working-class communities, or indigenous peoples – can be arrested and detained without proper trial. Historically, the *habeas corpus* act was an important precursor to the Declaration of Human Rights. Because of this declaration, which just celebrated its 70th anniversary in 2018, no-one can be arbitrarily arrested, and everyone, regardless of their provenance, race or social status, has the right to a fair trial. Conversely, the time before the Declaration of Human Rights appears to us, in retrospect, as the dark ages of a world without ethical recourse.

What does it mean, then, to reframe the topic of assisted suicide in terms of the human rights debate? It could be argued to mean that the current moment described in the article, in which no clear guideline for assisted suicide exists as yet, parallels the time before the institutionalization of human rights. Implicitly, then, the equation of the legal regulation on assisted suicide with the declaration of human rights frames medical practitioners as potentially holding arbitrary or at least unjustified power over patients who are powerless to resist their authority. Regardless of whether we are in favor of or against assisted suicide, it may therefore be important for us to note that introducing the metaphorical link of assisted suicide to human rights strikes a powerful emotional and affective chord. To the extent to which we may tend to identify with or at least accept the authority of the speaker who makes this connection, then – an identification

which may be enabled by the fact that this speaker is the only one whose ideas are represented to us in direct speech – this emotional influence may be all the more powerful.

## The Impact of the Narrative Framing on the Student Responses to the CORA Task

How can this qualitative narrative analysis of the information source that students most commonly used when solving the CORA task be linked back to their responses to this task? On the basis of the content analysis outlined above, we now return to the students' written statements.

One particularly remarkable aspect here is that none of the students question the expertise of Prof. Gerhardt, indicating the cognitive heuristic "authority bias" among all test participants (Metzger et al., 2015). They did not, as they could have done, wonder whether there are other experts on the topic of assisted suicide, and they did not look for other source materials. Rather, their short essay responses suggest that they invariably followed the "lead" (discussed in the section "Analyses of Student Responses") provided by the online source.

Moreover, students' responses indicated that they did not recognize how and why their trust in the expert's knowledge was established. As to the reasons for this conviction, almost all of the students referred to the credibility of the *Süddeutsche Zeitung* as a representative and unbiased source of information. Yet, this too may fall short of the actual complexity of the information landscape in the Internet age. While the *Süddeutsche Zeitung* is considered a trustworthy source, the choice of experts featured in their articles may nonetheless be biased in one direction or another.

With regard to the impact of the narrative framing of the used information in the student responses, three patterns among participants have been identified, which differ in terms of their search and reasoning approaches as well as in the extent to which the given information and arguments were recognized or neglected. The findings indicate that the participants within these groups evaluated the *credibility, trustworthiness and relevance* of the sources and incorporated arguments differently, whereby most of the students, however, did not weigh or compile the information and arguments provided, but rather selected information – most likely related to their own (prior) knowledge and beliefs, indicating the "confirmation bias" (Metzger et al., 2015; Zollo, 2019).

In particular, the students with the *pattern "Web Credibility-Evaluating,"* who had established the credibility of the *Süddeutsche Zeitung*, did not even remotely suspect that an article in this trustworthy newspaper might steer their opinion in a certain direction. This pertains to findings from prior research (outlined in section "Research Background"). For instance, none of the students picked up on the fact that Volker Gerhardt's opinion was given in direct speech, while other experts' opinions were only referred to indirectly. In reporting their own search behavior, students may thus not have recognized that narrative perspective and linguistic patterns (direct vs indirect speech) can have an emotional impact on their information processing and

reasoning. In journalistic writing, for example, direct speech can serve to establish an identification between the reader and the person who is being quoted. This identification can occur on the level of content as well as its emotional impact.

Finally, none of the students picked up on the metaphors that Volker Gerhardt used in his defense of assisted suicide (assisted suicide as a human right). Students may thus not have recognized the role of metaphors not only in guiding their reasoning and decision-making, but in having an emotional impact on their reaction to the expert's statement. By linking the students' responses to a narrative analysis of the source they most commonly used, we can thus point to the lacunae in students' COR.

## Discussion

These lacunae can then specifically be targeted in instructional interventions (see section "Narrative Analysis in Educational and Learning Research"). One of the aims of such an intervention would be to enable students to make the best possible use of the Internet as a tool for critical reasoning. Most importantly, such interventions should enable learners to continue searching even after they have arrived at a simple, unambiguous answer (as in the pattern "Unambiguous Fact-Checking"). In students' responses to the CORA task, this became especially manifest in their reaction to the expert. They questioned neither Gerhardt's expertise on the subject at hand (assisted suicide) nor the metaphors he used to steer readers' emotive reaction to support his own opinion.

An instructional intervention can equip students with the skills they need to continue searching even after they have arrived at a simple, unambiguous answer (Berliner, 2020). Through such interventions, students can learn to deal with *ambiguity*, which may lead them to a much more complex grasp of the topic. The Internet may then prove to be the ideal tool to foster their ability to devise complex, multi-faceted responses: it puts them in a position to continue searching for more complex responses. In the CORA task, this would have meant that the students do not stop at one expert (Volker Gerhardt), but rather look for other, alternative experts, and for other, alternative disciplines: from theology to law and medical ethics.

Literary and linguistic analysis may thus be a useful tool to teach students to understand how a given text (as in our case in a newspaper by the *Süddeutsche Zeitung* or elsewhere) may affect them. In the source in question, students are able to relate to the expert (in this case, a professor of philosophy) more directly and in a more personal and possibly, a more affective manner, since all other sources are only referred to in indirect speech. Once students understand this potential bias, they may then search for sources with alternative experts, and their final judgment and decision may differ significantly from the results of the CORA study (see also Nagel et al., 2020).

Our qualitative narrative analysis clearly emphasizes that students' COR is essential when learning with the help of the Internet. This is highly important for our consideration of the Internet as a tool for learning in the information age. In the CORA study, potentially, students would have had a wide variety of source materials available. As our analysis shows, however, that since they followed only the "lead" that had been laid out

for them in the one source of information, they did not use the other materials that they could theoretically have consulted. This is where "wise interventions" may be necessary to train students in COR skills that would allow them to make the best possible use of the Internet as a learning resource, and to enhance their learning and knowledge acquisition.

We illustrate in this article that to design a "wise instructional intervention," it is essential to combine learning data, such as from the CORA study and the methodology of narrative economics and medicine. For an intervention of this kind to be instructionally effective, a qualitative narrative analysis of the material and its content that students use for learning is required. As illustrated in the narrative analysis in this article (see section "Analysis of the Narrative Framing of the Most Commonly Used Online Source"), the extent to which abstract arguments are conveyed through human interest narratives needs to be especially focused: for instance, what (personal) stories are used in the text? By means of which linguistic or rhetorical features is affect achieved? On the basis of the narrative analysis, we can derive a set of hypotheses about how the narrative framing of learning materials can impact students' information processing and their reasoning. The idea which underlies this assumption is that the (learning) source establishes some "leads" to guide their readers' reasoning approaches and decision-making. Our findings from the analyses of students' actual responses, which were provided in short essays in the CORA study (see section "Analyses of Student Responses") suggest one particular hypothesis in this context, namely that most students tend to follow this "lead," since they did not recognize the strategies used in the text to elicit precisely the respective response.

As our study demonstrated, the methodologies of narrative medicine and narrative economics can be used not only for qualitative analysis in educational and learning research but also for teaching interventions. Both methodologies acknowledge that narrative framing is inseparable from content in medicine and economics. When linking this consideration to teaching and learning research, narrative knowledge is seen as a concept which explores how domain-specific content is influenced by the narratives through which it is conveyed (Dettori and Paiva, 2009; Clark, 2010; Goodson et al., 2010). Thus, language is not a neutral "tool" through which (domain-specific) content is conveyed, and it can significantly affect the presentation of content. It is therefore noteworthy that research into the role of affective influence is increasingly being conducted in a number of disciplines and academic fields in recent years, such as, for instance, in law (Bandes and Blumenthal, 2012) or narrative physics (Braid, 2006). Moreover, the attention paid to emotional influence on learning is in line with recent studies in brain research which have addressed the "cognitive emotional brain" (Pessoa, 2013). Prior research indicates that our ways of reasoning may not be guided by rationality alone but by complex processes involving both emotional and rational reasoning (e.g., Damasio, 2000).

In this context, we consider narrative medicine and narrative economics models for teaching intervention. We argue, that when used effectively, they may lead to a modification in students' online search behavior and the increase of their COR

skills. As the analysis of students' responses to the CORA tasks (in section "Analyses of Student Responses") indicate, students may already have a certain degree of critical reasoning when approaching online source material. We suggest, however, that narrative medicine and narrative economics can serve as "wise interventions" and as a practicable teaching tool in higher education which can substantially enhance students' COR skills (see section "Implications for Teaching and Learning in Higher Education in the Internet Age").

## CONCLUSION

### Narrative Analysis in Educational and Learning Research

As illustrated, the quantitative analyses of the CORA data and narrative analysis can be mutually complementary. This points to the fact that linking qualitative and quantitative research is essential when it comes to assessing and explaining students' ability to reason critically in the Internet age. Despite their brevity in short essays, the students' responses are in fact highly complex, and hence need to be evaluated through both qualitative and quantitative analysis.

As a further step along the way in this development, we may thus want to enhance the qualitative and quantitative research outlined in this article through teaching *interventions*. How might students be enabled to understand the role of narrative, and even more importantly, the affective impact created by these narratives? Just as Shiller (2017) stresses the role of affect in *understanding* (and, we would like to add here, in teaching) economic history, affect may also be crucial to consider in one more respect: in the Internet age, students must be able not only to assess, for instance, the trustworthiness of a scientific expert, but also the affective dimension which may accompany the framing of a certain concept or state of affairs by this expert.

In higher education, we should talk to students not only about what sources they use in understanding, for example, certain economic developments, and what they think about the trustworthiness of the sources, but should also teach them to understand how, on the level of narrative structure, these sources "work" and how they shape students' reasoning about a particular subject. By retracing their own reasoning and decision-making process with tools based on the methodologies of narrative economics and narrative medicine, students can enter into a dialog with themselves, as if interrogating an alter ego, about the impact of these sources on their own thinking, and the reasons for this impact (e.g., Sánchez-Martí et al., 2018). Instructional interventions of this kind can be developed by linking qualitative narrative and quantitative empirical research (see section "Methods and Analysis").

In the course of an instructional intervention using the narrative analysis, students can then reflect on their attitude to a given source both before and after using narrative research as a tool to unpack how the argument of a given source "worked." Thus, even prior to actually reading the article on assisted suicide, they may have said that the *Süddeutsche Zeitung* is certainly a reliable and credible source of information. After

conducting a narrative analysis of the article they used, however, they may understand that the article might nonetheless steer their attention in a given direction and could have its own agenda. Understanding such an agenda may be more relevant than ever given the recent scandal of the German news magazine *Spiegel*. In November 2018, it turned out that *Spiegel*, widely credited as one of Germany's major news magazines, had been duped by a journalist who had been fabricating his data for years. In this way and because of narrative research as a method for understanding both, for instance, economic data and its underlying *narratives*, students are no longer at the mercy of the sources but can enter into a dialog with them. In fact, one of the students' responses discussed above indicated his understanding of precisely this dilemma.

The preliminary qualitative analysis presented here suggests that an investigation of students' online search behavior is actually highly complex. How do we begin to tackle this complexity? What happens once students delve deeper into alternative sources? Do they understand that some of these other "experts" have an authority in the debate that may equal or even surpass that of the professor of philosophy whose opinion shapes the *Süddeutsche Zeitung* article? Here, a follow-up empirical study (discussed below) might be conceived of in which students do not search the Internet randomly looking for additional information and in which their search is instead guided by the parameters established in a previous narrative analysis of the original source.

### Limitations and Future Perspectives

There is one particular aspect which this article has addressed in the context of the methodologies of narrative medicine and narrative economics: it related this qualitative methodology to empirical quantitative research and to instructional interventions to promote students' COR skills in higher education in the Internet age. In this context, the CORA tasks were designed to assess students' skills in critically evaluating online sources and reasoning using evidence on contentious issues (Nagel et al., 2020).

Being a newly emerging research field, however, narrative medicine and narrative economics as methodologies have not yet been related to empirical research. In this article, we have proposed that this linkage between empirical research and the methodology of narrative analysis involves the following aspects in particular: it is essential to link the idea of students' COR to the concept of narrative medicine and narrative economics. If indeed, as Shiller (2017) proposes for instance, knowledge in economics is generated through narratives, then narratives can be potentially misleading. They can provide false "leads," or they can even manipulate students into subscribing to certain theories. This may particularly be the case in the Internet age. This article also related the latter aspect to student learning in higher education. It has thus established a link between narrative medicine and economics, student learning in an online environment, and students' COR skills.

In particular, we propose to link narrative medicine as a teaching methodology to Walton's concept of "wise intervention." So far, the relevance of narrative medicine for

students' information-seeking behavior and critical reasoning has not been empirically analyzed. Practitioners of narrative medicine have only argued that, after a narrative medicine intervention, students will approach the clinical setting in a new way (Arntfeld et al., 2013). Going beyond this approach, we want to explore how narrative medicine can change students' online information-seeking behavior, their reasoning and their decision-making, taking into account both content and narrative framing of the information they use. We hypothesize that since narrative medicine enhances students' understanding of the multi-perspective nature of a given medical problem (e.g., all the factors and information that must be considered in diagnosing and treating Alzheimer's disease), their advanced information-seeking behavior – after the intervention – will mirror this understanding. For instance, students may not stop searching after they have located the definition of Alzheimer's disease on Wikipedia, but will continue to search, for instance for patients' experiences or the relationship between Alzheimer's and the social environment, and to evaluate and critically reflect on the different pieces of information. In terms of economics, a similar process of a more critical information-seeking behavior may result from the use of narrative economics as a model for teaching economics, which may potentially be transferred to other domains (e.g., sociology).

Based on our results presented here, we suggest that we are only at the beginning of a new universe of research which is only beginning to take shape. In future research, it is essential not only to study the role of narratives for the acquisition of domain-specific knowledge in economics or medicine, but also to anchor such narrative research in domain-specific teaching and learning *per se*. Thus, narrative scholars have to collaborate with researchers and instructors from the respective domains. These latter experts also need to act as “fact checkers”: while knowledge may be narratively constructed, we still have to subscribe to the idea of a verifiable and warranted knowledge base in a given domain. Despite narrative variation, for instance economists will then generally agree on the veracity of a certain idea. For education research at a university level, it is essential that we do not jettison the belief in domain-specific knowledge. Rather, the relationship between domain-specific knowledge on the one hand and “narrative knowledge” on the other is at stake. Students hence need to be equipped with certain skills, COR being the most important among them. It is to the assessment of students' COR skills in the Internet age that this article has sought to contribute by linking narratives and certain domains in the field of higher education research.

As a future perspective, the findings about narrative knowledge in one domain (e.g., economics) need to be mapped onto another domain (e.g., medicine). Research of this kind seems highly promising, as outlined in this article. In this context, one gap in narrative medicine may be discerned: while narrative medicine has already been fruitfully linked to the didactics of medicine, few studies have empirically tested narrative medicine interventions in the medical classroom and their impact on students' learning (e.g., McAllister, 2015). For further investigation of narrative medicine in this context, two steps are required: first, narrative medicine

must be reconceptualized with regard to concepts such as “deeper learning” (Pellegrino and Hilton, 2012). Through this reconceptualization, narrative medicine would be linked to both education and learning research, which has not been the case so far. Second, empirical studies should be conducted, which would again combine qualitative and quantitative research, focusing for instance on the condition of dementia as it is currently being taught in the biomedical classroom. An empirical study similar to CORA here would gauge the way in which students' understanding of dementia is shaped by narrative, and would give medical students the task of defining dementia through the use of Internet sources. Through short questions and essay answers, researchers could assess students' ability to critically evaluate information about dementia, from its biomedical definition to its societal and ethical challenges.

When conducting follow-up empirical studies, some limitations of the present study, such as the limited representativeness of the sample, should be overcome to increase the generalizability of the findings. For instance, the identified patterns and profiles may vary depending on students' personal characteristics (e.g., for year of study and advanced education, see, Togia and Korobili, 2014). Based on a more balanced sample, for instance in terms of gender, study year and study domain, the possible relationship between these profiles and students' characteristics needs to be investigated. In particular, there might be domain-related differences in the identified student profiles, which did not become evident in this study due to the low proportion of medical students in this sample. However, as shown in another article (Nagel et al., 2020), we did not find any significant differences, neither in the students' task performance level nor in their response processes (based on the log file analyses), between students from the two domains. This finding is in line with other existing studies, where no significant differences in information-seeking between students from different domains were reported (e.g., Stover and Mabry, 2020). Therefore, we could hardly assume substantial differences in these profiles depending on the study domain. However, this question needs to be systematically investigated in a follow-up study.

## Implications for Teaching and Learning in Higher Education in the Internet Age

What would the narrative analysis mean for “wise interventions”? What happens if, in teaching, some of the narrative strategies used in the source were discussed with the students? So far, it might be argued that there is a sense in which their own reasoning and decision-making process would to some extent be a black box even to themselves. On the one hand, as their responses discussed in section “Analyses of Student Responses” indicated, most students were able to critically evaluate their sources, and they were able to distinguish the *Süddeutsche Zeitung* from other print sources which may be less reliable in terms of information, such as the German tabloid newspaper *Bildzeitung*. On the other hand, however, they may not (fully) recognize the underlying narratives or their affective cues which may lie beneath the respectability of the source in which the article is contained.



Seen from this perspective, the source may be both credible and reliable; yet, on the narrative and affective level, it may nonetheless steer the reader's reasoning and decision-making in a particular direction.

As a part of an instructional intervention, a narrative analysis of the source most commonly used by students in the CORA task, for instance, could enable students to overcome the "authority bias" that many of their responses implied. In this context, narrative 'competence' might also have provided them with a tool to interrogate the expert's credentials. The point here would have been not so much that Volker Gerhardt is a professor (with some of the students referring to him only as "the professor"), but to ask why his discipline would make him an expert on the topic of assisted suicide.

Once the students are able to understand the narrative tools and metaphors which the source uses to evoke a particular affect, then they may be able to retrace their own reasoning and decision-making process. This way, students may be enabled to see how narratives inevitably guide their understanding and may maneuver them into a certain reasoning direction. Through this understanding, they would be able to resist and critically evaluate these maneuvers. What other sources or perspectives, they may ask, has the text omitted? For instance, once they recognize, through a combination of linguistic and narrative analyses, the affective impact of direct over reported speech, they may question their identification with one particular speaker in the source. They may then look up some of the other experts who were only referred to in passing in the text, and whose opinion was reported only in reported speech.

There is a challenge here which emerges for teaching and learning research in particular. Research has to look into models for intervention: how could interventions be designed that equip students for COR? From what models could researchers draw in order to develop effective instructional interventions? (Berliner, 2020). To tackle these issues, we argue that it is necessary to develop multi-disciplinary models that are able to link empirical quantitative research and qualitative content analysis. In this article, we explore the role of narrative qualitative analysis in this context. Within the framework of education as well as learning research, content analysis is commonly used (e.g., McQuiggan et al., 2008; Kessler and Guenther, 2016). However, it is important to note that "content" is mostly analyzed from the perspective and through the methodology of a particular discipline (e.g., economics) as well as based on its cognitive components (following the established taxonomies, e.g., Anderson and Krathwohl, 2001). While highly productive on a number of levels, studies and analyses of this kind provide little insight into the role of the emotional factors of learning from the texts analyzed or the narrative structures through which they are constituted.

In the Internet age, students are confronted with a wide variety of source material. To solve a given task, they will "automatically" not only draw on the source material provided by the instructor, but will rely on multiple online sources (e.g., Maurer et al., 2020). One of the challenges for teaching models in higher education thus depends on providing students with the COR skills to successfully navigate online environments and to assess the nature and quality of the sources they find online.

We have proposed in this article, however, that tackling the issue of learning in online environments is not only related to the use of multiple sources (Britt and Rouet, 2012), and also requires students to be able to decode the ways in which a given text can affect them emotionally and hence shape their reasoning and decision-forming process in a particular way. Our findings from the narrative analyses of the CORA data material suggest that the role of narrative and its potential emotional impact is central not only to the ways we teach, as studies on "deeper learning" have shown (Pellegrino and Hilton, 2012), but that assessing the emotional impact of source material is also a key skill in student learning in the Internet age.

This has profound implications for the relevance of narratives and narrative analysis in medicine and economics. In conceptualizing interventions aimed at enhancing students' COR, we turn to narrative medicine and narrative economics not only for qualitative analysis but also as a basis for teaching interventions. Both fields stress the importance of narrative knowledge (Kreiwirth, 2000), domain-specific content, and the affective influence of narrative framing on learning.

Through the methodology of narrative economics, for instance, Shiller (2017) essentially argues that we tend to best remember domain-specific content – the factors which led to the Great Depression, for example – when it is told as a human-interest narrative. We relate to "narratives of other humans," as Shiller (2017) puts it. This can be related back to students' responses to the CORA task: their written essays indicated that students were able to relate not so much to the abstract topic of assisted suicide – which may seem remote from their own life-worlds – but to the "story" told by Volker Gerhardt, the expert. They related to Gerhardt as a person through whom the entire topic was framed, including the affect which accompanied his narrative.

Narrative medicine and narrative economics thus stress two facts in particular that are essential for the instructional interventions that we are proposing in this article: first, these fields state that no linguistic representation of domain-specific content is ever "neutral." Rather, it is conveyed through "story-telling," through narratives which have specific features, such as narrative perspective (first-person or third-person narration), metaphors, structure and mode of speech.

Second, narrative medicine in particular emphasizes that medical students and physicians need to understand how such narratives and their emotional impact may shape or even guide their information-seeking behavior and reasoning. It is important, narrative medicine argues, to understand the emotional impact of narratives on our decision-making and actions in order not to be "manipulated" by them. For this reason, narrative medicine is increasingly becoming a key component of medical education: its aims to teach medical students to understand the potential impact of narrative representation on their own recognition of domain-specific content (e.g., McAllister, 2015). This is especially pronounced when it comes to medical metaphors: as we illustrated with the example of takotsubo, the linguistic representation of a given medical condition can have a direct impact on how physicians diagnose this condition. Thus, in the case of the broken-heart syndrome, the metaphor

“backfired,” physicians failed to take patients’ symptoms seriously and hence underestimated the fact that this could be a somatic, and not just a psychosomatic, condition. As this example shows, narrative medicine is a potential approach as an instructional intervention to teaching medical students to understand how language and narrative can guide their information seeking behavior, the acquisition of domain-specific knowledge and their diagnostic competence.

It is for this reason that we argue in this article that narrative economics may well follow the example of narrative medicine when it comes to developing teaching interventions in higher education. Just as narrative medicine is meant to teach medical students to pay attention to language and narrative, narrative economics may serve to enhance the narrative competence of economics students. In both cases, the lack of understanding of narratives can lead to shallow processing or insufficient reasoning, misconceptions and erroneous knowledge and beliefs (Stanovich, 2003, 2016; Song, 2011).

So far, however, research in narrative economics has been located at the level of basic research. Its aim has been to study the relevance of narratives about a given economic development both at the level of memory and of emotional impact (e.g., Delafield-Butt and Adie, 2016). In this article, we argue that both processes – memory and affect – are also key to *teaching* economics; and they may be essential for “wise interventions” aimed at enhancing students’ COR skills with regard to the critical use of online sources when solving domain-specific tasks in economics.

In this way, the integration of qualitative and quantitative analysis can significantly contribute to understanding and explaining students’ information processing and their COR. Once students are equipped with a methodology to understand, both on the level of content and of its emotional impact, how the source guides their search path and reasoning, they may recognize that they did not find these sources randomly, but that their search path was itself shaped by the twist which the source gave to the question – in this case, assisted suicide – both on the level of content and in terms of its emotional impact. Bringing together narrative qualitative analysis and quantitative research may therefore be highly fruitful, and exceeds the capacity of each of the individual approaches used (Shiller, 2017).

In this medical context as in the context of narrative economics, empirical studies of this kind are key in linking the acquisition of domain-specific knowledge to students’ COR skills. It can be argued that including narrative knowledge as a concept in education and learning research can be an important contribution to investigating this link. In this context, narrative economics and narrative medicine could also be used as an instructional intervention. Once students have been made familiar with the methodology of narrative economics and

medicine, they could be given an Internet-based task. Researchers would then be able to assess students’ understanding of the narrative functions of the sources used, for instance to define dementia in a medical classroom. In this vein, in the domain of medicine as much as economics, students would be “inoculated” to manipulation by Internet sources, or they would at least be able to understand the potential emotional impact these sources can create. A methodological tool using narrative analysis, we have suggested in this article, is essential both for the acquisition of domain-specific knowledge and for COR in the information age.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was ensured in this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

MB provided the idea for the study, conducted the analyses, and co-wrote the manuscript. OZ-T developed the assessment, supported the analyses, and co-wrote the manuscript. JR was involved in the data collection and in preparing and reviewing the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was part of the Rhine-Main Universities Alliance (RMU) project, which was funded by the RMU fund.

## ACKNOWLEDGMENTS

We would like to thank the two reviewers and the editor who provided constructive feedback and helpful guidance in the revision of this manuscript. We would like to thank all students from the Medical Faculty of Goethe University Frankfurt and from the Faculty of Law and Economics at Johannes Gutenberg University Mainz who participated in this study.

## REFERENCES

- Alexander, P., Murphy, K., and Sun, Y. (2018). “Knowledge and belief change in academic development,” in *The Model of Domain Learning: Understanding the Development of Experts*, eds H. Fives and D. Dinsmore (New York, NY: Routledge).
- Anderson, L. W., and Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives*. New York, NY: Longman.
- Armstrong, S., and Brunskill, P. (2018). *Information Literacy: Separating Fact From Fiction*. Huntington Beach, CA: Teacher Created Materials Inc.

- Arntfeld, S., Slesar, K., Dickson, J., and Charon, R. (2013). Narrative medicine as a means of training medical students towards residency competence. *Patient Educ. Couns.* 91, 280–286. doi: 10.1016/j.pec.2013.01.014
- Bandes, S., and Blumenthal, J. (2012). Emotion and the law. *Annu. Rev. Law Soc. Sci.* 8, 161–181.
- Banerjee, M. (2018). *Medical Humanities in American Studies: Life Writing, Narrative Medicine and the Power of Autobiography*. Heidelberg: Universitätsverlag Winter.
- Berliner, D. (2020). “The role of modeling for ‘seeking truth’ in an educational policy classroom,” in *Frontiers and Advances in Positive Learning in the Age of Information*, ed. O. Zlatkin-Troitschanskaia (Wiesbaden: Springer), 21–39. doi: 10.1007/978-3-030-26578-6\_3
- Braasch, J. L. G., Bråten, I., and McCrudden, M. T. (2018). *Handbook of Multiple Source Use*. London: Routledge.
- Braid, D. (2006). ‘Doing good physics’: narrative and innovation in research. *J. Folk Res.* 43, 149–173. doi: 10.1353/jfr.2006.0012
- Brand-Gruwel, S., Kammerer, Y., van Meeuwen, L., and van Gog, T. (2017). Source evaluation of domain experts and novices during Web search. *J. Comput. Assist. Learn.* 33, 234–251. doi: 10.1111/jcal.12162
- Brand-Gruwel, S., Wopereis, I., and Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Comput. Educ.* 53, 1207–1217. doi: 10.1016/j.compedu.2009.06.004
- Britt, A., and Rouet, J.-F. (2012). “Learning with multiple documents: component skills and their acquisition,” in *Enhancing the Quality of Learning*, eds J. R. Kirby and M. J. Lawson (Cambridge: Cambridge University Press), 276–314.
- Bromme, R., and Thomm, E. (2016). Knowing who knows: Laypersons’ capabilities to judge experts’ pertinence for science topics. *Cogn. Sci.* 40, 241–252. doi: 10.1111/cogs.12252
- Carroll, N. (2001). “On the narrative connection,” in *New Perspectives on Narrative Perspective*, eds W. van Peer and S. Chatman (Albany, NY: SUNY Press).
- Charon, R. (2008). *Narrative Medicine: Honoring the Studies of Illness*. Oxford: Oxford University Press.
- Charon, R., Dasgupta, S., Herman, N., Irvine, C., Marcus, E. E., Colón, E. R., et al. (2017). *The Principles and Practice of Narrative Medicine*. Oxford: Oxford University Press.
- Chiu, Y.-L., Liang, Y.-C., and Tsai, C.-C. (2013). Internet-specific epistemic beliefs and self-regulated learning in online academic information searching. *Metacogn. Learn.* 8, 235–260. doi: 10.1007/s11409-013-9103-x
- Clark, M. C. (2010). Narrative learning: its contours and its possibilities. *N. Direct. Adult Contin. Educ.* 2010, 3–11. doi: 10.1002/ace.367
- Collins-Thompson, K., Soo Young, R., Haynes, C. C., and Syed, R. (2016). Assessing learning outcomes in web search: a comparison of tasks and query strategies. *CHIIR 2016*, 163–172. doi: 10.1145/2854946.2854972
- Craig, I., and Charon, R. (2017). “Deliver us from certainty: thinking for narrative ethics,” in *The Principles and Practice of Narrative Medicine*, eds R. Charon, et al. (Oxford: Oxford University Press), 110–135.
- Damasio, A. (2000). *The Feeling of What Happens: Body, Emotion, and the Making of Consciousness*. New York, NY: Vintage.
- de los Santos, T. M., and Nabi, R. L. (2019). Emotionally charged: exploring the role of emotion in online news information seeking and processing. *J. Broadcast. Electron. Media* 63, 39–58. doi: 10.1080/08838151.2019.1566861
- Delafeld-Butt, J. T., and Adie, J. (2016). The embodied narrative nature of learning: nurture in school. *Mind Brain Educ.* 10, 117–131. doi: 10.1111/mbe.12120
- Dettori, G., and Paiva, A. (2009). “Narrative learning in technology-enhanced environments,” in *Technology-Enhanced Learning*, eds N. Balacheff, S. Ludvigsen, T. de Jong, A. Lazonder, and S. Barnes (Dordrecht: Springer), 55–69. doi: 10.1007/978-1-4020-9827-7\_4
- Efferth, T., Banerjee, M., and Paul, N. W. (2017). Broken heart, tako-tsubo or stress cardiomyopathy? Metaphors, meanings and their medical impact. *Int. J. Cardiol.* 230, 262–268. doi: 10.1016/j.ijcard.2016.12.129
- Gibbs, R. (1994). *The Poetics of Mind: Figurative Thought, Language, and Understanding*. New York, NY: Cambridge University Press.
- Gierth, L., and Bromme, R. (2020). Attacking science on social media: how user comments affect perceived trustworthiness and credibility. *Publ. Understand. Sci.* 29, 230–247. doi: 10.1177/0963662519889275
- Goldman, J. (2014). *Can You Die of a Broken Heart?*. Available online at: <https://www.bbc.com/future/article/20140331-can-you-die-of-a-broken-heart> (accessed on 31 May 2020)
- Goodson, I., Biesta, G., Tedder, M., and Adair, N. (2010). *Narrative Learning*. London: Routledge.
- Gould, S. J. (1980). *The Panda’s Thumb: More Reflections on Natural History*. New York, NY: Norton.
- Hahnel, C., Kroehne, U., Goldhammer, F., Schoor, C., Mahlow, N., and Artelt, C. (2019). Validating process variables of sourcing in an assessment of multiple document comprehension. *Br. J. Educ. Psychol.* 89, 524–537. doi: 10.1111/bjep.12278
- Halln, F. (ed.) (2000). *Metaphors, and Analogy in the Sciences*. Wiesbaden: Springer.
- Harrison, N., and Luckett, K. (2019). Experts, knowledge and criticality in the age of ‘alternative facts’: re-examining the contribution of higher education. *Teach. High. Educ.* 24, 259–271. doi: 10.1080/13562517.2019.1578577
- Hartley, J. (2017). *The Uses of Digital Literacy*. New York, NY: Routledge.
- Hendriks, F., Kienhues, D., and Bromme, R. (2015). Measuring Laypeoples’ trust in experts in a digital age: the Muenster epistemic trustworthiness inventory (METI). *PLoS One* 10:e0139309. doi: 10.1371/journal.pone.0139309
- Hoppe, A., Holtz, P., Kammerer, Y., Yu, R., Dietze, S., and Erwerth, R. (2018). “Current challenges for studying search as learning processes,” in *Proceedings of 7th Workshop on Learning and Education with Web Data*, Amsterdam.
- Hsu, C.-H., Tsai, M.-J., Hou, H.-T., and Tsai, C.-C. (2014). Epistemic beliefs, online search strategies, and behavioral patterns while exploring socioscientific issues. *J. Sci. Educ. Tech.* 23, 471–480. doi: 10.1007/s10956-013-9477-1
- Johnson, F., Shaffi, L., and Rowley, J. (2016). Students’ approaches to the evaluation of digital information: insights from their trust judgments. *Br. J. Educ. Technol.* 47, 1243–1258. doi: 10.1111/bjet.12306
- Kahne, J., and Bowyer, B. (2017). Educating for democracy in a partisan age: confronting the challenges of motivated reasoning and misinformation. *Am. Educ. Res. J.* 54, 3–34. doi: 10.3102/0002831216679817
- Kay, L. (2000). *Who Wrote the Book of Life: A History of the Genetic Code*. Stanford: Stanford University Press.
- Kessler, S. H., and Guenther, L. (2016). Eyes on frame. *Int. Res.* 27, 303–320. doi: 10.1108/IntR-01-2016-0015
- Kreiswirth, M. (2000). Merely telling stories? Narrative and knowledge in the human sciences. *Poet. Today* 21, 293–318. doi: 10.1215/03335372-21-2-293
- Kunda, Z. (1990). The case for motivated reasoning. *Psychol. Bull.* 108, 480–498. doi: 10.1037/0033-2909.108.3.480
- List, A., and Alexander, P. A. (2017). Cognitive affective engagement model of multiple source use. *Educ. Psychol.* 52, 182–199. doi: 10.1080/00461520.2017.1329014
- List, A., and Alexander, P. A. (2018). “Cold and warm perspectives on the cognitive affective engagement model of multiple source use,” in *Handbook of Multiple Source Use*, eds J. L. G. Braasch, I. Bråten, and M. T. McCrudden (New York, NY: Routledge), 34–54. doi: 10.4324/9781315627496-3
- Lucassen, T., and Schraagen, J. M. (2011). Factual accuracy and trust in information: the role of expertise. *J. Am. Soc. Inf. Sci. Technol.* 62, 1232–1242. doi: 10.1002/asi.21545
- Luong, K. T., Moyer-Gusé, E., and McKnight, J. (2020). Let’s go to the movies... for science! The impact of entertainment narratives on science knowledge, interest, and information-seeking intention. *J. Media Psychol.* 12, 1–16. doi: 10.1027/1864-1105/a000272
- Malpas, J., and Zabala, S. (eds) (2010). *Consequences of Hermeneutics: Fifty Years After Gadamer’s Truth and Method*. Evanston: Northwestern University Press.
- Maurer, M., Schemer, C., Zlatkin-Troitschanskaia, O., and Jitomirski, J. (2020). “Positive and negative media effects on university students’ learning: preliminary findings and a research program,” in *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*, ed. O. Zlatkin-Troitschanskaia (New York, NY: Springer), 109–119. doi: 10.1007/978-3-030-26578-6\_8
- McAllister, M. (2015). Connecting narrative with mental health learning through discussion and analysis of selected contemporary films. *Int. J. Ment. Health Nurs.* 24, 304–313. doi: 10.1111/inm.12134
- McGrew, S., Breakstone, J., Ortega, T., Smith, M., and Wineburg, S. (2018). Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory Res. Soc. Educ.* 46, 165–193. doi: 10.1080/00933104.2017.1416320
- McQuiggan, S. W., Rowe, J. P., Lee, S., and Lester, J. C. (2008). “Story-based learning: the impact of narrative on learning experiences and outcomes,” in



- Intelligent Tutoring Systems*, eds B. P. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie (Berlin: Springer), 530–539. doi: 10.1007/978-3-540-69132-7\_56
- Metzger, M. J., and Flanagin, A. J. (2015). Credibility and trust of information in online environments: the use of cognitive heuristics. *J. Prag.* 59, 210–220. doi: 10.1016/j.pragma.2013.07.012
- Metzger, M. J., Flanagin, A. J., Markov, A., Pure, R., and Bulger, M. (2015). Believing the unbelievable: Understanding young people's information literacy beliefs and practices in the United States. *J. Child. Media* 9, 325–348.
- Metzger, M. J., Flanagin, A. J., and Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *J. Commun.* 60, 413–439. doi: 10.1111/j.1460-2466.2010.01488.x
- Molerov, D., Zlatkin-Troitschanskaia, O., and Schmidt, S. (2019). *Adapting the Civic Online Reasoning Assessment for Cross-National Use*. Toronto: American Education Research Association (AERA).
- Nagel, M.-T., Schäfer, S., Zlatkin-Troitschanskaia, O., Schemer, C., Maurer, M., Molerov, D., et al. (2020). How do university students' web search behavior, website characteristics, and the interaction of both influence students' critical online reasoning? *Front. Edu.* doi: 10.3389/educ.2020.565062
- Neuenhaus, N., Artelt, C., and Schneider, W. (2013). The impact of cross-curricular competences and prior knowledge on learning outcomes. *Int. J. High. Educ.* 2, 214–227.
- Oser, F. K., and Biedermann, H. (2020). “A three-level model for critical thinking: critical alertness, critical reflection, and critical analysis,” in *Frontiers and Advances in Positive Learning in the Age of informaTiOn (PLATO)*, ed. O. Zlatkin-Troitschanskaia (Cham: Springer), 89–106. doi: 10.1007/978-3-030-26578-6\_7
- Pellegrino, J., and Hilton, M. (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: National Research Council.
- Pessoa, L. (2013). *The Cognitive Emotional Brain: From Interactions to Integration*. Cambridge, MA: MIT Press.
- Pinker, S. (2007). *The Stuff of Thought: Language as a Window in Human Nature*. New York, NY: Penguin Books.
- Powell, T. E., Boomgaarden, H. G., de Swert, K., and de Vreese, C. H. (2019). Framing fast and slow: a dual processing account of multimodal framing effects. *Media Psychol.* 22, 572–600. doi: 10.1080/15213269.2018.1476891
- Rumelhart, D. E. (1979). “Some problems with the notion of literal meanings,” in *Metaphor and Thought*, ed. A. Ortony (Cambridge: Cambridge University Press).
- Sánchez-Martí, A., Sabariego Puig, M., Ruiz-Bueno, A., and Anglés Regós, R. (2018). Implementation and assessment of an experiment in reflective thinking to enrich higher education students' learning through mediated narratives. *Think. Skills Creat.* 29, 12–22. doi: 10.1016/j.tsc.2018.05.008
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. (2019). Assessment of university students' critical thinking: next generation performance assessment. *Int. J. Test.* 19, 337–362. doi: 10.1080/15305058.2018.1543309
- Shiller, R. (2017). *Economics and the Human Instinct for Storytelling*. Available online at: <https://review.chicagobooth.edu/economics/2017/article/economics-and-human-instinct-storytelling> (accessed on 30 December 2019).
- Shiller, R. (2019). *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*. Princeton, NJ: Princeton University Press.
- Song, M. (2011). *Effects of Background Context and Signaling on Comprehension Recall and Cognitive Load: The Perspective of Cognitive Load Theory*. Unpublished doctoral dissertation. Lincoln, NE: University of Nebraska.
- Spencer, D. (2020). *Metagnosis: Revelatory Narratives of Health and Identity*. Oxford: Oxford University Press.
- Spiegel, M., and Spencer, D. (2017). “This is what we do, and these things happen: literature, experience, emotion, and relationality in the classroom,” in *The Principles and Practice of Narrative Medicine*, eds R. Charon, et al. (Oxford: Oxford University Press), 37–59.
- Stanovich, K. E. (2003). “The fundamental computational biases of human cognition: heuristics that (sometimes) impair decision making and problem solving,” in *The Psychology of Problem Solving*, eds J. E. Davidson and R. J. Sternberg (Cambridge: Cambridge University Press), 291–342. doi: 10.1017/CBO9780511615771.011
- Stanovich, K. E. (2016). The comprehensive assessment of rational thinking. *Educ. Psychol.* 51, 23–34. doi: 10.1080/00461520.2015.1125787
- Stanovich, K. E. (2018). Miserliness in human cognition: the interaction of detection, override and mindware. *Think. Reason.* 24, 423–444. doi: 10.1080/13546783.2018.1459314
- Stanovich, K. E., West, R. F., and Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Curr. Dir. Psychol. Sci.* 22, 259–264. doi: 10.1177/0963721413480174
- Thibodeau, P. H., and Boroditsky, L. (2011). Metaphors we think with: the role of metaphor in reasoning. *PLoS One* 6:e16782. doi: 10.1371/journal.pone.0016782
- Trzebinski, J. (1995). “Narrative self, understanding, and action,” in *The Self in European and North American Culture: Developments and Processes*, eds A. Oosterwegel and R. Wicklund (Wiesbaden: Springer), 73–88. doi: 10.1007/978-94-011-0331-2\_7
- Ulyshen, T. Z., Koehler, M. J., and Gao, F. (2015). Understanding the connection between epistemic beliefs and internet searching. *J. Educ. Comput. Res.* 53, 345–383. doi: 10.1177/0735633115599604
- van Strien, J. L. H., Brand-Gruvel, S., and Boshuizen, H. P. A. (2014). Dealing with conflicting information from multiple nonlinear texts: effects of prior attitudes. *Comput. Hum. Behav.* 32, 101–111. doi: 10.1016/j.chb.2013.11.021
- van Strien, J. L. H., Kammerer, Y., Brand-Gruvel, S., and Boshuizen, H. P. A. (2016). How attitude strength biases information processing and evaluation on the web. *Comput. Hum. Behav.* 60, 245–252. doi: 10.1016/j.chb.2016.02.057
- Walton, G. M. (2014). The new science of wise psychological interventions. *Curr. Dir. Psychol. Sci.* 23, 73–82. doi: 10.1177/0963721413512856
- Wineburg, S., Smith, M., and Breakstone, J. (2018). What is learned in college history classes? *J. Am. Hist.* 104, 983–993. doi: 10.1093/jahist/jax434
- Yu, R., Gadiraju, U., Holtz, P., Rokicki, M., Kemkes, P., and Dietze, S. (2018). *Predicting User KNOWLEDGE GAIN in Informational Search Sessions*. Available online at: <http://arxiv.org/pdf/1805.00823v1> (accessed on 16 May 2020)
- Zhou, M., and Ren, J. (2016). “Use of cognitive and metacognitive strategies in online search: an eye-tracking study,” in *Proceedings of the International Conferences on Internet Technologies & Society (ITS), Education Technologies (ICEDuTECH), and Sustainability, Technology and Education (STE)*, eds P. Kommers, I. Tomayess, I. Theodora, E. McKay, and P. Isias Melbourne.
- Zlatkin-Troitschanskaia, O., Beck, K., Fischer, J., Braunheim, D., Schmidt, S., and Shavelson, R. J. (2020). The role of students' beliefs when critically reasoning from multiple contradictory sources of information in performance assessments. *Front. Psychol.* 11:2192. doi: 10.3389/fpsyg.2020.02192
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., and Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *Br. J. Educ. Psychol.* 89, 468–484. doi: 10.1111/bjep.12286
- Zollo, F. (2019). Dealing with digital misinformation: a polarised context of narratives and tribes. *EFSA J.* 17:e170720. doi: 10.2903/j.efsa.2019.e170720

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Banerjee, Zlatkin-Troitschanskaia and Roeper. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Test-Taking Motivation in Education Students: Task Battery Order Affected Within-Test-Taker Effort and Importance

Anett Wolgast<sup>1\*</sup>, Nico Schmidt<sup>2</sup> and Jochen Ranger<sup>2</sup>

<sup>1</sup> Department of Psychology, University of Applied Sciences Hannover, Hanover, Germany, <sup>2</sup> Institute of Psychology, Martin-Luther-University Halle-Wittenberg, Halle, Germany

## OPEN ACCESS

### Edited by:

Olga Zlatkin-Troitschanskaia,  
Johannes Gutenberg University  
Mainz, Germany

### Reviewed by:

Marion Händel,  
University of Erlangen Nuremberg,  
Germany  
Martin Hecht,  
Humboldt University of Berlin,  
Germany

### \*Correspondence:

Anett Wolgast  
anett.wolgast@gmail.com

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Psychology

**Received:** 06 May 2020

**Accepted:** 30 October 2020

**Published:** 25 November 2020

### Citation:

Wolgast A, Schmidt N and  
Ranger J (2020) Test-Taking  
Motivation in Education Students:  
Task Battery Order Affected  
Within-Test-Taker Effort  
and Importance.  
Front. Psychol. 11:559683.  
doi: 10.3389/fpsyg.2020.559683

Different types of tasks exist, including tasks for research purposes or exams assessing knowledge. According to expectation-value theory, tests are related to different levels of effort and importance within a test taker. Test-taking effort and importance in students decreased over the course of high-stakes tests or low-stakes-tests in research on test-taking motivation. However, whether test-order changes affect effort, importance, and response processes of education students have seldomly been experimentally examined. We aimed to examine changes in effort and importance resulting from variations in test battery order and their relations to response processes. We employed an experimental design assessing  $N = 320$  education students' test-taking effort and importance three times as well as their performance on cognitive ability tasks and a mock exam. Further relevant covariates were assessed once such as expectancies, test anxiety, and concentration. We randomly varied the order of the cognitive ability test and mock exam. The assumption of intraindividual changes in education students' effort and importance over the course of test taking was tested by one latent growth curve that separated data for each condition. In contrast to previous studies, responses and test response times were included in diffusion models for examining education students' response processes within the test-taking context. The results indicated intraindividual changes in education students' effort or importance depending on test order but similar mock-exam response processes. In particular effort did not decrease, when the cognitive ability test came first and the mock exam subsequently but significantly decreased, when the mock exam came first and the cognitive ability test subsequently. Diffusion modeling suggested differences in response processes (separation boundaries and estimated latent trait) on cognitive ability tasks suggesting higher motivational levels when the cognitive ability test came first than vice versa. The response processes on the mock exam tasks did not relate to condition.

**Keywords:** expectation-value theory, diffusion modeling, latent growth curve modeling, perspective-taking, exam test-taking motivation

## INTRODUCTION

Researchers analyzing data from the Programme for International Student Assessment (PISA) concerning the relations between motivation and test-taking achievement in mathematics reported that motivation explained 1–29% of the variance in achievement-test results (Kriegbaum et al., 2014). Further findings suggested item position effects on test performance (Weirich et al., 2016; Nagy et al., 2018, 2019; Rose et al., 2019; Liu and Hau, 2020). A problem found was decreased test performance over the course of taking a computer-assisted achievement test (List et al., 2017). That raised the question if motivation similarly decreased over the course of taking a computer-assisted achievement test. Researchers found low test-taking effort related to low test performance, discussed and tested several strategies for test takers' high effort levels, for example, incentives, integration into grading systems, or explaining test takers the relevance and importance of PISA test results (Baumert and Demmrich, 2001; Finn, 2015; Schüttpeitz-Brauns et al., 2020). Without applying any strategy to increase test-takers' effort, researchers found decreased intraindividual effort over the course of taking a test, this time effort of apprentices (technicians, clerks, and lab assistants, Lindner et al., 2018).

Decreasing test-taking effort is a serious problem since (computer-assisted) test-taking performance reflects an unknown amount of the tested ability in this case and threatens validity for the examined sample (Penk and Richter, 2017; Nagy et al., 2018). Furthermore, decreasing test-taking effort in a mock exam might affect achievement related choices (e.g., respond vs. not respond on a computer-assisted task) and the subsequent learning behavior in preparation of the exam. Achievement related choices in computer-assisted tasks regard test-takers' information processing. Undergraduate students in higher education often have the choice, if they respond on a computer-assisted task in a mock exam, and how much effort they spend on different types of task.

Effort is usually described as a component of achievement motivation (Eccles et al., 1984; Eccles and Wigfield, 1995; Wigfield and Eccles, 2000) or test-taking motivation (Wise and DeMars, 2005; Knekta and Eklöf, 2015; Knekta, 2017). Test-taking motivation is the engagement and effort that a person applies to a goal in order to achieve the best possible result in an achievement test (Wise and DeMars, 2005). Invested effort is conceptualized as relevant predictor on test performance according to the expectancy-value model applied to a test situation (Knekta and Eklöf, 2015, p. 663).

Knekta and Eklöf (2015) investigated adolescents' test-taking motivation and academic achievement, alongside further motivational components such as test-taking importance, expectancies, anxiety, and interest. A great deal of evidence supports the relevance of these motivational components for test performance (Eklöf and Hopfenbeck, 2018) and academic achievement (e.g., Nagy et al., 2010). For example, self-reported expectancies, test-taking effort, and test-taking importance have been found to determine performance in high-stakes tests (e.g., Knekta, 2017; Eklöf and Hopfenbeck, 2018); as one would

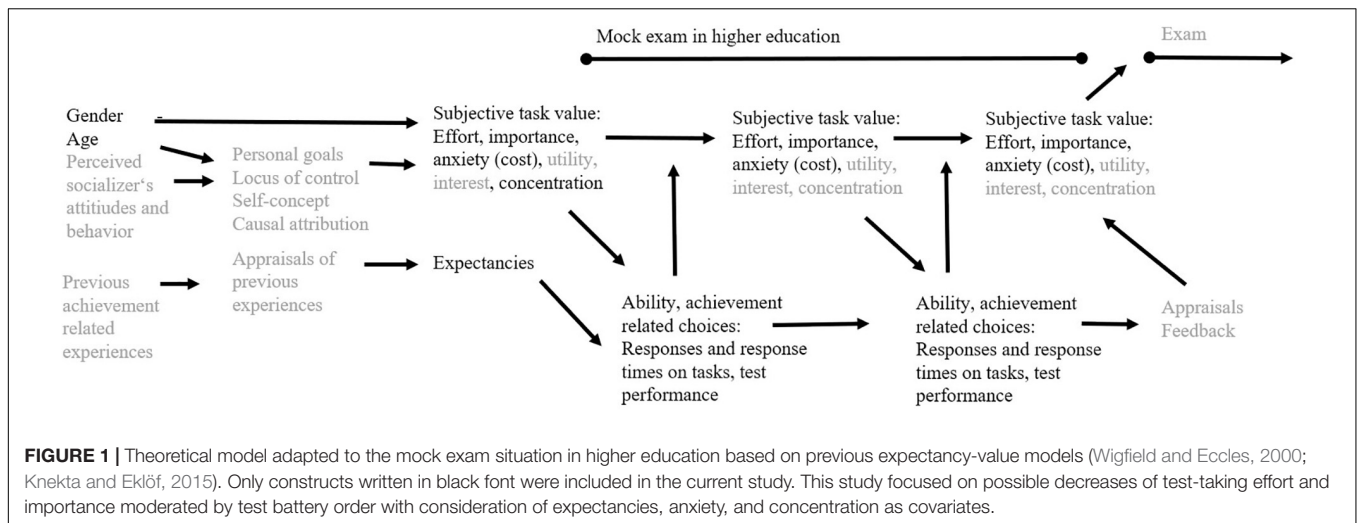
expect, high levels of these motivation components predicted higher performance levels. In low-stakes achievement tests, the relations between test-taking effort or test-taking importance (assessed by self-reports) and performance were inconsistently at zero (Sundre and Kitsantas, 2004; Penk and Richter, 2017) to low levels (Knekta, 2017; Eklöf and Hopfenbeck, 2018; Myers and Finney, 2019).

High-stakes achievement tests usually refer to ability assessments for selection purposes (e.g., enrollment in a type of school, internship, study program, or exams to complete a study program, e.g., Knekta, 2017). Low-stakes tests considered in previous research have included tests to practice high stakes-tests (e.g., mock exams), tests to evaluate educational programs (Brunner et al., 2007; Butler and Adams, 2007), tests to develop or update standardized achievement test inventories (e.g., standardization in a representative sample), and cognitive ability tests for research purposes (McHugh et al., 2004; Erle and Topolinski, 2015; Gorges et al., 2016; Goldhammer et al., 2017). Cognitive ability tests for research purposes and tests of subject content knowledge are often used in international large-scale assessments including students in school (Baumert and Demmrich, 2001; Butler and Adams, 2007; Kunina-Habenicht and Goldhammer, 2020) or standardized achievement tests in higher education in the United States (Silm et al., 2020).

The motivation at the end of a computer-assisted mock exam possibly determines undergraduate students' exam preparation. The theoretical model in **Figure 1** describes both test-taking effort and importance as significant determinants of the upcoming exam. The current study focused on test-taking effort and importance of *education students in higher education in Germany* at three measurement points *during a computer-assisted mock exam moderated by the experimentally varied order of a cognitive ability test and a battery of mock exam tasks*. Hence, the current study aimed to (1) examine whether education students' test-taking effort and importance decrease over the course of a computer-assisted cognitive ability test and subsequent computer-assisted mock-exam tasks, or vice versa, mock-exam tasks and a subsequent cognitive ability test considering further motivational components as covariates and (2) analyze differences in education students' information processing and response processes for these two task types depending on their order. The theoretical background of the model in **Figure 1** is outlined in the next section.

## THEORY AND ASSESSMENT OF TEST-TAKING EFFORT AND IMPORTANCE AS MOTIVATIONAL COMPONENTS

One way to disentangle the contributions of ability and motivation is to measure motivation in addition to the ability being tested (e.g., Baumert and Demmrich, 2001; Bensley et al., 2016), based on expectancy-value theory that has its roots in part in motive, expectancy, and incentive as determinants of "aroused motivation to achieve" proposed by Atkinson (1957, p. 362).



## Expectancy-Value Theory

Expectancy-value theory (Atkinson, 1957; Wigfield and Eccles, 2000) summarizes the relations among a number of individual background variables, that are in brief: gender, age, aptitudes, perceived socializer's beliefs and behavior, subjective appraisal of previous achievement related experiences, affective memories, and self-concept which help explain variance in learners' achievement related choices. Most of these background variables are out of the current study's scope except gender and age since Knekta and Eklöf (2015) presented their further developed expectancy-value model applied to a test situation (Knekta and Eklöf, 2015, p. 663). Both theoretical approaches (Wigfield and Eccles, 2000; Knekta and Eklöf, 2015) posit that a person's achievement related choices are in part explained by gender, age, and the subjective task value which includes a number of motivational components, namely effort, importance, expectancies, and anxiety. Expectancy of success, and subjective task value (i.e., incentive and attainment value, utility, interest).

## Evidence for the Expectancy-Value Theory

A large body of evidence supports the assumptions made in expectancy-value theory. For example, gender consistently explained variance in test-taking effort with females having an advantage over males in a review of literature (DeMars and Bashkov, 2013). The size of the gender gap seems to vary over age groups (DeMars and Bashkov, 2013). Other review results suggested that test-taking effort decreased with increasing years of age (Silm et al., 2020). Another study included undergraduates from 18 to 69 years of age with 56% being 35 years of age or less for investigating their test-taking behavior (Rios and Liu, 2017). The authors discussed the results and limitations of their study as follows: "we evaluated the comparability of proctored groups by gender, ethnicity, language, age, and GPA [grade point average]. We found no significant group differences across all variables except gender and age" (Rios and Liu, 2017, p. 11).

## Assessment of Test-Takers' Motivation by a Questionnaire

To assess levels of motivational components, researchers typically use well-established motivation inventories which were developed to measure motivation as a trait (Midgley et al., 1998; Simzar et al., 2015) or state (Vollmeyer and Rheinberg, 2006; Freund and Holling, 2011; Freund et al., 2011). Simzar et al. (2015) and other researchers (Arvey et al., 1990) have found inconsistent relations between trait motivation and test performance (Sundre and Kitsantas, 2004). Indeed, motivation while taking an achievement test is also conceptionally related to a person's motivational state in that situation. One questionnaire measuring current motivational state is the Questionnaire on Current Motivation (QCM) (e.g., Vollmeyer and Rheinberg, 2006), which several studies (e.g., Penk and Richter, 2017) have used to disentangle the relationship between current motivation, including the dimensions of anxiety, challenge, interest, and probability, and test performance (Freund et al., 2011). Findings from studies using the QCM indicate relations at moderate levels between interest and test scores (Freund et al., 2011). However, one at least partial limitation is that the QCM asks about current motivational state in a general manner. A measurement method closer to the test situation is to ask test takers how they estimate their current motivation before and after taking an achievement test (e.g., Baumert and Demmrich, 2001; Eklöf, 2006).

Eklöf (2006) developed the Test-Taking Motivation Questionnaire (TTMQ), which includes motivational components in line with the expectancy-value theory of achievement motivation (e.g., Wigfield and Eccles, 2000). The relations between these components and test performance were at low to moderate levels, indicating inconsistent findings (Wise and DeMars, 2005; Knekta and Eklöf, 2015; Penk and Schipolowski, 2015; Penk and Richter, 2017; Stenlund et al., 2018). Moreover, Penk and Richter (2017) identified changes in the motivational component of test-taking effort during test taking. Test takers' self-reported test-taking effort decreased from the beginning to the end of the test in this study (Penk and Richter, 2017) and in other studies (Attali, 2016;

Lindner et al., 2018). Test takers may easily recognize that the TTMQ items are intended to capture their motivational state and might thus respond in socially desirable ways. Hence, it is valuable to increase the validity of the TTMQ results by employing less subjective measures (AERA et al., 2014).

### Time on Task and Response Times as Indicators of Test-Takers' Motivation

Test-taking effort has been investigated by different measures, for example, response times (Wise and Kong, 2005), time on task (Attali, 2016), or self-reports (Knekta and Eklöf, 2015). A study compared test-taking effort (measured by time on task) and performance in a high-stakes achievement test vs. subsequent low-stakes achievement test with the result that the majority of test takers replicated their high stakes performance in the low-stakes condition with little effort (Attali, 2016).

Some researchers have used response times to test the assumption of low test-taking motivation reflected in low effort (Wise and Kong, 2005; Hartig and Buchholz, 2012; Debeer et al., 2014; Rios et al., 2014), examining persistence levels in terms of response times on puzzle tasks, or response times on anagram tasks (e.g., Gignac and Wong, 2018). Other studies included changes in response times over the course of an achievement test as indicators for test-taking motivation (Hartig and Buchholz, 2012; Goldhammer et al., 2017). Meta-analytic results suggested higher correlations between test-taking response time effort and test performance than self-reported effort assessed mainly by the Student Opinion Scale (Sundre and Moore, 2002) and test performance (Silm et al., 2020). Test-taking effort estimated using response times decreased over the course of test taking in these studies (Hartig and Buchholz, 2012; Debeer et al., 2014).

In summary, changes in self-reported effort over the course of test taking suggest decreased effort, which raises the question of potential strategies for keeping test-taking effort levels. The TTMQ, based on expectancy-value theory, captures current test-taking motivation (state), and is a widely used measure in large-scale assessments including students at school. Researchers examined and proposed strategies with the intention to increase German school students' test-taking motivation but examined relatively seldom changes in test-taking motivation or strategies to keep the level of test-taking motivation in education students in Germany (Silm et al., 2020). Based on expectancy-value theory and above-mentioned evidence, we focused on two motivational components among test takers: (1) the test-taking effort invested and (2) the subjective test-taking importance of the respective task (value component), while also considering the other components that are expectancies, concentration, and anxiety, as well as gender and age, as described below in the method section. Test-taking effort and importance are probably at higher levels when test takers are working on mock exam tasks than on cognitive ability tasks.

## THE PRESENT RESEARCH

We aimed to extend the findings on changes in test-taking motivation presented in the previous section (Baumert and

Demmrich, 2001; Debeer et al., 2014; Bensley et al., 2016; Knekta, 2017; Penk and Richter, 2017) by employing a computer-assisted experimental design with repeated motivational measures (test-taking effort, test-taking importance) in order to examine changes in these motivational components over experimental variations in task type order, and whether achievement related choices, information processing and response processes are affected by the electronically varied task type order. The purpose was to obtain new insights into possible changes in test-taking effort and test-taking importance across variations in task type order. Test-taking effort and importance were assessed before and after a computer-assisted cognitive ability test and mock exam to obtain insights into intraindividual changes in effort over test taking in a new context (i.e., education students in a computer-assisted environment in higher education) using different measures than in previous studies (e.g., Freund et al., 2011). Moreover, finding different levels of test-taking effort and importance in these conditions would conceptually replicate findings from previous studies on test-taking motivation in other contexts (e.g., Eklöf, 2006; Knekta, 2017). This would extend the validity of test-taking effort and/or importance scores to further test conditions and samples (Knekta and Eklöf, 2015; Penk and Schipolowski, 2015; Knekta, 2017).

Our hypotheses were as follows: (1) Test-taking effort and test-taking importance decrease across three measurement points during the test situation moderated by task type order (first cognitive ability tasks, second mock exam tasks vs. first mock exam tasks, second cognitive ability tasks) and with consideration of the five relevant covariates test expectancies, test anxiety, concentration, gender, and age. (2) Response processes on the ability tests differ depending on the task type order (first cognitive ability tasks, second mock exam tasks vs. first mock exam tasks, second cognitive ability tasks). We included the five relevant covariates in the analyses with regard to Hypothesis 1 since they are considered in the theoretical model (see **Figure 1**), previous research suggested them as relevant covariates as introduced above (DeMars and Bashkov, 2013; Rios and Liu, 2017; Silm et al., 2020), and covariates are commonly included into experimental designs to reduce variance for increasing statistical power.

To examine our assumptions, we adapted and used measures from previous research (Arvey et al., 1990; Butler and Adams, 2007; Erle and Topolinski, 2015; Knekta and Eklöf, 2015), with the exception of the mock exam tasks. We used items from the Test-Taking Motivation Questionnaire (TTMQ) developed by Eklöf (2006) that has previously been employed in large-scale surveys (e.g., Knekta and Eklöf, 2015), cognitive ability tasks (e.g., Erle and Topolinski, 2015; 10 further tasks for other research purposes, McHugh et al., 2004), and mock exam tasks in the two test order conditions. Similar to other researchers, we analyzed the changes in test-taking motivation over the course of an exam by structural equations, in particular, latent growth curve modeling (e.g., Penk and Richter, 2017). The term "latent" refers to constructs or processes which are not observable. The advantage of measuring factors and their relationships at a latent level is that measurement errors have been separated out.

We additionally analyzed the responses and the response times of the test takers in the cognitive ability tasks and the



mock exam tasks with a psychometric diffusion model. The psychometric diffusion model is capable to separate motivational parts from achievement parts of a test taker's test performance. This provides a more objective basis for the analysis of test takers' motivation. Psychometric diffusion modeling for these tasks has not yet been undertaken in previously published work. Hence, the current study, with its experimental design, extends previous research on test-taking effort with regard to response processes for different task types.

## MATERIALS AND METHODS

### Participants

The current study involved  $N = 320$  undergraduate education students (77% female,  $M_{\text{age}} = 21$ ,  $SD_{\text{age}} = 3.13$  at T1, seven missing values in gender, one missing value in age) who voluntarily attended an electronic mock exam at a German University. The sample size is sufficient for detecting moderate group differences and changes using latent growth curve modeling as simulation studies suggested (Fan, 2003). The electronic mock exam included questions concerning test-taking motivation (presented up to three times), cognitive ability tasks (less personally important tasks), mock exam tasks (personally important tasks), and demographic questions. The mock exam was computerized using the software package PsychoPy (Peirce et al., 2019) and presented on laptops in an e-exam hall. Each undergraduate student used one laptop on a desk with sight protection.

After welcoming, one of three supervisors read a standardized oral instruction in German aloud for the participants. For example, the instruction involved (in English for the current purposes): "We offer the mock exam for the first time and would like to know how you like it. Therefore, other tasks and a few questions about your motivation are included in addition to the mock exam tasks. Please answer all tasks and questions conscientiously so that the results are meaningful." The participants further received the information that they may expect 20 mock exam tasks. They could individually decide when to finish a mock exam task and proceed with the next one. There was no time limit. **Figure 2** presents the study design. All measures, task descriptions, and tasks were implemented in the programmed experiment using PsychoPy.

The data collection was completely anonymized by assigning the participants electronically generated IDs. There was no deception. All steps of the study followed international ethical standards (AERA et al., 2014). Data from 11 participants were invalid due to technical problems, such as system aborts, and had to be excluded. Thirty-four undergraduate students participated in interventions for other research purposes than presented here, leaving data from  $n = 275$  participants remaining for current analyses.

### Measures

The motivational measures employed had already been used in international large-scale surveys (e.g., Arvey et al., 1990; Knekta and Eklöf, 2015). To test the theoretical model introduced (see **Figure 1**) with the focus on education students' test-taking

effort and importance, we adapted some items to the current study as detailed below. Test-takers' expectancies, test anxiety, and concentration are included as covariates and assessed once (interest for other research purposes than presented here). Test-taking effort and importance are assessed three times (see Motivation T1, T2, and T3, in **Figure 2**). Measures only available in English were translated into German using standard cross-translation procedures. All items were presented in German during the mock exam but example items will be translated to English here. Participants indicated their concentration, expectancies, test anxiety, test-taking effort, and importance on rating scales ranging from  $-1.5$  (*strongly disagree*) to  $1.5$  (*strongly agree*).

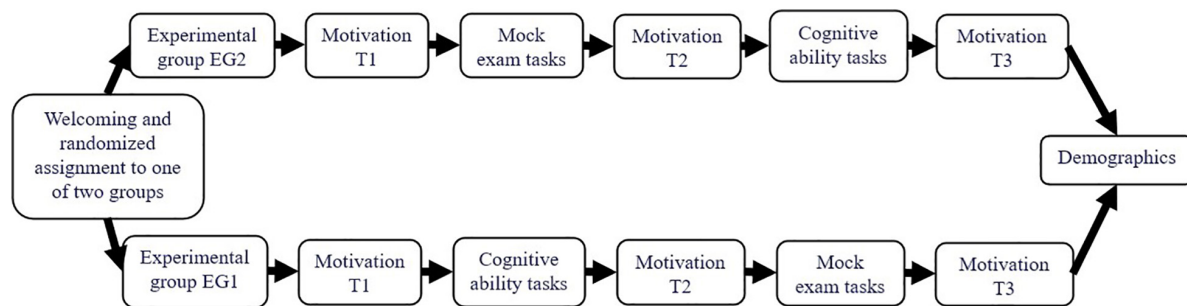
We used McDonald's  $\omega$ , instead of Cronbach's  $\alpha$ , to estimate the internal consistency of test-taking motivation and each of its dimensions test-taking effort, test-taking importance, expectancies, anxiety, and concentration simultaneously (Dunn et al., 2014). For example, Dunn et al. (2014) argued for McDonald's  $\omega$  since it is a point estimate that makes few and realistic assumptions, requires congeneric variables rather than tau-equivalent variables (Zinbarg, 2006; Revelle and Zinbarg, 2008; Hayes et al., 2020). Furthermore, inflation and attenuation of internal consistency estimation are less likely (see Dunn et al., 2014, for further advantages over Cronbach's  $\alpha$ ). McDonald's coefficient can be calculated within the R environment (R Development Core Team, 2009) using the R package *psych* (Revelle, 2019) and interpreted by the same levels as Cronbach's  $\alpha$  (Schweizer, 2011). Note the increasing number of publications about Cronbach's  $\alpha$  vs. McDonald's  $\omega$  which consistently suggest McDonald's  $\omega$  (Zinbarg, 2006; Revelle and Zinbarg, 2008; Hayes et al., 2020).

### Motivational Factors

*Test-taking effort* (Knekta and Eklöf, 2015) with regard to the current test situation was measured three times (T1–3) during the mock exam with five items: in the baseline assessment (T1), after the first task battery, and after the second task battery (T3). An example item is "I am doing my best on these tasks." McDonald's  $\omega_{\text{total}} = 0.95$  suggested good internal consistency for the three-factor solution and each factor (T1  $\omega = 0.87$ , T2  $\omega = 0.89$ , T3  $\omega = 0.89$ ). Subsequently, items assessing test-taking importance were presented.

*Test-taking importance* (Knekta and Eklöf, 2015) was measured three times (T1–3) with the same three items: in the baseline assessment (T1), after the first task battery and test-taking effort items as well as after the second task battery and effort items (T3). An example item is "The tasks are important to me." McDonald's  $\omega_{\text{total}} = 0.93$  suggested good internal consistency for the three-factor solution as well as the factors test-taking importance at T1 and T3 each except T2 with only acceptable internal consistency (T1  $\omega = 0.87$ , T2  $\omega = 0.60$ , T3  $\omega = 0.97$ ).

Moreover, McDonald's  $\omega_{\text{total}} = 0.87$  suggested good internal consistency for the motivational five-factors solution incl. expectancies ( $\omega = 0.63$ ), anxiety ( $\omega = 0.74$ ), concentration ( $\omega = 0.69$ ), test-taking effort ( $\omega = 0.87$ ), test-taking importance ( $\omega = 0.87$ ) at T1 and acceptable internal consistency of these



**FIGURE 2 |** The study design. This study focused on possible changes in test-taking effort and importance. Motivation at Time 1 (T1) included the five factors test-taking effort, test-taking importance, expectancies, anxiety, and concentration according to the theoretical model (see **Figure 1**). Motivation at Time 2 (T2) and Time 3 (T3) included the factors test-taking effort and test-taking importance.

factors each. These five motivational variables were included according to the introduced theoretical model (see **Figure 1**). Subsequent to the test-taking importance items, *expectancies* (Knehta and Eklöf, 2015) were assessed with three items adapted to the current study (T1). An example item is “Compared with other students, I think I am doing well on the tasks.” *Test anxiety* was assessed with three items and presented before the first set of tasks. An example item is “I am so nervous when I take the tasks that I forget things I usually know” (adapted from Knehta and Eklöf, 2015, p. 666). *Concentration* (Arvey et al., 1990) was assessed with four items at the end of the baseline assessment (T1). An example item is “It is hard to keep my mind on this test.” Expectancies, anxiety, and concentration were included as manifest covariates only to consider their effects on the criterion variables test-taking effort and test-taking importance at T3 since the theoretical model and previous findings suggested such relations (Knehta and Eklöf, 2015; Silm et al., 2020).

### Cognitive Ability Tasks and Mock Exam Tasks

Pioneers of psychology already tested and described cognitive abilities such as perception (James, 1884), reasoning (Piaget, 1928), and visuo-spatial perspective-taking (Flavell et al., 1978). Since perspective-taking is highly important for education students’ social interactions with children, adolescents, and adults (Wolgast et al., 2019), we chose proven cognitive ability tasks as typical tasks for research purposes in psychology. These *cognitive ability tasks* were considered as personally less important low-stakes tasks because they were not part of the lecture or module curriculum and irrelevant for the exam the students had to take in order to finish the course. Sixteen tasks assessed the cognitive ability *visuo-spatial social perspective-taking* that is seeing what another person sees by putting oneself mentally in the target’s spatial position (Kessler and Thomson, 2010; Erle and Topolinski, 2015).

Erle and Topolinski (2015) used the visuo-spatial perspective-taking paradigm developed by Kessler and Thomson (2010). Each of the first 16 tasks involved a photograph (with friendly permission from Thorsten M. Erle for using the photographs in further research). The photograph showed a female or male target person sitting at a round table (arms on the table) from a bird’s-eye perspective. A book and a banana lay on the table

close to the person’s left arm and right arm, respectively, or vice versa. The person’s position at the table rotated from photograph to photograph between 120, 160, 200, and 240° from the participant’s point of view. Previous research has found perspective-taking to be difficult at these angles (Janczyk, 2013). Each position was presented with a female target person in eight photographs and a male target person in further eight photographs (16 tasks). Participants indicated whether the book was lying closer to the target person’s left or right arm by pressing “n” (*left*) or “m” (*right*) on the keyboard (“n” is located to the left of “m” on German keyboards). There was no time limit. All cognitive ability tasks were presented in German. McDonald’s  $\omega = 0.98$  suggested almost perfect internal consistency.

Twenty single-choice *mock exam tasks* were developed to coincide with a lecture for undergraduate education students entitled “Educational Psychology.” McDonald’s  $\omega = 0.61$  suggested acceptable internal consistency for these tasks. The mock exam tasks were considered as individually important low-stakes achievement test because the students’ upcoming module exam consisted of tasks of this type with similar content. Hence, the undergraduates had the opportunity to practice this type of task in order to be well prepared for the module exam. An example mock exam task is “Which phenomenon related to a child’s reasoning did Piaget and colleagues investigate with the three-mountain task? (a) object permanence, (b) centering, (c) egocentrism, (d) logical contradictions.” The tasks were presented in German; the example has been translated into English for current purposes.

### Procedure

Participants were randomly assigned to two conditions: EG1 responded first to the cognitive ability tasks and then to the mock exam tasks, while the order was vice versa for EG2 (first mock exam tasks, then cognitive ability tasks). All participants had the opportunity to take the mock exam tasks and subsequently receive automatically generated feedback on how many tasks they solved. The respondents participated voluntarily and gave consent to analyze their data, which was anonymously collected. Taking the tests lasted less than 1 h in total, including initial instruction.

## Statistical Analyses

### Latent Growth Curve Modeling

We used latent growth curve modeling (R Development Core Team, 2009; Rosseel, 2010) and weighted least squares with mean and variance adjustment estimation (WLSMV) (Rosseel, 2019) to test for within-test-takers' changes and differences in responses on motivational items depending on condition. We set the significance-level at  $\alpha = 0.05$ . The variables included in the modeling were grand mean centered.

First, we conducted a confirmatory factor analysis (CFA) and tested the theoretical six-factor model (test-taking effort and importance at T1, T2, T3) by the data. The unstandardized effort factor-loading of the fourth item (Item "E4," Knekta and Eklöf, 2015, p. 666, adapted to university: *I spent more effort on this test than I do on other tests we have in university.*) was  $\lambda = 0.21$  and statistically not significant with  $p = 0.15$  in the EG1. Consequently, we excluded Item E4 from further analyses. The final two factor CFA model included the latent factor *test-taking effort* measured by the respective four items and their residuals at T1, T2, and T3, and the latent factor *test-taking importance* measured by the respective three items and their residuals over the three measurement points. Scalar measurement invariance is a prerequisite for latent growth curve modeling. Measurement invariance was tested using the two factor CFA model in a multi-group analysis across groups and time points. This CFA model including constrained factor loadings suggested scalar invariance, Delta Comparative Fit Index ( $\Delta$  CFI) = 0.004; Delta Root Mean Square Error of Approximation ( $\Delta$  RMSEA) = 0.003, according to recommended cutoffs (Hu and Bentler, 1999; Svetina and Rutkowski, 2014). The factor structure and intercepts found for EG1's data were equivalent to the factor structure and intercepts found for the EG2's data and at T1, T2, and T3. This CFA model is depicted in the **Supplementary Figure S1**. A simplified version of the second order latent growth curve model for the analysis of individual within-test-takers' change in test-taking effort and importance is depicted in **Figure 3** (see **Supplementary Figure S2** for a technical version). The second order latent growth curve model was specified including random intercepts and random slopes by extending the CFA model as follows: At first order latent level, the variance of the factors *test-taking effort* and *importance* each at T1, T2, and T3 has been constrained to the same value for compound symmetry covariance structure (Rosseel, 2019, 2020). At second order latent level, *test-taking effort intercept* has been specified with the three latent factors *test-taking effort* at T1, T2, and T3 with each path fixed to one. The latent factor *test-taking effort slope* has been specified with these three factors and the paths fixed to 0, 1, 2 respectively. The means of expectancies, anxiety, and concentration from the baseline assessment as well as gender and age were included to predict the factor *effort intercept* because these covariates should explain different *effort intercepts* between the education students. The covariates' category each existed before the study such as gender and age or were assessed before assessing test-taking effort and importance. Gender and age are included as covariates in SEM (Mutz and Pemantle, 2015) since previous findings consistently suggested their relations to test-taking motivation in educational contexts (e.g., DeMars and Bashkov, 2013; Silm et al., 2020). The

covariates are included in SEM to consider their anticipated effects on the criterion variables test-taking effort and importance at T3 (see Mutz and Pemantle, 2015, for standards in experimental research).

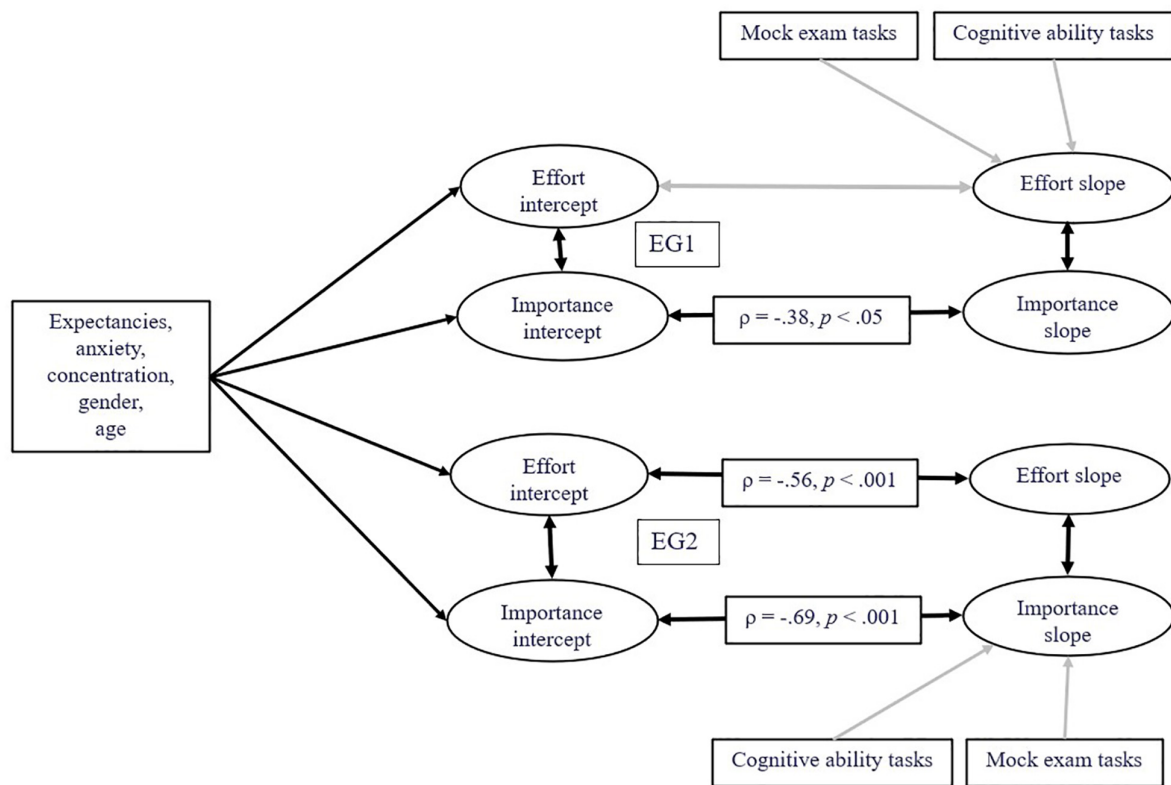
"A mean structure is automatically assumed, and the observed intercepts are fixed to zero by default, while the latent variable intercepts/means are freely estimated" (Rosseel, 2020, p. 28). The sum scores of the cognitive ability tasks and exam tasks each were included to predict the *test-taking effort slope* (instead of *test-taking effort intercept*) because we experimentally varied the exam tasks' position. We used the means and sum scores of predictor variables instead of measuring latent factors to keep the number of parameters as low as possible. Random intercepts and slopes for each latent factor were specified with correlations to themselves and to each other. The second-order latent factors *test-taking importance intercept* and *test-taking importance slope* were analogously specified and the same covariates included. Correlations between some test-taking effort indicators were allowed according to modification indices (see **Supplementary Figure S2**).

The model specification considered EG1's and EG2's data separately, so EG1's data were analyzed without EG2's data and vice versa. Criterion variables were *latent test-taking effort intercept* and *slope* as well *latent test-taking importance intercept* and *slope*.

### Latent Diffusion Modeling

To analyze participants' response processes, we included the responses and response times in a latent trait diffusion model. The diffusion model allows researchers to examine response process components in binary decision tasks (Voss et al., 2013). Binary decision tasks are, for example, the presented cognitive ability tasks where test takers have to choose one of two response options or the questions in the mock exam where the test takers have to decide between the correct and the incorrect response (Molenaar et al., 2015). The diffusion model is based on the assumption that test takers continuously accumulate evidence for the two response options. A momentary preference is formed by weighting the evidence for the two response options against each other. As soon as the momentary preference exceeds a critical level, the test taker responds by selecting the more preferred option.

Diffusion modeling involves defining three process parameters. (1) Information accumulation is a measure of one's mental simulation of two possible outcomes using the available information (*drift rate*,  $\nu$ ). (2) The amount of information required for a response is reflected in the decision threshold (*boundary separation parameter*,  $a$ ). (3) The response time includes time for reactions (e.g., moving one's finger to the keyboard in a computer-based task) and/or other sensory, mental or motor responses aside from the time needed to make a decision (*non-decision time parameter*,  $t_{er}$ ) (Ratcliff and McKoon, 2008; Voss et al., 2013). The drift rate ( $\nu$ ) provides insights into information uptake latency, with high uptake speed reflecting high performance, and is a manifestation of a test taker's capability. The lower the drift rate, the more difficult the task is in relation to a given individual latent trait (e.g., ability



**FIGURE 3 |** Simplified scheme of the latent growth curve model.  $\rho$  = standardized latent correlation coefficient (see **Table 4** for standard errors and confidence intervals). Gray arrows represent statistically not significant relations. Latent factors test-taking effort and importance at T1, T2, and T3, indicators and residuals are not depicted in favor of clarity (see **Supplementary Figure S2** for the technical model depiction).

or attitude, see Voss et al., 2013). Low drift rates are reflected in low response accuracy and long response times. The boundary separation reflects the response caution. It is assumed to be a manifestation of a test taker's effort or importance, reflecting their carefulness when responding. Low levels of the boundary separation are reflected in low response accuracy and short response times, two typical signs of rapid guessing.

In addition to the parameters in manifest diffusion modeling, the psychometric diffusion modeling under the item response theory allows to estimate the latent person contribution and task contribution to the response process. The person contribution refers to information processing (latent trait  $\theta$ ) and response caution (latent trait  $\omega$ ) of a person as well as investigate relationships between these latent traits ( $\theta$ ,  $\omega$ ) and constructs such as test-taking effort or importance. The task contribution refers to the task difficulty.

Two model types are distinguished, the D-diffusion (Tuerlinckx and De Boeck, 2005) and the Q-diffusion model (Van der Maas et al., 2011). They basically differ in their parameterization. In the D-diffusion model, the effective drift rate is the difference of the latent trait of a test taker and the corresponding intercept. This parameterization allows task probabilities from zero to one. In the Q-diffusion model, the effective drift rate is the quotient of latent ability and

the corresponding intercept. The Q-diffusion model requires task solving probabilities of at least 50% for calculating the diffusion parameters. In case of solving probabilities lower than 50%, the D-diffusion model can be used with consideration of its predictions.

We applied diffusion modeling within the R environment (R Development Core Team, 2009) using the R package *diffIRT* (Molenaar et al., 2015). The non-decision time (*ter*) was constrained to control delays resulting from the different laptops we used on the non-response times. The R code can be obtained from the authors.

## RESULTS

Descriptive results are summarized in **Tables 1A,B**. Product-moment correlations between the variables used are detailed in **Table 2**. The correlation coefficients suggest zero to low not significant correlations of test-taking effort (T1, T2, T3), test-taking importance (T1, T2, T3), expectancies, anxiety, and concentration with the cognitive ability tasks, and mock exam tasks. Means and standard errors of test takers' responses on test-taking effort and importance items are depicted in **Figure 4**. These line diagrams suggested changes in education students' test-taking effort and importance. For examining these changes at



latent level and with consideration of the covariates *expectancies*, *anxiety*, and *concentration*, we investigated within test-taker effects by structural equations and diffusion modeling.

## Within Test-Taker Effects

First, we employed latent growth curve modeling to disclose changes in education students' test taking *effort* and *importance* over T1, T2, and T3 moderated by condition (EG1: cognitive ability tasks first vs. EG2: mock exam tasks first). The simplified model structure is depicted in **Figure 3** (without depiction of residuals and indicators). The goodness of fit between the theoretical model and data was good (Hu and Bentler, 1999; Svetina and Rutkowski, 2014),  $\chi^2(641) = 731.04$ ,  $p = 0.008$ , CFI = 0.99, RMSEA = 0.033, 95% CI [0.018, 0.044], SRMR = 0.078 (WLSMV-estimation).

**Figure 3** provides results such as standardized latent correlation coefficients and statistical significance levels.

In **Table 3**, these standardized latent correlation coefficients are presented with standard errors, significance levels and confidence intervals each. **Table 3** provides furthermore variances of effort intercept, effort slope, importance intercept, and importance slope. The results suggested significantly decreased test-taking effort ( $\rho = -0.56$ ,  $p < 0.001$ ) and importance ( $\rho = -0.69$ ,  $p < 0.001$ ) in EG2 over test-taking time supporting Hypothesis 1.

However, the EG1's test-taking effort did not decrease over time ( $\rho = 0.63$ ,  $p = 0.27$ ), only their test-taking importance decreased ( $\rho = -0.38$ ,  $p < 0.05$ ) but less than the EG2's test-taking importance. No significant relations existed between cognitive

ability task performance or mock exam task performance and the factors *test-taking effort slope*, and *test-taking importance slope*.

This model explained 28% of variance in the latent factor effort intercept, 30% in the latent factor effort slope, 83% of variance in the latent factor test-taking effort at T3, 22% of variance in the latent factor importance intercept, 1% in the latent factor importance slope, and 83% in the latent factor importance at T3 in the EG1. In the EG2, this model explained 19% of variance in the latent factor effort intercept, 0.4% in the latent factor effort slope, 86% of variance in the latent factor test-taking effort at T3, 13% of variance in the latent factor importance intercept, 0.2% in the latent factor importance slope, and 84% in the latent factor importance at T3.

The theoretical model adapted to the current study (see **Figure 1**) implies that achievement related choices can involve decisions in response processes on tasks. We examined the EG1's vs. EG2's responses and response times on the cognitive ability tasks and mock exam tasks by diffusion modeling as described next.

## Education Students' Response Processes for the Tasks

The responses for the cognitive ability tasks as well as response times were included in a Q-diffusion model to analyze the achievement related choices and response processes according to the theoretical model in **Figure 1**. We investigated the goodness of fit between the theoretical and observed response time distribution with QQ-plots which suggested good fit for both groups (see **Supplementary Figure S3** for examples). The average intercept of the boundary separation and the average intercept of the drift rate over the items are summarized in **Table 4** for the EG1 and the EG2. A Wald test for the equivalence of the boundary intercepts in the two groups was significant ( $X^2 = 42.00$ ,  $df = 16$ ,  $p < 0.01$ ). A *post hoc* comparison of the intercepts in the single items revealed that parameters deviated in two of the 16 items on  $\alpha = 0.05$ . The corresponding Wald test for the equivalence of the drift intercepts was also significant ( $X^2 = 37.83$ ,  $df = 16$ ,  $p < 0.01$ ). The drift rates differed in three of the 16 items on  $\alpha = 0.05$ . This implies that neither the average response caution, usually considered as a motivational aspect of the response process, nor the average rate of information accumulation, usually considered as an aspect of a test taker's performance, differed between the groups in most items.

Investigating the D-diffusion model response fit using QQ-plots for the 20 exam tasks suggested acceptable fit between expected and observed distributions in both groups (see **Supplementary Figure S3** for examples). A Wald test for the equivalence of the boundary intercepts in the two groups was significant ( $X^2 = 33.90$ ,  $df = 20$ ,  $p = 0.03$ ). A *post hoc* comparison of the intercepts in the single items revealed that parameters deviated in 5 of the 20 items on  $\alpha = 0.05$ . The corresponding Wald test for the equivalence of the drift intercepts was insignificant ( $X^2 = 22.81$ ,  $df = 20$ ,  $p = 0.29$ ).

We included the latent information processing  $\omega$  (speed) and response caution  $\theta$  (trait) from diffusion modeling as criterion variables in general linear modeling to examine whether they

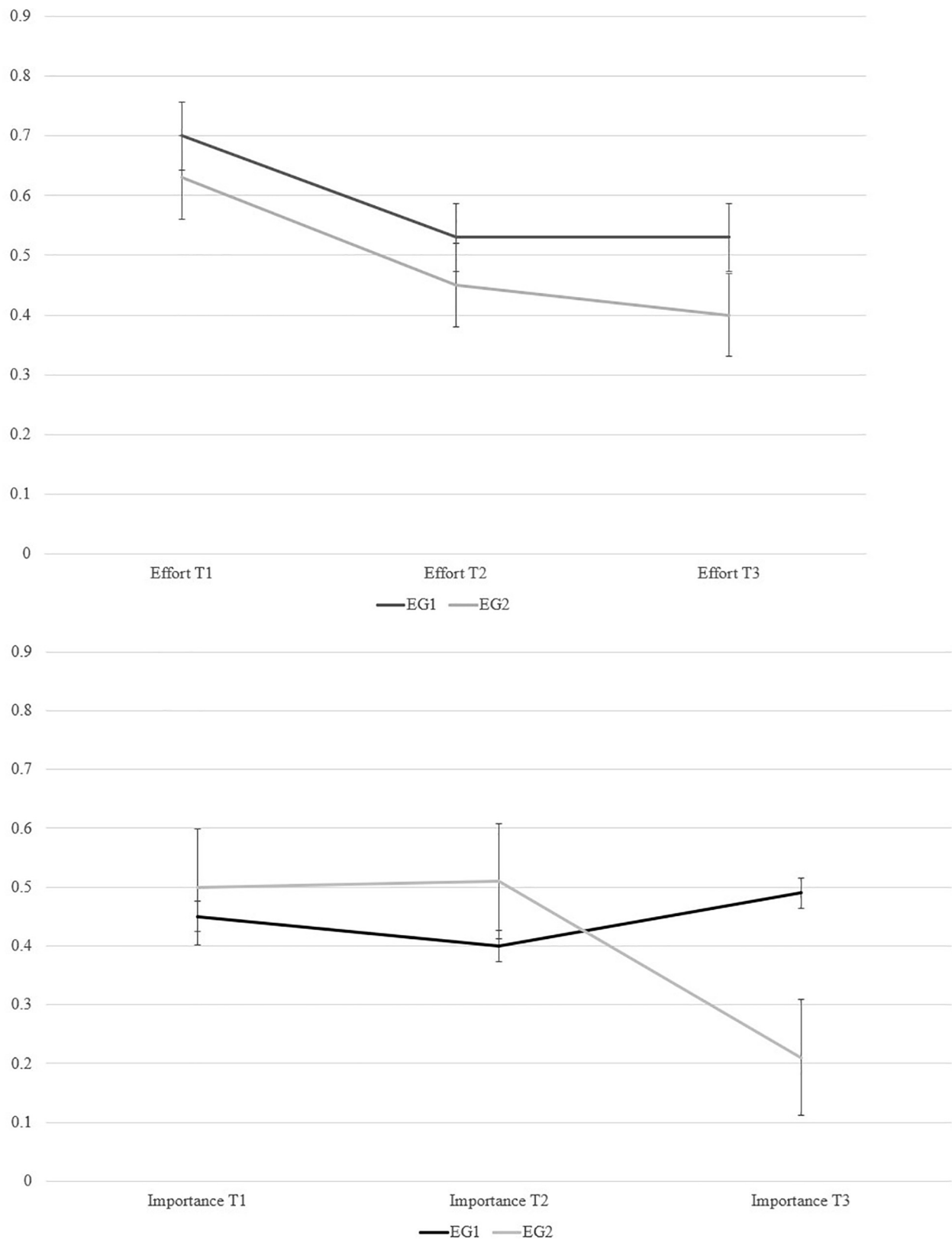
**TABLE 1A |** Means and standard deviations of the motivational components test-taking effort and test-taking importance in the EG1 ( $n = 125$ ) and EG2 ( $n = 150$ ) at T1, T2, and T3.

	Time 1		Time 2		Time 3	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
<b>EG1</b>						
Effort <sup>a</sup>	0.70	0.38	0.53	0.45	0.53	0.40
Importance <sup>a</sup>	0.45	0.49	0.40	0.54	0.49	0.44
<b>EG2</b>						
Effort <sup>a</sup>	0.63	0.37	0.45	0.42	0.16	0.57
Importance <sup>a</sup>	0.50	0.47	0.51	0.44	0.21	0.58

<sup>a</sup>Test-taking effort and test-taking importance means of test takers' responses ( $-1.5 =$  strongly disagree,  $-0.5 =$  disagree,  $0.5 =$  agree,  $1.5 =$  strongly agree).

**TABLE 1B |** Mean accuracy and response times on cognitive ability tasks and mock-exam tasks in Experimental Group EG1 ( $n = 125$ ) vs. EG2 ( $n = 150$ ).

	Cognitive ability tasks		Mock exam tasks	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
<b>EG1</b>				
Mean accuracy	94%	17%	57%	88%
Mean rt (s)	33.63	13.62	607.65	163.53
<b>EG2</b>				
Mean accuracy	92%	19%	54%	91%
Mean rt (s)	33.97	15.88	692.37	207.04



**FIGURE 4 |** Line diagrams of changes in test-taking effort (above) and test-taking importance (below), means of test takers' responses ( $-1.5 = \text{strongly disagree}$ ,  $-0.5 = \text{disagree}$ ,  $0.5 = \text{agree}$ ,  $1.5 = \text{strongly agree}$ , error bars represent standard errors).

**TABLE 2 |** Correlations among test-taking effort at T1–3, test-taking importance at T1–3, expectancies, test-taking anxiety, concentration, cognitive ability tasks, mock exam tasks, and age in Experimental Group EG1 vs. EG2.

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1 Effort T1		<b>0.89</b>	<b>0.56</b>	<b>0.79</b>	<b>0.77</b>	0.48	<b>−0.64</b>	0.08	0.13	<0.01	0.04	−0.21
2 Effort T2	<b>0.73</b>		<b>0.68</b>	<b>0.80</b>	<b>0.86</b>	<b>0.59</b>	<b>−0.79</b>	0.02	0.13	−0.02	0.05	−0.22
3 Effort T3	<b>0.82</b>	<b>0.83</b>		0.52	<b>0.59</b>	<b>0.91</b>	<b>−0.64</b>	−0.27	0.27	−0.16	−0.28	−0.25
4 Importance T1	<b>0.70</b>	<b>0.84</b>	<b>0.81</b>		<b>0.88</b>	0.54	<b>−0.64</b>	−0.04	0.07	0.04	0.07	−0.05
5 Importance T2	0.54	<b>0.83</b>	<b>0.78</b>	<b>0.94</b>		<b>0.65</b>	<b>−0.71</b>	0.01	0.13	0.07	−0.02	−0.25
6 Importance T3	<b>0.65</b>	<b>0.79</b>	<b>0.88</b>	<b>0.92</b>	<b>0.93</b>		<b>−0.64</b>	−0.18	0.14	−0.14	−0.22	−0.26
7 Concentration	<b>−0.68</b>	<b>−0.58</b>	<b>−0.74</b>	<b>−0.58</b>	−0.48	<b>−0.61</b>		−0.11	0.04	−0.15	−0.18	0.19
8 Expectancies	0.18	0.15	0.06	0.17	0.11	0.12	−0.23		−0.41	−0.20	0.28	0.10
9 Anxiety	−0.02	0.12	0.07	0.31	0.32	0.26	0.15	−0.47		−0.09	−0.48	−0.33
10 Cogn. ability	−0.31	−0.35	−0.19	−0.44	−0.37	−0.35	−0.03	−0.39	−0.35		−0.19	−0.26
11 Mock exam	0.09	−0.12	−0.04	−0.10	−0.22	−0.16	−0.30	0.43	−0.47	−0.18		−0.14
12 Age	−0.09	−0.25	−0.13	−0.25	−0.31	−0.17	−0.26	−0.14	−0.08	0.09	0.06	

EG1 ( $n = 125$ ) below the diagonal EG2 ( $n = 150$ ) above the diagonal,  $p < 0.05$  in bold.

Cogn. Ability, cognitive ability tasks; T1, Time 1.

related to the condition (i.e., cognitive ability tasks or mock exam tasks first), baseline variables (i.e., expectancies, anxiety, concentration, test-taking effort, test-taking importance) gender, and age. For the cognitive ability tasks, the latent information processing  $\omega$  and response caution  $\theta$  did not relate to the baseline variables expectancies, anxiety, concentration, test-taking effort, test-taking importance, gender, and age with all eight regression coefficients close to zero ( $B_{\omega} = -0.04$  to  $0.01$ ,  $B_{\theta} = -0.01$  to  $0.02$ ) except condition that related to response caution ( $B_{\theta} = 0.15$ ,  $p = 0.05$ ) with higher response caution in the EG1 than EG2.

Including latent information processing  $\omega$  and response caution  $\theta$  with the baseline variables into the model to predict test-taking effort and importance at T3 yielded significant relations between concentration ( $B_{ei} = -0.32$ ,  $p < 0.001$ ), test-taking effort ( $B_{ei} = 0.48$ ,  $p < 0.001$ ), and importance ( $B_{ei} = 0.74$ ,  $p < 0.001$ ) at T1 as well as condition ( $B_{ei} = 2.80$ ,  $p < 0.001$ ) to test-taking effort and importance at T3 with EG1 having an advantage. While concentration levels were higher in the EG2 than EG1, test-taking effort and importance levels were higher in the EG1 than EG2. This model explained 36% of variance in test-taking effort and importance,  $F(10, 252) = 15.82$ ,  $p < 0.001$ .

**TABLE 3A |** Standardized latent regression coefficients and correlation coefficients, standard errors, confidence intervals, from latent growth curve modeling with intercepts and slopes of test-taking effort and test-taking importance in Experimental Group EG1 and EG2.

EG1, $n = 125$	$\beta$	SE	$p$	CI <sub>lower</sub>	CI <sub>upper</sub>
Effort slope regressed on					
Cognitive ability tasks	0.53	0.38	0.86	−0.51	0.57
Exam tasks	−0.10	0.66	0.88	−0.39	0.19
Importance slope regressed on					
Cognitive ability tasks	0.07	0.08	0.44	−0.10	0.23
Exam tasks	−0.05	0.14	0.72	−0.33	0.23
EG2, $n = 150$					
Effort slope regressed on					
Cognitive ability tasks	0.03	0.07	0.68	−0.11	0.17
Exam tasks	−0.05	0.09	0.54	−0.23	0.12
Importance slope regressed on					
Cognitive ability tasks	0.03	0.07	0.69	−0.11	0.16
Exam tasks	−0.03	0.09	0.76	−0.20	0.14
EG1	$\rho$	SE	$p$	CI <sub>lower</sub>	CI <sub>upper</sub>
Effort intercept~slope	0.63	0.57	0.27	−0.48	1.74
Importance intercept~slope	−0.38	0.18	0.03	−0.73	−0.04
EG2					
Effort intercept~slope	−0.56	0.12	<0.001	−0.79	−0.33
Importance intercept~slope	−0.69	0.12	<0.001	−0.91	−0.46

**TABLE 3B |** Effort intercept, effort slope, importance intercept, and importance slope: Variances at latent level, and explained variances from latent growth modeling.

EG1	Va	SE	$p$
Effort intercept variance	0.27	0.14	0.046
Effort slope variance	0.01	0.08	0.95
Importance intercept variance	0.41	0.10	<0.001
Importance slope variance	0.07	0.03	0.04
EG2			
Effort intercept variance	0.39	0.10	<0.001
Effort slope variance	0.12	0.03	<0.001
Importance intercept variance	0.54	0.12	<0.001
Importance slope variance	0.12	0.04	0.001
EG1	$R^2$		
Effort intercept	0.28		
Effort slope	0.30		
Importance intercept	0.22		
Importance slope	0.01		
EG2			
Effort intercept	0.19		
Effort slope	0.004		
Importance intercept	0.13		
Importance slope	0.002		

**TABLE 4 |** Mean boundary separations ( $a$ ), mean drift rates ( $v$ ), and standard deviations from diffusion modeling including the cognitive ability tasks or mock exam tasks in Experimental Group EG1 vs. EG2.

Tasks	Parameter	EG1		EG2	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Cognitive ability	$a$	0.27	0.03	0.25	0.02
	$v$	0.95	0.14	1.04	0.15
Exam	$a$	0.10	0.02	0.09	0.02
	$v$	−0.04	0.10	−0.03	0.09

$a$ , boundary separation,  $v$ , drift rate,  $t_{er}$  was constrained to same values.

$Q$ -diffusion model has been used for the cognitive ability tasks,  $D$ -diffusion model for the exam tasks due to solving probabilities below 50% which the driftIRT-function cannot handle.

For the mock exam tasks, the latent information processing  $\omega$  and response caution  $\theta$  did not relate to the baseline variables expectancies, anxiety, concentration, test-taking effort, test-taking importance, gender, and age with all eight regression coefficients close to zero ( $B_{\omega} = -0.02$  to  $0.01$ ,  $B_{\theta} = -0.05$  to  $0.01$ ). Including latent information processing  $\omega$  and response caution  $\theta$  with the baseline variables into the model to predict test-taking effort and importance at T3 yielded significant relations between concentration ( $B_{ci} = -0.33$ ,  $p = 0.002$ ), test-taking effort ( $B_{ei} = 0.47$ ,  $p < 0.001$ ), and importance ( $B_{ei} = 0.74$ ,  $p < 0.001$ ) at T1 as well as condition ( $B_{ei} = 2.73$ ,  $p < 0.001$ ) to test-taking effort and importance at T3 with EG1 having an advantage. This model explained 36% of variance in test-taking effort and importance,  $F(10, 252) = 15.63$ ,  $p < 0.001$ . While concentration levels were higher in the EG2 than EG1, test-taking effort and importance levels were higher in the EG1 than EG2. This model explained 36% of variance in test-taking effort and importance at T3,  $F(10, 252) = 15.63$ ,  $p < 0.001$ .

## DISCUSSION AND CONCLUSION

Based on expectancy-value theory (Wigfield and Eccles, 2000; Knekt and Eklöf, 2015), the current study sought to examine (1) whether test-taking effort and test-taking importance decrease across three measurement points during the computer-assisted test situation moderated by test-battery order and with consideration of the five covariates test expectancies, test anxiety, concentration, gender, and age and (2) whether response processes on the ability tests and mock exam differ depending on the task type order (EG1: first cognitive ability tasks, second mock exam tasks vs. EG2: first mock exam tasks, second cognitive ability tasks). The response processes refer to the achievement related choices depicted in Figure 1 and regard information processing. Thus, both hypotheses focus on the education students' test-taking behavior in a computer-assisted environment. Self-reported test-taking effort and importance provide subjective information about test-taking behavior in the computer-assisted environment. Information processing and response processes in tasks involve responses and response

times that provide rather objective information about test-taking behavior than self-reported test-taking effort and importance.

The results from latent growth curve modeling suggested that test-taking effort and importance in EG2 significantly decreased among the education students over different task type orders (first mock exam tasks, then cognitive ability tasks) in the computer-assisted environment. Test-taking effort significantly decreased almost linearly from T1 over the mock exam tasks and cognitive ability tasks to T3 in EG2. These declines are in accordance with Hypothesis 1. Test-taking importance significantly decreased in EG1 (moderate effect) and EG2 (strong effect) (Cohen, 1988). Previous findings suggested that test-taking effort and importance changed even over the course of low-stakes testing (e.g., Penk and Richter, 2017). The decline in test-taking effort and test-taking importance in EG2 conceptually replicates similar findings from previous studies including students in school and different tasks (e.g., Knekt, 2017; Penk and Richter, 2017).

However, test-taking effort in the current study did not decrease when cognitive ability tasks were presented first (EG1) in the computer-assisted environment. The results support Hypothesis 1 in part since test-taking effort and importance significantly decreased in EG2 with higher levels when working on mock exam tasks than on the subsequent cognitive ability tasks. Test-taking importance also decreased in EG1 but without an advantage for the mock exam tasks. EG1's not decreased test-taking effort contradicts Hypothesis 1. Note that test-taking effort and importance were assessed by computer-assisted self-report measures. Diffusion modeling allows more objective insights into response processes in the sense of achievement related choices and information processing while working on tasks.

For the mock exam tasks, boundary intercepts suggesting response caution  $\theta$  (latent trait and motivational aspect) were similar in both groups. This result implies similar motivational levels in both groups while working on the mock exam tasks. For the cognitive ability tasks, boundary intercepts suggesting response caution  $\theta$  significantly differed between the groups. Response caution  $\theta$  related to the condition with higher response caution in EG1 than EG2. Thus, EG1's motivational levels were higher than EG2's motivational levels from a more objective point of view than self-reports. This difference is in accordance with Hypothesis 2 and with the result that test-taking effort did not decrease in EG1.

The new finding here is that the education students invested similar test-taking effort in the cognitive ability tasks as in the subsequent mock exam tasks (EG1). In EG1, test-taking effort did not decrease over the task types. Latent diffusion modeling (Van der Maas et al., 2011; Voss et al., 2013) suggested similar response processes on mock exam tasks but differences in the response processes (boundary intercepts) on cognitive ability tasks suggesting higher objective motivational levels in EG1 than EG2.

Latent diffusion modeling has not been undertaken in previous research on test-taking motivation in low-stakes tests. The high accuracy on the computer-assisted cognitive ability tasks might be one explanation approach for the similar response processes on the computer-assisted mock exam tasks between conditions. Another explanation might be that the mock



exam tasks predominantly required recalling subject content knowledge of educational psychology and rarely knowledge transfer to educational practice. Changes in education students' test-taking effort might affect tasks in other computer-assisted environments than presented here which require knowledge transfer to contexts in practice because such transfer is known as cognitively difficult. Alternatively, the testing time of about 30 min was too short for a test-taking effort decline related to a cognitive performance decline.

We concluded from the results, the education students were able to keep their self-reported test-taking effort levels during computer-assisted cognitive tasks and subsequent computer-assisted mock exam tasks. Diffusion modeling suggested objectively measured higher motivational levels during the cognitive tasks when they were presented first (EG1) than when mock exam tasks were presented first (EG2). The education students were not able to keep their test-taking effort during the computer-assisted cognitive ability tasks following the computer-assisted mock exam tasks.

Weak or non-existing relations between the motivational components (assessed by self-reports) and performance in low-stakes achievement tests are already known from other studies that presented relations inconsistently at zero (Sundre and Kitsantas, 2004; Penk and Richter, 2017) to low levels (Knehta, 2017; Eklöf and Hopfenbeck, 2018; Myers and Finney, 2019). The weak or non-existing relations between the motivational components on the cognitive ability tasks and mock exam tasks might have resulted from low to moderate task difficulties which not require to be motivated for performing equally in both conditions.

## Limitations and Implications for Future Research

Participants in the current study were education students and a self-selected sample tested in an e-exam hall (one person per laptop); however, each participant was randomly assigned to one of two experimental conditions. Gender was not equally distributed in the study. The computer-assisted cognitive ability tasks were tasks for research purposes rather than widely used standardized inventories. The computer-assisted mock exam was developed based on the participating students' educational psychology curriculum, including somewhat broad fundamentals of cognitive psychology, developmental psychology, and social psychology. Consequently, the relatively low internal consistency measured by McDonald's  $\omega$  might reflect the curriculum's broad content. Despite these limitations, however, the present study contributes to the understanding of motivation-performance patterns during computer-assisted test taking in higher education. This finding might be also relevant for motivation, information processing and responses on online exam tasks and online self-assessment tasks. The described differences between the conditions might be considered in the development of new computer-assisted (online) task batteries for exams or self-assessments, especially their order.

Future research might include a within-subject design and computer-assisted (online) tests accompanied by measures

assessing test-taking effort and importance before and after the respective test (e.g., effort with regard to Test 1 assessed before Test 1 and subsequently to Test 1, then effort with regard to Test 2 assessed before Test 2 and subsequently to Test 2). It is important for further studies to examine test-taking effort and importance during different ability tasks, because responses on tasks other than those presented here may differently stimulate information processing and differently relate to test-taking effort and importance. Hence, future research should examine the relations between test-taking effort, test-taking importance, and responses on different computer-assisted (online) ability tasks to increase the validity of the presented results (AERA et al., 2014). The current study increased its validity by using test-taking-effort and importance measures, because the changes found support the hypothesis that states should be measured rather than traits (Eklöf, 2006; AERA et al., 2014).

The main contribution of the empirical work presented here is that test-taking effort and importance were assessed three times over an experimentally varied task battery order considering information processing and response processes in a computer-assisted environment in higher education. Roughly 20 years ago, Baumert and Demmrich (2001) presented insignificant findings on strategies to increase students' test-taking motivation in PISA. However, from this study's perspective, the more important question is how to keep test-taking effort and importance relatively stable and avoid declines, rather than discussing how to increase test-taking motivation, as has been the case in previous research (e.g., Baumert and Demmrich, 2001).

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: [https://osf.io/nmxq7/?view\\_only=f150703b9a664d648c772055aa3335b3](https://osf.io/nmxq7/?view_only=f150703b9a664d648c772055aa3335b3).

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

AW, NS, and JR contributed in collaboration to the design, conception, and data collection of the study. NS wrote a test documentation including sample and measure descriptions as well as descriptive results presented in the current report. AW analyzed the data by structural equations and latent diffusion models, and wrote the first draft of the manuscript. JR reviewed the manuscript contributing significantly to improve it, in particular by also calculating diffusion models and writing parts of the

section about diffusion modeling. All authors contributed to the manuscript revision, reread and approved the current version.

## FUNDING

This research was funded by the German Federal Ministry of Education (Grant No. 01PL17065) including open access publication fees.

## ACKNOWLEDGMENTS

We thank Vivien Eichhoff for her support when collecting the data.

## REFERENCES

- AERA, APA, and NCME (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Arvey, R. D., Strickland, W., Drauden, G., and Martin, C. (1990). Motivational components of test taking. *Person. Psychol.* 43, 695–716. doi: 10.1111/j.1744-6570.1990.tb00679.x
- Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychol. Rev.* 64, 22–32. doi: 10.1037/h0043445
- Attali, Y. (2016). Effort in low-stakes assessments: What does it take to perform as well as in a high-stakes setting? *Edu. Psychol. Measur.* 76, 1045–1058. doi: 10.1177/0013164416634789
- Baumert, J., and Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *Eur. J. Psychol. Edu.* 16, 441–462. doi: 10.1007/BF03173192
- Bensley, D. A., Rainey, C., Murtagh, M. P., Flinn, J. A., Maschiochi, C., Bernhardt, P. C., et al. (2016). Closing the assessment loop on critical thinking: The challenges of multidimensional testing and low test-taking motivation. *Think. Skills Creat.* 21, 158–168. doi: 10.1016/j.tsc.2016.06.006
- Brunner, M., Artelt, C., Krauss, S., and Baumert, J. (2007). Coaching for the PISA test. *Lear. Instruct.* 17, 111–122. doi: 10.1016/j.learninstruc.2007.01.002
- Butler, J., and Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *J. Appl. Measur.* 8, 279–304.
- Cohen, J. (1988). Set correlation and contingency tables. *Appl. Psychol. Measur.* 12, 425–434. doi: 10.1177/014662168801200410
- Debeer, D., Buchholz, J., Hartig, J., and Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *J. Edu. Behav. Stat.* 39, 502–523. doi: 10.3102/1076998614558485
- DeMars, C. E., and Bashkov, B. M. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Res. Pract. Asses.* 8, 69–82.
- Dunn, T. J., Baguley, T., and Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *Br. J. Psychol.* 105, 399–412. doi: 10.1111/bjop.12046
- Eccles, J. S., and Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personal. Soc. Psychol. Bull.* 21, 215–225. doi: 10.1177/0146167295213003
- Eccles, J. S., Midgley, C., and Adler, T. (1984). Grade-related changes in the school environment. *Develop. Achiev. Motiv.* 3, 238–331.
- Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Edu. Psychol. Measur.* 66, 643–656. doi: 10.1177/0013164405278574
- Eklöf, H., and Hopfenbeck, T. N. (2018). "Self-reported effort and motivation in the PISA test," in *International Large-Scale Assessments in Education: Insider Research Perspectives*, ed. B. Maddox (London: Bloomsbury), 121–136.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.559683/full#supplementary-material>

**Supplementary Figure 1** | CFA-model including test-taking effort and test-taking importance at T1, T2, and T3.

**Supplementary Figure 2** | Latent growth curve model.

**Supplementary Figure 3** | QQ-plots to investigate goodness of fit in diffusion modeling including (A) cognitive ability tasks and (B) mock exam tasks.

**Supplementary Table 1** | Standardized regression coefficients, standard errors, confidence intervals, and correlations from latent growth curve modeling with intercepts and slopes of test-taking effort and test-taking importance, and covariates in both groups.

- Erle, T. M., and Topolinski, S. (2015). Spatial and empathic perspective-taking correlate on a dispositional level. *Soc. Cog.* 33, 187–210. doi: 10.1521/soco.2015.33.3.187
- Fan, X. (2003). Power of latent growth modeling for detecting group differences in linear growth trajectory parameters. *Struct. Equ. Modeling* 10, 380–400. doi: 10.1207/S15328007SEM1003
- Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Res. Rep. Ser.* 2015, 1–17. doi: 10.1002/ets2.12067
- Flavell, J. H., Omanson, R. C., and Latham, C. (1978). Solving spatial perspective-taking problems by rule versus computation: A developmental study. *Dev. Psychol.* 14, 462–473. doi: 10.1037/0012-1649.14.5.462
- Freund, P. A., and Holling, H. (2011). Who wants to take an intelligence test? Personality and achievement motivation in the context of ability testing. *Personal. Individ. Differ.* 50, 723–728. doi: 10.1016/j.paid.2010.12.025
- Freund, P. A., Kuhn, J.-T., and Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personal. Individ. Differ.* 51, 629–634. doi: 10.1016/j.paid.2011.05.033
- Gignac, G. E., and Wong, K. K. (2018). A psychometric examination of the anagram persistence task: More than two unsolvable anagrams may not be better. *Assessment* 27, 1198–1212. doi: 10.1177/1073191118789260
- Goldhammer, F., Martens, T., and Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. *Large Scale Asses. Edu.* 5:18. doi: 10.1186/s40536-017-0051-9
- Gorges, J., Maehler, D. B., Koch, T., and Offerhaus, J. (2016). Who likes to learn new things: measuring adult motivation to learn with PIAAC data from 21 countries. *Large Scale Assess. Educ.* 4:9. doi: 10.1186/s40536-016-0024-4
- Hartig, J., and Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychol. Test Assess. Modeling* 54, 418–431.
- Hayes, A. F., Coutts, J. J., and But, R. (2020). Use omega rather than Cronbach's alpha for estimating reliability. *Commun. Methods Measures* 1, 1–24. doi: 10.1080/19312458.2020.1718629
- Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118
- James, W. (1884). On some omissions of introspective psychology. *Mind* 33, 1–26.
- Janczyk, M. (2013). Level 2 perspective taking entails two processes: Evidence from PRP experiments. *J. Exp. Psychol.* 39, 1878–1887. doi: 10.1037/a0033336
- Kessler, K., and Thomson, L. A. (2010). The embodied nature of spatial perspective taking: Embodied transformation versus sensorimotor interference. *Cognition* 114, 72–88. doi: 10.1016/j.cognition.2009.08.015
- Knekta, E. (2017). Are all pupils equally motivated to do their best on all tests? Differences in reported test-taking motivation within and between tests with

- different stakes. *Scand. J. Educ. Res.* 61, 95–111. doi: 10.1080/00313831.2015.1119723
- Knekta, E., and Eklöf, H. (2015). Modeling the test-taking motivation construct through investigation of psychometric properties of an expectancy-value-based questionnaire. *J. Psychoeduc. Assess.* 33, 662–673. doi: 10.1177/0734282914551956
- Kriegbaum, K., Jansen, M., and Spinath, B. (2014). Motivation: A predictor of PISA's mathematical competence beyond intelligence and prior test achievement. *Learn. Individ. Diff.* 43, 140–148. doi: 10.1016/j.lindif.2015.08.026
- Kunina-Habenicht, O., and Goldhammer, F. (2020). ICT Engagement: a new construct and its assessment in PISA 2015. *Large Scale Assess. Educ.* 8:6. doi: 10.1186/s40536-020-00084-z
- Lindner, C., Nagy, G., and Retelsdorf, J. (2018). The need for self-control in achievement tests: Changes in students' state self-control capacity and effort investment. *Soc. Psychol. Educ.* 21, 1113–1131. doi: 10.1007/s11218-018-9455-9
- List, M. K., Robitzsch, A., Lüdtke, O., Köller, O., and Nagy, G. (2017). Performance decline in low-stakes educational assessments: different mixture modeling approaches. *Large Scale Assess. Educ.* 5:15. doi: 10.1186/s40536-017-0049-3
- Liu, Y., and Hau, K.-T. (2020). Measuring motivation to take low-stakes large-scale test: New model based on analyses of "Participant-Owned-Defined" missingness. *Educ. Psychol. Measur.* 2020:0013164420911972. doi: 10.1177/0013164420911972
- McHugh, L., Barnes-Holmes, Y., and Barnes-Holmes, D. (2004). Perspective-taking as relational responding: A developmental profile. *Psychol. Rec.* 54, 115–144. doi: 10.1007/BF03395465
- Midgley, C., Kaplan, A., Middleton, M., Maehr, M. L., Urdan, T., Anderman, L. H., et al. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contempor. Educ. Psychol.* 23, 113–131. doi: 10.1006/ceps.1998.0965
- Molenaar, D., Tuerlinckx, F., and Van der Maas, H. L. J. (2015). Package diffIRT. *J. Statistic. Software* 4:66.
- Mutz, D. C., and Pemantle, R. (2015). Standards for experimental research: Encouraging a better understanding of experimental methods. *J. Exp. Politic. Sci.* 2, 192–215. doi: 10.1017/XPS.2015.4
- Myers, A. J., and Finney, S. J. (2019). Change in self-reported motivation before to after test completion: Relation with performance. *J. Exp. Educ.* 1, 1–21. doi: 10.1080/00220973.2019.1680942
- Nagy, G., Nagengast, B., Becker, M., Rose, N., and Frey, A. (2018). Item position effects in a reading comprehension test: An IRT study of individual differences and individual correlates. *Psychol. Test Assess. Modeling* 60, 165–187.
- Nagy, G., Nagengast, B., Frey, A., Becker, M., and Rose, N. (2019). A multilevel study of position effects in PISA achievement tests: student- and school-level predictors in the German tracked school system. *Assess. Educ.* 26, 422–443. doi: 10.1080/0969594X.2018.1449100
- Nagy, G., Watt, H. M. G., Eccles, J. S., Trautwein, U., Lüdtke, O., and Baumert, J. (2010). The Development of Students' Mathematics Self-Concept in Relation to Gender: Different Countries, Different Trajectories? *J. Res. Adol.* 20, 482–506. doi: 10.1111/j.1532-7795.2010.00644.x
- Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., et al. (2019). PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* 51, 195–203.
- Penk, C., and Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educ. Assess., Eval. Account.* 29, 55–79. doi: 10.1007/s11092-016-9248-7
- Penk, C., and Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learn. Individ. Differ.* 42, 27–35. doi: 10.1016/j.lindif.2015.08.002
- Piaget, J. (1928). La causalité chez l'enfant. *Br. J. Psychol.* 18, 276–301. doi: 10.1111/j.2044-8295.1928.tb00466.x
- R Development Core Team. (2009). *R: A language and environment for statistical computing [Computer software manual]*. Vienna: R Development Core Team.
- Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Comput.* 20, 873–922. doi: 10.1162/neco.2008.12.06.420
- Revelle, W. (2019). *Using R and the psych package to find ω*. Version 2.0.9. 1–20.
- Revelle, W., and Zinbarg, R. E. (2008). Coefficients alpha, beta, omega and the glb: comments on Sijtsma. *Psychometrika* 1, 1–14.
- Rios, J. A., and Liu, O. L. (2017). Online proctored versus unproctored low-stakes internet test administration: Is there differential test-taking behavior and performance? *Am. J. Dis. Educ.* 2017:1258628. doi: 10.1080/08923647.2017.1258628
- Rios, J. A., Liu, O. L., and Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Direct. Institut. Res.* 161, 69–82.
- Rose, N., Nagy, G., Nagengast, B., Frey, A., and Becker, M. (2019). Modeling multiple item context effects with generalized linear mixed models. *Front. Psychol.* 10:248. doi: 10.3389/fpsyg.2019.00248
- Rosseel, Y. (2010). lavaan: an R package for structural equation modeling and more. *J. Statistic. software* 48, 1–36.
- Rosseel, Y. (2019). *Structural equation modeling with lavaan*. Available online at: [https://personality-project.org/r/tutorials/summerschool.14/rosseel\\_sem\\_intro.pdf](https://personality-project.org/r/tutorials/summerschool.14/rosseel_sem_intro.pdf).
- Rosseel, Y. (2020). *The lavaan tutorial*, Department of Data Analysis. Belgium: Ghent University.
- Schüttelz-Brauns, K., Hecht, M., Hardt, K., Karay, Y., Zupanec, M., and Kämmer, J. E. (2020). Institutional strategies related to test-taking behavior in low stakes assessment. *Adv. Health Sci. Educ.* 25, 321–335. doi: 10.1007/s10459-019-09928-y
- Schweizer, K. (2011). On the changing role of cronbach's  $\alpha$  in the evaluation of the quality of a measure. *Eur. J. Psychol. Assess.* 27, 143–144. doi: 10.1027/1015-5759/a000069
- Silm, G., Pedaste, M., and Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educ. Res. Rev.* 31:335. doi: 10.1016/j.edurev.2020.100335
- Simzar, R. M., Martinez, M., Rutherford, T., Domina, T., and Conley, A. M. M. (2015). Raising the stakes: How students' motivation for mathematics associates with high- and low-stakes test achievement. *Learn. Individ. Differ.* 39, 49–63. doi: 10.1016/j.lindif.2015.03.002
- Stenlund, T., Lyrén, P. E., and Eklöf, H. (2018). The successful test taker: exploring test-taking behavior profiles through cluster analysis. *Eur. J. Psychol. Educ.* 33, 403–417. doi: 10.1007/s10212-017-0332-2
- Sundre, D. L., and Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contempor. Educ. Psychol.* 29, 6–26. doi: 10.1016/S0361-476X(02)00063-2
- Sundre, D. L., and Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assess. Update* 14, 8–9.
- Svetina, D., and Rutkowski, L. (2014). Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments. *Meredith* 1993, 1–17.
- Tuerlinckx, F., and De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika* 70, 629–650. doi: 10.1007/s11336-000-0810-3
- Van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., and Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychol. Rev.* 118, 339–356. doi: 10.1037/a0022749
- Vollmeyer, R., and Rheinberg, F. (2006). Motivational effects on self-regulated learning with different tasks. *Educ. Psychol. Rev.* 18, 239–253. doi: 10.1007/s10648-006-9017-0
- Voss, A., Nagler, M., and Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Exp. Psychol.* 60, 385–402. doi: 10.1027/1618-3169/a000218
- Weirich, S., Hecht, M., Penk, C., Roppelt, A., and Böhme, K. (2016). Item position effects are moderated by changes in test-taking effort. *Appl. Psychol. Measur.* 41, 115–129. doi: 10.1177/0146621616676791
- Wigfield, A., and Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contempor. Educ. Psychol.* 25, 68–81. doi: 10.1006/ceps.1999.1015
- Wise, S. L., and DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educ. Assess.* 10, 1–17. doi: 10.1207/s15326977ea1001\_1

- Wise, S. L., and Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Appl. Measur. Educ.* 18, 163–183.
- Wolgast, A., Tandler, N., Harrison, L., and Umlauf, S. (2019). Adults' dispositional and situational perspective-taking: a systematic review. *Educ. Psychol. Rev.* 32, 353–389. doi: 10.1007/s10648-019-09507-y
- Zinbarg, R. E. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for  $\omega_h$ . *Appl. Psychol. Measur.* 30, 121–144. doi: 10.1177/0146621605278814

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Wolgast, Schmidt and Ranger. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Undergraduate Students' Critical Online Reasoning—Process Mining Analysis

Susanne Schmidt<sup>1\*</sup>, Olga Zlatkin-Troitschanskaia<sup>1</sup>, Jochen Roeper<sup>2</sup>, Verena Klose<sup>2</sup>, Maruschka Weber<sup>2</sup>, Ann-Kathrin Bültmann<sup>1</sup> and Sebastian Brückner<sup>1</sup>

<sup>1</sup> Department of Business and Economics Education, Johannes Gutenberg University, Mainz, Germany, <sup>2</sup> Department of Neurophysiology, University Hospital of the Goethe University, Frankfurt, Germany

## OPEN ACCESS

### Edited by:

Raquel Gilar-Corbi,  
University of Alicante, Spain

### Reviewed by:

Damanjit Sandhu,  
Punjabi University, India  
Gonzalo Lorenzo Lledo,  
University of Alicante, Spain

### \*Correspondence:

Susanne Schmidt  
susanne.schmidt@uni-mainz.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Psychology

**Received:** 25 June 2020

**Accepted:** 02 November 2020

**Published:** 30 November 2020

### Citation:

Schmidt S,  
Zlatkin-Troitschanskaia O, Roeper J,  
Klose V, Weber M, Bültmann A-K and  
Brückner S (2020) Undergraduate  
Students' Critical Online  
Reasoning—Process Mining Analysis.  
Front. Psychol. 11:576273.  
doi: 10.3389/fpsyg.2020.576273

To successfully learn using open Internet resources, students must be able to *critically search, evaluate and select online information, and verify sources*. Defined as critical online reasoning (COR), this construct is operationalized on two levels in our study: (1) the *student level* using the newly developed Critical Online Reasoning Assessment (CORA), and (2) the *online information processing level* using event log data, including gaze durations and fixations. The written responses of 32 students for one CORA task were scored by three independent raters. The resulting score was operationalized as “task performance,” whereas the gaze fixations and durations were defined as indicators of “process performance.” Following a person-oriented approach, we conducted a process mining (PM) analysis, as well as a latent class analysis (LCA) to test whether—following the dual-process theory—the undergraduates could be distinguished into two groups based on both their process and task performance. Using PM, the process performance of all 32 students was visualized and compared, indicating two distinct response process patterns. One group of students (11), defined as “strategic information processors,” processed online information more comprehensively, as well as more efficiently, which was also reflected in their higher task scores. In contrast, the distributions of the process performance variables for the other group (21), defined as “avoidance information processors,” indicated a poorer process performance, which was also reflected in their lower task scores. In the LCA, where two student groups were empirically distinguished by combining the process performance indicators and the task score as a joint discriminant criterion, we confirmed these two COR profiles, which were reflected in high vs. low process and task performances. The estimated parameters indicated that high-performing students were significantly more efficient at conducting strategic information processing, as reflected in their higher process performance. These findings are so far based on quantitative analyses using event log data. To enable a more differentiated analysis of students' visual attention dynamics, more in-depth qualitative research of the identified student profiles in terms of COR will be required.

**Keywords:** online information processing, Critical Online Reasoning Assessment, person-oriented approach, event logs, eye-tracking, process mining, latent class analysis, response process patterns

## INTRODUCTION

### Study Background

The Internet and social media are among the most frequently used sources of information today. University students also often prefer online information to more traditional teaching materials such as textbooks (Brooks, 2016; List and Alexander, 2017; McGrew et al., 2019; Maurer et al., 2020). Learning by using freely accessible online resources offers many opportunities, but it also poses novel challenges. Conventional teaching material provided by universities is usually carefully curated by experts and tailored to maximize learning for students at clearly defined stages of their respective curricula. In contrast, the difficulty in using online resources lies in having not only to find suitable resources but also evaluate them without expert guidance. While the information stated on these websites may be correct and may simply not be well suited for the students' learning aims, there are also more problematic scenarios. It has become evident that not only "fake news" but also "fake science" characterized by scientifically incorrect information circulates on the Internet (Ciampaglia, 2018). Therefore, to successfully learn by using open Internet resources, students must be able to critically search, evaluate, select, and verify online information and sources. In addition to a critically minded attitude, students need the ability to distinguish reliable from unreliable or manipulative information and to question and critically examine online information so they can build their knowledge and expertise on reliable information.

In recent years, however, there has been increasing evidence that university students struggle to critically assess and evaluate information from the Internet and are often influenced by unreliable sources (McGrew et al., 2018; Wineburg et al., 2018). Although current research assumed that search context and strategies were related to the information seeking and evaluation processes, not much is known about the specific search strategies and activities of students on the web, especially with regard to learning using the Internet (Walraven et al., 2009; Collins-Thompson et al., 2016).

To investigate how university students deal with online information and what influences their information processing, we used the newly developed online test "Critical Online Reasoning Assessment" (CORA), which is based on the Civic Online Reasoning assessment developed by the Stanford History Education Group (Wineburg et al., 2016). During the assessment, the test takers are presented with novel performance tasks; while working, they are asked to freely browse the Internet to find and select reliable information relevant for solving the given tasks within the relatively short time frame of 10 min. As part of their written response, they are asked to justify their task solutions by using arguments based on their online research. For the study presented here, university students from three disciplines (medicine, economics, and education) at two universities in two German federal states were tested using CORA (for details, see the section "Study Design").

Recently, the Standards for Educational and Psychological Testing (American Educational Research Association [AERA] et al., 2014) have emphasized the particular importance of test "validity evidence" based on response processes. Response

processes refer to psychological operations, approaches and behaviors of test takers when they carry out tasks and create solutions. They are revealed through response process data, i.e., verbalizations, eye movements, response times, or computer clicks (Ercikan and Pellegrino, 2017). Therefore, American Educational Research Association [AERA] et al. (2014) introduced the response processes as one of the central criteria of "validity," i.e., "the degree to which a test score can be interpreted as representing the intended underlying construct" (Cook and Beckman, 2006). These processes need to be distinguished from construct-irrelevant processes, i.e., not defined by the construct (e.g., guessing as a task processing strategy).

Following this paradigm in educational assessment, which focuses on validation of scores by means of analyzing response processes (Ercikan and Pellegrino, 2017), the aim of this study was to obtain validity evidence about response processes. This entails (i) describing the processes underlying students' critical online reasoning (COR) when solving the CORA tasks, as well as (ii) analyzing the relationship between the task scores ("critical online reasoning") and the response processes of the CORA participants to define the extent to which the response processes typically reflect information processing and solving strategies associated with the COR construct (see the section "Construct Definition and Fundamental Assumptions").

### Research Objectives

As recent research on online information processing indicated (see the section "Theoretical Framework"), web searches entail cognitive and metacognitive processes influenced by individual differences. Therefore, students with similar cognitive and/or metacognitive abilities tend to develop very different search and information processing strategies (e.g., Kao et al., 2008; Zhou and Ren, 2016; Brand-Gruwel et al., 2017). This is particularly true for CORA, where students' spontaneous web searches in a natural online environment were measured *in vivo*. To precisely describe and analyze the different patterns in students' CORA task solution processes and online information processing, we introduced a new method of process mining (PM).

To comprehensively examine students' response processes in detail when dealing with online information, their web search activity while solving the CORA tasks was recorded for analysis. The collected data included (i) the entire log data collected throughout the task-solving process, including all websites the students accessed during their search, as well as the eye movements of test takers recorded during their CORA task processing and online search to gain additional insights into the participants' cognitive processes while using online media; and (ii) the videos of their task-solving behavior recorded by the eye tracker, to more directly observe what students do and do not do while solving CORA tasks (for details, see the section "Study Design").

Starting from a theoretical process model (see the section "Theoretical Framework"), the main research objective of this article is (i) to describe the response processes of students dealing with online information while they are working on the CORA tasks, and (ii) to investigate the relationship between the students' online information processing strategies and their performance

on the CORA tasks, in particular, by comparing students with high vs. low process and task performance.

Based on event logs including eye-tracking (ET) data (see the section “Materials and Methods”), as indicators of students’ response processes, log files and eye movements while solving the CORA tasks were considered. This provided information about the numbers, types, and orders of the main process steps (such as reading the instruction, Google searches, selecting websites, reading webpages, writing response), and the durations and fixations per individual process step (for details, see the section “Materials and Methods”).

Using the PM approach, we investigated the following research questions (RQs):

RQ1: How can the students’ response processes, related to their COR while solving CORA tasks, be visualized and precisely described?

RQ2: What are similar and distinct patterns in the students’ response processes related to their COR based on

- their process performance variables: fixations per individual process steps as well as the duration, number, type, and order of the individual process steps?
- how are they related to the (overall) task performance, i.e., the score of the written responses?

To answer these RQs, the log file data using ET data were prepared, and patterns contained in event logs recorded in the eye tracker while the participants solved the CORA tasks were identified and analyzed using PM (see the sections “Study Design” and “Materials and Methods”). In the next step, the raw ET and log data were processed and transformed into a newly generated and aggregated data set, which was used in the subsequent model-based statistical analyses [latent class analyses (LCA)] to test the research hypotheses (see the section “Results”).

## THEORETICAL FRAMEWORK

### State of Research and Conceptual Foundation

There is currently no unified framework for and definition of COR in the literature. Existing studies about students’ information-seeking behavior (Hargittai et al., 2010) and the underlying processes (e.g., “search as learning,” Hoppe et al., 2018) have been based on different frameworks and research traditions: recent research on *online* information processing and behavior was primarily based on frameworks developed in the context of “multiple source use/comprehension” (e.g., “multiple documents literacy,” Anmarkrud et al., 2014; for an overview on MSU, see Braasch et al., 2018) and “information problem solving” (Brand-Gruwel et al., 2005). These approaches required a decomposition of different (meta-)cognitive subskills and solving strategies such as “defining the problem” to deal with information problems. One research strand has been particularly focused on the credibility of web-based information as an explanatory factor of online information-seeking behavior (e.g., approaches on “web credibility,” Fogg et al., 2003; Metzger and Flanagin,

2015; “credibility evaluation,” Metzger et al., 2010; “information trust,” Lucassen and Schraagen, 2011). Recent studies indicate the particular role and influence of cognitive heuristics that information users employ when evaluating the credibility of online information and sources. Moreover, credibility evaluations appear to be primarily due to website characteristics.

With regard to “reasoning,” the research refers to well-established traditions, focusing in particular on generic reasoning and scientific reasoning (e.g., Fischer et al., 2014; Stanovich, 2016) and (corresponding) reasoning biases and heuristics (Kahneman et al., 1982; Gigerenzer, 2008; Hilbert, 2012) of information processing and decision making (for an overview, see, e.g., Stanovich, 2003; Toplak et al., 2007). More specifically, for instance, recent results indicate that most students routinely applied (meta-)cognitive heuristics (e.g., self-confirmation heuristics) to process and to evaluate the credibility of online information and sources (Metzger et al., 2010).

Research with a particular focus on students’ web search behavior and their navigation of online environments and online information resources, however, is still relatively scarce. The existing studies focusing on process analysis of students’ use of multiple online sources and information, which we described as “online reasoning” (McGrew et al., 2019), are based on an amalgamation of several theories and frameworks of cognition and learning; in particular: (i) development of expertise in (meta-)cognitive and affective information processing with different media (Mayer, 2002; Alexander, 2003; List and Alexander, 2019); and (ii) critical use of multimodally represented information from multiple sources (Shaw, 1996; Wiley et al., 2009; Braasch and Bråten, 2017; List and Alexander, 2017; Yu et al., 2018). For instance, List and Alexander (2017) found four different navigation profiles of satisfying approaches when students navigate information without time limitations (e.g., the limited navigation profile, and the distributed navigation profile); these profiles were differently correlated with task performance.

Over the last few years, process research on (online) learning behavior (Ercikan and Pellegrino, 2017; Zumbo and Hubley, 2017) using verbal data (Leighton, 2017) and computer-generated data (Goldhammer and Zehner, 2017; Li et al., 2017; Oranje et al., 2017; Russell and Huber, 2017) (e.g., log data, ET, dwell times) has increasingly been developed to gain insights into students’ (meta-)cognitive information processing, indicating that response process evidence (“cognitive validity”) is required to validate claims regarding the cognitive complexity of performance assessments such as CORA.

### Construct Definition and Fundamental Assumptions

When developing the conceptual framework for this study, we used the broad definition of COR, which describes the ability to effectively search, verify (i.e., to prove the accuracy), and evaluate (i.e., to draw conclusions from examining) online information (McGrew et al., 2018), as a starting point. We claim, therefore, that COR expresses itself in the ability to identify the author and/or the organization behind a source of information and

to make an informed assessment regarding their motives and their trustworthiness, to verify their claims by consulting other (reliable) sources, and finally to come to a conclusive decision about the utility of the source.

To capture the response processes underlying COR, the construct is further defined by including more distinct facets of online information processing and information problem-solving strategies. First, the individual's distinct phases within the online information processing are described in more detail. For systematizing and classifying these processes, we applied a descriptive approach (e.g., Gerjets et al., 2011) based on the Information Problem-Solving (IPS-I) model (Brand-Gruwel and Stadler, 2011). This model distinguishes constitutive basic abilities of problem solving when using online information, which are activated by regulatory and conditional skills, for instance, searching, scanning, processing, and evaluating (online) information.

Second, to qualitatively describe how these skills manifest, the subprocesses involved in processing online information are classified into heuristic and systematic processes in accordance with the dual-process theory (Chen and Chaiken, 1999; for the heuristic analytic process model, see Evans, 2003). This theory has already been established in several research domains, including inference and reasoning (De Neys, 2006; Evans, 2006), the evaluation of credibility (Metzger et al., 2010), as well as in the context of ET studies (e.g., Horstmann et al., 2009; Wu et al., 2019). Horstmann et al. (2019) claimed, for instance, a relation between (meta-)cognitive heuristics and the different components of visual search such as skipping (distracters), dwelling, and revisiting.

Originally, this theory was developed to explain the prevalence of cognitive bias in argumentation tasks (Evans, 1984, 1989). According to different research perspectives, heuristic and systematic cognitive processes can occur either simultaneously ("Parallel Models," Sloman, 1996) or sequentially ("Default Intervention Models," Evans and Stanovich, 2013). Heuristic processes are mostly experience-based (Gronchi and Giovannelli, 2018) and are assumed to occur fast, unconsciously, automatically, and with low cognitive effort (Horstmann et al., 2009; for more details about heuristics, see Stanovich, 2012; for schema theory, see Anderson et al., 1978). Systematic processes require a higher cognitive effort, make use of complex mental models, and activate deliberative reasoning; they are goal- and rule-driven, analytical, precise, and based on weighing up the positive and negative aspects of various options (Chen and Chaiken, 1999; Evans and Stanovich, 2013).

The existing studies that focus on the heuristic and systematic processes of information seeking indicate the common use of cognitive heuristics instead of deliberate strategies while evaluating and comparing online information (Horstmann et al., 2009). The perceived ranking of web search results has an important impact on the evaluation and judgment of online information (Hargittai et al., 2010). For instance, students rely immensely on the ranking provided by search engines and mostly access only the first few websites presented (e.g., Walraven et al., 2009; Gerjets et al., 2011; Zhang et al., 2011). Observing domain experts while solving an information problem and comparing

them to novices, Brand-Gruwel et al. (2009) showed that experts spend more time on the process steps of information problem solving, and they metacognitively evaluated their solving process more often. Experts were also more likely to alternate between searching and viewing webpages and decided to leave webpages to return to the hit list faster than novices. Collins-Thompson et al. (2016) showed that the amount of time web searchers spend on one document during the searching process was positively correlated with their higher-level cognitive learning scores. Likewise, Anmarkrud et al. (2014) indicated the positive relationship of students strategic processing while reading the documents and source awareness in multiple document use of undergraduates. Wineburg and McGrew (2017) found that professional "fact checkers" read laterally and leave a website after a quick scan to initially gain more insight into the credibility of the website through external sources, whereas less experienced students read vertically and judge the website according to its own attributes.

Based on prior research, we assume that COR should be based on strategic information processing, i.e., a combination of both experience-based (meta-)cognitive heuristics to efficiently process online information, which can be applied flexibly in the context of information problem-solving in certain (task-related) situations, as well as systematic processes to activate the deliberative ("critical") reasoning; strategic information processing may relate to a better "process performance" and "task performance."

## Eye Movements as an Indicator of Cognitive Processes

To provide insights into students' cognitive information processing, in particular, operationalized through gaze durations and fixations, ET is increasingly used in related research (Horstmann et al., 2009; Gerjets et al., 2011). Accordingly, sequences of eye movements are identified that can be used to operationalize students' depth of processing and thus draw inferences on their cognitive effort. The focus of cognitive information processing is generally on attention-related processes (Orquin and Loose, 2013). Therefore, fixations on so-called "areas of interest" (AOIs) are often chosen as an indicator of cognitive effort (Gerjets et al., 2011; Raney et al., 2014). Fixations are periods of stabilized eye positioning (fixed gaze), during which a small AOI in the visual field (about the size of the moon in the sky) is presented on the fovea, the part of the retina with the highest visual acuity (Duchowski, 2007). Complex visual scenes are analyzed by a sequence of fixations under attentional control and are separated by jump-like eye movements every few hundred milliseconds (saccades). Given that fixations give access to highly resolved visual information, they are also indicators of cognitive processing depth (Holmqvist et al., 2011) and might therefore be considered as surrogate markers for "process performance." Thus, a basic assumption of ET is that increased processing demands are associated with increased processing time and/or changes in the patterns of fixations. Increased processing time may be reflected by longer fixations and/or larger numbers of fixations (forward and regressive) (Raney et al., 2014).



For instance, in an ET study, Zhou and Ren (2016) examined the cognitive strategies of students, focusing on their searching process when looking for specific online information. They found that high-performing students revised search queries more often and spent more time reading and assessing the information in the selected webpages for its relevance. Moreover, high performers switched more frequently between search results and webpages before staying on a certain webpage, which might indicate a more critical metacognitive engagement.

As eye movements are influenced by numerous factors (e.g., Reitbauer, 2008; Cyr and Head, 2013), conclusions from ET data on information processing and online reasoning—in accordance with the dual process theory—constitute a model-based method of complexity reduction to gain first insights into these processes and their relationships with students' CORA task performance. We assume that students' eye movements while solving CORA tasks reflect their process performance. More specifically, fixations per individual process step, as well as the duration, number, type, and order of the individual process steps—in accordance with the IPS Model—are used as indicators of students' process performance (see the section “Materials and Methods”) and are related to their task performance (i.e., the score of the written responses). In addition, following the dual process theory and the aforementioned findings from related ET research, we assume two distinct patterns in the students' process performance related to their COR.

## STUDY DESIGN

Critical online reasoning was measured using the newly developed CORA (Molero et al., 2019). Originally, it contained five tasks, which we reduced to two representative tasks for this PM study, i.e., analysis of log events based on ET data. In the tasks, participants were presented with URLs that directed them to a website (published by a company focusing on online marketing) about vegan protein sources (Task 1) and a tweet (published by a market-liberal organization) about German state revenue (Task 2), respectively. These sources of online information were to be evaluated by the participants with regard to credibility in a free-answer format. They were also asked to give evidence by providing URLs to websites/sources that supported the argumentation in their task response (written statement).

The two CORA tasks were implemented into an ET test environment using Tobii Pro Lab software and hardware. To facilitate the subsequent extraction and comparisons of the test participants' data, especially with regard to how they used the websites linked in the URLs, a web stimulus presentation was used instead of a more open screen recording. To integrate the two tasks into a web stimulus, an online writing document was generated using the web-based text editor EduPad (EduPad is a collaborative text environment based on the Etherpad software<sup>1</sup>). Participants were asked to formulate their argumentative task response in the same EduPad document (hereafter referred to as “Task Editor”). During the CORA, they could use any

information on the provided website/tweet and were also asked to use other websites and search engines, to stimulate a naturalistic online information processing and problem-solving behavior. The students could switch between websites and the document containing the tasks and reread the task prompt or edit their written task response.

Participants were given 10 min per task to complete the two CORA tasks (excluding the time required for calibrating the eye tracker). A short instruction (hereafter referred to as “Reading Instruction”) at the beginning of the task informed the participants about the study procedure as well as about suggested approaches to solving the task, i.e., using all information on the linked websites as well as other websites, using search engines, and how they could return to the task editor. After reading the instruction, students had to actively start the task by activating the task editor EduPad.

Process data, which comprise log events including eye movements during the CORA, were recorded with a Tobii Pro X3-120 eye tracker using a sampling rate of 120 Hz. The recorded data provided further details about participants' response processes and how they processed and interacted with information online, for instance, through mouse clicks, query streams, or weblogs. To visualize the information processing behavior of each student, we used PM, where fixations were counted for all task process steps, including for every single webpage. Thus, the sum of fixations for each webpage was calculated and considered an additional indicator of the participants' process performance while working on the CORA tasks (for details, see the section “Data Transformation”).

After the CORA, raw event log data that contained the fixations and web search events of all participants at the millisecond level were exported, prepared, and transformed to exploratively determine the response process steps and students' “process performance” using a PM approach (see the section “Process Mining”). Based on the PM analysis, we gained insights into the process steps as well as the distinct process steps of each student. “Distinct,” for the purposes of this study, referred to the sum of identical (but potentially repeatedly executed) process steps. For instance, if a student opened a website, this was probably followed by some activity on that particular website, like reading or scrolling, and then the student would typically return to the task editor, to take, for instance, several notes about the contents on this webpage or to copy the hyperlink as a reference for the written response. However, the student then may have returned to that same website, which the event log would record as a next process step with the same name but with a different timestamp. Therefore, the event log data list at least three process steps (website → “Task Editor” → website), but the number for *distinct* activities would be only two as the student repeated one unique process step (visiting the same website twice).

This transformed data were then used for exploratory PM analyses, as well as statistical model-based analyses to test the formulated hypotheses (see the section “Discussion of Process Mining Results and Research Hypotheses”). The process data were also aggregated (such as the average number of fixations per process step of each student) and combined with the “task performance” data, which included the dependent variable of

<sup>1</sup><https://edupad.ch/#about>

the score of the students' written responses to the CORA tasks. The students' responses (written statements) were scored by three independent (trained) raters using a developed and validated rating scheme ranging from 0 to a maximum of 2 points (as a 5-point Likert scale: 0, 0.5, 1, 1.5, 2; for details, see Molerov et al., 2019).

Based on a person-oriented analysis approach, the hypotheses on the expected significant differences between the students in terms of their "process performance," which was calculated based on the (i) assessed and aggregated process-related variables using the number, of total process steps as well as the number of distinct processes performed, the average number of fixations as well as the average duration of each process step for every student, all of which were analyzed in relation to (ii) the students' task performance (CORA task score), were investigated by means of an LCA (see the section "Latent Class Analyses").

## Sample

In total, 32 undergraduate students from the fields of medicine (9 participants), economics (9 participants), and business and economics education (14 participants) took part in the CORA. In the context of three obligatory lectures or seminars (in economics, education, and medicine) at two German universities, all students attending these courses were asked to complete a paper-pencil survey to assess their domain-specific knowledge and other personal characteristics (e.g., fluid intelligence, media use). Subsequently, the students who had participated in this survey were asked to take part in the CORA as well, to recruit approximately 10 test persons from each of the three groups, in accordance with the purposeful sampling method (for details, see Palinkas et al., 2015).

For this article, a purposefully selected sample of participants was used, as the amount of data for an unselected number of participants would have been too large and not feasible for practical research purposes. When selecting this sample, we included students from all study semesters, and we selected students from two disciplines to initially control for domain specificity. Another important criterion for the sampling was the students' central descriptive characteristics such as gender, age, migration background, and prior education, which may influence their web search behavior and COR task performance.

For the first task, the data on all 32 participants could be used, whereas for the second CORA task, the sample was reduced to  $N = 30$ , as the survey had to be terminated prematurely because of technical problems for two participants. Additional background information on the participants, such as gender, age, and study semester, was assessed. The average age for the sample was 22.37 years ( $SD = 4.1$  years); 71% of the 32 participants were female. Most of the participants were in their first or second year of study. With regard to the distribution of the descriptive characteristics such as age and gender, no significant deviations from the overall student population in these study domains were found. As participation in this study was voluntary, however, a bias in the sampling cannot be ruled out (see the section "Limitations").

## MATERIALS AND METHODS

### Process Mining

A unique aspect of the CORA is that the assessment of students' spontaneous web searches took place in a naturalistic online environment, with a freely accessible world wide web without any restrictions (besides a limited task processing time of 10 min). At the same time, this makes data analysis particularly demanding and requires a precise and differentiated description of the data in its sequential process structure in the first step. To be able to answer the two RQs, RQ1, and RQ2, which aim to visualize, describe, and discover patterns in the students' CORA task response behavior, we focused on the method of PM to analyze process-related student data as we tracked in detail how each student approached the CORA task. For PM data transformation and visualization, we used PAFnow<sup>2</sup>.

Process mining is an aggregative, visualization methodology to gain insights and acquire knowledge about the test-taking process behavior of individual participants, recorded in event logs. In educational research, this is typically based on data collected in computer-based assessments (Tóth et al., 2017). Using the visualized data collected through PM and representing students' processing behavior in a process graph can reveal information about the homogeneity and heterogeneity of students' "process performance" (which can then be related to their task performance in the next step, see the section "Latent Class Analyses"). By comparing the *number* and *kind* of process steps, the *order* of these process steps or the time spent on each process step, i.e., duration, we explored whether there were similarities or differences (at the process level) between or within students.

In the present study, the number of fixations within each process step was also included in the analyses. Hereby, we distinguished between the analyses at the process level or at the student level. The number of process steps, for instance, related to the student level. The number of fixations, however, can also refer to the process level, as these data were recorded for every single activity of the participants on a millisecond level and with high temporal resolution. For the PM analysis, the number of fixations was aggregated for each meaningful process step. Accordingly, the conducted analyses referred to different levels: the PM approach refers to the process level, whereas the LCA refers to the student level (according to a person-oriented approach, see the following section).

### Person-Oriented Approach and Latent Class Analyses

To investigate RQ2, the additional methodological focus for the analyses of students' response process data was—in accordance with Bergman and El-Khoury (2001)—on the person-oriented approach, which states "that interindividual differences and group differences need not be added to the error variance, and that they are worthy of being made the object of investigation" (von Eye, 2006, pp. 11–12). As prior research on

<sup>2</sup><http://www.pafnow.com/>

(online) information processing indicated, the way information is processed—in our case, while solving the CORA task—is (at least partially) (sub)group-specific (Kao et al., 2008; Zhou and Ren, 2016; Brand-Gruwel et al., 2017). Given this assumption (see the section “Construct Definition and Fundamental Assumptions”), we aim to empirically uncover the existence of possible subgroups in the student population, i.e., to investigate latent subgroups. The term “subgroups” is defined as a set of participants in the sample who are more similar to each other in terms of task and process performance than others. The division of the student population into subgroups focuses on the individual differences between the students, as these differences have a decisive influence on and characterize the information processing (Sterba and Bauer, 2010). These group differences or the affiliation of the participants to certain subgroups was not known *a priori* in the present study. This was taken into account in our study by examining whether the students were divided into latent subgroups based on their task score and the indicators of their process performance, i.e., the duration of the distinct process steps, the number of performed (distinct) process steps, and the number of average fixations per process step. The person-oriented approach allowed for an adequate testing of the research hypotheses against the backdrop of strong heterogeneity and different information processing approaches of the participants.

Based on the “dual process” model (see the section “Construct Definition and Fundamental Assumptions”), it was assumed that depending on both their task performance as well as on their process performance, the students population can be empirically divided into two subgroups: (1) high performers and (2) low performers. It was supposed that the process performance not only leads to an according test performance (CORA task score), but it is also reflected in process performance indicators that can provide a comprehensive picture about the cognitive information processing strategies the students used to achieve a (higher) test score, such as duration, fixations, and number of (distinct) process steps. Therefore, the investigation of high and low performers not only accounted for the task score, but also for process-relevant variables, using an LCA as an appropriate empirical model to test whether the supposed multigroup structure can be empirically determined.

The probability that a person  $s$  has a certain response pattern is the same for all persons; therefore,  $p(a_s) = \prod_{b=1}^5 p(y_{ib})$  (Rost and Eid, 2009, p. 490), with  $i$  as CaseId and  $b = \{\text{task score, total fixations, total duration, number of process steps, number of distinct process steps}\}$ . Taking into account the latent class belonging to one of the two classes  $c1$  (high performer) or  $c2$  (low performer), the results in the conditional probability are  $p(a_s|c) = \prod_{b=1}^5 p(y_{ib}|c)$ . These two response pattern probabilities are required to determine the conditional class belonging to probability  $p(c|a_s)$ . The goal of a model of LCA is therefore to predict the probability that a person  $s$  conditionally belongs to a certain class  $c$  based on his/her “process performance” vector  $a_s$ . The model is as follows (Gollwitzer, 2012, p. 302):

$$p(c|a_s) = \pi_c \frac{p(a_s|c)}{p(a_s)},$$

where  $\pi_c = p(c)$  stands for the unconditional class belonging to probability (relative class size), with  $\pi_{c1} + \pi_{c2} = 1$  (i.e., each person in the sample belongs to exactly one latent class  $c$ ). The unconditional “process performance” indicators can thus be defined as a discrete mixture of the conditional probabilities of performance patterns, and the following holds for both classes  $c1$  and  $c2$  (Rost and Eid, 2009, p. 490; Gollwitzer, 2012, p. 302; Masyn, 2013, p. 558):

$$p(a_s) = \pi_{c1} \cdot p(a_s|c1) + \pi_{c2} \cdot p(a_s|c2).$$

The LCA was conducted using Stata 16<sup>3</sup> with an identity link and reporting the conditional classification for each student and the predicted category of being a high or low performer. The global LCA model fit was evaluated using Akaike's information criterion (AIC) and Bayesian information criterion (BIC) where the two-class model was benchmarked versus a one-class model to test whether the hypothesis of having two groups in the sample (high- and low-performing students) can be confirmed (see the section “Latent Class Analyses”).

## Data Transformation

To conduct a PM analysis to visualize the response processes (RQ1) and to perform the LCA to investigate the latent subgroups based on their process performance indicators to find similar and distinct patterns in both their task performance as well as in their response processes (RQ2), process data are required. To collect the process data of each student while working on the CORA task, the eye-tracker Tobii X3-120 (120 Hz) was used. We gathered ET information on both gaze-related data, such as eye movements and fixations (e.g., on the webpages) and additional process-related data, such as the different events during the web search, including the URL of the visited website and the keyboard events. For instance, when the response to the CORA task was written, the data were recorded and stored by the eye tracker in a tsv-formatted table (for an example of the raw data, see **Supplementary Table S1**). The durations of these events were also documented.

The raw data from the eye tracker were calibrated on the milliseconds level: each key stroke and each gaze point were stored in the data set, so that eye movements were temporally aligned to other process data. For the PM approach to reproduce the process behavior of the web search and task performance, however, a higher time level is required. The following steps of data processing and aggregation were conducted to create a meaningful event log that allowed the interpretation of students' response processes regarding the CORA task:

## Evaluation of Event Occurrences and Analysis of Visited Websites

In the eye-tracker data set, three different types of events were recorded: keyboard events, mouse events, and URL events. Keyboard events comprised typical writing events, such as pressing letter or number keys, as well as other events, such as changing window ([alt] + [tab]) or copy/pasting events ([ctrl] + c/[ctrl] + v). Mouse events are typically just scrolling or clicking activities. Regarding URL events, each and every

<sup>3</sup><https://www.stata.com/>



single URL that was accessed by a student was recorded (see **Supplementary Table S2**).

### Aggregation of Single Events to Distinct Process Steps

Based on the findings from the analyses of visited websites (incl. webpages), we aggregated these results for occurrences between students. For defined typical events, such as watching a YouTube video, conducting a web search using Google, gathering information from a newspaper website or other common URLs, the occurrences were translated into a comparable process step. If student A opened YouTube and watched a video X and student B watched video Y, the differences between the videos in terms of content were not further considered as it was not relevant for the PM analysis. This analysis does not aim to explain the differences between students who opened different YouTube videos or who gathered information from different newspapers. Therefore, the commonly visited URLs' events were aggregated in the meaningful categories, and then into (distinct) process steps. URLs that were visited by only very few students were not aggregated (for a detailed overview of the URLs' events and their aggregation to process steps, see **Supplementary Table S2**).

Moreover, keyboard events were also aggregated using a similar procedure as described in (1). All activities related to writing were summarized under "Keyboard Event—Writing" and all other keyboard events were summarized under "Keyboard Events—Other." Similarly, all mouse events were also aggregated in one process steps. In addition to URL events, mouse and keyboard events, the process steps of reading the instructions (welcome text and general description of CORA), and task editor (where the CORA task itself is shown and where the students write their responses) were also distinguished. In addition, the event log also consisted of further (less informative but still process-relevant) process steps, such as "Eye-Tracker Calibration," "Recording Start and End," or "Web Stimulus Start and End" (for the full event log, see **Supplementary Table S3**).

For the PM analyses, we focused only on the most important process steps to visualize the construct-relevant web search and task-solving process (for an example of a shortened event log for the PM analyses, see **Supplementary Table S4**). For instance, mouse events were excluded from the PM analysis, as they did occur at any time and were associated with most processes (e.g., while reading the instructions, during web searches, or while writing the response).

### Summarization of Fixations and Working Time for the Particular Process Steps, as Indicators of "Process Performance"

For all process steps described in 2, duration, i.e., time spent on each process step as well as the number of fixations recorded in each process step, was calculated. When aggregating the raw data as described above, both indicators of "process performance," the number of fixations and the duration, were summarized for each process step (for details, see the section "Process Mining"). Based on this strategy, a comparison of duration and fixations of each student and for each single process step was possible, as well as a comparison of the same "process performance"

indicators (such as duration and the number of fixations) between different students.

### Building a Data Model With Process (Event Log Table) and Student (Case Attributes Table) Related Variables

To conduct PM analyses and investigate process behavior on the process level, a data set with process-related variables was required in which the sequence of the process steps was in the correct temporal order. This table was called an event log table, as it comprises all relevant variables on the event level. An event log showed the unique identifier for each student (CaseId) and the process steps each student executed, as well as the according timestamps. Furthermore, the event log consisted of the number of fixations for each process step (for an example of an event log of one student, see **Supplementary Table S3**).

To explain "process performance" between students, variables on the student level, such as the CORA task score, were required. Based on the unique student identifier, the CaseId, a separate table for process-related variables aggregated on the student level, was added to the data model. This table was called case attributes table and consisted of the CORA task score and all aggregated process variables for each student: the total number of fixations and the total time spent on the task (total visit duration), as well as the total number of process steps (e.g., the count of the rows of the event log table as aggregated process steps, see **Supplementary Table S4**) and the number of distinct process steps. For the latter, for instance, while the total number of process steps for one student was 30, within these 30 steps, he/she read the instructions, opened the task editor, and immediately started to write the response, so that the number of distinct process steps would only be three in this case.

Based on these four major steps regarding data preparation, a new transformed and aggregated data set was created that allowed for the following analyses as described in the section "Results."

## RESULTS

### Process Mining

Using PM as an explorative approach to visualize and precisely describe the students' response processes while solving the CORA tasks (RQ1), first, the processes students applied while working on the CORA tasks were analyzed for the entire sample of 32 students. The process steps while working on CORA task 1, which included an unrestricted web search, were visualized and analyzed by first comparing the structure of the entire process graph for all students, including all events that were recorded by the eye tracker (see the section "Data Transformation"), to evaluate the students' task-solving behavior and process performance. The aim was to reveal potential common patterns in the students' performance variables, such as fixations, duration, and process steps (RQ2), which go along with the COR construct definition of searching, evaluating, and refining information before or while formulating a response (see the section "State of Research and Conceptual Foundation"). However, a visualization of all the process steps in one process graph did not provide the type of information that would



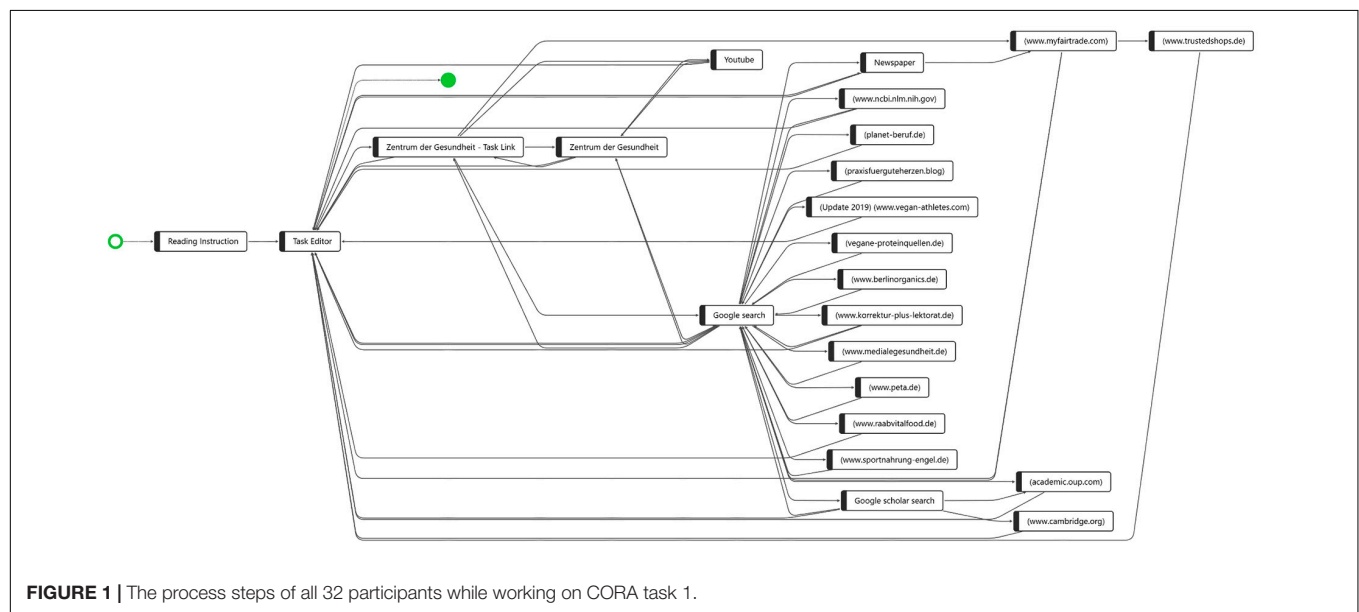
have allowed for meaningful interpretation and for answering the two RQs, as a precise revelation of distinct or common patterns in the students' response processes was hardly possible (see **Supplementary Figure S1**). For the following PM analysis, only the process steps of "Reading Instruction," "Task Editor," as well as all URL events were included (for an example of short event log for PM analysis, see **Supplementary Table S4**), so that the visualization of the process graph was readable, and the focus was on the interpretation of the web search activities. "Keyboard Events" and "Mouse Events" were also excluded from the process graph, as they could occur at any time during each process.

**Figure 1** shows the process variants of all 32 students while working on CORA task 1 combined in one graph. The starting point (on the left side marked as a green hollow circle) has only one arrow pointing at the process step "Reading Instruction," indicating that all students started by reading the instruction. In the second process step, all students opened the "Task Editor," which had two functions in the computer-based assessment

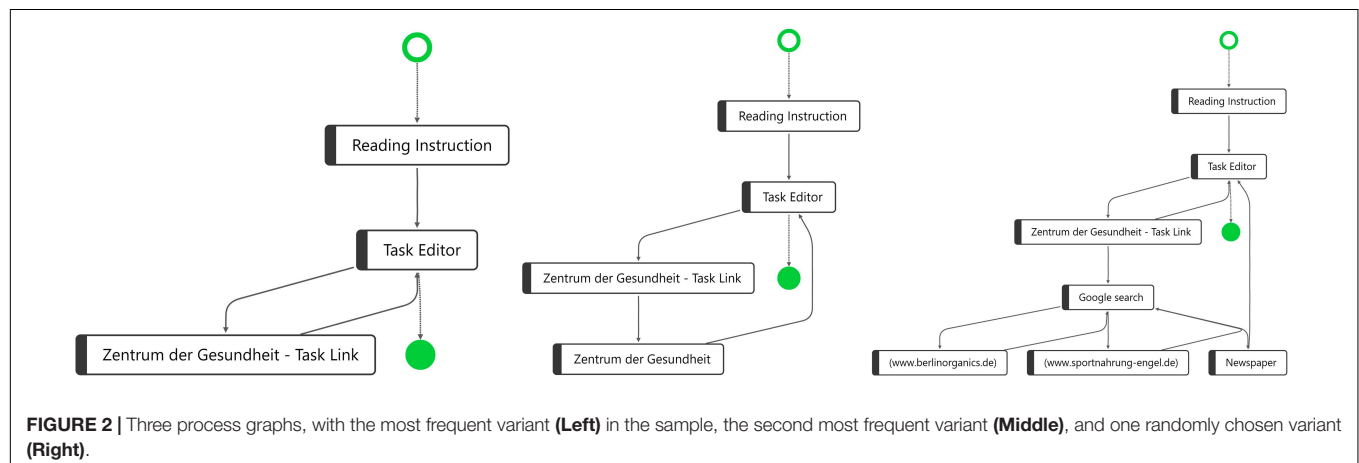
format used in the CORA: (1), it contained the task description as well as the task prompt, and (2) it was the text editor where the students wrote and submitted their responses to the CORA task (see the section "Study Design").

The process sequence (visualized by the arrows in the graph) at the beginning of the CORA task was identical for all students of our sample: "Reading Instruction"? "Task Editor." After the process step of opening the task editor and reading the task prompt, and perhaps already starting to write the response, however, the process graph became less clear. The arrowheads enable a differentiation between whether a sequence flows *from* the "Task Editor" to, for instance, the website linked in the CORA task "Zentrum der Gesundheit—Task Link" or back from any process step *to* the "Task Editor," indicating that the sequences around the "Task Editor" to or from any other process step are manifold.

Another important visualization is the green, filled circle, which indicates the end point of task processing; i.e., the "Task Editor" is the only ending point in the process indicating that



**FIGURE 1** | The process steps of all 32 participants while working on CORA task 1.



**FIGURE 2** | Three process graphs, with the most frequent variant (**Left**) in the sample, the second most frequent variant (**Middle**), and one randomly chosen variant (**Right**).

all students ended the task with writing their response. While performing a Google search appears to be a common behavior among the 32 participants, after the process step “Google search,” a total of 16 different websites were visited by the participants. Based on the visualization of the order of the sequences of the process steps for all students, however, we do not know whether these 16 websites were visited by all 32 students (or only by one), and we cannot yet identify any common patterns in this process graph. Between the starting point and the end point, the visualization of all 32 participants' processes in one process graph does not allow for a precise description of the processes performed and, for instance, observing whether there is a common process variant regarding the complete task processing and where the distinct differences in information processing between the students are. Therefore, in the next step, the individual variants were explored separately, indicating that of 32 students, almost every student navigated the CORA task differently. However, we also managed to find some common patterns (**Figure 2**). In contrast to **Figure 1**, **Figure 2** shows a top-to-bottom visualization of the process graph; i.e., the starting point is now at the top and the end point is at the bottom.

One frequent process variant, the process path for four of the students, is shown on the left side of **Figure 2**. Another frequent variant, applied by seven students, is shown in the middle of **Figure 2**. The difference between the two variants is that in variant 2, the students opened another webpage by “Zentrum der Gesundheit” instead of directly returning back to the “Task Editor.” This behavior can be defined as an “avoidance strategy,” as these seven students did not conduct a web search at all. They completely skipped searching and evaluating additional online information—two processes that are a crucial part of COR according to our construct definition (see the section “Construct Definition and Fundamental Assumptions”). In another variant, two students showed a similar behavior as in variant 2, but they additionally opened a YouTube video directly after visiting “Zentrum der Gesundheit” (see **Supplementary Figure S2**). Similarly, we found a variant in which three students, in addition to the behavior in variant 2, opened Google and entered a search term, but did not proceed to access a website; instead, they returned directly back to the task prompt (see **Supplementary Figure S3**). Following our COR construct definition, this behavior can also be interpreted as an “avoidance strategy,” as it cannot be assumed that these students applied critical reasoning to evaluate the trustworthiness of the task link.

Through this visualization, we identified that for 16 students (i.e., half of the sample), the response processes while working on the CORA task (despite the individual differences) can be classified as an “avoidance strategy.” At the same time, the explorative PM analysis revealed that there are also students who conducted a web search after opening the task prompt, as required by the COR framework. On the right side of **Figure 2**, a randomly chosen process for one single student—is shown. This student who conducted a Google search opened three different websites: (a) [www.berlinorganics.de](http://www.berlinorganics.de), (b) [www.sportnahrung-engel.de](http://www.sportnahrung-engel.de), and (c) a news website. Regarding this process behavior, we can assume that this student researched, verified, and refined

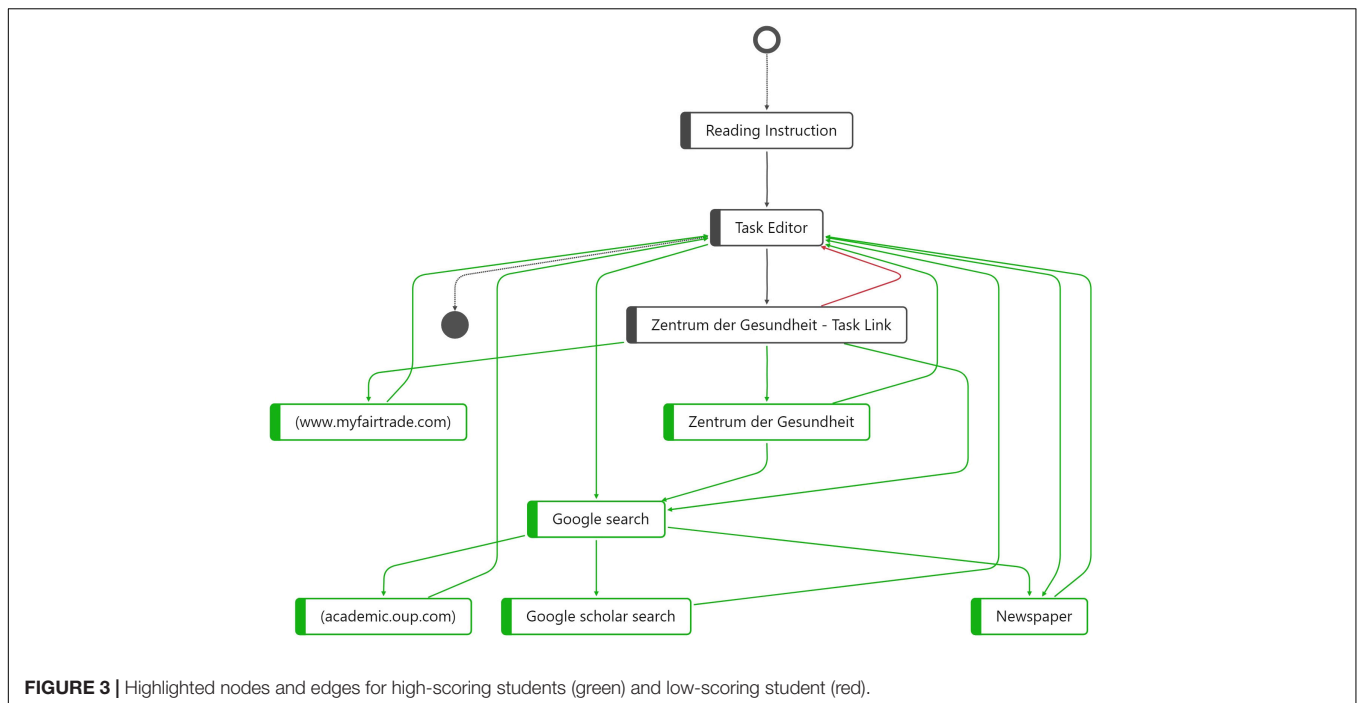
the results before formulating her/his response. Exploring the process graphs for the other 15 students showed a similar process behavior, but with a variation in the number of distinct process steps. Some students opened only a few websites in addition to the website linked in the task prompt; others, like the student shown on the right side in **Figure 2**, performed several process steps and visited many websites. According to our assumption (see the section “Construct Definition and Fundamental Assumptions”), behavior of this kind among students in this group indicates a “strategic information processing strategy.”

In the next step, taking into consideration the results on task performance at the student level, we found with regard to the three variants shown in **Figure 2** that the variants on the left and in the middle show students who employed “avoidance strategies” and mostly achieved scores between 0 and 0.5 on the task, whereas students who used a strategic information processing strategy achieved scores between 1 and 2 on the CORA task.

Overall, **Figures 1** and **2** indicated that despite visualizing either all process variants or only the most frequent variants in the sample, it was still not possible to generate precise descriptions of students' information processing. As demonstrated in the process graph, the search behavior as such cannot explain precisely why students achieved lower or higher scores. Although the number of process steps appears to be related to the task score, we need to include additional indicators of the process performance that goes along with the task performance. While all participants needed a similar amount of time to complete the task, it can be assumed that the distribution of the time spent on the individual process steps is also related to their process behavior and that it is therefore an important indicator of process performance. For instance, when students used the strategic information processing strategy instead of an avoidance strategy, the time they spent on each process step must be shorter compared to students who performed only three or four different activities. Similarly, the number of fixations while looking at different websites can provide first indications as to whether students only scrolled and quickly “skimmed” a webpage or whether they read the contents on a particular page.

To gain more insights into the task behavior and the individual process steps within the process graphs, taking into account the fixations and durations (in seconds) of the separate process steps, two students who had performed their process steps in a very typical order were selected: one from the group of students who used the “avoidance strategy” and one from the group of students who used the “strategic information processing strategy,” respectively. We selected one student with a low score and one with the high score, to investigate how and to what extent the process performance differs between two typical representatives of both groups.

In the following section, the processes of these two students (one with a high and one with a low CORA task score) are described and compared. To this end, using the PM approach to identify similarities, differences and distinct patterns in the students' response processes, the task-solving processes of one low-scoring student (ID 26) and one high-scoring student (ID 16) were first combined in one graph to facilitate a comparison (**Figure 3**).



## Distinct Process Steps

To visualize the *distinct process steps*, in **Figure 3**, a comparison between the two students is shown in a single process graph, which allows us to quantitatively and qualitatively describe the number and types of process steps the students performed while working on the CORA task. The process graph once again shows the differences between students who employ an avoidance strategy and those who employ a strategic information processing strategy: The colors indicate which process steps (nodes) and sequences (edges) were performed only by the low-scoring (red) or only by the high-scoring student (green). Gray nodes and edges describe process steps and edges that were the same for both students, i.e., starting the process of solving the CORA task by reading the instruction and then opening the task editor. The next process step was also identical, with both students opening the link mentioned in the task, “Zentrum der Gesundheit—Task Link,” as in the task, the respondents were asked to follow precisely this hyperlink.

From this task link onward, however, **Figure 3** reveals distinctive differences between high- and low-scoring students. For instance, the low-scoring student opened the task editor again directly after accessing the task link (red edge), which indicates that the low-scoring student did not perform a web search at all—even though it was explicitly mentioned in the task prompt, again indicating the use of an avoidance strategy. In accordance with our COR construct definition, searching, evaluating, and selecting online information was considered an essential facet of COR. Without conducting a web search, it is hardly possible to evaluate the trustworthiness and reliability of online sources presented in the CORA task. This is particularly true for “Zentrum der Gesundheit—Task Link.”

The green nodes and edges show a completely different picture for the high-scoring student. After opening the task editor, the high-scoring student opened another website by “Zentrum der Gesundheit” that was not the same as the one the task link referred to, which can be considered a necessary process step to evaluate the credibility of the website as required in the task prompt. Subsequently, the high-scoring student started a Google search. The aggregated visualization in **Figure 3** shows the similarities and differences in the distinct process steps, but the order of the sequences is hardly visible. Therefore, we additionally evaluated the entire underlying event log for the high-scoring student with ID 16. Regarding the order of the process steps, the data reveal the following sequences:

(1) Reading Instruction → (2) Task Editor → (3) Zentrum der Gesundheit → (4) Task Link → (5) Google search → (6) academic.oup.com → (7) Task Editor → (8) Zentrum der Gesundheit - Task Link → (9) Zentrum der Gesundheit → (10) Google search → (11) Newspaper → (12) Task Editor → (13) Zentrum der Gesundheit → (14) Task Link → (15) Zentrum der Gesundheit → (16) Task Editor → (17) Newspaper → (18) Task Editor → (19) Zentrum der Gesundheit → (20) Task Link → (21) www.myfairtrade.com → (22) Task Editor → (23) Google search → (24) Google scholar search → (25) Task Editor

This sequence of 25 process steps indicates that the first Google search was concluded by accessing the website academic.oup.com. Subsequently, the high-scoring student returned to the task editor. Following this, the task link was opened once again, followed by another webpage of “Zentrum der Gesundheit,” after which the high-scoring student conducted a second Google search. During this second Google search, the

high-scoring student accessed a news site and, subsequently, returned to the task editor once again. Afterward, the high scorer accessed the news site again and then returned to the editor. The high-scoring student then conducted an additional Google search, before finally submitting the solution in the task editor. This task response behavior and these process steps can be interpreted as strategic information processing based on the definition of the COR construct measured here. The high-scoring student gathered additional information online by searching and selecting information to evaluate the reliability of the website “Zentrum der Gesundheit” before writing a response to the task.

Regarding the number of process steps, if we count the gray nodes for the low-scoring student (ID 26), the process graph reveals only three distinct process steps. In contrast, for the high-scoring student (ID 16), as shown by the gray nodes as well as the green nodes, which represent process steps unique to the high scorer, nine distinct process steps were determined. Thus, the response behaviors of low- and high-scoring student differ substantially from one another in terms of the number, kind, and order of distinct process steps.

## Fixations per Distinct Process Step

Next, to visualize the differences in the number of fixations per process step, the combined process graph from **Figure 3** was split into two separate process graphs for the low- and high-scoring student (**Figure 4**). On the left side of **Figure 4**, we see the graph for the low-scoring student, and on the right side, the one for the high-scoring student. The colors of the nodes show the process steps with the highest (red) and lowest (blue) number of fixations in relation to the fixations of each student.

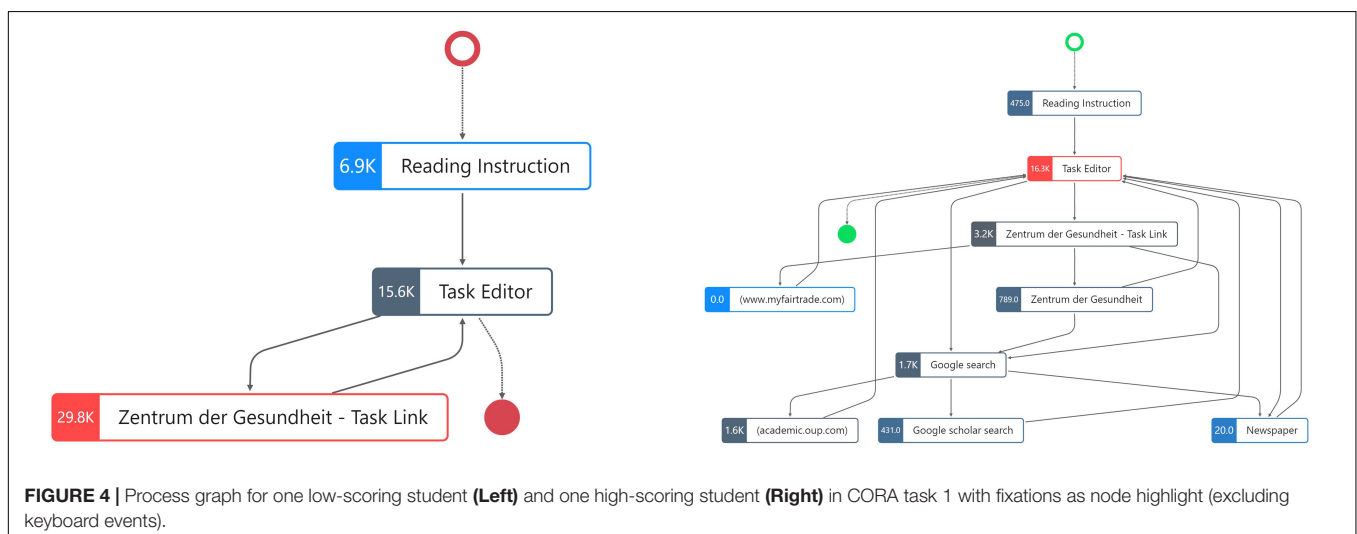
Comparing these two process graphs in **Figure 4**, it becomes evident that the high-scoring student's fixations are distributed among eight of the nine distinct process steps, with 3,200 fixations on the website that was linked in the CORA task (“Zentrum der Gesundheit—Task Link”) and 20 to 1,600 fixations while conducting the web search. In contrast, the low-scoring student's fixations are distributed only among the three distinct

process steps, with the largest number of fixations being recorded while the low-scoring student read the task link website. However, for both students, the similar number of fixations was determined while they were actively using the task editor (15,600 for the low-scoring student and 16,300 for the high-scoring student). This means that they generated the majority of fixations while reading the task prompt and while writing their responses.

For an interpretation of when the fixations occurred within the distinct process steps such as “Task Editor,” for instance, rather while reading the task prompt or rather while writing a response, we conducted an analysis of the videos recorded by the eye tracker (**Table 1**). For instance, when summarizing the fixations for the process steps “Task Editor—Reading Task,” the high-scoring student read the task with 1,784 fixations, whereas the low-scoring student read it with 4,670 fixations. As the videos in combination with the event log data indicate, while the high-scoring student read the task twice (first during the initial access to the task editor and the second time after reading the “Zentrum der Gesundheit—Task Link” website and conducting the first Google search; see left side of **Table 1**), the low-scoring student read the task three times. The first time was also during the initial accessing of the “Task Editor,” the second time also after reading the task link website, and then the third time after he/she started to write his/her response and then returned to read the task prompt again (see right side of **Table 1**). This indicates that the low-scoring student based his/her response (statement on the trustworthiness of the web source) only on reading the task prompt, as well as on the task link website, which can also be considered part of the task prompt, again indicating the use of the avoidance strategy.

## Duration per Distinct Process Step

Regarding the duration of the identified distinct process steps (**Figure 5**), the low-scoring student spent more time on reading the instruction (1.3 min) compared to the high-scoring student who spent only 5 s. The low-scoring student spent the most time on reading the website linked in the task (“Zentrum der





**TABLE 1 |** Process steps and fixations in time related order of the *high* scorer (left) and the *low* scorer (right) extended by the separation in the “Task Editor” by reading and writing.

Process step name	Fixations
Reading Instruction	475
Task Editor— <b>Reading Task</b>	1,144
Zentrum der Gesundheit—Task Link	2,779
Google search	1,317
Academic.oup.com	1,582
Task Editor— <b>Reading Task</b>	640
Task Editor— <b>Writing Response</b>	1,159
Zentrum der Gesundheit—Task Link	413
Zentrum der Gesundheit	692
Google search	13
Newspaper	258
Task Editor— <b>Writing Response</b>	3,800
Zentrum der Gesundheit—Task Link	0
Zentrum der Gesundheit	97
Task Editor Reading	179
Newspaper	20
Task Editor— <b>Writing Response</b>	3,571
Zentrum der Gesundheit—Task Link	0
www.myfairtrade.com	0
Task Editor— <b>Writing Response</b>	3,021
Google search	375
Google scholar search	431
Task Editor— <b>Writing Response</b>	2,740
Reading Instruction	6,896
Task Editor— <b>Reading Task</b>	438
Zentrum der Gesundheit—Task Link	26,014
Task Editor— <b>Reading Task</b>	2,634
Task Editor— <b>Writing Response</b>	2,591
Task Editor— <b>Reading Task</b>	1,598
Task Editor— <b>Writing Response</b>	1,341
Zentrum der Gesundheit—Task Link	2,087
Task Editor— <b>Writing Response</b>	3,076
Zentrum der Gesundheit—Task Link	1,704
Task Editor— <b>Writing Response</b>	3,907

**TABLE 2 |** Relevant process-related variables of the high-scoring student and the low-scoring student.

	Low-scoring student	High-scoring student
Number of process steps (full event log)	42	248
Number of distinct process steps (full event log)	10	17
Number of process steps (process mining event log)	8	22
Number of distinct process steps (process mining event log)	3	9
Average duration per step (s)	20.929	2.484
Average fixations per step	1,379.214	129.133
Score task 1	0	2

students using an avoidance strategy and students using strategic information processing is summarized in **Table 2**.

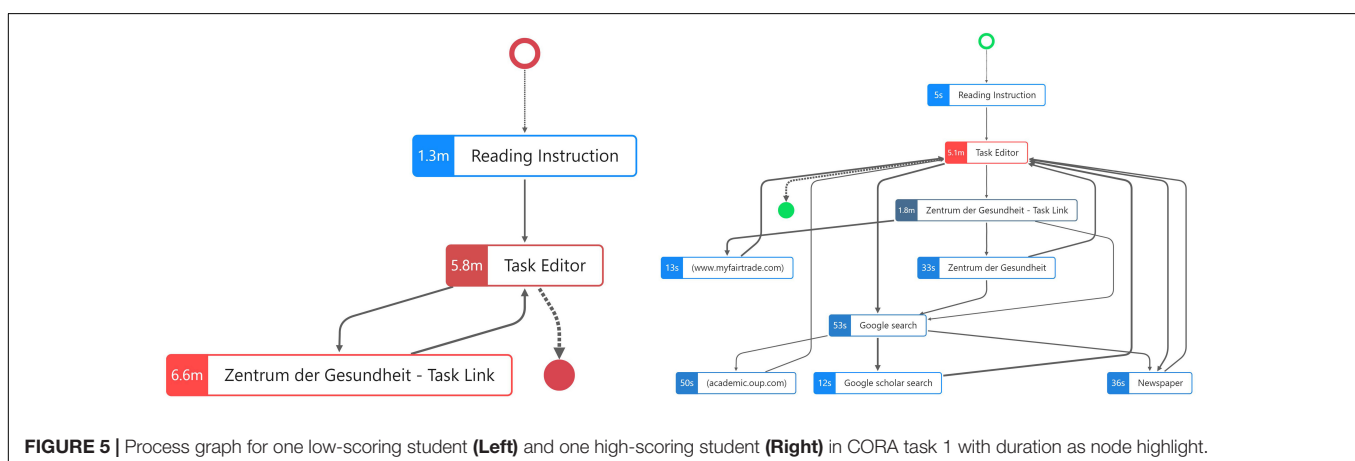
In summary, based on the PM data, it became evident that the response processes, i.e., the way high- and low-scoring students as typical representatives of the two identified response patterns process the CORA task and online information, differ substantially with regard to all process-related variables measured in this study (**Table 2**). These are, in particular, the total number of process steps, the distinct process steps, the number of fixations in these steps, and duration per (distinct) process step. In **Table 2**, we additionally distinguish between the number of (distinct) process steps between the PM data and the full event log. The full event log consists of all events that were omitted for the PM analysis (such as keyboard and mouse events, see the section “Data Transformation”). To examine whether individual differences in these response patterns can be found in the data set using the full event log, the consecutive model-based analysis building on these initial exploratory analyses was conducted (see the section “Latent Class Analyses”).

Gesundheit—Task Link” with 6.6 min), whereas the high-scoring student spent 1.8 min (**Figure 5**).

To conclude the PM analyses, the process performance of the high- and low-scoring student from the two groups of

## Discussion of Process Mining Results and Research Hypotheses

To answer RQ1 (see the section “Research Objectives”), we chose a PM approach to visualize and describe the students’ response



processes while solving the CORA tasks. As demonstrated in the process graphs, however, simply visualizing and exploratively analyzing the raw event log data recorded by the Tobii ET tool did not lead to a satisfying answer to RQ1. The process data were too detailed in terms of time and event granularity. Therefore, the event log data were aggregated so that the PM results led to answers in RQ1. For RQ2, which focused on finding similar and distinct patterns in the students' responses, we were able to identify two different patterns of students' response processes while solving the CORA tasks. In this context, we followed a stepwise approach. First, the process-related variables, including the number of fixations per individual process step, as well as the duration, number, type, and order of the individual process steps, were taken into account. Second, the task performance scores, which were performed by three independent raters, were integrated into the process data and the integrative analyses. In accordance with our COR construct definition and the theoretical assumptions (see the section "Construct Definition and Fundamental Assumptions"), the revealed patterns were defined as "avoidance strategy" vs. "strategic information processing."

Remarkably, students from the first group show both a lower process performance and a lower task performance, in contrast to the latter group who showed a higher performance. More specifically, with regard to all assessed process-related indicators (Table 2), the students from the latter group with the higher scores process online information differently than the students from the first group with the low scores. In particular, high-scoring students process online information more intensively as indicated by a larger number of distinct process steps and total process steps, as well as more efficiently as indicated by a distribution of total fixations in these different steps, and shorter durations for each step, indicating again the use of strategic information processing according to our theoretical assumption. In contrast, the distributions of these process variables for the low-scoring students indicate the much poorer process performance as identified in the process graph for all students in this group.

To summarize the answer to RQ2, the results indicate that students with a higher process performance have significantly higher scores than students with low scores, suggesting a significant relationship between students' process performance and their task performance. Thus, in the subsequent statistical analyses, the following two hypotheses will be tested:

**H1:** Two empirically separable student groups (high vs. low performers) can be identified based on both (i) the students' process-related data, i.e., number and duration of the (distinct) process steps (such as searching for information, writing response) they carry out while processing the CORA tasks and the distribution of total fixations on these different steps, as well as (ii) the students' CORA task performance (test score).

**H2:** Students who had a higher process performance, i.e., more fixations within certain (distinct) process steps and more process steps (i.e., spending less time on single task-related activities), have a higher probability to be a high performer (i.e., a higher task score), while the opposite process performance data indicate a low-performing student.

## Latent Class Analyses

To investigate the two research hypotheses H1 and H2, which are based on the empirical results to RQ2, we conducted an LCA using the same indicators as in the PM analysis and aggregating them at the student level. Before performing the latent group analyses, distributions of all assessed process-related variables ["Number of Process Steps in total"; "Number of Distinct Process Steps"; "Average Duration per Process Step" (seconds); "Average Number of Fixations per Process Step"] and the task scores, which were included in the LCA, were calculated for the entire sample (see **Supplementary Figure S1**).

To test H1 and analyze whether the two distinct groups can be empirically identified among the participants, an LCA was conducted using Stata 16. The LCA classified the students with regard to both their task scores as well as the four further process-related variables concerning their entire task processing, including web search behavior (number of process steps, distinct process steps, processing duration per step, and number of fixations per step).

As the fit indices for the two-class LCA models indicate, log-likelihood, AIC, and BIC are lower in the two-class model than in the one-class model (log-likelihood with  $-553.371$  in the one class model and  $-536.643$  in the two-class model; AIC with  $1,126.742$  in the one-class model and  $1,105.287$  in the two-class model; BIC with  $1,141.399$  in the one-class model and  $1,128.739$  in the two-class model); all class means predicted in this LCA model are significant (Table 3). Two empirically separable groups of students (low and high performers) could be distinguished that differ significantly with regard to the measured process-related variables (process performance) and the task performance (test score).

In summary, low performers perform fewer (total and distinct) process steps, spend more time on each process step, and have more fixations per process step and a lower task score. Thus, H1 can be confirmed, indicating a significant positive relationship between the process performance and task performance in both of the two empirically distinct groups of students in this sample.

To further determine whether the differentiation between high and low performer for all participants in the sample is meaningful (H2), the posterior probability for both classes was predicted for each student based on the two-class model (Table 4).

As shown in Table 4, the probability of belonging to one of the two classes is higher than 70% for almost every student

**TABLE 3 |** Predicted class means for the two groups of high- and low-performing students.

	Group of low performers	Group of high performers
<i>n</i>	21	11
No. of process steps in total	96.480***	159.407***
No. of distinct process steps	12.250***	14.591***
Duration per step (s)	9.033***	5.340***
Fixations per step	518.823***	346.352***
Score item 1	0.309**	0.882***

\*\*\* $p < 0.01$ , \*\* $p < 0.05$ ,  $p < 0.1$ .

**TABLE 4 |** Variables for classification and posterior probabilities for both classes for each participant.

Participant ID	Process steps	Distinct process steps	Average duration per step	Average fixations per step	Score task 1	Posterior probability class 2	Posterior probability class 1
16	248	17	2.483871	129.1331	2	1	0
18	230	14	3.221739	256.9348	1	0.99996	0.00004
22	152	16	5.473684	262.6711	1	0.99954	0.00046
17	124	15	4.33871	341.9677	1	0.99096	0.00904
3	140	16	7.121428	517.8857	0.75	0.98744	0.01256
13	170	14	5.511765	262.3412	0.5	0.98329	0.01671
31	142	14	4.78169	378.4859	0.75	0.95424	0.04576
4	113	15	6.238938	499.6283	1	0.91722	0.08278
32	205	12	5.434146	297.5805	0.25	0.88659	0.11341
6	125	14	6.064	418.616	1	0.87108	0.12892
8	128	14	6.96875	381.4063	0.5	0.59231	0.40769
28	116	14	8.094828	353.9138	0.5	0.31101	0.68899
29	123	13	5.723577	415.0732	0.5	0.26231	0.73769
9	112	14	8.830358	613.9107	0.75	0.14116	0.85884
7	102	12	6.5	407.8235	1	0.05434	0.94566
10	91	13	9.527472	779.033	1.75	0.05142	0.94858
11	124	13	6.354839	545.9597	0	0.03728	0.96272
14	127	13	8.669291	444	0	0.02075	0.97925
19	117	13	8.717949	417.1966	0	0.01276	0.98724
5	118	12	6.779661	352.7373	0	0.00963	0.99037
30	99	11	6.242424	509.9192	1	0.00703	0.99297
15	76	14	8.947369	384.7763	0	0.00611	0.99389
1	97	13	8.329897	444.5464	0	0.00439	0.99561
12	100	11	8.86	429.89	0.5	0.00066	0.99934
23	109	11	7.651376	457.6147	0	0.00046	0.99954
21	87	12	12.2069	407.1035	0.25	0.00015	0.99985
20	68	11	8.073529	303.25	0.25	0.00013	0.99987
2	90	12	10.35556	595.9222	0	0.00011	0.99989
24	80	11	9.1	261.9375	0	0.00009	0.99991
27	72	12	9.861111	768.3333	0	0.00002	0.99998
25	59	12	10.66102	668.678	0	0.00001	0.99999
26	42	10	20.92857	1,379.214	0	0	1

**TABLE 5 |** Number of visited websites for high and low performers.

	High performer ( <i>n</i> = 11)	Low performer ( <i>n</i> = 21)
No. of distinct websites	56	33
Total no. of visited websites	265	316
Average of total no. of visited websites	24	15

(see columns Posterior Probability Class 1 and 2: for class 1 participant ID's 16–8; for class 2 participant ID's 28–26). Each class also comprises at least one student with 100% probability (ID 26 for class 1 and ID 16 for class 2). Thus, *H2* can be confirmed.

As an additional indicator of the “process performance” of the two groups, the number of websites visited by students was analyzed. This article does not aim to analyze individually visited websites; instead, the different types of websites (e.g., newspapers, Wikipedia articles, Twitter blogs, YouTube videos) were evaluated and aggregated into meaningful categories, building distinct websites. In total, the 32 students in the sample

visited 89 distinct websites for task 1. Most of these were visited by only one student (e.g., [ncbi.nlm.nih.gov](http://ncbi.nlm.nih.gov)); only few were visited by almost all students (e.g., Google for conducting the web search). The total number of visited websites and that of distinct websites are shown in **Table 5**, indicating significant differences between the two groups and thus further supporting *H2*.

## DISCUSSION

Explorative PM provided first insights into the response processes involved in students' CORA task solving and dealing with online information, and indicated a relationship between students' process performance and their task performance. Existing studies already revealed differences regarding specific groups (e.g., Zhou and Ren, 2016), for instance, researchers vs. students (Wineburg and McGrew, 2017) or experts vs. novices (Brand-Gruwel et al., 2009). Based on prior research, this article distinguished groups according to a performance criterion, i.e., “process performance” and “task performance” (the CORA task score), and therefore exposes further possible distinct characteristics of response

processes when dealing with online information that lead to better performance ("high performer") or worse performance ("low performer") in the task on COR.

Using PM analyses as an approach to visualizing and precisely describing the students' response processes while solving the CORA tasks (RQ1), two distinct process patterns were identified among the 32 participants. In RQ2, we focused on identifying commonalities and differences in these patterns. The two identified patterns were defined as "avoidance" vs. "strategic information processing" according to our COR construct definition and the underlying theoretical framework (see the section "Theoretical Framework"). When selecting two typical representatives from the both groups, the response process of a high-scoring student (i.e., higher test score) was characterized by a higher process performance (i.e., more total process steps as well as more distinct process steps, while at the same time he/she spent less time on single process steps, e.g., on a specific webpage). Subsequently, the student from the high-scoring group distributed his/her time as well as fixations according to his/her wider range of process steps, which resulted in shorter durations per process step. In contrast, the student from the low-scoring group (i.e., lower test score) showed a lower process performance, i.e., spent most of his/her time on only one website, which led to many fixations, all of which, however, were focused on this one specific distinct process step (i.e., visiting the webpage linked in the task).

The student from the high-scoring group started writing his/her response only after conducting a web search, indicating that they weighed up different pieces of information and options, which may relate to a more analytical response process (Chen and Chaiken, 1999; Evans and Stanovich, 2013). The student from the low-scoring group started writing an answer immediately after visiting the webpage linked in the task, which indicates a tendency toward cognitive heuristics (Walther and Burkell, 2002; Metzger, 2007) and a solution behavior characterized by using an avoidance strategy, i.e., lower cognitive effort by judging a website without searching and evaluating additional online information (as was required in the CORA task). However, these are only initial indications for a rather heuristic or strategic task processing behavior, supporting our theoretical assumptions. To be able to make more accurate statements about the actual (meta-)cognitive heuristics and information processing strategies that lead to higher vs. lower process and performance on the CORA task, a comprehensive analysis of eye movements, particularly within previously defined AOIs, would be required.

Based on the results of PM determined when answering RQ1 and the empirical findings determined when answering RQ2, two research hypotheses were formulated, and further statistical analyses were conducted to examine the response process behavior patterns in the sample. To test H1, which required us to determine each student's probability of belonging to one of the defined subgroups of high- or low-performing students, both the process-related variables and the test score were included in an LCA. First, the sample could be divided into two distinct groups of students ("low performers" and "high performers") by means of an LCA; here, too, the groups differed significantly in terms of task scores, as well as the process variables that

had already been identified as relevant in PM, supporting two distinct process patterns: "avoidance" vs. "strategic information processing." The LCA indicated that all of the 32 students belong to one of the two groups with a statistically high probability. As a result, H1 cannot be rejected. The results of the LCA also support that H2 cannot be rejected, as the group of high-performing students met the postulated assumptions [higher task performance (score) and higher process performance, see the section Latent Class Analyses]. However, the generalizability of these response process profiles for the overall student population requires further investigation in replication studies, including a random sampling of participants (see the sections "Discussion" and "Limitations"). It would be of particular interest to analyze in a longitudinal design how and to what extent these online information patterns may be developed over a course of study in higher education. The identified patterns in the response process behavior of students solving the CORA task should also be investigated in an experimental research design that explicitly triggers different information problem-solving strategies and (distinct) process steps (e.g., web search, evaluation of different websites) in an experimental group and with different stimuli.

In terms of contributions to the research field, our results are in line with findings from existing ET studies on web search behavior. First, as already revealed in many studies, most students have not yet developed a sufficient level of the abilities and skills (such as selecting and evaluating online information) (e.g., Walraven et al., 2009; Wineburg et al., 2018; McGrew et al., 2019) that constitute the construct of COR. Second, on the basis of ET and web search log data, we identified two groups of students who differed significantly in terms of both their test performance and all assessed response process indicators such as process steps, fixations, and duration (i.e., process performance). This finding is also in line with previous research that determined such different profiles of evaluation behavior with regard to online sources (e.g., Brand-Gruwel et al., 2017; List and Alexander, 2017).

More specifically, and also in line with previous research (e.g., Zhou and Ren, 2016), in this study, we identified substantial differences between high- and low-performing students in relation to the number, kind, and order of the distinct process steps, in particular during a web search as well as with regard to both duration and distribution of fixations per distinct process step. Using PM, we identified two very different patterns in the response processes and in particular the online search behavior of two groups of students with higher and lower CORA task scores, which were confirmed by means of an LCA. The significant differences in terms of both duration and fixations per individual step also suggest differences with regard to visual attention and eye-movement patterns between the two student groups. For instance, PM analyses indicate that students from the high-scoring group have a significantly larger number of (Google) search queries and processing activities with regard to the selected websites (reading and selecting information), indicating a strategic processing profile. In contrast, students from the low-scoring group showed only limited or even no search activity, indicating an avoidance processing profile. Combined with results regarding fixations and durations, which can be interpreted as indicators of processing new information



(Holmqvist et al., 2011), these findings also initially indicate differences regarding the (meta-)cognitive activity of both student groups that need further in-depth investigation (see the section “Future Perspectives”).

## Limitations

Even though the PM analysis provided many conclusive insights into students' task response behavior and online information processing, indicating two distinct profiles that were confirmed through LCA, it also has certain limitations. Because of the purposeful sampling based on certain defined selection criteria (Palinkas et al., 2015) but not random sampling, the representativeness of the sample is questionable, as it might affect the students' response behavior. Thus, the generalizability of the results is limited. Moreover, as participation in this study was not mandatory and the CORA did not have any positive or negative effects on the students' regular study progress (e.g., in the form of grades), i.e., it was a low-stakes test, the students' motivation—which can strongly impact their task scores—is questionable.

Because we followed a person-oriented approach, variable-oriented analyses (such as regression models) were not conducted in this study. The aim was to identify subgroups and not to explain potential differences with further external criteria apart from the construct-relevant process variables. Although other contextual factors (such as the course of study and the semester) were surveyed, they were not included in the analyses so far because of the already high complexity of the analysis design. Similarly, no control was carried out on the measured personal factors (such as intelligence, expertise or previous knowledge on the topic of the CORA tasks). However, as these variables play a significant role in the handling of online information (Willoughby et al., 2009; Gadiraju et al., 2018), it cannot be ruled out that effects biased the results on COR (for implications, see the section “Latent Class Analyses”).

By using ET methodology, time-related, accurate, and exact data about the students' solution processes were collected; however, typical ET measures were only used in a highly aggregated form for the PM analyses. High-resolution ET metrics that could have provided more detailed insights into the students' eye movements, and therefore their gaze behavior, were not included in the PM analysis. On the one hand, this article did not focus on analyzing and interpreting ET data in terms of metrics such as fixation duration/dwell times and saccades to make inferences on visual attention and eye movements in relation to defined AOIs due to the extremely high complexity of this event log data set. On the other hand, the determination of AOI is always subject to substantial errors (Orquin et al., 2015), as it is influenced by the test designers' opinions. Thus, in follow-up studies, an automatic determination of AOIs on the level of complete webpages will be implemented (as suggested by Hienert et al., 2019), so that an often arbitrary AOI determination cannot negatively influence the analyses.

Furthermore, as process steps were primarily analyzed quantitatively (e.g., number of total and distinct process steps), there were few qualitative differentiations between the distinct process steps, in particular regarding the qualitative characteristics of the accessed websites (such as difficulty,

complexity, etc.). Nevertheless, such qualitative aspects were considered in the rating of the students' written responses in the CORA, as one criterion for the scoring was the quality (e.g., scientific or non-scientific) of the URLs provided in the students' responses.

Overall, the described findings emphasize the high importance of examining the processes involved in students' ability to comprehensively deal with online information, as the level of ability to critically reflect on online information seems to be rather low among students (reflected in both the process performance and the task score distribution in this study). In this study, the students showed either an avoidance strategy or a strategic processing strategy. Although the latter led to a significantly higher CORA task performance, this strategy does not necessarily cover all main processes of COR according to our construct definition. In fact, as the distributions of the task scores indicated (see **Supplementary Figure S1**), only two students from this group achieved the maximum score of two points. Hence, further research is required to understand and explain processes that lead to this low COR skill level among many students and, consequently, to deduce how a critically reflective handling of online information could be promoted in higher education.

## Future Perspectives

In the next research step, a more in-depth qualitative analysis of the identified groups and response process patterns would be required to build a solid basis for the formulation of hypotheses with regard to theoretically expected gaze patterns (Pifarré et al., 2018). For further studies, therefore, the students' (meta-)cognitive processes when dealing with online information should be investigated in more depth and in a longitudinal design. Experimental between-subject design, for instance, regarding the search behavior with and without prior instruction, could be implemented here. Eye-movement diagnostics should also be brought more into focus to enable more specific descriptions of students' visual attention and indications of cognitive load.

The person-oriented methodological approach applied here should be expanded and combined with a variable-oriented approach to take into account any contextual or personal factors that may influence gaze behavior (such as complexity of presented information and general cognitive ability), as many studies indicate (Horstmann et al., 2009; Raney et al., 2014). In subsequent studies, therefore, potential explanatory variables of the various processes must also be included in a variable-oriented approach to test for discriminant validity as well. For instance, the question arises as to whether different response processes are to be expected among students with a certain academic subject, certain educational indicators, or in different familial and social contexts. In particular, the possible effects of different domains and possible curricular and/or instructional specifics in the study programs that may impact students' COR should also be considered in follow-up research to test for instructional validity (Pellegrino et al., 2016).

Qualitative studies, for instance, rating the different websites used by each student, need to be conducted as well to enable further qualitative analyses of the sequences of information

processing. Because of the open nature of the CORA, the primary question is how to find the best way to deal with the task, which is particularly useful for instruction and teaching in higher education. For example, the question is whether a certain process step or a sequence of process steps is decisive for a successful solution of the CORA task. In the past, there have been frequent studies on backward or forward reasoning, which find ideal task solvers in various test environments (Norman et al., 1999). A similar question needs to be researched on the basis of the available findings: can ideal solution patterns be found in (partial) sequences, and can instructional settings be developed based on these patterns, for instance, by indicating to the learner that a text should be read first, and the web search carried out subsequently and with a certain term specification? Additional explanations by the test takers, for instance, through concurrent verbal protocols, could provide further insights into the students' causal decision contexts (Leighton and Gierl, 2007) and be combined and evaluated in parallel to the time-sequential recordings of the ET data (Maddox et al., 2018) and predefined process steps.

In this study, we used the purposeful sampling method (Palinkas et al., 2015) and focus on specific characteristics in our sampling to identify differences in the construct and the students' processes while responding to the CORA tasks. Using this sampling approach, we identified distal indicators to analyze the breadth of possible response processes. However, the criteria for purposeful sampling can be expanded in future studies, to, e.g., include other indicators such as intelligence or domain-specific prior knowledge. The effects of these kinds of additional indicators on COR processes need to be sufficiently analyzed in follow-up research. This should include adding further domains in replication studies.

Overall, it would be of great value for further experimental and longitudinal studies to consider the students' handling of online information in a differentiated way with regard to additional contextual factors and personal factors (at different levels of analysis) to control for intercorrelations, in an integrated person-variable-oriented approach (Rauthmann and Sherman, 2016).

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## REFERENCES

- Alexander, P. A. (2003). The development of expertise. The journey from acclimation to proficiency. *Educ. Res.* 32, 10–14. doi: 10.3102/0013189X032008010
- American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the State Officer for Data Protection and Freedom of Information Rhineland-Palatinate. The participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

SS co-developed the assessment, conducted the analyses, and co-wrote the manuscript. OZ-T provided the idea for the study, co-developed the assessment, supervised the analyses, and co-wrote the manuscript. JR was involved in the data collection and in preparing and reviewing the manuscript. VK and MW were involved in the data collection, and supported the analyses. A-KB co-implemented the ET test environment, and was involved in the data collection, and in preparing the manuscript, and supported the analyses. SB co-developed the ET test environment and was involved in the analyses, and in preparing the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was part of an RMU project, which was funded by the RMU fund.

## ACKNOWLEDGMENTS

We would like to thank the editor and the reviewers who provided constructive feedback and helpful guidance in the revision of this manuscript. We would like to thank all students from the Medical Faculty of Goethe University Frankfurt, and from the Faculty of Law and Economics at Johannes Gutenberg University who participated in this study.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.576273/full#supplementary-material>

- Anderson, R. C., Spiro, R. J., and Anderson, M. C. (1978). Schemata as scaffolding for the representation of information in connected discourse. *Am. Educ. Res. J.* 15, 433–440. doi: 10.3102/00028312015003433
- Anmarkrud, Ø., Bråten, I., and Strømsø, H. I. (2014). Multiple-documents literacy: strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learn. Individ. Differ.* 30, 64–76. doi: 10.1016/j.lindif.2013.01.007
- Bergman, L. R., and El-Khoury, B. M. (2001). Developmental processes and the modern typological perspective. *Eur. Psychol.* 6, 177–186. doi: 10.1027//1016-9040.6.3.177

- Braasch, J. L. G., and Bråten, I. (2017). The discrepancy-induced source comprehension (D-ISC) model: basic assumptions and preliminary evidence. *Educ. Psychol.* 52, 167–181. doi: 10.1080/00461520.2017.1323219
- Braasch, J. L. G., Bråten, I., and McCrudden, M. T. (2018). *Handbook of Multiple Source Use*. New York, NY: Routledge.
- Brand-Gruwel, S., Kammerer, Y., van Meeuwen, L., and van Gog, T. (2017). Source evaluation of domain experts and novices during Web search. *J. Comput. Assist. Learn.* 33, 234–251. doi: 10.1111/jcal.12162
- Brand-Gruwel, S., and Stadler, M. (2011). Solving information-based problems: evaluating sources and information. *Learn. Instr.* 21, 175–179. doi: 10.1016/j.learninstruc.2010.02.008
- Brand-Gruwel, S., Wopereis, I., and Vermetten, Y. (2005). Information problem solving by experts and novices: analysis of a complex cognitive skill. *Comput. Hum. Behav.* 21, 487–508. doi: 10.1016/j.chb.2004.10.005
- Brand-Gruwel, S., Wopereis, I., and Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Comput. Educ.* 53, 1207–1217. doi: 10.1016/j.compedu.2009.06.004
- Brooks, D. C. (2016). *ECAR Study of Undergraduate Students and Information Technology*, 2016. Louisville, CO: ECAR.
- Chen, S., and Chaiken, S. (1999). “The heuristic-systematic model in its broader context,” in *Dual-Process Theories in Social Psychology*, eds S. Chaiken, and Y. Trope, (New York, NY: Guilford Press), 73–96.
- Ciampaglia, G. L. (2018). “The digital misinformation pipeline,” in *Positive Learning in the Age of Information*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel, (Wiesbaden: Springer), 413–421. doi: 10.1007/978-3-658-19567-0\_25
- Collins-Thompson, K., Soo Young, R., Haynes, C. C., and Syed, R. (2016). “Assessing learning outcomes in web search: a comparison of tasks and query strategies,” in *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, New York, NY, 163–172. doi: 10.1145/2854946.2854972
- Cook, D. A., and Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *Am. J. Med.* 119, 166.e7–166.e16. doi: 10.1016/j.amjmed.2005.10.036
- Cyr, D., and Head, M. (2013). The impact of task framing and viewing time on user website perceptions and viewing behavior. *Int. J. Hum. Comput. Stud.* 7, 1089–1102. doi: 10.1016/j.ijhcs.2013.08.009
- De Neys, W. D. (2006). Dual processing in reasoning: two systems but one reasoner. *Psychol. Sci.* 17, 428–433. doi: 10.1111/j.1467-9280.2006.01723.x
- Duchowski, A. T. (2007). *Eye Tracking Methodology: Theory and Practice*. Berlin: Springer. doi: 10.1007/978-3-319-57883-5
- Ercikan, K., and Pellegrino, J. W. (2017). “Validation of score meaning using examinee response processes for the next generation of assessments,” in *Validation of Score Meaning for the Next Generation of Assessments*, eds K. Ercikan, and J. W. Pellegrino, (New York, NY: Routledge), 1–8. doi: 10.4324/9781315708591-1
- Evans, J. S. B. (1984). Heuristic and analytic processes in reasoning. *Br. J. Psychol.* 75, 451–468. doi: 10.1111/j.2044-8295.1984.tb01915.x
- Evans, J. S. B. (1989). *Bias in Human Reasoning: Causes and Consequences*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends Cogn. Sci.* 7, 454–459. doi: 10.1016/j.tics.2003.08.012
- Evans, J. S. B. (2006). The heuristic-analytic theory of reasoning: extension and evaluation. *Psychon. Bull. Rev.* 13, 378–395. doi: 10.3758/bf03193858
- Evans, J. S. B., and Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/174569161246068
- Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. *Frontline Learn. Res.* 2, 28–45.
- Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., and Tauber, E. R. (2003). “How do users evaluate the credibility of Web sites?” in *Proceedings of the 2003 Conference on Designing for User Experiences - DUX '03*, (New York, NY: ACM Press), 1–15. doi: 10.1145/997078.997097
- Gadiraju, U., Yu, R., Dietze, S., and Holtz, P. (2018). “Analyzing knowledge gain of users in informational search sessions on the web,” in *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval*, New Brunswick, NJ, 2–11. doi: 10.1145/3176349.3176381
- Gerjets, P., Kammerer, Y., and Werner, B. (2011). Measuring spontaneous and instructed evaluation processes during Web search: integrating concurrent thinking-aloud protocols and eye-tracking data. *Learn. Instr.* 21, 220–231. doi: 10.1016/j.learninstruc.2010.02.005
- Gigerenzer, G. (2008). Why heuristics work. *Perspect. Psychol. Sci.* 3, 20–29. doi: 10.1111/j.1745-6916.2008.00058.x
- Goldhammer, F., and Zehner, F. (2017). What to make of and how to interpret process data. *Measurement* 15, 128–132. doi: 10.1080/15366367.2017.1411651
- Gollwitzer, M. (2012). “Latent-class-analysis,” in *Testtheorie und Fragebogenkonstruktion*, eds H. Moosbrugger, and A. Kelava (Springer, Berlin: Springer), doi: 10.1007/978-3-642-20072-4\_12
- Gronchi, G., and Giovannelli, F. (2018). Dual process theory of thought and default mode network: a possible neural foundation of fast thinking. *Front. Psychol.* 9:1237. doi: 10.3389/fpsyg.2018.01237
- Hargittai, E., Fullerton, L., Menchen-Trevino, E., and Thomas, K. Y. (2010). Trust online: young adults' evaluation of web content. *Int. J. Commun.* 4, 468–494.
- Hienert, D., Kern, D., Mitsui, M., Shah, C., and Belkin, N. J. (2019). “Reading protocol: understanding what has been read in interactive information retrieval tasks,” in *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, Glasgow, 73–81.
- Hilbert, M. (2012). Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychol. Bull.* 138, 211–237. doi: 10.1037/a0025940
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and van de Weijer, J. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford: Oxford University Press.
- Hoppe, A., Holtz, P., Kammerer, Y., Yu, R., Dietze, S., and Ewerth, R. (2018). “Current challenges for studying search as learning processes,” in *Proceedings of Learning and Education with Web Data*, Amsterdam.
- Horstmann, G., Becker, S. I., and Grubert, A. (2019). Dwelling on simple stimuli in visual search. *Atten. Percept. Psychophys.* 82, 607–625. doi: 10.3758/s13414-019-01872-8
- Horstmann, N., Ahlgrim, A., and Glockner, A. (2009). How distinct are intuition and deliberation? An eye tracking analysis of instruction-induced decision modes. *Judgm. Decis. Mak.* 4, 335–354.
- Kahneman, D., Slovic, S. P., Slovic, P., and Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge university press.
- Kao, G. Y.-M., Lei, P.-L., and Sun, C.-T. (2008). Thinking style impacts on web search strategies. *Comput. Hum. Behav.* 24, 1330–1341. doi: 10.1016/j.chb.2007.07.009
- Leighton, J. P. (2017). *Using Think-Aloud Interviews and Cognitive Labs in Educational Research*. New York, NY: Oxford University Press.
- Leighton, J. P., and Gierl, M. J. (2007). “Verbal reports as data for cognitive diagnostic assessment,” in *Cognitive Diagnostic Assessment for Education: Theory and Applications*, eds J. P. Leighton, and M. J. Gierl, (Cambridge: Cambridge University Press), 146–172. doi: 10.1017/CBO9780511611186.006
- Li, Z., Banerjee, J., and Zumbo, B. D. (2017). “Response time data as validity evidence: has it lived up to its promise and, if not, what would it take to do so,” in *Understanding and Investigating Response Processes in Validation Research*, eds B. D. Zumbo, and A. M. Hubley, (Cham: Springer International Publishing), 159–178. doi: 10.1007/978-3-319-56129-5\_9
- List, A., and Alexander, P. A. (2017). Analyzing and integrating models of multiple text comprehension. *Educ. Psychol.* 52, 143–147. doi: 10.1080/00461520.2017.1328309
- List, A., and Alexander, P. A. (2019). Toward an integrated framework of multiple text use. *Educ. Psychol.* 54, 20–39. doi: 10.1080/00461520.2018.1505514
- Lucassen, T., and Schraagen, J. M. (2011). Factual accuracy and trust in information: the role of expertise. *J. Am. Soc. Inf. Sci. Technol.* 62, 1232–1242. doi: 10.1002/asi.21545
- Maddox, B., Bayliss, A. P., Fleming, P., Engelhardt, P. E., Edwards, S. G., and Borgonovi, F. (2018). Observing response processes with eye tracking in international large-scale assessments: evidence from the OECD PIAAC assessment. *Eur. J. Psychol. Educ.* 33, 543–558. doi: 10.1007/s10212-018-0380-2
- Masyn, K. E. (2013). “Latent class analysis and finite mixture modeling,” in *Oxford Library of Psychology. The Oxford Handbook of Quantitative Methods: Statistical Analysis*, ed. T. D. Little (Oxford University Press), 551–611.



- Maurer, M., Schemer, C., Zlatkin-Troitschanskaia, O., and Jitmirski, J. (2020). "Positive and negative media effects on university students' learning: preliminary findings and a research program," in *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)*, ed. O. Zlatkin-Troitschanskaia, (Cham: Springer), 109–119. doi: 10.1007/978-3-030-26578-6\_8
- Mayer, R. E. (2002). "Multimedia learning," in *Psychology of Learning and Motivation. Advances in Research and Theory*, Vol. 41, ed. B. H. Ross, (Amsterdam: Elsevier), 85–139. doi: 10.1017/cbo9780511811678.006
- McGrew, S., Breakstone, J., Ortega, T., Smith, M., and Wineburg, S. (2018). Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory Res. Soc. Educ.* 46, 165–193. doi: 10.1080/00933104.2017.1416320
- McGrew, S., Smith, M., Breakstone, J., Ortega, T., and Wineburg, S. (2019). Improving university students' web savvy: an intervention study. *Br. J. Educ. Psychol.* 89, 485–500. doi: 10.1111/bjep.12279
- Metzger, M. J. (2007). Making sense of credibility on the web: models for evaluating online information and recommendations for future research Miriam. *J. Am. Soc. Inf. Sci. Technol.* 58, 2078–2091. doi: 10.1002/asi.20672
- Metzger, M. J., and Flanagin, A. J. (2015). Credibility and trust of information in online environments: the use of cognitive heuristics. *J. Pragmat.* 59, 210–220. doi: 10.1016/j.pragma.2013.07.012
- Metzger, M. J., Flanagin, A. J., and Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *J. Commun.* 60, 413–439. doi: 10.1111/j.1460-2466.2010.01488.x
- Molerov, D., Zlatkin-Troitschanskaia, O., and Schmidt, S. (2019). *Adapting the Civic Online Reasoning Assessment for Cross-National Use*. Toronto: American Education Research Association.
- Norman, G. R., Brooks, L. R., Colle, C. L., and Hatala, R. M. (1999). The benefit of diagnostic hypotheses in clinical reasoning: experimental study of an instructional intervention for forward and backward reasoning. *Cogn. Instr.* 17, 433–448. doi: 10.1207/S1532690XCI1704\_3
- Oranje, A., Gorin, J., Jia, Y., and Kerr, D. (2017). "Collecting, analyzing, and interpreting response time, eye-tracking, and log data," in *Validation of Score Meaning for the Next Generation of Assessments. The Use of Response Processes*, eds K. Ercikan, and J. W. Pellegrino, (New York, NY: Routledge), 39–51. doi: 10.4324/9781315708591-4
- Orquin, J., and Loose, S. (2013). Attention and choice: a review on eye movements in decision making. *Acta Psychol.* 144, 190–206. doi: 10.1016/j.actpsy.2013.06.003
- Orquin, J. L., Ashby, N. J. S., and Clarke, A. D. F. (2015). Areas of interest as a signal detection problem in behavioral eye-tracking research. *J. Behav. Decis. Mak.* 29, 103–115. doi: 10.1002/bdm.1867
- Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., and Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Adm. Policy Ment. Health* 42, 533–544. doi: 10.1007/s10488-013-0528-y
- Pellegrino, J. W., DiBello, L. V., and Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educ. Psychol.* 51, 1–23. doi: 10.1080/00461520.2016.1145550
- Pifarré, M., Jarodzka, H. M., Brand Gruwel, S., and Argelagos, E. (2018). Unpacking cognitive skills engaged in web-search: how can log files, eye movements, and cued-retrospective reports help? An in-depth qualitative case study. *Int. J. Innov. Learn.* 24, 152–175. doi: 10.1504/ijil.2018.10014361
- Raney, G. E., Campbell, S. J., and Bovee, J. C. (2014). Using eye movements to evaluate the cognitive processes involved in text comprehension. *J. Vis. Exp.* 83:e50780. doi: 10.3791/50780
- Rauthmann, J. F., and Sherman, R. A. (2016). Situation change: stability and change of situation variables between and within persons. *Front. Psychol.* 6:1938. doi: 10.3389/fpsyg.2015.01938
- Reitbauer, M. (2008). Keep an eye on information processing: eye tracking evidence for the influence of hypertext structures on navigational behaviour and textual complexity. *LSP Prof. Commun.* 8, 15–38.
- Rost, J., and Eid, M. (2009). "Mischverteilungsmodelle," in *Enzyklopädie der Psychologie*, ed. H. Holling, (Göttingen: Hogrefe), 483–524.
- Russell, L. B., and Huber, U. (2017). "Some thoughts on gathering response process validity evidence: in the context in online measurement and digital revolution," in *Understanding and Investigating Response Processes in Validation Research*, eds B. D. Zumbo, and A. M. Hubley, (Cham: Springer International Publishing), 229–250. doi: 10.1007/978-3-319-56129-5\_13
- Shaw, V. F. (1996). The cognitive processes in informal reasoning. *Think. & Reason.* 2, 51–80. doi: 10.1080/135467896394564
- Slovan, S. A. (1996). The empirical case for two systems of reasoning. *Psychol. Bull.* 119, 3–22. doi: 10.1037/0033-2909.119.1.3
- Stanovich, K. E. (2003). "The fundamental computational biases of human cognition: heuristics that (sometimes) impair decision making and problem solving," in *The Psychology of Problem Solving*, eds J. E. Davidson, and R. J. Sternberg, (Cambridge: Cambridge University Press), 291–342. doi: 10.1017/CBO9780511615771.011
- Stanovich, K. E. (2012). "On the distinction between rationality and intelligence: implications for understanding individual differences in reasoning," in *The Oxford Handbook of Thinking and Reasoning*, eds K. Holyoak, and R. Morrison, (New York, NY: Oxford University Press), 343–365.
- Stanovich, K. E. (2016). The comprehensive assessment of rational thinking. *Educ. Psychol.* 51, 23–34. doi: 10.1080/00461520.2015.1125787
- Sterba, S. K., and Bauer, D. J. (2010). Matching method with theory in person-oriented developmental psychopathology research. *Dev. Psychopathol.* 22, 239–254. doi: 10.1017/s0954579410000015
- Toplak, M. E., Liu, E., MacPherson, R., Toneatto, T., and Stanovich, K. E. (2007). The reasoning skills and thinking dispositions of problem gamblers: a dual process taxonomy. *J. Behav. Decis. Mak.* 20, 103–124. doi: 10.1002/bdm.544
- Tóth, K., Rölke, H., Goldhammer, F., and Barkow, I. (2017). "Educational process mining: new possibilities for understanding students' problem-solving skills," in *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*, eds B. Csapó, and J. Funke, (Paris: OECD), 193–210. doi: 10.1787/9789264273955-14-en
- von Eye, A. (2006). "Variablen- und personenorientierte Forschung," in *Veränderungsmessung und Längsschnittstudien in der Empirischen Erziehungswissenschaft*, eds A. Ittel, and H. Merckens, (Wiesbaden: Verlag für Sozialwissenschaften), 9–26. doi: 10.1007/978-3-531-90502-0\_2
- Walraven, A., Brand-Gruwel, S., and Boshuizen, H. P. A. (2009). How students evaluate information and sources when searching the World Wide Web for information. *Comput. Educ.* 52, 234–246. doi: 10.1016/j.compedu.2008.08.003
- Walthen, C. N., and Burkell, J. (2002). Believe it or not: factors influencing credibility on the web. *J. Am. Soc. Inf. Sci. Technol.* 53, 134–144. doi: 10.1002/asi.10016
- Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., and Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. *Am. Educ. Res. J.* 46, 1060–1106. doi: 10.3102/0002831209333183
- Willoughby, T., Anderson, A. S., Wood, E., Mueller, J., and Ross, C. (2009). Fast searching for information on the internet to use in a learning context: the impact of domain knowledge. *Comput. Educ.* 52, 640–648. doi: 10.1016/j.compedu.2008.11.009
- Wineburg, S., Breakstone, J., McGrew, S., and Ortega, T. (2018). "Why Google can't save us: the challenges of our post-Gutenberg moment," in *Positive Learning in the Age of Information*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel, (Wiesbaden: Springer), 221–228. doi: 10.1007/978-3-658-19567-0\_13
- Wineburg, S., and McGrew, S. (2017). *Lateral Reading: Reading Less and Learning more when Evaluating Digital Information*. Stanford History Education Group Working Paper No. 2017-A1. Stanford, CA: Stanford Graduate School of Education. doi: 10.2139/ssrn.3048994
- Wineburg, S., McGrew, S., Breakstone, J., and Ortega, T. (2016). *Evaluating Information: The Cornerstone of Civic Online Reasoning*. New York, NY: Stanford Digital Repository.
- Wu, D., Huang, H., Liu, N., and Miao, D. (2019). Information processing under high and low distractions using eye tracking. *Cogn. Process.* 20, 11–18. doi: 10.1007/s10339-018-0876-3
- Yu, R., Gadiraju, U., Holtz, P., Rokicki, M., Kemkes, P., and Dietze, S. (2018). *Predicting User Knowledge Gain in Informational Search Sessions. ACM Sigir*. Available at: <http://arxiv.org/pdf/1805.00823v1> (accessed May 16, 2020).
- Zhang, X., Cole, M., and Belkin, N. (2011). "Predicting users' domain knowledge from search behaviors," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*,



- eds W.-Y. Ma, J.-Y. Nie, R. Baeza-Yates, T.-S. Chua, and W. B. Croft, (New York, NY: ACM Press), 1225–1226.
- Zhou, M., and Ren, J. (2016). "Use of cognitive and metacognitive strategies in online search: an eye-tracking study," in *Proceedings of the International Conferences on Internet Technologies & Society (ITS), Education Technologies (ICEduTECH), and Sustainability, Technology and Education (STE)*, eds P. Kommers, I. Tomayess, I. Theodora, E. McKay, and P. Isias, (Melbourne: IADIS Press), 347–349.
- Zumbo, B. D., and Hubley, A. M. (2017). *Understanding and Investigating Response Processes in Validation Research*. Cham: Springer International Publishing.
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Schmidt, Zlatkin-Troitschanskaia, Roeper, Klose, Weber, Bültmann and Brückner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Constraints and Affordances of Online Engagement With Scientific Information—A Literature Review

Friederike Hendriks<sup>1\*</sup>, Elisabeth Mayweg-Paus<sup>2</sup>, Mark Felton<sup>3</sup>, Kalypso Iordanou<sup>4</sup>, Regina Jucks<sup>1</sup> and Maria Zimmermann<sup>2</sup>

<sup>1</sup> Institute for Psychology in Education and Instruction, Department of Psychology and Sport Studies, University of Münster, Münster, Germany, <sup>2</sup> Institute of Educational Studies, Faculty of Humanities and Social Sciences, Humboldt University of Berlin, Einstein Center Digital Future, Berlin, Germany, <sup>3</sup> Department of Teacher Education, Lurie College of Education, San Jose State University, San Jose, CA, United States, <sup>4</sup> School of Sciences, University of Central Lancashire, Larnaka, Cyprus

## OPEN ACCESS

### Edited by:

Patricia A. Alexander,  
University of Maryland, United States

### Reviewed by:

Stefan Fries,  
Bielefeld University, Germany  
Byeong-Young Cho,  
Hanyang University, South Korea

### \*Correspondence:

Friederike Hendriks  
f.hendriks@uni-muenster.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Psychology

**Received:** 15 June 2020

**Accepted:** 16 November 2020

**Published:** 08 December 2020

### Citation:

Hendriks F, Mayweg-Paus E,  
Felton M, Iordanou K, Jucks R and  
Zimmermann M (2020) Constraints  
and Affordances of Online  
Engagement With Scientific  
Information—A Literature Review.  
Front. Psychol. 11:572744.  
doi: 10.3389/fpsyg.2020.572744

Many urgent problems that societies currently face—from climate change to a global pandemic—require citizens to engage with scientific information as members of democratic societies as well as to solve problems in their personal lives. Most often, to solve their epistemic aims (aims directed at achieving knowledge and understanding) regarding such socio-scientific issues, individuals search for information online, where there exists a multitude of possibly relevant and highly interconnected sources of different perspectives, sometimes providing conflicting information. The paper provides a review of the literature aimed at identifying (a) constraints and affordances that scientific knowledge and the online information environment entail and (b) individuals' cognitive and motivational processes that have been found to hinder, or conversely, support practices of engagement (such as critical information evaluation or two-sided dialogue). Doing this, a conceptual framework for understanding and fostering what we call *online engagement with scientific information* is introduced, which is conceived as consisting of individual engagement (engaging on one's own in the search, selection, evaluation, and integration of information) and dialogic engagement (engaging in discourse with others to interpret, articulate and critically examine scientific information). In turn, this paper identifies individual and contextual conditions for individuals' goal-directed and effortful online engagement with scientific information.

**Keywords:** epistemic cognition, argumentation, scientific literacy, digital literacy, multiple documents literacy, online engagement with scientific information

## INTRODUCTION

Socio-scientific issues—from climate change to the ongoing COVID-19 pandemic (we will use the latter issue as an example in this article)—hold many consequences for personal, social, and civic life (Feinstein and Waddington, 2020). For such issues, defining the problem as well as coming up with possible solutions often rests on knowledge and evidence from the natural but also from the social sciences, which are well-beyond most citizens expertise (Zeidler, 2014). Nonetheless, most citizens want and need to stay informed and will likely seek information online, as searching for information on specific science-related issues is usually done on the Internet (National Science Board, 2018). In recent years, the percentage of people who use the Internet to learn about science has substantially

increased, and there, they encounter a wide variety of digital media formats, including social media (Pavelle and Wilkinson, 2020). In this article, we review literature on the cognitive and motivational processes underlying *online engagement with scientific information* (OESI) that individuals employ in order to utilize the affordances and overcome the challenges of searching for and dealing with scientific information in online information environments.

“Engagement” is an elusive concept but has been conceptualized as a behavioral manifestation of motivation or productive participation in a learning activity (e.g., Eccles and Wang, 2012; Bråten et al., 2018). Similar to previous models of engagement (Guthrie and Klauda, 2016), we understand OESI as *goal-directed* (that is, directed at achieving epistemic aims) and *effortful* activity in dealing with scientific information in online information environments, where this activity can be both *individual* and *dialogic*; is supported by *cognitive*, but also *motivational processes*; and leads to the individual arriving at *epistemic ends* (the target of epistemic aims). In the following, we describe our heuristic model in more detail (see **Figure 1** for a graphical representation).

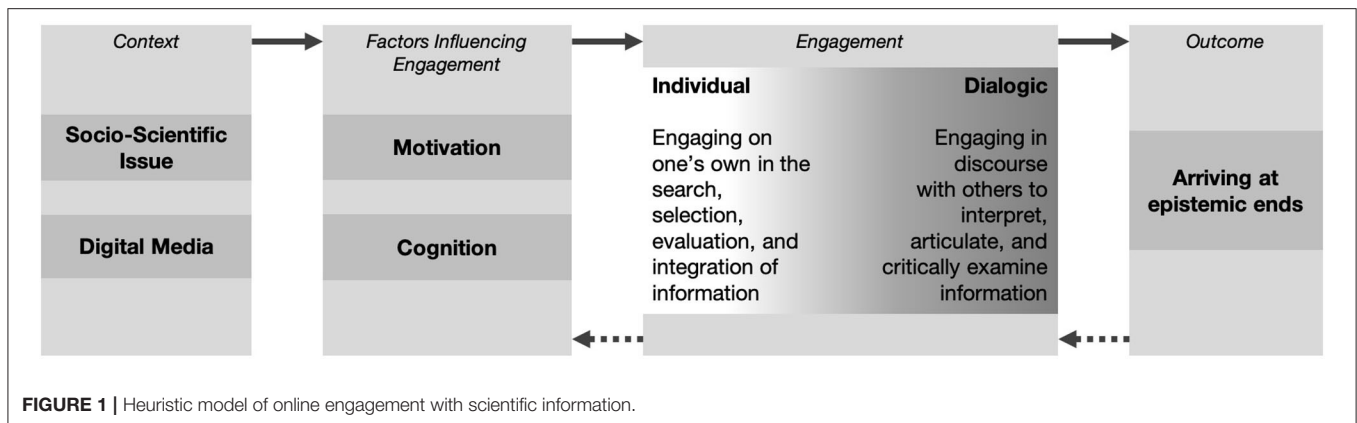
Central to our understanding of OESI is individuals’ adoption of epistemic aims. In their AIR model of epistemic cognition, Chinn et al. (2014; see also, Chinn et al., 2011), identify epistemic aims, ideals and reliable processes that individuals apply to achieve epistemic ends. We describe all three components here briefly, before spelling out their relation to our notion of OESI. First, *epistemic aims* are “a subset of the goals people adopt, specifically those goals related to inquiry and finding things out” (Chinn et al., 2011; p. 142), and they are directed at achieving epistemic ends, for example, gathering “true” facts about a topic, avoiding misinformation on the topic, or acquiring a deeper understanding. Second, how much an epistemic end is valued will affect the selection of epistemic ends. An information seeker will review the success of an information search along her *epistemic ideals*, which could be described as the standard that determines whether a person has achieved her epistemic end; such a standard might be whether the information comes from a highly authoritative source or whether it is based on empirical evidence (Chinn et al., 2014; see also section Epistemic (meta-)cognition). And, third, to achieve epistemic ends, *reliable processes* are applied, which specify the conditions and cognitive operations to achieve reliable knowledge. Importantly, which processes are deemed reliable depends on the context and the individual’s knowledge about the processes. For example, while observation is usually a reliable process to find things out about the (natural) world, individuals may overestimate the reliability of this process, which may lead to misconceptions (Chinn et al., 2014).

Epistemic aims underlie OESI and moderate transitions from stage to stage in our heuristic model (see **Figure 1**). First, when an individual is confronted with a socio-scientific topic in online media environments, which harbor specific constraints and affordances (see section Constraints and Affordances Entailed in the Context of OESI), this elicits cognitive and motivational processes, possibly leading the individual to form (an) epistemic aim(s). If so, these processes become more goal directed (as the individuals strives to arrive at an epistemic end). For

example, if the individual adopts the epistemic aim of avoiding misinformation, she might consider more reliable processes in her search for information, such as referring to fact-check websites, which allow her to compare her achievements with her epistemic ideals (e.g., that accepted information must be evidence based). However, to adequately deal with context constraints and affordances (e.g., the amount of misinformation present in social media), the employed (reliable) processes must also be effortful. Such goal-directed and effortful engagement is what we describe as OESI, and we further differentiate individual engagement (engaging on one’s own in the search, selection, evaluation, and integration of information) and dialogic engagement (engaging in discourse with others to interpret, articulate and critically examine scientific information). We assume that individuals will not follow a specific sequential order when engaging in these two types of engagement and their associated processes, but instead, depending on the situation and the individual’s epistemic aim, any process could be the beginning of an episode of engagement and could lead to any other of the processes—within and between the two parts –, whereby the individual may even switch back and forth, commit to two processes at the same time, or skip a process. Finally, it is also possible for individuals to move back to previous stages: Practices of engagement may, in turn, motivate cognitive and motivational processes (e.g., if the individual feels self-efficient during critical information evaluation, she might be more motivated to achieve her epistemic aims). Furthermore, when the individuals arrives at her epistemic ends—or, instead, is partially or entirely unsuccessful in achieving her aim—she might reconsider her initial epistemic aims and enter another episode of engagement.

However, OESI may not lead to similar (and similarly measurable) achievements as does engagement in formal education settings. By defining outcomes as arriving at one’s epistemic ends, we aim to highlight a central dilemma. Defining a successful outcome largely depends on which standards define achievement: personal (e.g., being content with a personal decision; relieving anxiety) or normative (e.g., achieving full understanding of a concept in alignment with the current scientific state of knowledge). We are aware that these aims require very different cognitive and motivational processes; consequently, we focus on engagement that is moderated by individuals’ epistemic aims and we review research to find out which reliable processes are beneficial for achieving such aims, and for dealing with context constraints and affordances in the process [in contrast, Greene et al. (2020) recently focused on incidental learning in online environments]. Thus, the purpose of this article is to review the literature in several related fields in educational science and educational psychology to identify aspects of the context, and of individual’s cognitive and motivational prerequisites that are especially beneficial or detrimental to effortful and productive OESI. Only when it is goal-directed and effortful can OESI lead to an individual successfully arriving at their respective epistemic ends.

Educational researchers and educational psychology researchers have long investigated individuals’ reasoning and engagement with scientific and online information, and have posited educational implications; these researchers have



delved much deeper into specific aspects relevant to our heuristic framework (e.g., Alexander and The Disciplined Reading and Learning Research Laboratory, 2012; Leu et al., 2013; Fischer et al., 2014; Tabak, 2015; Cho and Afflerbach, 2017; Breakstone et al., 2018; Britt et al., 2019; Coiro, 2020). Taking past conceptualizations into account, we use the term “online engagement with scientific information” not to introduce an entirely new concept or to replace any related concept; instead, here we review this literature, specifically to provide a comprehensive overview of OESI—focusing its context and on cognitive and motivational processes that support it—to derive implications for education and instruction.

## Constraints and Affordances Entailed in the Context of OESI

Information that we consider relevant for OESI is acquired in online environments and (a) contains an elaborate claim on a socio-scientific issue, or (b) is detailed enough to serve as evidence, or (c) both. For example, we would consider as relevant any text, audio, and video sources, as well as images and graphical representations (e.g., a tweet featuring a graph, a YouTube video, an open access scientific article), but we would not consider as relevant a meme consisting only of a photograph and some text, which is only meant to entertain. For individuals to deal with such information to achieve their epistemic aims, they must overcome the constraints and utilize the affordances that is entailed in the respective contexts (Barzilai and Chinn, 2019). We will briefly outline these in **Table 1**.

Two characteristics of scientific knowledge are especially challenging for laypeople to deal with (Bromme and Goldman, 2014; Hendriks and Kienhues, 2019). First, scientific knowledge is characterized by complexity (Keil, 2008) as scientific theories vary in depth (deep causal complexity) and breadth (interrelatedness with other theories or concepts) (Bromme and Goldman, 2014). Consequently, full understanding of scientific phenomena requires both highly specialized knowledge in one field (e.g., virology) and related background knowledge from many other disciplines (e.g., biology, chemistry). For many questions in socio-scientific issues, the complexity of (natural) scientific knowledge is further amplified by manifold interrelations with the social sciences. This is especially the case

when issues entail risk, which can exist both on a personal level (e.g., health risks) and on a societal level (e.g., economic risks). Second, scientific knowledge is intrinsically uncertain (Friedman et al., 1999), whereby uncertainty arises not only during evidence gathering processes (e.g., measurement error, inadequacies of measurement), but also from lack of knowledge or expert disagreement (van der Bles et al., 2019). Scientific uncertainty is becoming increasingly apparent to a larger public as the COVID-19 pandemic progresses, because evidence is rapidly accumulated and published online (sometimes before peer-review), such that public debates often involve highly uncertain scientific knowledge.

Both the complexity and uncertainty of scientific knowledge are amplified in online information environments. Online, there are many possibly relevant information sources that vary in format (e.g., text, video), in genre (e.g., scientific, journalistic, opinion, entertainment), and in explanatory power (e.g., relevant to the topic and founded in evidence). Moreover, sources are highly interconnected; that is, online documents not only embed and interlink diverse formats and genres (Alexander and The Disciplined Reading and Learning Research Laboratory, 2012; Goldman and Scardamalia, 2013), but interconnectedness is also established when sources cite and embed sources of different quality (e.g., when a scientist is interviewed by conspiracy-affiliated news sites), or when scientific arguments are disputed by industry stakeholders. To the individual, this amplifies the complexity of an already complex scientific topic. But also, scientific uncertainty can be amplified, especially as new and yet uncertain results are highly accessible online. In particular (digital), media pieces often display disagreement between experts (Boykoff and Boykoff, 2004), such as when scientists openly disagreed with statements by the WHO about the effectiveness of wearing face masks to protect against COVID-19 (Howard, 2020). Furthermore, around publicly contested issues like climate change and vaccination, skeptics have been especially strategic about utilizing uncertainty to manufacture doubt around scientific knowledge on the issue (Oreskes and Conway, 2011) and attack scientific evidence especially in digital media (e.g., Elgesem et al., 2015; Mercer, 2018).

As a result of these constraints, laypeople find it challenging to engage with scientific knowledge online to achieve epistemic



**TABLE 1** | Some context constraints and affordances of Online Engagement with Scientific Information (OESI).

	Scientific knowledge		Online information environment	
	Constraints and affordances	Examples	Constraints and affordances	Examples
Complexity	Complexity of knowledge in depth and breadth	Full understanding of the transmission of the SARS-Cov-19 virus requires knowledge from a variety of disciplines (e.g., infectology, virology, epidemiology) and relevant background knowledge from other disciplines (e.g., biology, chemistry). Above this, when deciding whether to re-open schools during a pandemic social science knowledge is required (e.g., from educational sciences)	Interconnected and embedded sources	A Wikipedia page includes hyperlinks to other Wikipedia pages. A science blog entry consisting of mainly text embeds pictures and graphs (embedded formats). A science-skeptic social media entry embeds a video of an interview with a scientist (hierarchical structure of formats and credibility cues).
Uncertainty	Uncertainty of evidence	A scientific measurement is imprecise. A scientific study cannot be replicated. It is yet unknown which long-term health effects remain after an infection with SARS-Cov-19. Scientists disagree about the effectiveness of a treatment.	Use of uncertainty to discredit science	Social Media entry advising against wearing cloth face masks, citing uncertainty about their effectiveness and uncertainty about adverse effects. An online newspaper article using balance reporting (devoting the same space to both sides of the issue) even though there is consensus within science.
Risks	Entailed risks on the personal and societal level	Health consequences of infection with SARS-Cov-19. Economic repercussions of the pandemic. Psychological effects of isolation during the pandemic. Educational effects of digital-only schooling.	Disinformation, misinformation and “fake news”	Disinformation: a member of the far-right deliberately posts on Facebook that scientists in China created the new Coronavirus. Misinformation: Someone shares this post considering it to be credible. “Fake news”: mimicking the layout of “real news” and sensationalizing (scientific) news to draw attention and promote sharing.
Level of Gatekeeping	High editorial gatekeeping, highly authoritative sources, limited access	Scientific journal articles authored by scientific experts (sometimes published as pre-print or open access). Reports by a selected group of experts (e.g., initiated institutions like the WHO or within scientific academies)	Low editorial gatekeeping, high diversity of sources, easy access	Science blog authored by a scientist. Journalistic article published on a newspaper's website. Youtube video by a person with a doctoral degree. Facebook entry by a layperson.
Communicative habits	Scientific genre	Almost all scientific publications are journal articles (often enhanced with representations). Most scientific journal articles follow a specific structure, style, format, and use of scientific jargon.	Variety of formats (e.g., text, video, representations), and genre (e.g., informational, narrative)	A scientist blogs about her study using comprehensible language. A Youtuber uses personalized language. A narrative video about virus reproduction.
Agency	Relevance to everyday- and societal questions	Immediate relevance of questions to behavior (e.g., washing hands, wearing a face mask), social and family life (e.g., visiting grandparents) and civic life (e.g., voting, protesting).	User agency	Users can decide what information to consider (e.g., by ordering an Email-newsletter or following certain social media accounts), but also how to consume it (e.g., free surfing, deciding to watch a video instead of reading a text).
Social Affordances	Argumentation as intrinsic to science	Social practices of science (e.g., conferences, peer-review, consensus building). Public Engagement with Science (Citizen Science, engaging members the public in generating research questions or funding decisions).	Social affordances, interactivity	Digital media entail affordances for immediate audience feedback and users' own active contributions: e.g., Like-button, comment section, discussion forum, creating own content.

aims; yet, the context of OESI also entails affordances that individuals can utilize. Socio-scientific issues may motivate individuals to purposefully engage with scientific information, because the scientific questions are highly relevant, and are often contextualized in everyday life and societal questions (Feinstein and Waddington, 2020). Science fundamentally rests on the active dialogue about and the critique of scientific claims (Osborne, 2010), and members of the public can now contribute more to this dialogue through efforts such as the movement toward Public Engagement with Science (Leshner, 2003). Furthermore, increased access to scientific information via digital media creates even more opportunities for individuals to connect with science (Brossard and Scheufele, 2013). Especially because scientific knowledge is often communicated in very formalized ways in terms of formats and language use, digital media platforms grant laypeople the opportunity to learn about science in various different formats and in much more accessible and engaging language; for example, YouTube videos often use an entertaining and narrative style to communicate quality informational content. However, because individuals can access such a wide variety of sources, they must be able to identify not only trustworthy sources, but also communicative intentions to distinguish, for example, institutional public relations information from critical science journalism, and even from science-related entertainment. Moreover, online, individuals must be especially aware of messages that are deliberately posted to disseminate false information, called *disinformation* or “fake news” (a term that has also been weaponized in political contexts; Molina et al., 2019). In contrast, *misinformation* is spread without malevolent intentions (Molina et al., 2019; Scheufele and Krause, 2019), but it is still a threat toward an individual’s engagement with scientific claims and evidence.

The requirement to effortfully seek out credible information represent the downsides of individuals’ ability to be active agents in using and interacting with online digital media platforms (Evans et al., 2016), where they can deliberately choose to engage with certain technologies, media, and content. Furthermore, individuals may even create their own content and—utilizing digital media’s social affordances (Hopkins, 2016)—interact and engage in dialogue with other users.

In the article, we refer to research that describes which cognitive and motivational processes people employ to deal with these context constraints and affordances. While we do differentiate some constraints and affordances for the two contexts, some individuals may perceive an aspect that we introduced as constraint to be more of an affordance, and vice versa. For example, a comment section to a blog entry might initially be an affordance, but dealing with a high number of reader comments may hinder individuals’ evaluation of information, thus making it a constraint.

## INDIVIDUAL ENGAGEMENT

Searching for information to achieve epistemic aims is an iterative and dynamic process. To make sense of scientific

information in order to achieve their epistemic aims on their own—to form “true” beliefs or understanding—individuals must employ reliable processes. To describe the necessary cognitive processes during an information search, we will first describe the MD-Trace (Multiple Documents-Task-based Relevance Assessment and Content Extraction) model (Rouet and Britt, 2011). According to this model, a search is initialized by an individual’s mental representation of the searching task in a task model (see also, Rouet et al., 2017). Further, her task model also involves considering available knowledge and resources, such as prior topic knowledge and knowledge about search strategies (Rouet et al., 2017). As a result of these processes, the individual determines whether further information is needed to fulfill task demands and against what standard the search result should be compared. Having initiated the search process, she tests whether the sought information is relevant to her task model and *selects* documents accordingly. To *process* and evaluate the selected documents, the individual mentally represents them in an intertext model, which links contents of the documents to their meta-information (information about, e.g., the source, date, or rank of the search result), and includes intertext predicates (e.g., possible conflicts). *Integrating* information into the mental model allows the individual to coherently represent her acquired understanding of the issue. Finally, she may compare this integrated mental model against her initial task model to decide whether to redo certain steps of the search task or to go ahead with creating a search product (e.g., write an essay or make notes next to search results to further concretize a search task). However, at each step, individuals face several challenges (Rouet and Britt, 2011). In this section, we will summarize research on how searching, selecting, processing, and integrating scientific information are supported or hindered by aspects of the context and the individual’s cognitive and motivational processes.

## Constraints and Affordances of the Online Information Environment to Individual Engagement

When searching for information, media affordances determine how specific technologies are used. That is, while users may deliberately choose to use technologies or digital media for the potential features they offer; at the same time, such features also determine the ways in which users can engage with the technology. For example, when acquiring (scientific) information, individuals tend to use only one type of search engine, which might be enforced by the default use of digital assistants commonly installed on smartphones and computers (Kammerer et al., 2018). Additionally, characteristics of a search engine result page (SERP), such as the algorithm it uses to present search results, the interface it offers for users to manually filter search results, or the sparsity of information it displays (i.e., a title, short excerpt of the web page, and the URL) may influence whether an individual selects any of search results and whether they perform any further search queries. Research indicates that individuals would rather view the highest-ranked search results within a SERP (e.g., Salmerón et al., 2013; Haas and Unkel, 2017), even if those results are less relevant (Pan et al., 2007). Further,

younger users in particular might select search results based on superficial cues like the search result's title (Lai and Farbroth, 2014), or boldface or capitalization (Rouet et al., 2011). Also, the number of documents that individuals select seems to vary by task: When individuals are asked to find a discrete answer to a question (instead of answering in an open-ended way), they select more documents (List et al., 2016a). Furthermore, individuals do not use all features of a search engine that perhaps would allow them to conduct more appropriate search inquiries. Kammerer and Gerjets (2014) found that interfaces displaying the results in a three-by-three grid more often led users to select and view search results according to their trustworthiness than according to their search rank. Similarly, Salmerón et al. (2010) found that individuals had more efficient reading times and displayed more explorative search behavior when using a graphical-overview interface (i.e., indicating the semantic relationships between the search results) instead of a standard list interface. Prior knowledge about the search topic may further benefit an individual during an information search when the search engine interface allows it: Experts performed faster and more accurate searches than laypersons when the interface was semantically structured (Salmerón et al., 2005).

Second, the interconnectivity and embeddedness of information sources—both hierarchically (documents that are interlinked), and horizontally (one document that is embedded within another)—may be challenging for information seekers to deal with (Cho and Afflerbach, 2017; Goldman and Scardamalia, 2013). These features call for flexibility in how individuals access information (Shapiro and Niederhauser, 2004), namely they have to access information in a non-sequential, non-linear way. This might require some specific aspect of digital literacy: Although expert searchers (fact checkers) were found to perform lateral reading, that is, opening several browser tabs during a search to check the reliability of a search result, this was not done by topic experts (historians) or students (Wineburg and McGrew, 2019).

The goal-directed and effortful evaluation of online information may further be constrained by several context features of scientific information in digital media environments (Breakstone et al., 2018; Forzani, 2019), such as genre, presentation of information (such as the use of distracting imagery), or other users' endorsements. Unfortunately, individuals often use only superficial or unreliable indicators for determining the credibility of online information (Coiro et al., 2015; McGrew et al., 2018). For example, individuals may not be able to distinguish sponsored news content from unbiased news stories or to identify the verified social media accounts of public organizations (McGrew et al., 2018). Furthermore, the extent to which adolescents use social media sites for entertainment purposes can be negatively related to their ability to discriminate reliable from unreliable online information (Macedo-Rouet et al., 2019b). Some online platforms, and especially social media, seem not to be regarded as trustworthy by individuals in general. Wikipedia is sometimes dismissed as information source without considering its inherent quality control (Breakstone et al., 2018). Evidence suggests that individuals deem Twitter and blog entries less trustworthy than (for example) newspaper articles and

refrain from citing them, even if they entail relevant first-hand information about an issue (List et al., 2017).

Further, the communicative design of scientific information appears to affect its evaluation. Using a more "scientific" language style, such as including descriptions of scientific methods and in-text citations, leads readers to judge the information as more "scientific" and believable overall (Thomm and Bromme, 2012). Over a series of studies, Scharrer and colleagues (e.g., Scharrer et al., 2012, 2017) found that when a scientific text was written in a comprehensible fashion (compared to when the text contained technical terms and was, thus, incomprehensible for laypeople), readers were more easily persuaded by the text's arguments and less inclined to consult further expert advice. Furthermore, when individuals are engaged online in argumentation, presenting a piece of information in the form of question and answer rather than in the context of a traditional text may be a more effective way to promote the acquisition of factual knowledge (Iordanou et al., 2019a). The question-and-answer format appears to have facilitated learning, possibly by highlighting the potential use of a particular piece of information.

Another feature of online environments is that not only social media and blogs but also many online news sites allow for user comments, which might influence how users evaluate the content of the main article. For example, attitudes about a scientific issue may be influenced by the perceived consensus among other readers expressed through blog comments (Anderson et al., 2014; Lewandowsky et al., 2019). Furthermore, in some instances recommendations and social endorsements might play a role in evaluation and could reflect on evaluations of the credibility of health messages and of the expertise of the author (Jucks and Thon, 2017). In one study, when Facebook posts were shared by a close friend, this only raised the credibility of otherwise distrusted news sources (participants rated their trust in several news sources prior to reading the posts) but not of trusted sources (Oeldorf-Hirsch and DeVoss, 2020).

To sum up, during the first steps of searching for and selecting relevant information, characteristics of the online environment [e.g., (social) affordances of SERPs and digital media, communicative habits in digital media] may constrain, but also inspire effortful cognitive processes when searching, selecting and evaluating information. Dual-process theories propose that—unless task or person characteristics require it—individuals will default to heuristic processing instead of effortful and systematic processing (Salmerón et al., 2013). In an online information search, a variety of heuristic cues determine whether a search result is credible or relevant to the task at hand (Hilligoss and Rieh, 2008; Sundar, 2008; Metzger and Flanagin, 2013). Taraborelli (2008) stated that research has mainly focused on predictive judgments of credibility evaluation instead of evaluative judgments; this means that individuals may often engage in a first selection phase to sort out low-quality information in which superficial cues guide information selection, whereas in a second step they might engage in more effortful evaluation (Hilligoss and Rieh, 2008). In fact, in one study, individuals' first selection of search results relied on the order of appearance in a SERP, but they bookmarked more relevant pages to examine further (Salmerón et al., 2013). In

another study, individuals did first select links by their titles, but on second glance they considered cues more indicative of information quality, like URLs and snippets with brief descriptions (Hautala et al., 2018).

However, the activity of online searching itself may lead to a feeling of knowing—the case when an individual perceives to possess knowledge but cannot actually retrieve it from memory (Pintrich, 2000; Koriati, 2012). Such an overestimation of acquired knowledge (Fisher et al., 2015) may result from representing the Internet as transactive memory (an external, collective memory system), leading one to better remember where a previously learned item is stored than to recall the item itself (Sparrow et al., 2011). Similarly, searchers might experience a “feeling of findability,” where they overestimate the availability of information online (Risko et al., 2016). These problematic assumptions may stem from a failure to distinguish “what is known” from “how was this knowledge acquired” (Kuhn, 1999). Such knowledge illusions may bias the integrated mental model of search results and thus, may negatively influence the integration of information into a coherent representation of the issue. As such, when misrepresenting acquired knowledge as a result of an online search, the individual might give up on an epistemic aim prematurely due to the assumption that it has been already resolved.

## Emotion and Motivation

Central to our understanding of OESI is identifying when individuals process information more effortfully instead of heuristically; importantly, the process of formulating epistemic aims and following through to resolve them might be strongly influenced by emotion and motivation. Referring back to dual-process theories, Griffin et al. (1999) identified several motivators for more systematic processing of information about risk. First, they found that the central motivators of information seeking were *information insufficiency*—when a person experiences a large gap between current knowledge and her personal sufficiency threshold (Griffin et al., 1999)—and a perceived normative pressure to be informed. Information insufficiency can follow affective responses to perceived risks (Dunwoody and Griffin, 2015). In fact, Yang and Kahlor (2013) found that while positive affect about climate change (e.g., hope) was related to information avoidance, negative affect (e.g., worry) was related to higher information insufficiency and the intention to seek information. Further, feeling personally threatened could bias how search terms are generated in an online search: Participants who were asked to reflect about a threat in their personal life generated more positive search terms in an unrelated Internet search than participants who were not instructed to think about personal problems (Greving and Sassenberg, 2015).

Similar notions and empirical evidence can be found in the literature on epistemic emotions, which are emotions directed at achieving epistemic ends (Muis et al., 2015). For example, enjoyment and curiosity may be positively related to the belief that justifying a knowledge claim requires critical evaluation, and anxiety and frustration may be lower when individuals believe that knowledge is uncertain (Muis et al., 2015). As such, different epistemic emotions may follow an experience of inconsistent or

conflicting information. In fact, when individuals were surprised by incorrect answers in a trivia task (especially when their answers were given with high confidence) they had—as mediated by curiosity—more motivation to seek out explanations for these answers and request further information (Vogl et al., 2020).

In even more fundamental ways, the Cognitive Affective Engagement Model (CAEM) of multiple source use (List and Alexander, 2017b) addresses “learners’ affective, cognitive, and behavioral involvement in multiple text use” (List and Alexander, 2017b, p. 184). Both situational and individual interest (Schiefele, 2009) have been found to promote learning and behavior (see also, Deci, 1992). Situational interest is a state that might be triggered by a single text (for example, when it is very easily comprehensible or coherent), while individual interest in a domain or topic is a trait-like personal characteristic (Schiefele, 2009). In consequence, the CAEM specifies an affective engagement dimension, which refers to an information seeker’s interest and motivational involvement in the task at hand (also affected by topic-specific attitudes and prior beliefs), whereas the second dimension, behavioral dispositions, refers to the skills and strategies necessary for selecting, evaluating, and integrating information and documents at hand. By crossing these two dimensions, the CAEM states that learners fall into one of four default stances that guide their multiple-document comprehension: A “disengaged learner” selects and uses information without engaging much in evaluating and integrating. An “affectively engaged learner” accumulates information while engaging only in limited integration of multiple documents. An “evaluative learner” scrutinizes documents for relevance and credibility, but, due to limited motivational engagement, is less willing or able to fully integrate selected documents. A “critical analytic learner” possesses similar skills as the “evaluative learner” regarding source evaluation and verification, but since the critical analytical learner is highly motivated to engage in effortful and elaborate processing, they are able to succeed in integrating information into a coherent representation of the issue and, thus, might produce the most successful search result.

In sum, central motivators of goal-directed and effortful OESI are both personal relevance and topic specific risk perceptions (both affordances of socio-scientific topics). Furthermore, experienced information insufficiency may not be the only motivator to formulate epistemic aims; this may also be motivated by situational interest and epistemic emotions such as curiosity. Beyond individuals’ skills to engage in reliable processes in dealing with scientific information, effortful evaluation and integration of information may also be fostered or constrained by emotions (both topic specific, e.g., hope or worry; and epistemic, i.e., directed at learning and understanding) and motivational involvement in the task.

## Epistemic (Meta-)Cognition

Epistemic beliefs have long been investigated as part of reasoning and arguing about scientific information. Such beliefs about the nature of knowledge and knowing (e.g., holding beliefs about scientific knowledge being uncertain, complex, or needing expert justification) may incite the use of reliable processes and



strategies during OESI. Several studies in which students were asked to think aloud during an online search have demonstrated that students use their epistemic beliefs to define standards for learning and accordingly select their strategies (Hofer, 2004; Mason et al., 2010a,b, 2011; Barzilai and Zohar, 2012). For example, beliefs about the complexity of an issue led individuals to reflect on the need to compare several documents and collect contrasting views (Mason et al., 2011), and the belief that knowledge is given and stable did co-occur with less use of strategies to actively construct knowledge from texts (Bråten and Strømsø, 2006). A person's epistemological understanding ties in with her metacognitive processes and strategies (Kuhn, 1999; Muis, 2007; Barzilai and Zohar, 2016), as it may directly influence the standards she sets for acquiring knowledge and understanding (Muis, 2007). As such, Barzilai and Zohar (2016) have argued that epistemic metacognitive knowledge (as a specific part of metacognition) may "guide the execution of cognitive-level epistemic strategies as well as their selection, monitoring, and evaluation" (Barzilai and Zohar, 2016, p. 414).

Furthermore, epistemic beliefs may also affect how effortfully individuals execute practices of OESI. Evidence from studies using the think-aloud technique shows that epistemic beliefs influence individuals' abilities to engage in evaluating information both while navigating the web—e.g., identifying argumentative fallacies (Mason et al., 2010b)—and while reading (Ferguson et al., 2012; Iordanou et al., 2019b). Further, viewing knowledge as tentative enhances meaning-making as one deals with multiple documents (Bråten and Strømsø, 2010) and supports credibility assessment of newspaper articles, for example when they present simplified accounts of an issue (Strømsø et al., 2011). Individuals with evaluativist epistemic beliefs engage more often in evaluating the credibility of evidence presented in texts and use scientific research as their standard for judgment; for example, they might consider the number of scientific studies supporting a particular piece of evidence (Iordanou et al., 2019b). Besides supporting the evaluation of single pieces of information, adequate epistemic beliefs also support the evaluation and integration of multiple pieces of information presented in different sources (Bråten et al., 2011; Barzilai and Eshet-Alkalai, 2015). Empirical evidence shows that adequate epistemic beliefs support the integration of information during online learning (Barzilai and Zohar, 2012) and during reading of multiple texts (Ferguson and Bråten, 2013), where comprehension mediates the relationship between epistemic perspectives and information-source integration (Barzilai and Eshet-Alkalai, 2015).

In sum, beliefs about the nature of scientific knowledge may directly influence which strategies and practices are employed during OESI (Muis, 2007; Barzilai and Zohar, 2016), and may also affect the epistemic ideals by which epistemic ends are evaluated (Chinn et al., 2014). That is, in addition to an individual's scientific literacy (see section Evidence Evaluation and Scientific Literacy), her epistemic beliefs may inform how she assesses the uncertainty and complexity of scientific information, and these beliefs may also guide the selection and metacognitive regulation of reliable processes for achieving her epistemic aims.

## Source Evaluation

Due to limited gatekeeping of scientific information online (vs. editorial gatekeeping in scientific journals or traditional media), evaluating the source of scientific information is an especially important process within OESI, as it underlies the selection, evaluation, and integration of credible information. When retrospectively justifying document selection, students used epistemic criteria (e.g., source type, author) less often than non-epistemic criteria (e.g., order in the search list, relevance), but the more epistemic justifications were made, the more arguments and citations they presented in an open-ended search result (List et al., 2016b). However, individuals prefer authors of information to have good reputations (Rieh, 2002; Hilligoss and Rieh, 2008; Winter and Krämer, 2012); more specifically, readers tend to select blog posts by experts who possess relevant expertise on the topic in question (Winter and Krämer, 2014) and prefer disciplinary relatedness of search results to mere lexical similarity with search terms (Keil and Kominsky, 2013).

Diverse research findings suggest a variety of cues that individuals consider during source evaluation. First, the experts' language use seems to affect how trustworthy she is perceived. Individuals develop expectations about what constitutes appropriate language in different social and cultural contexts, and, thus, language accommodation or non-accommodation by speakers (reflecting their intentions and motives) may influence how individuals evaluate a speaker (Dragojevic et al., 2016). For example, an expert's use of technical language in scientific information may lead to her being ascribed higher expertise (Thon and Jucks, 2017) as well as higher integrity and benevolence when her use of (technical) language is appropriate to the context, e.g., when she uses less technical language when addressing laypeople (vs. experts) in online health forums (Zimmermann and Jucks, 2018), or less aggressive language in an online video (König and Jucks, 2019). Furthermore, the perception of a communicator in an online video as being comprehensible and entertaining also led to higher ascriptions of trustworthiness (Reif et al., 2020). Individuals also take an expert's motives into account when evaluating trustworthiness; for example, readers were more inclined to trust a scientist when they believed the scientist intended to inform rather than persuade them (Rabinovich et al., 2012), when the scientist provided a two-sided stance (instead of a one-sided stance) (Mayweg-Paus and Jucks, 2018) or mentioned the ethical aspects of a scientific issue (Hendriks et al., 2016). Furthermore, people perceived a source to be less trustworthy when the source had a vested interest in a claim (König and Jucks, 2019; Gierth and Bromme, 2020); this even sometimes motivated people to engage in effortful processing of complex evidence (Gierth and Bromme, 2020).

While these findings suggest that individuals are often able to adequately judge the trustworthiness of sources, research on "sourcing" (referring to when individuals pay attention to and use source features, such as the author, but also publication date) in multiple-document comprehension has found that students often fail to pay attention to source information (Britt and Aglinskias, 2002; Sandoval et al., 2016; for a review see, Brante and

Strømsø, 2017). In fact, when evaluating multiple documents, individuals may not attend to author competence at all, and younger individuals (in elementary and middle-school) even failed to do so when explicitly prompted to evaluate sources (Macedo-Rouet et al., 2019a; Paul et al., 2019).

However, interacting with online information might not hinder successful sourcing *per se*. For example, reading an online document (instead of its printed-out version) increased memory for sources, which helped readers construct coherent interpretations of the issue at hand (Salmerón et al., 2017); that is, it helped them integrate information. Further, interacting with multiple sources is more effective than reading a single source for text comprehension and establishing source and content integration (e.g., Le Bigot and Rouet, 2007; Stadtler et al., 2013; Stang Lund et al., 2019); that is, individuals seem to have increased awareness about source information and create stronger content-source links when a conflict cannot be resolved by content information alone (Braasch et al., 2012; Strømsø et al., 2013; Stadtler and Bromme, 2014) or when information conflicts with prior beliefs about a topic (Bråten et al., 2016). As such, conflicts within single or multiple texts, as well as conflicts between newly acquired information and prior knowledge, might promote more effortful and strategic evaluation of sources (Braasch and Bråten, 2017). Further, relevant prior topic knowledge seems to benefit individuals' sourcing abilities (Stang Lund et al., 2019), whereas individuals with low prior knowledge may even prefer untrustworthy information sources (Bråten et al., 2011). In sum, while individuals use many different cues to determine source trustworthiness, encountering conflicting information about socio-scientific issues online seems to motivate individuals to engage in more effortful (source) evaluation and integration of information.

## Evidence Evaluation and Scientific Literacy

Evaluating the strength of evidence (or even its inner-scientific significance) should be central to individuals' consideration of information from a normative standpoint, but this is challenging for laypeople considering the uncertainty and complexity of scientific information and their own bounded understanding of science (Bromme and Goldman, 2014). One possibility to rate the credibility of scientific claims would be to assess argument strength and structure, for example whether a claim is backed by evidence. While laypeople adequately assess argument strength to be greater when it is supported by a greater amount of evidence (Corner and Hahn, 2009; Hendriks et al., 2020), they may sometimes not take prior studies into account when assessing the probability of an effect to be true (Thompson et al., 2020). Individuals might assume that the tentativeness included in scientific information means that the scientific results have limited credibility (Flemming et al., 2015); however, in one study that gave readers a refutation text alerting them that this assumption is wrong, the assumption was successfully reduced (Flemming et al., 2020). Similarly, a stronger epistemic belief regarding the uncertainty of science might alleviate the adverse effects of scientific tentativeness on the credibility of information (Rabinovich and Morton, 2012; Kimmerle et al., 2015). However, when making inferences from evidence, people may follow a

causality bias, such as when interpreting correlational data (Shah et al., 2017). That is, new evidence may be rejected if it does not fit within a broader single causal framework (Koslowski et al., 2008). Further, it is unclear which type of evidence individuals consider to be informative. Although some studies have indicated that statistical evidence (citing a study), expert statements, and causal evidence are perceived to be more persuasive than anecdotal evidence (Hornikx, 2008), adding anecdotal stories into scientific news articles decreased the extent to which participants engaged in scientific reasoning about the evidence (Rodríguez et al., 2016). Moreover, individuals often do not take multivariate causality into consideration (Kuhn, 2020). Thus, successful online information behavior on complex topics is constrained by individuals' tendencies to think simplistically about complex issues instead understanding that most phenomena are caused by multiple contributing factors or, for judgments of a non-causal nature, taking multiple considerations into account (Kuhn and Iordanou, 2020).

Basic scientific literacy will also likely help individuals successfully evaluate and integrate scientific evidence. Internationally, educational frameworks for scientific literacy (e.g., OECD, 2017; National Research Council, 2012) have emphasized that a central aim of science education should be to familiarize students with processes of scientific inquiry, evidence evaluation, and argumentation. Scientific literacy has been ascribed three core dimensions: content knowledge (about a few core scientific concepts), procedural knowledge, and epistemic knowledge (Kind and Osborne, 2017). As such, it is important to consider how individuals understand not only the processes of doing science but also the modes by which it achieves reliable knowledge, such as expert epistemic practices (Golan Duncan et al., 2018). Kienhues et al. (2018) recently argued that "science-based arguments can be understood and judged by criteria on three layers of scientific knowledge: (1) the ontology, (2) the methods and sources, and (3) the social practices required for the generation and justification of the argument" (Kienhues et al., 2018, p. 253). They argue that everyday evaluation of scientific arguments may benefit from switching between these layers. For example, when it is not feasible to come to a conclusion about a scientific issue based on reliable evidence (maybe due to conflicting pieces of evidence), the individual may switch to investigating which scientific processes were used, which will help them identify which argument is backed by stronger evidence. If that is not feasible, the individual might judge whether the conflicting positions might be partly due to the complexity of the topic or the motivations of the involved experts behind the conflicting positions (Dieckmann et al., 2017; Thomm et al., 2017). Even if someone has limited content knowledge, they can still be successful in assessing a scientific issue online by determining, for example, whether there is consensus among scientists about an issue (a social practice of science; Oreskes, 2007; van der Linden et al., 2015) and then adopting the consensus view.

To summarize the two previous sections, individuals themselves often cannot adequately evaluate the credibility of a provided scientific claim, and some have argued that in such

a case it is instead more feasible to evaluate the trustworthiness of the information source (Bromme and Goldman, 2014; Hendriks and Kienhues, 2019). That is, holding epistemic ideals regarding the justification of knowledge in consensus, or by a highly trustworthy source might be more beneficial for deciding whether to accept online information as provisionally true. Hence, instead of asking “What is true?” individuals can rather solve the problem by asking “Whom do I believe?” (Bromme et al., 2010; Stadler and Bromme, 2014). Hendriks et al. (2015) define epistemic trust as the willingness of a person to depend on an information source for knowledge; this trust is not blind, however, but relies on a person’s epistemic vigilance toward cues that indicate whether an information source might be deceptive or ignorant (Sperber et al., 2010). In digital settings, evaluations of epistemic trustworthiness of expert sources rely on considering an expert’s expertise (possessing relevant knowledge), integrity (adhering to the rules of their profession), and benevolence (having the interest of others at heart) (Hendriks et al., 2015).

### (Prior) Attitudes and Beliefs

Prior topic knowledge and attitudes can affect processes of individual engagement from the start of setting up a task model search to the (internal) formulation of a solution. On the one hand, prior topic knowledge and attitudes can result in individuals using more appropriate keywords and selecting more relevant information (e.g., MaKinster et al., 2002), on the other hand, they may also bias the information search. Selective exposure to information (sometimes referred to as *confirmation bias*) means that an individual is more likely to select attitude-consistent information (Fischer et al., 2005; Rothmund et al., 2015; Knobloch-Westerwick et al., 2020), and also evaluate that information more favorably (van Strien et al., 2016; Strømsø et al., 2017). An explanation for selective exposure during an information search might be defense goals, whereby an individual ignores or dismisses counter-attitudinal information to preserve their own worldview (Cappella et al., 2015; Winter et al., 2016). Nevertheless, those information seekers with high need for cognition are more likely to select two-sided information (e.g., suggested by the link title) for further reading (Winter and Krämer, 2012). Prior knowledge and attitudes may also detrimentally affect the evaluation and integration of scientific information online. Arguably, prior beliefs are internal representation with which newly acquired information has to be integrated. Richter (2015) assumes a “text-belief consistency effect” for integrating information into mental (situation) models. In fact, research shows that prior beliefs and attitudes might affect the way a person evaluates information and integrates new evidence into their internal representation of an issue. Chinn and Brewer (1998) showed that only in very few cases did anomalous evidence (evidence inconsistent with individuals’ already established theories) result in careful consideration and adaptation of individuals’ theories; often, such evidence was just ignored or discounted.

Motivated reasoning is also an important drive for rejecting information that is not consistent with the dominant belief in

an individual’s social group (Kahan, 2013). For example, group identity may cause individuals to apply defensive motivations when reading about scientific issues and, in consequence, might further strengthen the text-belief consistency effect (Maier et al., 2018). In one study, Nauroth et al. (2015) showed that people who self-identified with the social group of gamers devalued identity-threatening scientific information (e.g., playing video games increases violence in youth) that was presented in a science blog, and, when allowed to post a comment, they criticized the methodology of the scientific study. Further, in another study identity-threatening information affected how reputable and competent participants perceived the scientist authors to be (Nauroth et al., 2017). However, biased evaluation of scientific evidence may not only arise from an identity threat but also from a threat to one’s general values. For example, the more central a person held non-violence to be in their self-concept, the more positively they evaluated a scientific study that claimed video games promote violence (Bender et al., 2016). Also, expert sources may be considered more credible when the ethical stance of the reader aligns with that of the source, leading to higher agreement with the source’s claims (Scharrer et al., 2019).

In sum, prior beliefs and attitudes may play a central—and often detrimental—role in establishing a task model for searching for scientific information, as well as evaluating and integrating information. However, sometimes, prior beliefs may motivate effortful processing and evaluation of documents (Rouet and Britt, 2011; List and Alexander, 2017b; Rouet et al., 2017)—for example by eliciting curiosity by being unexpected (see section Emotion and Motivation) or evoking situational interest—allowing individuals to switch from belief protection to belief reflection (List and Alexander, 2017b). By judging the plausibility (“the potential truthfulness of a claim”; Sinatra and Lombardi, 2020, p. 5) individuals may utilize their prior knowledge by allowing them to select the most likely alternative, especially when an issue is contradictory and uncertain. Lombardi et al. (2016) provided a theoretical framework for plausibility judgments, which entail (a) alignment with prior knowledge and beliefs, (b) complexity of and (c) perceived conjecture within novel information, (d) judgments of source trustworthiness, and (e) the individual’s heuristic processing and possible biases. Plausibility judgments may be guided by different degrees of evaluation. While most judgments are implicit (due to a preference for heuristic processing, see above), individuals’ epistemic dispositions and motives (e.g., need for cognition) may lead to more effortful processing. Further, if motivated (e.g., if they are interested and self-efficient), individuals may also reappraise their original judgments, guided by more explicit processing and increased effort in reasoning. In consequence, Sinatra and Lombardi (2020) suggested that fostering individuals’ capabilities to quickly make plausibility judgments about information—by efficiently employing their prior beliefs and knowledge—may be more fruitful in “post-truth” contexts (similar to the contexts we previously described for OESI) than training effortful strategies to evaluate information and its sources.



## DIALOGIC ENGAGEMENT

Besides seeking and evaluating information independently to form beliefs, OESI includes engaging in discourse with others to share, interpret and critically examine scientific information. In this sense, social media platforms have emerged not only as an important source of information (Head and Eisenberg, 2010; Kim et al., 2014), but also as a public forum for engaging with science (Baram-Tsabari and Schejter, 2019). In fact, we perceive individual and dialogic engagement as reciprocal processes. For example, individually forming an understanding of an issue is immediately beneficial for constructing arguments when engaging in dialogue with others, and, conversely, dialogue and deliberation with others might one to revise their original understanding (see section Reciprocity of Dialogic and Individual Engagement).

When we consider OESI as a social process, it involves the overlapping processes of interpreting information, building arguments from that information and contrasting those arguments with competing arguments. Berland and Reiser (2009) propose that these processes, which they refer to as sensemaking, articulation and persuasion, respectively, form the foundation of scientific argumentation. Although scientific argumentation can be an individual process, as a dialogic process it presents a unique set of affordances and constraints. In the following sections, we explore these affordances and constraints and propose ways in which scientific argumentation as a social process can be leveraged to focus the epistemic aims and outcomes of OESI.

### Constraints and of Affordances of the Online Information Environment to Dialogic Engagement

Many different social media platforms exist, and their functions range from social networking and community building to collaborative knowledge construction and sharing (Leonardi, 2015; Krancher et al., 2018). Building on these potential functions, social media platforms may benefit the motivations and outcomes of OESI (Gao et al., 2012). However, to understand and to exploit the full potential of communication for using online information successfully with others on social media, we need to consider the role that a social media platform's characteristics play in users' abilities to select and establish network connections and to interact with other users (DeNardis, 2014). Following Ariel and Avidar (2015), the degree of interactivity is thereby not primarily determined by the technical features of a platform (interactivity as a medium characteristic) but rather by the actual aims and behaviors of its users (interactivity as a process-related variable). In other words, social networks such as Facebook, Twitter and Instagram do not necessarily produce interactive communication behavior *per se*, but rather they provide opportunities for different ways of communicative exchange.

In this regard, Rafaeli (1988) interactivity model suggests three possible types of messages in communication. The first type refers to one-way communication between a sender and a

receiver, and messages are characterized by low responsiveness. The second type allows for two-way directional communication, as the receiver of a message becomes a sender and is, therefore, responsive to the information provided (or posted). However, only the third type enables real interactivity in a two-way flow of messages between users and is, therefore, highly responsive. Here, such interactive messages encourage the interaction to continue back and forth. Consequently, the transmission of information can be seen as the center of interaction, and interactivity seems to be a central attribute of the process of communication itself (Rafaeli and Ariel, 2007; Ariel and Avidar, 2015).

### Types and Goals of Dialogue

When we think about using information to communicate with others online, we should also think about the purpose of such communication. Two-way communication, or dialogue, can be divided into different types, each with a particular set of epistemic aims (Rapanta and Christodoulou, 2019). Walton (2010) identifies seven dialogue types that apply to communication in both face-to-face and online settings. These are information-seeking, discovery, inquiry, deliberation, negotiation, persuasion, and eristic dialogue (or "irrational dispute"). All are argumentative insofar as speakers posit how information can be brought to bear on claims, but they differ in their initial state and intended outcomes. For example, both inquiry and persuasion involve making claims with evidence, but inquiry focuses on collecting evidence to test claims, while persuasion focuses on citing claims and evidence to defend a conclusion. Dialogue types can also be distinguished by their social-emotional goals. To get at the role that personal stakes can play in dialogue, Asterhan (2013) proposes a distinction between *competitive interpersonal goals* and *collaborative interpersonal goals*. The former are competitive in that speakers take an adversarial stance on what they perceive to be zero-sum outcomes, and the latter are collaborative in that speakers take a cooperative stance on what they see as a shared enterprise. It is important to note that these interpersonal goal states are distinguishable from dialogue types. Some dialogue types may be more likely than others to trigger competitive interpersonal goals (persuasion, negotiation, and eristic, for example), while others may tend toward collaborative interpersonal goals (information-seeking, inquiry, and deliberation). However, interpersonal goals represent social-emotional outcomes that are distinct from the competitive or collaborative epistemic aims used to define dialogue types (except perhaps for eristic argument, which is primarily driven by interpersonal conflict). For example, negotiations can be conducted either collaboratively or competitively, depending on the stance, strategies, and dialogic moves chosen by each party (Lewicki et al., 2001). Likewise, although deliberations aim at group consensus, they may unfold as either collaborative or adversarial exchanges depending on the ways in which interpersonal dynamics emerge and are negotiated during dialogue (Tuler, 2000).

Based upon these considerations, we now focus on the potential benefits of argumentative dialogue as a two-way communication mode for addressing OESI in the context



of dealing with (conflicting) scientific knowledge and socio-scientific issues. Numerous studies point out that dealing with complex content within argumentative dialogue has a positive effect on reasoning about information in online contexts [an overview is given in a meta-analysis by Noroozi et al. (2012)]. In order to successfully co-construct an elaborated understanding of an issue (e.g., Teasley, 1997; Chi, 2009), users need to apply “reasoning that operates on the reasoning of another” (transactivity, Berkowitz and Gibbs, 1983, p. 402). In this sense, transactive dialogue as a specific form a two-way argumentative requires coherent reference and mutual elaboration of each other’s contributions by aiming at the integration of different knowledge backgrounds and perspectives (Asterhan and Schwarz, 2016). However, before well-elaborated consensus building is achieved, each contribution needs to be scrutinized critically (conflict-oriented consensus building; Fischer et al., 2002). Accordingly, an important feature of this type of consensus building is that individuals do not accept contributions of their partners as they are. In this context, efficient communication comprises strategies that directly address and challenge the argumentative structure and content (e.g., scientific evidence) of the other’s contributions (Mayweg-Paus and Macagno, 2016; Mayweg-Paus et al., 2016a). In particular, critical questioning seems to be a strong argumentative strategy given its capacity to address deeper grounds of disagreement, bringing into light background knowledge and knowledge beliefs that might otherwise escape attention. In such cases, a goal is to avoid pseudo-agreements or pseudo-disagreements (Jucks and Paus, 2013) and to focus the discussion on the true source of differences in opinion. Consequently, asking critical questions seems to play a pivotal role in the context of knowledge construction (Chinn and Osborne, 2008) and for developing insights into not only science-related issues (Mayweg-Paus et al., 2016b; Thiebach et al., 2016) but also history (Wissinger and De La Paz, 2016) and public policy (Song and Ferretti, 2013).

When individuals hold one another accountable to standards for accurately collecting and interpreting information and validly using information as evidence, two-way communication offers distinct advantages over one-way communication. However, a two-way discussion can also undermine the quality of reasoning about evidence. The same set of forces that drive motivated reasoning when individuals think alone [see section (Prior) Attitudes and Beliefs] can also compromise reasoning when we engage in dialogue. Critical discussions, particularly those that polarize views on a topic (Kuhn and Lao, 1996), can prompt individuals to both overvalue confirming evidence and discount disconfirming evidence (Schulz-Hardt et al., 2002). This phenomenon is particularly concerning in Internet forums that attract users with polarized views on public issues [Baek et al., 2012; see also section (Prior) Attitudes and Beliefs]. Dialogue can also elicit adversarial behaviors that undermine the potential benefits of two-way communication. Thus, under some conditions, the competitive epistemic goals of persuasion can trigger competitive interpersonal goals that foreclose transactive dialogue (Asterhan, 2013; Felton et al., 2019). When speakers confuse the two goals, they tend to repeat themselves without

elaborating their arguments, disagree without explaining why, and advance a barrage of arguments without addressing each other’s counterarguments (Felton et al., 2015b). On the other hand, two-way communication can also trigger collaborative interpersonal goals that undermine dialogue. Several studies suggest that face threat can lead speakers to avoid critical discussion (See, e.g., Asterhan, 2013; Felton et al., 2019). The phenomenon may be particularly problematic when speakers encounter disagreement unexpectedly during in-group dialogue. In these circumstances, speakers are more likely to prioritize group or interpersonal cohesion over engaging in critical discussion and transactive dialogue (Concannon et al., 2015).

## Diverging Opinions and Dialogic Engagement

Collaboratively achieving epistemic aims in dialogic argument depends substantially on the discourse partners’ efforts to deeply elaborate on and challenge their partner’s knowledge and arguments (e.g., Kuhn and Udell, 2003). In this context, the dialogic character (or two-way mode) of argumentation can support OESI through (a) enhancing the quality of argumentation and the use of evidence (Crowell and Kuhn, 2014; Mayweg-Paus and Macagno, 2016) and (b) the evaluation and reconciliation of diverging claims (Nussbaum and Edwards, 2011; Felton et al., 2015a). In an argumentative dialogue, a person is subject to the interlocutor’s scrutiny of her own position, which enhances her need to be more critical not only toward her own position but also the opposing one. In such dialogues, the reasons for preferring one point of view or one piece of evidence over another must be analyzed by taking a critical stance toward the presented evidence (Osborne et al., 2004). This challenge can only be addressed by drawing on more sufficient evidence and elaborating more and in greater depth on the different viewpoints and their backings (burden of proof, Walton and Macagno, 2007; Macagno and Walton, 2012).

There are several ways to address these potential challenges to effortful two-way, critical discussion. One effective strategy is to mitigate or de-emphasize competitive interpersonal goals by focusing attention on the epistemic aims of discourse. In the context of one-way communication, giving individuals specific instructions to generate reasons (Ferretti et al., 2000) or counter-arguments and rebuttals (Nussbaum and Kardash, 2005) have reduced my-side bias in writing when compared with instructions to persuade the audience. In two-way communication, focusing on collaborative epistemic aims (deliberation) as opposed to competitive epistemic aims (persuasion) in dialogue can lead to decreased interpersonal competitive behaviors and an increase in transactive dialogue (Felton et al., 2015b). These same conditions can also mitigate confirmation bias (Villarroel et al., 2016). However, it is important to note that in these studies, speakers were paired with someone who disagreed with them on the topic of discussion, and, therefore, the studies were designed to elicit the critical dialogue. But also, explicit expressions of disagreement in Youtube comment sections have the potential foster collaborative interaction (Dubovi and Tabak, 2020). What emerges in studies

that compare competitive and collaborative epistemic aims is an optimization problem. Dissent is a valuable component in overcoming motivated reasoning, particularly when measures are taken to reduce the risk of losing face (Schulz-Hardt et al., 2006). Thus, dialogue can be structured to explicitly make room for dissent (Schulz-Hardt et al., 2006). However, cognitive engagement is an important ingredient in such conversations (Kuhn and Lao, 1996), and focusing on transactive dialogue aimed at epistemic aims enhances the quality of reasoning. Individuals must hold themselves accountable to expressing disagreement when it arises to avoid quick consensus while simultaneously focusing on collaborative interpersonal goals to promote transactive dialogue (Asterhan, 2013; Thiebach et al., 2016). Ultimately, collaboratively achieving epistemic aims involves focusing dialogue on epistemic aims while threading the needle of interpersonal goals to produce a social-emotional context that fosters critical discussion.

## Reciprocity of Dialogic and Individual Engagement

Collaboratively dealing with diverging (or even conflicting) claims might hold potential for the development of individual epistemological understanding, as it brings into light the existence of multiple perspectives and can promote a more balanced integration of pro and counter arguments in one's line of reasoning. Empirical evidence shows that individuals—after engaging in intervention studies that allowed them to engage in both argumentation with peers through the computer and in reflective activities for an extended period of time—showed improvements in their ability to evaluate others' arguments and the evidence that supported their arguments (Iordanou and Constantinou, 2015; Mayweg-Paus et al., 2016b). Further, engaging in an online discourse with peers who held an opposing view (vs. the same view) led to different inquiry behavior during online discussions and to different gains in terms of argument skills. In particular, in Iordanou and Kuhn (2020) study, individuals who engaged online in discussions with peers holding an opposing view chose to search for information regarding the opposing alternative first when given the opportunity. In contrast, individuals who engaged in online discussions with same-side individuals preferred to seek information related to their own position. Differences were also observed in the prevalence and types of functional evidence-based argumentative idea units in individual final essays, and they favored the students who engaged online in discussions with peers holding an opposing view. Here, engagement in online discussions with individuals holding opposing or same-side views may have fostered an epistemological understanding of recognizing that the other is reasoning from a perspective different from one's own, but that this perspective is still worth examining (Iordanou and Kuhn, 2020), or that it is important to take a step back and re-evaluate one's own understanding (Forzani, 2019). However, most people typically show difficulties with being able “to construct fully justified dual-position arguments and to explain and reconcile differences between accounts” (Barzilai and Ka'adan, 2017, p. 223). Apparently,

recognizing multiperspectivity does not automatically mean one can apply sophisticated strategies when evaluating opposing views or arguments. Based on several empirical findings, Kuhn (2019) addresses this point by suggesting that understanding multivariable causality is a link toward evaluating and integrating multiple perspectives. Following this approach, OESI should include information- (or knowledge-) seeking activities for identifying and negotiating the multiple factors that can cause a phenomenon and to bring them into ongoing discussion.

In sum, dialogic engagement can take a number of forms depending on interactivity (one-way, two-way bounded, two-way unbounded), epistemic purposes (information seeking, discovery, inquiry, deliberation, negotiation, persuasion, eristic), and interpersonal dynamics of communication (collaborative, competitive). When we combine these variables, a complex array of permutations results. When individuals engage with others online about information, they gain access to critical dialogue that can enhance reasoning by focusing attention on the epistemic aims, ideals and reliable processes governing the use of information (Chinn et al., 2014). These epistemic concerns, when combined with critical dialogue, enhance reasoning about information and may even promote growth that transfers to individual engagement. However, to be successful in this endeavor, individuals must work collaboratively with others to examine their reasoning, even when epistemic aims are competitive.

## CONCLUSION AND IMPLICATIONS

In this paper, we have addressed conditions that may benefit, but also hinder effortful online engagement with scientific information (OESI). The Internet offers users immediate access to a wide variety of information on socio-scientific issues, and also allows for user agency and interactivity. Coiro (2015) argued that, in theory, the Internet is an ideal place to engage with (scientific) information to achieve deeper learning and understanding and—from a reading perspective—she presents strategies learners need to achieve such epistemic aims. However, it is not feasible to assume that readers will allocate their cognitive and motivational resources to the systematic processing of all information they find online regarding an issue of interest (Stadtler, 2017). As such, our review collects literature on cognitive and motivational processes that may help individuals overcome constraints and utilize affordances of scientific information in the online environment. Before concluding this literature review with a discussion of our heuristic model of OESI, we elaborate on how to foster individual and dialogic OESI in (higher) education.

We have discussed several context factors that may both constrain and motivate effortful OESI. Dealing with the uncertainty and complexity of scientific knowledge (emphasized in online information environments) is a challenge that might be hard to overcome (Kienhues et al., 2020). In consequence, individuals might often encounter conflicts—of newly acquired information with their own beliefs, between information sources, or between their beliefs and those of their dialogue partner.

As this review has shown, critical and deliberative scrutiny of information is central to OESI. That is, engagement should be directed at achieving epistemic aims while holding oneself and others accountable to appropriate epistemic ideals, at applying reliable processes in information search, selection, evaluation, integration, and in engaging in dialogue with others (in line with apt epistemic performance, Barzilai and Chinn, 2018). However, as we have discussed, cognitive biases (such as confirmation bias, motivated cognition, and competitive interpersonal goals) constrain otherwise reliable processes and may sometimes emerge under the guise of “critical thinking” (e.g., being critical toward experts’ claims also has become a rhetorical device of science skeptics). As such, in order to counter one-sided reasoning and argumentation, open-minded thinking is directly beneficial to effortful OESI, because it entails the willingness to hold up all views (including one’s own) under scrutiny of critical examination, even taking on the risk of identity threats, in order to follow through with epistemic aims (Taylor, 2016). Open-mindedness has been shown to not only benefit individual engagement with scientific information, such as knowledge about scientific issues and argument evaluation (Sinatra et al., 2003; Southerland and Sinatra, 2006; West et al., 2008), but also dialogic engagement in dialogue with others (Kuhn and Udell, 2003, 2007). We argue that it is through a balance of (individual or dialogic) critical examination of information and open-minded thinking that goal-directed and effortful OESI emerges. Sharon and Baram-Tsabari (2020) provide examples of several educational approaches to foster open-minded thinking, such as exposure to exemplars of virtues and practicing virtuous behaviors.

One environment that holds high potential for directly instructing critical and open-minded thinking by employing authentic search tasks is higher education classrooms. This environment is suitable for two reasons: First, students are already instructed to successfully deal with theories, models, evidence, and arguments within their discipline. Golan Duncan et al. (2018) identify that understanding experts’ evidentiary practices (how experts analyze, evaluate, interpret, and integrate evidence to derive and inform theories) and being able to rely on scientific evidence even though one’s own understanding of science is bounded (lay epistemic practices) are central for laypeople’s ability to deal with scientific evidence. Searching for information on socio-scientific topics (related somewhat to a learner’s area of expertise) is an ill-structured but solvable task, and it may also allow for reflection of the boundaries of students’ expertise, especially when they are granted the opportunity to engage in dialogue with students from different disciplinary backgrounds or with diverging views on the issue. Second, while scientific inquiry tasks, such as lab work, are important to achieve procedural knowledge in their own discipline, there is limited opportunity for learners to engage in understanding of the social processes that are used to create reliable knowledge; however, both scientific knowledge as well as digital media entail social affordances allowing for dialogic engagement in authentic search tasks.

We have previously argued that the two parts—individual and dialogic engagement—are reciprocal rather than separate

or sequential. While individual engagement might prepare the individual to engage in dialogue with others, such dialogic engagement might not only induce more individual engagement, but it may also foster skills and strategies needed for practices in individual engagement. Engaging learners in collaborative reasoning and argumentation about scientific information fosters individuals’ epistemic cognition (e.g., Iordanou, 2016; Fisher et al., 2017), but it also creates a space to collaboratively reflect and elaborate on online scientific information in two ways: First, individuals may discuss the quality of online information, and, second, they may critically reflect and reason collaboratively on the criteria that guided their evaluations (Barzilai and Chinn, 2018). Thus, dialogue with others entails the potential to reflect on both one’s own and others’ individual engagement practices (Mayweg-Paus et al., 2020). In particular, to promote the development of epistemological understanding in their students, educational scholars need to address searching to learn as an information-seeking activity within the process of argumentation as well as learning to search in the context of argumentative dialogue (Redfors et al., 2014), which works as a mechanism for critical reflection on sourcing strategies, information providers, and media, and may also serve knowledge co-construction (Dubovi and Tabak, 2020). In this way, online dialogue becomes not only a medium for the transfer of information but also a means by which we gain epistemological insight into the nature of information and its many uses in our communication with others.

In our heuristic model, two aspects are not discussed in further depth. First, we decided not to define the cognitive and behavioral manifestations of the practices of engagement. Several descriptive models and literature reviews exist that describe one or several of these practices and their interrelations in more detail (in the context of multiple documents comprehension: e.g., Rouet and Britt, 2011; List and Alexander, 2017a; 2017b; epistemic cognition: e.g., Chinn et al., 2011, 2014; digital literacy: e.g., Cho and Afflerbach, 2017; Coiro, 2020; functional scientific literacy: e.g., Tabak, 2015). Second, we have not described how individuals would achieve their epistemic aims (the outcome of engagement), and whether it is always feasible to assume that individuals would always achieve these through goal-directed and effortful OESI. While there are models outlining knowledge integration with prior information (Richter, 2015), integration of diverging sources (Braasch and Bråten, 2017), and knowledge co-construction through collaborative dialogue and argumentation (Asterhan and Schwarz, 2016; Iordanou et al., 2019a), further research should investigate how knowledge construction takes place in authentic online information search (in contrast to dealing with provided information—often in text form—in a research or classroom setting), especially taking into account online-specific constraints and affordances. Newer studies have increasingly included combinations of process and outcome variables to more comprehensively examine online engagement (e.g., Bråten et al., 2014a,b; List and Alexander, 2018; Kammerer et al., 2020), or even tested theoretical models linking cognition, motivation, and learning (e.g., Muis et al., 2015). Furthermore, goal-directed and effortful OESI requires metacognitive knowledge and skills, such as current updates of

the search task and monitoring of one's process (Barzilai and Chinn, 2018). A few studies have used think-aloud methods to investigate individuals' (epistemic) meta-cognition during online engagement (e.g., Mason et al., 2011; Barzilai and Zohar, 2012). While we think that such approaches should guide future empirical investigations into practices within OESI, our literature review also shows that there is ample research and evidence that future studies may build on.

Furthermore, our heuristic model of OESI could be extended in the future to include a larger variety of online information. Information technologies are constantly changing and with them users' access to information (e.g., on different devices, in different apps), information formats (e.g., interactive representations and video), information design (e.g., the use of nudges), and distribution (e.g., by algorithms, artificial intelligence). Hence, engagement with online information (and dealing with new and unique constraints and affordances) might already or will in the future involve even more steps, strategies, or skills (as well as many more variables mediating their effortful execution) than we have discussed in this review. Research on users' cognition and behavior in dealing with online scientific information—and especially on communication formats beyond informational text—is still sparse, but is growing in different disciplines

(e.g., psychology, educational sciences, communication science, information sciences). We hope that future research would strive toward further integration of theoretical ideas and models within and across disciplinary bounds.

## AUTHOR CONTRIBUTIONS

FH and EM-P conceived and presented the idea and structure of the article. FH took the lead in writing the manuscript. EM-P, MF, KI, and MZ each engaged in writing sections of the article. All authors provided feedback and ideas.

## ACKNOWLEDGMENTS

The idea for this article is the result of a workshop in June 2019 that received funding from two special interest groups (SIG) of the European Association for Research on Learning and Instruction (EARLI). We would like to thank the coordinators of the EARLI SIGs Inquiry Learning and Argumentation, Dialogue, and Reasoning, and the participants of the workshop. We furthermore thank Celeste Brenneka for language editing. We acknowledge support from the Open Access Publication Fund of the University of Muenster.

## REFERENCES

- Alexander, P. A., and The Disciplined Reading and Learning Research Laboratory (2012). Reading into the future: competence for the 21st century. *Educ. Psychol.* 47, 259–280. doi: 10.1080/00461520.2012.722511
- Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., and Ladwig, P. (2014). The “nasty effect”: online incivility and risk perceptions of emerging technologies. *J. Comput. Commun.* 19, 373–387. doi: 10.1111/jcc4.12009
- Ariel, Y., and Avidar, R. (2015). Information, interactivity, and social media. *Atlan. J. Commun.* 23, 19–30. doi: 10.1080/15456870.2015.972404
- Asterhan, C. S. C. (2013). “Epistemic and interpersonal dimensions of peer argumentation: conceptualization and quantitative assessment,” in *Affective Learning Together*, eds M. Baker, J. Andriessen and S. Jarvela (New York, NY: Routledge, Advances in Learning and Instruction series), 251–271
- Asterhan, C. S. C., and Schwarz, B. B. (2016). Argumentation for learning: well-trodden paths and unexplored territories. *Educ. Psychol.* 51, 164–187. doi: 10.1080/00461520.2016.1155458
- Baek, Y. M., Wojcieszak, M., and Delli Carpini, M. X. (2012). Online versus face-to-face deliberation: Who? Why? What? With what effects? *New Media Soc.* 14, 363–383. doi: 10.1177/1461444811413191
- Baram-Tsabari, A., and Schejter, A. M. (2019). “New media: a double-edged sword in support of Public Engagement with Science,” in *Learning In a Networked Society*, eds Y. Kali, A. Baram-Tsabari, and A. M. Schejter (Cham: Springer), 79–95. doi: 10.1007/978-3-030-14610-8\_5
- Barzilai, S., and Chinn, C. A. (2018). On the goals of epistemic education: promoting apt epistemic performance. *J. Learn. Sci.* 27, 353–389. doi: 10.1080/10508406.2017.1392968
- Barzilai, S., and Chinn, C. A. (2019). “Epistemic thinking in a networked society: contemporary challenges and educational responses,” in *Learning in a Networked Society*, eds Y. Kali, A. Baram-Tsabari, and A. M. Schejter (Cham: Springer). doi: 10.1007/978-3-030-14610-8\_4
- Barzilai, S., and Eshet-Alkalai, Y. (2015). The role of epistemic perspectives in comprehension of multiple author viewpoints. *Learn Instr.* 36, 86–103. doi: 10.1016/j.learninstruc.2014.12.003
- Barzilai, S., and Ka'adan, I. (2017). Learning to integrate divergent information sources: the interplay of epistemic cognition and epistemic metacognition. *Metacogn. Learn.* 12, 193–232. doi: 10.1007/s11409-016-9165-7
- Barzilai, S., and Zohar, A. (2012). Epistemic thinking in action: evaluating and integrating online sources. *Cogn. Instr.* 30, 39–85. doi: 10.1080/07370008.2011.636495
- Barzilai, S., and Zohar, A. (2016). “Epistemic (meta)cognition: ways of thinking about knowledge and knowing,” in *Handbook of Epistemic Cognition*, eds J. A. Greene, W. A. Sandoval, and I. Bråten (London: Routledge), 409–424
- Bender, J., Rothmund, T., Nauroth, P., and Gollwitzer, M. (2016). How moral threat shapes laypersons' engagement with science. *Pers. Soc. Psychol. Bull.* 42, 1723–1735. doi: 10.1177/0146167216671518
- Berkowitz, M. W., and Gibbs, J. C. (1983). Measuring the developmental features of moral discussion. *Merrill-Palmer Q.* 29, 399–410
- Berland, L. K., and Reiser, B. J. (2009). Making sense of argumentation and explanation. *Sci. Educ.* 93, 26–55. doi: 10.1002/sce.20286
- Boykoff, M. T., and Boykoff, J. M. (2004). Balance as bias: global warming and the US prestige press. *Glob. Environ. Chang.* 14, 125–136. doi: 10.1016/j.gloenvcha.2003.10.001
- Bråten, I., Anmarkrud, Ø., Brandmo, C., and Strømso, H. I. (2014a). Developing and testing a model of direct and indirect relationships between individual differences, processing, and multiple-text comprehension. *Learn. Instr.* 30, 9–24. doi: 10.1016/j.learninstruc.2013.11.002
- Bråten, I., Brante, E. W., and Strømso, H. I. (2018). What really matters: the role of behavioural engagement in multiple document literacy tasks. *J. Res. Read.* 41, 680–699. doi: 10.1111/1467-9817.12247
- Bråten, I., Ferguson, L. E., Anmarkrud, Ø., Strømso, H. I., and Brandmo, C. (2014b). Modeling relations between students' justification for knowing beliefs in science, motivation for understanding what they read in science, and science achievement. *Int. J. Educ. Res.* 66, 1–12. doi: 10.1016/j.ijer.2014.01.004
- Bråten, I., Salmerón, L., and Strømso, H. I. (2016). Who said that? Investigating the plausibility-induced source focusing assumption with norwegian undergraduate readers. *Contemp. Educ. Psychol.* 46, 253–262. doi: 10.1016/j.cedpsych.2016.07.004
- Bråten, I., and Strømso, H. I. (2006). Epistemological beliefs, interest, and gender as predictors of Internet-based learning activities. *Comput. Hum. Behav.* 22, 1027–1042. doi: 10.1016/j.chb.2004.03.026
- Bråten, I., and Strømso, H. I. (2010). Effects of task instruction and personal epistemology on the understanding of multiple texts about climate change. *Disc. Proc.* 47, 1–31. doi: 10.1080/01638530902959646



- Bråten, I., Strømsø, H. I., and Salmerón, L. (2011). Trust and mistrust when students read multiple information sources about climate change. *Learn. Instr.* 21, 180–192. doi: 10.1016/j.learninstruc.2010.02.002
- Braasch, J. L. G., and Bråten, I. (2017). The discrepancy-induced source comprehension (d-ISC) model: basic assumptions and preliminary evidence. *Educ. Psychol.* 52, 167–181. doi: 10.1080/00461520.2017.1323219
- Braasch, J. L. G., Rouet, J.-F., Vibert, N., and Britt, M. A. (2012). Readers' use of source information in text comprehension. *Mem. Cognit.* 40, 450–465. doi: 10.3758/s13421-011-0160-6
- Brante, E. W., and Strømsø, H. I. (2017). Sourcing in text comprehension: a review of interventions targeting sourcing skills. *Educ. Psychol. Rev.* 30, 773–799. doi: 10.1007/s10648-017-9421-7
- Breakstone, J., McGrew, S., Smith, M., Ortega, T., and Wineburg, S. (2018). Teaching students to navigate the online landscape. *Soc. Educ.* 82, 219–221. Available online at: <https://www.ingentaconnect.com/content/ncss/se/2018/00000082/00000004/art00010#expand/collapse>
- Britt, M. A., and Aglinskias, C. (2002). Improving students' ability to identify and use source information. *Cogn. Instr.* 20, 485–522. doi: 10.1207/s1532690XCI2004\_2
- Britt, M. A., Rouet, J. F., Blaum, D., and Millis, K. (2019). A reasoned approach to dealing with fake news. *Policy Insights Behav. Brain Sci.* 6, 94–101. doi: 10.1177/2372732218814855
- Bromme, R., and Goldman, S. R. (2014). The public's bounded understanding of science. *Educ. Psychol.* 49, 59–69. doi: 10.1080/00461520.2014.921572
- Bromme, R., Kienhues, D., and Porsch, T. (2010). "Who knows what and who can we believe? Epistemological beliefs are beliefs about knowledge (mostly) attained from others," in *Personal Epistemology in the Classroom: Theory, Research, and Implications for Practice*, eds L. D. Bendixen and F. C. Feucht (Cambridge: Cambridge University Press), 163–194. doi: 10.1017/CBO9780511691904.006
- Brossard, D., and Scheufele, D. A. (2013). Science, new media, and the public. *Science* 339, 40–41. doi: 10.1126/science.1232329
- Cappella, J. N., Kim, H. S., and Albarracín, D. (2015). Selection and transmission processes for information in the emerging media environment: psychological motives and message characteristics. *Media Psychol.* 18, 396–424. doi: 10.1080/15213269.2014.941112
- Chi, M. T. H. (2009). Active-constructive-interactive: a conceptual framework for differentiating learning activities. *Top. Cogn. Sci.* 1, 73–105. doi: 10.1111/j.1756-8765.2008.01005.x
- Chinn, C. A., and Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *J. Res. Sci. Teach.* 35, 623–654. doi: 10.1002/(SICI)1098-2736(199808)35:6<623::AID-TEA3>3.0.CO;2-O
- Chinn, C. A., Buckland, L. A., and Samarapungavan, A. (2011). Expanding the dimensions of epistemic cognition: arguments from philosophy and psychology. *Educ. Psychol.* 46, 141–167. doi: 10.1080/00461520.2011.587722
- Chinn, C. A., and Osborne, J. (2008). Student's questions: a potential resource for teaching and learning science. *Stud. Sci. Educ.* 44:1, 1–39. doi: 10.1080/030572607012828101
- Chinn, C. A., Rinehart, R. W., and Buckland, L. A. (2014). "Epistemic cognition and evaluating information: applying the AIR model of epistemic cognition," in *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences*, eds D. N. Rapp and J. L. G. Braasch (Cambridge, MA: MIT Press), 425–453
- Cho, B.-Y., and Afflerbach, P. (2017). "An evolving perspective of constructively responsive reading comprehension strategies in multilayered digital text environments," in *Handbook of Research on Reading Comprehension*, eds S. E. Israel (New York, NY: Guilford Press), 109–134.
- Coiro, J. (2015). "Purposeful, Critical, and Flexible. Vital Dimensions of Online Reading and Learning," in *Reading at a crossroads? Disjunctures and continuities in current conceptions and practice*, eds R. J. Spiro, M. DeSchrwyer, M. S. Morsink, P. M. Hagerman, and P. Thompson (New York, NY: Routledge), 53–64.
- Coiro, J. (2020). Toward a multifaceted heuristic of digital reading to inform assessment, research, practice, and policy. *Read. Res. Q.* doi: 10.1002/rrq.302. [Epub ahead of print].
- Coiro, J., Coscarelli, C., Maykel, C., and Forzani, E. (2015). Investigating criteria that seventh graders use to evaluate the quality of online information. *J. Adolesc. Adult Lit.* 59, 287–297. doi: 10.1002/jaal.448
- Concannon, S., Healey, P. G., and Purver, M. (2015). *Taking a Stance: A Corpus Study of Reported Speech*. Available online at: <https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/9041/CONCANNONShiftingopinions2015Published.pdf?sequence=2> (accessed April 19, 2020).
- Corner, A., and Hahn, U. (2009). Evaluating science arguments: evidence, uncertainty, and argument strength. *J. Exp. Psychol. Appl.* 15, 199–212. doi: 10.1037/a0016533
- Crowell, A., and Kuhn, D. (2014). Developing dialogic argumentation skills: a 3-year intervention study. *J. Cogn. Dev.* 15, 363–381. doi: 10.1080/15248372.2012.725187
- Deci, E. L. (1992). "The relation of interest to the motivation of behavior: a self-determination theory perspective," in *The Role of Interest in Learning and Development*, eds K. A. Renninger, S. Hidi, and A. Krapp (Mahwah, NJ: Lawrence Erlbaum Associates, Inc.), 43–70.
- DeNardis, L. (2014). *The Global War for Internet Governance*. New Haven, CT: Yale University Press.
- Dieckmann, N. F., Johnson, B. B., Gregory, R., Mayorga, M., Han, P. K. J., and Slovic, P. (2017). Public perceptions of expert disagreement: bias and incompetence or a complex and random world? *Public Underst. Sci.* 26, 325–338. doi: 10.1177/0963662515603271
- Dragojevic, M., Gasior, J., and Giles, H. (2016). "Accommodative Strategies as Core of the Theory," in *Communication Accommodation Theory*, ed. H. Giles (Cambridge: Cambridge University Press), 36–59. doi: 10.1017/CBO9781316226537.003
- Dubovi, I., and Tabak, I. (2020). An empirical analysis of knowledge co-construction in YouTube comments. *Comput. Educ.* 156:103939. doi: 10.1016/j.compedu.2020.103939
- Dunwoody, S., and Griffin, R. J. (2015). "Risk Information Seeking and Processing Model," in *The SAGE Handbook of Risk Communication*, eds H. Cho, T. Reimer, and K. A. McComas (Thousand Oaks, CA: SAGE Publications, Inc.), 102–116. doi: 10.4135/9781483387918.n14
- Eccles, J., and Wang, M.-T. (2012). "Part I Commentary: So what is student engagement anyway?," in *Handbook of Research on Student Engagement*, eds S. L. Christenson, A. L. Reschly, and C. Wylie (New York, NY: Springer US) 133–145. doi: 10.1007/978-1-4614-2018-7\_6
- Elgesem, D., Steskal, L., and Diakopoulos, N. (2015). Structure and content of the discourse on climate change in the blogosphere: the big picture. *Environ. Commun.* 9, 169–188. doi: 10.1080/17524032.2014.983536
- Evans, S. K., Pearce, K. E., Vitak, J., and Treem, J. W. (2016). Explicating affordances: a conceptual framework for understanding affordances in communication research. *J. Comput. Mediat. Commun.* 22, 35–52. doi: 10.1111/jcc4.12180
- Feinstein, N. W., and Waddington, D. I. (2020). Individual truth judgments or purposeful, collective sensemaking? Rethinking science education's response to the post-truth era. *Educ. Psychol.* 55, 155–166. doi: 10.1080/00461520.2020.1780130
- Felton, M., Crowell, A., Garcia-Mila, M., and Villarroel, C. (2019). Capturing deliberative argument: an analytic coding scheme for studying argumentative dialogue and its benefits for learning. *Learn. Cult. Soc. Interact.* doi: 10.1016/j.lcsi.2019.100350. [Epub ahead of print].
- Felton, M., Crowell, A., and Liu, T. (2015a). Arguing to agree: mitigating my-side bias through consensus-seeking dialogue. *Writ. Commun.* 32, 317–331. doi: 10.1177/0741088315590788
- Felton, M., Garcia-Mila, M., Villarroel, C., and Gilabert, S. (2015b). Arguing collaboratively: argumentative discourse types and their potential for knowledge building. *Br. J. Educ. Psychol.* 85, 372–386. doi: 10.1111/bjep.12078
- Ferguson, L. E., and Bråten, I. (2013). Student profiles of knowledge and epistemic beliefs: changes and relations to multiple-text comprehension. *Learn. Instr.* 25:49e61. doi: 10.1016/j.learninstruc.2012.11.003
- Ferguson, L. E., Bråten, I., and Strømsø, H. I. (2012). Epistemic cognition when students read multiple documents containing conflicting scientific evidence: a think-aloud study. *Learn. Instr.* 22, 103–120. doi: 10.1016/j.learninstruc.2011.08.002
- Ferretti, R. P., MacArthur, C. A., and Dowdy, N. S. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *J. Educ. Psychol.* 92, 694–702. doi: 10.1037/0022-0663.92.4.694

- Fischer, F., Bruhn, J., Gräsel, C., and Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. *Learn. Instr.* 12, 213–232. doi: 10.1016/S0959-4752(01)00005-6
- Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Dettmers, S., and Trautwein, U. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. *Front. Learn. Res.* 4, 28–45. doi: 10.14786/flr.v2i2.96
- Fischer, P., Jonas, E., Frey, D., and Schulz-Hardt, S. (2005). Selective exposure to information: the impact of information limits. *Eur. J. Soc. Psychol.* 35, 469–492. doi: 10.1002/ejsp.264
- Fisher, M., Goddu, M. K., and Keil, F. C. (2015). Searching for explanations: how the Internet inflates estimates of internal knowledge. *J. Exp. Psychol. Gen.* 144, 674–687. doi: 10.1037/xge0000070
- Fisher, M., Knobe, J., Strickland, B., and Keil, F. C. (2017). The influence of social interaction on intuitions of objectivity and subjectivity. *Cogn. Sci.* 41, 1119–1134. doi: 10.1111/cogs.12380
- Flemming, D., Feinkohl, I., Cress, U., and Kimmerle, J. (2015). Individual uncertainty and the uncertainty of science: the impact of perceived conflict and general self-efficacy on the perception of tentativeness and credibility of scientific information. *Front. Psychol.* 6:1859. doi: 10.3389/fpsyg.2015.01859
- Flemming, D., Kimmerle, J., Cress, U., and Sinatra, G. M. (2020). Research is tentative, but that's okay: overcoming misconceptions about scientific tentativeness through refutation texts. *Discl. Process.* 57, 17–35. doi: 10.1080/0163853X.2019.1629805
- Forzani, E. (2019). A three-tiered framework for proactive critical evaluation during online inquiry. *J. Adolesc. Adult Lit.* 63, 401–414. doi: 10.1002/jaal.1004
- Friedman, S. M., Dunwoody, S., and Rogers, C. L. (1999). *Communicating Uncertainty: Media Coverage of New and Controversial Science*. New York, NY: Routledge. doi: 10.4324/9781410601360
- Gao, F., Luo, T., and Zhang, K. (2012). Tweeting for learning: a critical analysis of research on microblogging in education published in 2008–2011. *Br. J. Educ. Technol.* 43, 783–801. doi: 10.1111/j.1467-8535.2012.01357.x
- Gierth, L., and Bromme, R. (2020). Beware of vested interests: epistemic vigilance improves reasoning about scientific evidence (for some people). *PLoS ONE* 15:e0231387. doi: 10.1371/journal.pone.0231387
- Golan Duncan, R., Chinn, C. A., and Barzilai, S. (2018). Grasp of evidence: problematizing and expanding the next generation science standards' conceptualization of evidence. *J. Reseach Sci. Teach.* 55, 907–937. doi: 10.1002/tea.21468
- Goldman, S. R., and Scardamalia, M. (2013). Managing, understanding, applying, and creating knowledge in the information age: next-generation challenges and opportunities. *Cogn. Instr.* 31, 255–269. doi: 10.1080/10824669.2013.773217
- Greene, J. A., Copeland, D. Z., and Deekens, V. M. (2020). A model of technology incidental learning effects. *Educ. Psychol. Rev.* doi: 10.1007/s10648-020-09575-5. [Epub ahead of print].
- Greving, H., and Sassenberg, K. (2015). Counter-regulation online: threat biases retrieval of information during Internet search. *Comput. Hum. Behav.* 50, 291–298. doi: 10.1016/j.chb.2015.03.077
- Griffin, R. J., Dunwoody, S., and Neuwirth, K. (1999). Proposed model of the relationship of risk information seeking and processing to the development of preventive behaviors. *Environ. Res.* 80, 230–245. doi: 10.1006/enrs.1998.3940
- Guthrie, J. T., and Klauda, S. L. (2016). "Engagement and Motivational Processes in Reading," in *Handbook of Individual Differences in Reading: Reader, Text, and Context*, ed. P. Afflerbach (New York, NY: Routledge), 41–53
- Haas, A., and Unkel, J. (2017). Ranking versus reputation: perception and effects of search result credibility. *Behav. Inf. Technol.* 36, 1285–1298. doi: 10.1080/0144929X.2017.1381166
- Hautala, J., Kiili, C., Kammerer, Y., Loberg, O., Hokkanen, S., and Leppänen, P. H. T. (2018). Sixth graders' evaluation strategies when reading Internet search results: an eye-tracking study. *Behav. Inf. Technol.* 37, 761–773. doi: 10.1080/0144929X.2018.1477992
- Head, A., and Eisenberg, M. B. (2010). Truth be told: how college students evaluate and use information in the digital age. *Proj. Inf. Lit. Prog. Rep.* 1–72. doi: 10.2139/ssrn.2281485
- Hendriks, F., and Kienhues, D. (2019). "Science understanding between scientific literacy and trust: contributions from psychological and educational research," in *Science Communication*, eds A. Leßmöllmann, M. Dascal, and T. Gloning (Berlin, Boston, MA: De Gruyter), 29–50. doi: 10.1515/978311025522-002
- Hendriks, F., Kienhues, D., and Bromme, R. (2015). Measuring laypeople's trust in experts in a digital age: the muenster epistemic trustworthiness inventory (METI). *PLoS ONE* 10:e0139309. doi: 10.1371/journal.pone.0139309
- Hendriks, F., Kienhues, D., and Bromme, R. (2016). Evoking vigilance: would you (dis)trust a scientist who discusses ethical implications of research in a science blog? *Public Underst. Sci.* 25, 992–1008. doi: 10.1177/0963662516646048
- Hendriks, F., Kienhues, D., and Bromme, R. (2020). Replication crisis = trust crisis? The effect of successful vs failed replications on laypeople's trust in researchers and research. *Public Underst. Sci.* 29, 270–288. doi: 10.1177/0963662520902383
- Hillgoss, B., and Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: construct, heuristics, and interaction in context. *Inf. Process. Manag.* 44, 1467–1484. doi: 10.1016/j.ipm.2007.10.001
- Hofer, B. K. (2004). Epistemological understanding as a metacognitive process: thinking aloud during online searching. *Educ. Psychol.* 39, 43–55. doi: 10.1207/s15326985ep3901\_5
- Hopkins, J. (2016). "The Concept of Affordances in Digital Media," in *Handbuch Soziale Praktiken und Digitale Alltagswelten*, eds H. Frieze, G. Rebane, M. Nolden, and M. Schreier (Wiesbaden: Springer Fachmedien Wiesbaden), 1–8. doi: 10.1007/978-3-658-08460-8\_67-1
- Hornikx, J. (2008). Comparing the actual and expected persuasiveness of evidence types: how good are lay people at selecting persuasive evidence? *Argumentation* 22, 555–569. doi: 10.1007/s10503-007-9067-6
- Howard, J. (2020). *Should You Wear a Mask? US Health Officials Re-examine Guidance Amid Coronavirus Crisis*. CNN. Available online at: <https://edition.cnn.com/2020/03/31/health/coronavirus-masks-experts-debate/index.html>.
- Iordanou, K. (2016). Developing epistemological understanding in scientific and social domains through argumentation. *Zeitschrift für Pädagogische Psychol.* 30, 109–119. doi: 10.1024/1010-0652/a000172
- Iordanou, K., and Constantinou, C. P. (2015). Supporting use of evidence in argumentation through practice in argumentation and reflection in the context of SOCRATES learning environment. *Sci. Educ.* 99, 282–311. doi: 10.1002/sce.21152
- Iordanou, K., and Kuhn, D. (2020). Contemplating the opposition: does a personal touch matter? *Discl. Process.* 57, 343–359. doi: 10.1080/0163853X.2019.1701918
- Iordanou, K., Kuhn, D., Matos, F., Shi, Y., and Hemberger, L. (2019a). Learning by arguing. *Learn. Instr.* 63, 101–207. doi: 10.1016/j.learninstruc.2019.05.004
- Iordanou, K., Muis, K. R., and Kendeou, P. (2019b). Epistemic perspective and online epistemic processing of evidence: developmental and domain differences. *J. Exp. Educ.* 87, 531–551. doi: 10.1080/00220973.2018.1482857
- Jucks, R., and Paus, E. (2013). Different words for the same concept: learning collaboratively from multiple documents. *Cogn. Instr.* 31, 227–254. doi: 10.1080/07370008.2013.769993
- Jucks, R., and Thon, F. M. (2017). Better to have many opinions than one from an expert? Social validation by one trustworthy source versus the masses in online health forums. *Comput. Hum. Behav.* 70, 375–381. doi: 10.1016/j.chb.2017.01.019
- Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgm. Decis. Mak.* 8, 407–424. doi: 10.2139/ssrn.2182588
- Kammerer, Y., Brand-Gruwel, S., and Jarodzka, H. (2018). The future of learning by searching the web: mobile, social, and multimodal. *Front. Learn. Res.* 6, 81–91. doi: 10.14786/flr.v6i2.343
- Kammerer, Y., and Gerjets, P. (2014). The role of search result position and source trustworthiness in the selection of web search results when using a list or a grid interface. *Int. J. Hum. Comput. Interact.* 30, 177–191. doi: 10.1080/10447318.2013.846790
- Kammerer, Y., Gottschling, S., and Bråten, I. (2020). The role of internet-specific justification beliefs in source evaluation and corroboration during web search on an unsettled socio-scientific issue. *J. Educ. Comput. Res.* doi: 10.1177/0735633120952731. [Epub ahead of print].
- Keil, F. C. (2008). Getting to the truth: grounding incomplete knowledge. *Brooklyn Law Rev.* 73, 1035–1052. Available online at: <https://brooklynworks.brooklaw.edu/blr/vol73/iss3/8>
- Keil, F. C., and Kominsky, J. F. (2013). Missing links in middle school: developing use of disciplinary relatedness in evaluating internet search results. *PLoS ONE* 8:e67777. doi: 10.1371/journal.pone.0067777

- Kienhues, D., Jucks, R., and Bromme, R. (2020). Sealing the gateways for post-truthism: reestablishing the epistemic authority of science. *Educ. Psychol.* 55, 144–154. doi: 10.1080/00461520.2020.1784012
- Kienhues, D., Thomm, E., and Bromme, R. (2018). “Specificity reloaded: how multiple layers of specificity influence reasoning in science argument evaluation,” in *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*, eds F. Fischer, C. A. Chinn, K. Engelmann, and J. Osborne (London: Taylor and Francis), 251–270. doi: 10.4324/9780203731826-14
- Kim, K.-S., Sin, S.-C. J., and Yoo-Lee, E. Y. (2014). Undergraduates’ use of social media as information sources. *Coll. Res. Libr.* 75:4. doi: 10.5860/crl.75.4.442
- Kimmerle, J., Flemming, D., Feinkohl, I., and Cress, U. (2015). How laypeople understand the tentativeness of medical research news in the media: an experimental study on the perception of information about deep brain stimulation. *Sci. Commun.* 37, 173–189. doi: 10.1177/1075547014556541
- Kind, P., and Osborne, J. (2017). Styles of scientific reasoning: a cultural rationale for science education? *Sci. Educ.* 101, 8–31. doi: 10.1002/sce.21251
- Knobloch-Westerwick, S., Mothes, C., and Polavin, N. (2020). Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. *Commun. Res.* 47, 104–124. doi: 10.1177/0093650217719596
- König, L., and Jucks, R. (2019). Hot topics in science communication: aggressive language decreases trustworthiness and credibility in scientific debates. *Public Underst. Sci.* 28, 401–416. doi: 10.1177/0963662519833903
- Koriat, A. (2012). “The subjective confidence in one’s knowledge and judgements: some metatheoretical considerations,” in *Foundations of Metacognition*, eds M. J. Beran, J. Brandl, J. Perner, and J. Proust (Oxford: Oxford University Press), 213–233. doi: 10.1093/acprof:oso/9780199646739.003.0014
- Koslowski, B., Marasia, J., Chelenza, M., and Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. *Cogn. Dev.* 23, 472–487. doi: 10.1016/j.cogdev.2008.09.007
- Krancher, O., Dibbern, J., and Meyer, P. (2018). How social media-enabled communication awareness enhances project team performance. *J. Assoc. Inf. Syst.* 19, 813–856. doi: 10.17705/1jais.00510
- Kuhn, D. (1999). A developmental model of critical thinking. *Educ. Res.* 28, 16–46. doi: 10.3102/0013189X028002016
- Kuhn, D. (2019). Critical thinking as discourse. *Hum. Dev.* 62, 146–164. doi: 10.1159/000500171
- Kuhn, D. (2020). Why is reconciling divergent views a challenge? *Curr. Dir. Psychol. Sci.* 29, 27–32. doi: 10.1177/0963721419885996
- Kuhn, D., and Iordanou, K. (2020). “Epistemology as a core dimension of cognitive development,” in *Reason, Bias, and Inquiry: New Perspectives from the Crossroads of Epistemology and Psychology*, eds D. Dunning and N. Ballantyne (Oxford: Oxford University Press)
- Kuhn, D., and Lao, J. (1996). Effects of evidence on attitudes: is polarization the norm? *Psychol. Sci.* 7, 115–120. doi: 10.1111/j.1467-9280.1996.tb00340.x
- Kuhn, D., and Udell, W. (2003). The development of argument skills. *Child Dev.* 74, 1245–1260. doi: 10.1111/1467-8624.00605
- Kuhn, D., and Udell, W. (2007). Coordinating own and other perspectives in argument. *Think. Reason.* 13, 90–104. doi: 10.1080/13546780600625447
- Lai, L., and Farbroth, A. (2014). What makes you click? The effect of question headlines on readership in computer-mediated communication. *Soc. Influ.* 9, 289–299. doi: 10.1080/15534510.2013.847859
- Le Bigot, L., and Rouet, J. F. (2007). The impact of presentation format, task assignment, and prior knowledge on students’ comprehension of multiple online documents. *J. Lit. Res.* 39, 445–470. doi: 10.1080/10862960701675317
- Leonardi, P. (2015). Ambient awareness and knowledge acquisition: using social media to learn “Who knows what” and “Who knows whom.” *MIS Quart.* 39, 747–762. doi: 10.2530/MISQ/2015/39.4.1
- Leshner, A. I. (2003). Public engagement with science. *Science* 299:977. doi: 10.1126/science.299.5609.977
- Leu, D. J., Kinzer, C. K., Coiro, J., Castek, J., and Henry, L. A. (2013). “New literacies: a dual-level theory of the changing nature of literacy, instruction, and assessment,” in *Theoretical Models and Processes of Reading*, eds D. E. Alvermann, N. J. Unrau, and R. B. Ruddell (Newark, DE: International Reading Association), 1150–1181. doi: 10.1598/0710.42
- Lewandowsky, S., Cook, J., Fay, N., and Gignac, G. E. (2019). Science by social media: attitudes towards climate change are mediated by perceived social consensus. *Mem. Cognit.* 47, 1445–1456. doi: 10.3758/s13421-019-00948-y
- Lewicki, R. J., Saunders, D. M., and Minton, J. W. (2001). *Essentials of Negotiation*. New York, NY: McGraw-Hill/Irwin.
- List, A., and Alexander, P. A. (2017a). Analyzing and integrating models of multiple text comprehension. *Educ. Psychol.* 52, 143–147. doi: 10.1080/00461520.2017.1328309
- List, A., and Alexander, P. A. (2017b). Cognitive affective engagement model of multiple source use. *Educ. Psychol.* 52, 182–199. doi: 10.1080/00461520.2017.1329014
- List, A., and Alexander, P. A. (2018). Corroborating students’ self-reports of source evaluation. *Behav. Inf. Technol.* 37, 198–216. doi: 10.1080/0144929X.2018.1430849
- List, A., Alexander, P. A., and Stephens, L. A. (2017). Trust but verify: examining the association between students’ sourcing behaviors and ratings of text trustworthiness. *Discl. Process.* 54, 83–104. doi: 10.1080/0163853X.2016.1174654
- List, A., Grossnickle, E. M., and Alexander, P. A. (2016a). Profiling students’ multiple source use by question type. *Read. Psychol.* 37, 753–797. doi: 10.1080/02702711.2015.1111962
- List, A., Grossnickle, E. M., and Alexander, P. A. (2016b). Undergraduate students’ justifications for source selection in a digital academic context. *J. Educ. Comput. Res.* 54, 22–61. doi: 10.1177/0735633115606659
- Lombardi, D., Nussbaum, E. M., and Sinatra, G. M. (2016). Plausibility judgments in conceptual change and epistemic cognition. *Educ. Psychol.* 51, 35–56. doi: 10.1080/00461520.2015.1113134
- Macagno, F., and Walton, D. (2012). Presumptions in legal argumentation. *Ratio Juris.* 25, 271–300. doi: 10.1111/j.1467-9337.2012.00514.x
- Macedo-Rouet, M., Potocki, A., Scharer, L., Ros, C., Stadler, M., Salmerón, L., et al. (2019a). How good is this page? Benefits and limits of prompting on adolescents’ evaluation of web information quality. *Read. Res. Q.* 54, 299–321. doi: 10.1002/rrq.241
- Macedo-Rouet, M., Salmerón, L., Ros, C., Pérez, A., Stadler, M., and Rouet, J. F. (2019b). Are frequent users of social network sites good information evaluators? An investigation of adolescents’ sourcing abilities / ‘Son los usuarios frecuentes de las redes sociales evaluadores competentes? Un estudio de las habilidades de los adolescentes par. *J. Study Educ. Dev.* 43, 101–138. doi: 10.1080/02103702.2019.1690849
- Maier, J., Richter, T., Nauroth, P., and Gollwitzer, M. (2018). For me or for them: How in-group identification and beliefs influence the comprehension of controversial texts. *J. Res. Read.* 41, S48–S65. doi: 10.1111/1467-9817.12132
- MaKinster, J. G., Beghetto, R. A., and Plucker, J. A. (2002). Why can’t I find newton’s third law? Case studies of students’ use of the Web as a science resource. *J. Sci. Educ. Technol.* 11, 155–172. doi: 10.1023/A:1014617530297
- Mason, L., Ariasi, N., and Boldrin, A. (2011). Epistemic beliefs in action: spontaneous reflections about knowledge and knowing during online information searching and their influence on learning. *Learn. Instr.* 21, 137–151. doi: 10.1016/j.learninstruc.2010.01.001
- Mason, L., Boldrin, A., and Ariasi, N. (2010a). Epistemic metacognition in context: evaluating and learning online information. *Metacogn. Learn.* 5, 67–90. doi: 10.1007/s11409-009-9048-2
- Mason, L., Boldrin, A., and Ariasi, N. (2010b). Searching the Web to learn about a controversial topic: are students epistemically active? *Instr. Sci.* 38, 607–633. doi: 10.1007/s11251-008-9089-y
- Mayweg-Paus, E., and Jucks, R. (2018). Conflicting evidence or conflicting opinions? Two-sided expert discussions contribute to experts’ trustworthiness. *J. Lang. Soc. Psychol.* 37, 203–223. doi: 10.1177/0261927X17716102
- Mayweg-Paus, E., and Macagno, F. (2016). How dialogic settings influence evidence use in adolescent students. *Zeitschrift für Pädagogische Psychol.* 30, 121–132. doi: 10.1024/1010-0652/a000171
- Mayweg-Paus, E., Macagno, F., and Kuhn, D. (2016a). Developing argumentation strategies in electronic dialogs: is modeling effective? *Discl. Process.* 53, 280–297. doi: 10.1080/0163853X.2015.1040323
- Mayweg-Paus, E., Thiebach, M., and Jucks, R. (2016b). Let me critically question this!—Insights from a training study on the role of questioning on argumentative discourse. *Int. J. Educ. Res.* 79, 195–210. doi: 10.1016/j.ijer.2016.05.017
- Mayweg-Paus, E., Zimmermann, M., Le, N.T., and Pinkwart, N. (2020). A review of technologies for collaborative online information seeking:



- on the contribution of collaborative argumentation. *Educ. Inf. Technol.* doi: 10.1007/s10639-020-10345-7. [Epub ahead of print].
- McGrew, S., Breakstone, J., Ortega, T., Smith, M., and Wineburg, S. (2018). Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory Res. Soc. Educ.* 46, 165–193. doi: 10.1080/00933104.2017.1416320
- Mercer, D. (2018). Why Popper can't resolve the debate over global warming: problems with the uses of philosophy of science in the media and public framing of the science of global warming. *Public Underst. Sci.* 27, 139–152. doi: 10.1177/0963662516645040
- Metzger, M. J., and Flanagin, A. J. (2013). Credibility and trust of information in online environments: the use of cognitive heuristics. *J. Pragmat.* 59, 210–220. doi: 10.1016/j.pragma.2013.07.012
- Molina, M. D., Sundar, S. S., Le, T., and Lee, D. (2019). "Fake News" is not simply false information: a concept explication and taxonomy of online content. *Am. Behav. Sci.* doi: 10.1177/0002764219878224. [Epub ahead of print].
- Muis, K. R. (2007). The role of epistemic beliefs in self-regulated learning. *Educ. Psychol.* 42, 173–190. doi: 10.1080/00461520701416306
- Muis, K. R., Pekrun, R., Sinatra, G. M., Azevedo, R., Trevors, G., Meier, E., et al. (2015). The curious case of climate change: testing a theoretical model of epistemic beliefs, epistemic emotions, and complex learning. *Learn. Instr.* 39, 168–183. doi: 10.1016/j.learninstruc.2015.06.003
- National Research Council (2012). *A Framework for K-12 Science Education. Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Research Council.
- National Science Board (2018). *Science and Engineering Indicators 2018. NSB-2018-1*. Alexandria, VA: National Science Foundation.
- Nauroth, P., Gollwitzer, M., Bender, J., and Rothmund, T. (2015). Social identity threat motivates science-discrediting online comments. *PLoS ONE* 10:e0117476. doi: 10.1371/journal.pone.0117476
- Nauroth, P., Gollwitzer, M., Kozuchowski, H., Bender, J., and Rothmund, T. (2017). The effects of social identity threat and social identity affirmation on laypersons' perception of scientists. *Public Underst. Sci.* 26, 754–770. doi: 10.1177/0963662516631289
- Noroozi, O., Weinberger, A., Biemans, H. J., Mulder, M., and Chizari, M. (2012). Argumentation-based computer supported collaborative learning (ABCSCCL). A synthesis of 15 years of research. *Educ. Res. Rev.* 7, 79–106. doi: 10.1016/j.edurev.2011.11.006
- Nussbaum, E. M., and Edwards, O. V. (2011). Critical questions and argument stratagems: a framework for enhancing and analyzing students' reasoning practices. *J. Learn. Sci.* 20, 443–488. doi: 10.1080/10508406.2011.564567
- Nussbaum, E. M., and Kardash, C. M. (2005). The effects of goal instructions and text on the generation of counterarguments during writing. *J. Educ. Psychol.* 97:157. doi: 10.1037/0022-0663.97.2.157
- OECD (2017). *PISA 2015 Assessment and Analytical Framework*. Paris: OECD Publishing.
- Oeldorf-Hirsch, A., and DeVoss, C. L. (2020). Who posted that story? Processing layered sources in facebook news posts. *J. Mass Commun. Q.* 97, 141–160. doi: 10.1177/1077699019857673
- Oreskes, N. (2007). "The scientific consensus on climate change: How do we know we're not wrong," in *Climate Change: What It Means for Us, Our Children, and Our Grandchildren*, eds J. F. C. DiMento and P. Doughman (Cambridge, MA: MIT Press), 65–99
- Oreskes, N., and Conway, E. M. (2011). *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. New York, NY: Bloomsbury Press.
- Osborne, J. (2010). Arguing to learn in science: the role of collaborative, critical discourse. *Science* 328, 463–466. doi: 10.1126/science.1183944
- Osborne, J., Erduran, S., and Simon, S. (2004). Enhancing the quality of argumentation in school science. *J. Res. Sci. Teach.* 41, 994–1020. doi: 10.1002/tea.20035
- Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., and Granka, L. (2007). In google we trust: users' decisions on rank, position, and relevance. *J. Comput. Commun.* 12, 801–823. doi: 10.1111/j.1083-6101.2007.00351.x
- Paul, J., Stadler, M., and Bromme, R. (2019). Effects of a sourcing prompt and conflicts in reading materials on elementary students' use of source information. *Discl. Process.* 56, 155–169. doi: 10.1080/0163853X.2017.1402165
- Pavelle, S., and Wilkinson, C. (2020). Into the digital wild: utilizing twitter, instagram, you tube, and facebook for effective science and environmental communication. *Front. Commun.* 5, 1–8. doi: 10.3389/fcomm.2020.575122
- Pintrich, P. R. (2000). "The role of goal orientation in self-regulated learning," in *Handbook of Self-Regulation*, eds M. Boekaerts, P. R. Pintrich, and M. Zeidner (San Diego, CA: Academic Press), 451–502
- Rabinovich, A., and Morton, T. A. (2012). Unquestioned answers or unanswered questions: beliefs about science guide responses to uncertainty in climate change risk communication. *Risk Anal.* 32, 992–1002. doi: 10.1111/j.1539-6924.2012.01771.x
- Rabinovich, A., Morton, T. A., and Birney, M. E. (2012). Communicating climate science: the role of perceived communicator's motives. *J. Environ. Psychol.* 32, 11–18. doi: 10.1016/j.jenvp.2011.09.002
- Rafaeli, S. (1988). "Interactivity: From new media to communication," in *Advancing Communication Science: Merging Mass and Interpersonal Process*, eds R. P. Hawkins, J. M. Wiemann, and S. Pingree (Newbury Park, CA: Sage), 110–134
- Rafaeli, S., and Ariel, Y. (2007). "Assessing interactivity in computer-mediated research," in *The Oxford handbook of Internet Psychology*, eds A. N. Joinson, K. Y. A. McKenna, T. Postmes, and U.-D. Reips (Oxford: Oxford University Press), 71–89.
- Rapanta, C., and Christodoulou, A. (2019). Walton's types of argumentation dialogues as classroom discourse sequences. *Learn. Cult. Soc. Interact.* doi: 10.1016/j.lcsi.2019.100352. [Epub ahead of print].
- Redfors, A., Hansson, L., Kyza, E. A., Nicolaidou, I., Asher, I., Tabak, I., et al. (2014). "CoReflect: web-based inquiry learning environments on socio-scientific Issues," in *Topics and Trends in Current Science Education*, eds C. Bruguère, A. Tiberghien, and P. Clément (Dordrecht: Springer), 553–566. doi: 10.1007/978-94-007-7281-6\_34
- Reif, A., Kneisel, T., Schäfer, M., and Taddicken, M. (2020). Why are scientific experts perceived as trustworthy? Emotional assessment within tv and youtube videos. *Media Commun.* 8, 191–205. doi: 10.17645/mac.v8i1.2536
- Richter, T. (2015). Validation and comprehension of text information: two sides of the same coin. *Discl. Process.* 52, 337–355. doi: 10.1080/0163853X.2015.1025665
- Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *J. Am. Soc. Inf. Sci. Technol.* 53, 141–161. doi: 10.1002/asi.10017
- Risko, E. F., Ferguson, A. M., and McLean, D. (2016). On retrieving information from external knowledge stores: feeling-of-findability, feeling-of-knowing and Internet search. *Comput. Human Behav.* 65, 534–543. doi: 10.1016/j.chb.2016.08.046
- Rodriguez, F., Rhodes, R. E., Miller, K. F., and Shah, P. (2016). Examining the influence of anecdotal stories and the interplay of individual differences on reasoning. *Think. Reason.* 22, 274–296. doi: 10.1080/13546783.2016.1139506
- Rothmund, T., Bender, J., Nauroth, P., and Gollwitzer, M. (2015). Public concerns about violent video games are moral concerns-how moral threat can make pacifists susceptible to scientific and political claims against violent video games. *Eur. J. Soc. Psychol.* 45, 769–783. doi: 10.1002/ejsp.2125
- Rouet, J.-F., and Britt, M. A. (2011). "Relevance processes in multiple document comprehension," in *Text Relevance and Learning from Text*, eds M. T. McCrudden, J. P. Magliano, and G. Schraw (Greenwich, CT: Information Age), 19–52
- Rouet, J.-F., Britt, M. A., and Durik, A. M. (2017). RESOLV: readers' representation of reading contexts and tasks. *Educ. Psychol.* 52, 200–215. doi: 10.1080/00461520.2017.1329015
- Rouet, J.-F., Ros, C., Goumi, A., Macedo-Rouet, M., and Dinet, J. (2011). The influence of surface and deep cues on primary and secondary school students' assessment of relevance in Web menus. *Learn. Instr.* 21, 205–219. doi: 10.1016/j.learninstruc.2010.02.007
- Salmerón, L., Cañas, J. J., and Fajardo, I. (2005). Are expert users always better searchers? Interaction of expertise and semantic grouping in hypertext search tasks. *Behav. Inf. Technol.* 24, 471–475. doi: 10.1080/0144329042000320018
- Salmerón, L., Gil, L., and Bråten, I. (2017). Effects of reading real versus printout versions of multiple documents on students' sourcing and integrated understanding. *Contemp. Educ. Psychol.* 52, 25–35. doi: 10.1016/j.cedpsych.2017.12.002
- Salmerón, L., Gil, L., Bråten, I., and Strømsø, H. (2010). Comprehension effects of signalling relationships between documents in search engines. *Comput. Human Behav.* 26, 419–426. doi: 10.1016/j.chb.2009.11.013



- Salmerón, L., Kammerer, Y., and García-Carrión, P. (2013). Searching the web for conflicting topics: page and user factors. *Comput. Hum. Behav.* 29, 2161–2171. doi: 10.1016/j.chb.2013.04.034
- Sandoval, W. A., Greene, J. A., and Bråten, I. (2016). Understanding and promoting thinking about knowledge: origins, issues, and future directions of research on epistemic cognition. *Rev. Res. Educ.* 40, 457–496. doi: 10.3102/0091732X16669319
- Scharrer, L., Bromme, R., Britt, M. A., and Stadler, M. (2012). The seduction of easiness: how science depictions influence laypeople's reliance on their own evaluation of scientific information. *Learn. Instr.* 22, 231–243. doi: 10.1016/j.learninstruc.2011.11.004
- Scharrer, L., Rupieler, Y., Stadler, M., and Bromme, R. (2017). When science becomes too easy: science popularization inclines laypeople to underrate their dependence on experts. *Public Underst. Sci.* 26, 1003–1018. doi: 10.1177/0963662516680311
- Scharrer, L., Stadler, M., and Bromme, R. (2019). Biased recipients encounter biased sources: effect of ethical (dis-)agreement between recipient and author on evaluating scientific claims. *Appl. Cogn. Psychol.* 33, 1165–1177. doi: 10.1002/acp.3563
- Scheufele, D. A., and Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proc. Natl. Acad. Sci. U.S.A.* 116, 7662–7669. doi: 10.1073/pnas.1805871115
- Schiefele, U. (2009). "Situational and individual interest," in *Handbook of Motivation at School*, eds K. Wentzel and A. Wigfield (New York, NY: Routledge), 197–222
- Schulz-Hardt, S., Brodbeck, F. C., Mojzisch, A., Kerschreiter, R., and Frey, D. (2006). Group decision making in hidden profile situations: Dissent as a facilitator for decision quality. *J. Pers. Soc. Psychol.* 91, 1080–1093. doi: 10.1037/0022-3514.91.6.1080
- Schulz-Hardt, S., Jochims, M., and Frey, D. (2002). Productive conflict in group decision making: genuine and contrived dissent as strategies to counteract biased information seeking. *Organ. Behav. Hum. Decis. Process.* 88, 563–586. doi: 10.1016/S0749-5978(02)00001-8
- Shah, P., Michal, A., Ibrahim, A., Rhodes, R., and Rodriguez, F. (2017). What makes everyday scientific reasoning so challenging? *Psychol. Learn. Motiv.* 66, 251–299. doi: 10.1016/bs.plm.2016.11.006
- Shapiro, A., and Niederhauser, D. (2004). "Learning from hypertext: research issues and findings," in *Handbook of Research on Educational Communications and Technology*, ed. D. H. Jonassen (Mahwah, NJ: Erlbaum), 605–620
- Sharon, A. J., and Baram-Tsabari, A. (2020). Can science literacy help individuals identify misinformation in everyday life? *Sci. Educ.* 104, 873–894. doi: 10.1002/sce.21581
- Sinatra, G. M., and Lombardi, D. (2020). Evaluating sources of scientific evidence and claims in the post-truth era may require reappraising plausibility judgments. *Educ. Psychol.* 55, 120–131. doi: 10.1080/00461520.2020.1730181
- Sinatra, G. M., Southerland, S. A., McConaughy, F., and Demastes, J. W. (2003). Intentions and beliefs in students' understanding and acceptance of biological evolution. *J. Res. Sci. Teach.* 40, 510–528. doi: 10.1002/tea.10087
- Song, Y., and Ferretti, R. P. (2013). Teaching critical questions about argumentation through the revising process: effects of strategy instruction on college students' argumentative essays. *Read. Writ.* 26, 67–90. doi: 10.1007/s11145-012-9381-8
- Southerland, S. A., and Sinatra, G. M. (2006). "The shifting roles of acceptance and dispositions in understanding biological evolution," in *Beyond Cartesian Dualism*, eds W. W. Cobern, K. Tobin, H. Brown-Acquay, M. Espinet, G. Irzik, O. Jegede, et al. (Berlin; Heidelberg: Springer), 69–78. doi: 10.1007/1-4020-3808-9\_6
- Sparrow, B., Liu, J., and Wegner, D. M. (2011). Google effects on memory: cognitive consequences of having information at our fingertips. *Science* 333, 776–778. doi: 10.1126/science.1207745
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., et al. (2010). Epistemic vigilance. *Mind Lang.* 25, 359–393. doi: 10.1111/j.1468-0017.2010.01394.x
- Stadler, M. (2017). The art of reading in a knowledge society: commentary on the special issue on models of multiple text comprehension. *Educ. Psychol.* 52, 225–231. doi: 10.1080/00461520.2017.1322969
- Stadler, M., and Bromme, R. (2014). "The content-source integration model: a taxonomic description of how readers comprehend conflicting scientific information," in *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and Educational Sciences*, eds D. N. Rapp and J. L. G. Braasch (Cambridge, MA: MIT Press), 379–402
- Stadler, M., Scharrer, L., Brummernhenrich, B., and Bromme, R. (2013). Dealing with uncertainty: Readers' memory for and use of conflicting information from science texts as function of presentation format and source expertise. *Cogn. Instr.* 31, 130–150. doi: 10.1080/07370008.2013.769996
- Stang Lund, E., Bråten, I., Brandmo, C., Brante, E. W., and Strømsø, H. I. (2019). Direct and indirect effects of textual and individual factors on source-content integration when reading about a socio-scientific issue. *Read. Writ.* 32, 335–356. doi: 10.1007/s11145-018-9868-z
- Strømsø, H. I., Bråten, I., and Britt, M. A. (2011). Do students' beliefs about knowledge and knowing predict their judgement of texts' trustworthiness? *Educ. Psychol.* 31, 177–206. doi: 10.1080/01443410.2010.538039
- Strømsø, H. I., Bråten, I., Britt, M. A., and Ferguson, L. E. (2013). Spontaneous sourcing among students reading multiple documents. *Cogn. Instr.* 31, 176–203. doi: 10.1080/07370008.2013.769994
- Strømsø, H. I., Bråten, I., and Stenseth, T. (2017). The role of students' prior topic beliefs in recall and evaluation of information from texts on socio-scientific issues. *Nord. Psychol.* 69, 127–142. doi: 10.1080/19012276.2016.1198270
- Sundar, S. S. (2008). "The MAIN model: a heuristic approach to understanding technology effects on credibility," in *Digital Media, Youth, and Credibility*, eds M. J. Metzger and A. J. Flanagin (Cambridge, MA: MIT Press), 73–100
- Tabak, I. (2015). "Functional scientific literacy: seeing the science within the words and across the web," in *Handbook of Educational Psychology*, eds L. Corno and E. M. Anderman (New York, NY: Routledge), 269–280.
- Taraborelli, D. (2008). "How the Web is changing the way we trust," in *Current Issues in Computing and Philosophy*, eds K. Waelbers, A. Briggle, and P. Brey (IOS Press, Amsterdam), 194–204.
- Taylor, R. M. (2016). Open-mindedness: an intellectual virtue in the pursuit of knowledge and understanding. *Educ. Theory* 66, 599–618. doi: 10.1111/edth.12201
- Teasley, S. D. (1997). "Talking about reasoning: how important is the peer in peer collaboration?" in *Discourse, Tools and Reasoning: Essays on Situated Cognition*, eds L. B. Resnick, R. Säljö, C. Pontecorvo, and B. Burge (Berlin: Springer), 361–384. doi: 10.1007/978-3-662-03362-3\_16
- Thiebach, M., Mayweg-Paus, E., and Jucks, R. (2016). Better to agree or disagree? The role of critical questioning and elaboration in argumentative discourse. *Zeitschrift für Pädagogische Psychologie* 30, 133–149. doi: 10.1024/1010-0652/a000174
- Thomm, E., Barzilai, S., and Bromme, R. (2017). Why do experts disagree? The role of conflict topics and epistemic perspectives in conflict explanations. *Learn. Instr.* 52, 15–26. doi: 10.1016/j.learninstruc.2017.03.008
- Thomm, E., and Bromme, R. (2012). "It should at least seem scientific!" textual features of "scientificness" and their impact on lay assessments of online information. *Sci. Educ.* 96, 187–211. doi: 10.1002/sce.20480
- Thompson, W. B., Garry, A., Taylor, J., and Radell, M. L. (2020). Is one study as good as three? College graduates seem to think so, even if they took statistics classes. *Psychol. Learn. Teach.* 19, 143–160. doi: 10.1177/1475725719877590
- Thon, F. M., and Jucks, R. (2017). Believing in expertise: how authors' credentials and language use influence the credibility of online health information. *Health Commun.* 32, 828–836. doi: 10.1080/10410236.2016.1172296
- Tuler, S. (2000). Forms of talk in policy dialogue: distinguishing between adversarial and collaborative discourse. *J. Risk Res.* 3, 1–17. doi: 10.1080/136698700376671
- van der Bles, A. M., van der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L., et al. (2019). Communicating uncertainty about facts, numbers and science. *R. Soc. Open Sci.* 6:181870. doi: 10.1098/rsos.181870
- van der Linden, S. L. D., Leiserowitz, A. A., Feinberg, G. D., and Maibach, E. W. (2015). The scientific consensus on climate change as a gateway belief: experimental evidence. *PLoS ONE* 10:e0118489. doi: 10.1371/journal.pone.0118489
- van Strien, J. L. H., Kammerer, Y., Brand-Gruwel, S., and Boshuizen, H. P. A. (2016). How attitude strength biases information processing and evaluation on the web. *Comput. Human Behav.* 60, 245–252. doi: 10.1016/j.chb.2016.02.057
- Villarroel, C., Felton, M., and Garcia-Mila, M. (2016). Arguing against confirmation bias: the effect of argumentative discourse goals on the use of

- disconfirming evidence in written argument. *Int. J. Educ. Res.* 79, 167–179. doi: 10.1016/j.ijer.2016.06.009
- Vogl, E., Pekrun, R., Murayama, K., and Loderer, K. (2020). Surprised–curious–confused: epistemic emotions and knowledge exploration. *Emotion* 20, 625–641. doi: 10.1037/emo0000578
- Walton, D. (2010). “Types of dialogue and burdens of proof,” in *Computational Models of Argument: Proceedings of COMMA*, eds P. Baroni, F. Cerutti, M. Giacomin, and G. R. Simari (Amsterdam: IOS Press), 13–24
- Walton, D., and Macagno, F. (2007). The fallaciousness of threats: character and ad baculum. *Argumentation* 21, 63–81. doi: 10.1007/s10503-006-9018-7
- West, R. F., Toplak, M. E., and Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: associations with cognitive ability and thinking dispositions. *J. Educ. Psychol.* 100, 930–941. doi: 10.1037/a0012842
- Wineburg, S., and McGrew, S. C. (2019). Lateral reading and the nature of expertise: reading less and learning more when evaluating digital information. *Teach. Coll. Rec.* 121. Available online at: <https://www.tcrecord.org/content.asp?contentid=22806>
- Winter, S., and Krämer, N. C. (2012). Selecting science information in Web 2.0: how source cues, message sidedness, and need for cognition influence users’ exposure to blog posts. *J. Comput.-Mediat. Commun.* 18, 80–96. doi: 10.1111/j.1083-6101.2012.01596.x
- Winter, S., and Krämer, N. C. (2014). A question of credibility - effects of source cues and recommendations on information selection on news sites and blogs. *Communications* 39, 435–456. doi: 10.1515/commun-2014-0020
- Winter, S., Metzger, M. J., and Flanagin, A. J. (2016). Selective use of news cues: a multiple-motive perspective on information selection in social media environments. *J. Commun.* 66, 669–693. doi: 10.1111/jcom.12241
- Wissinger, D. R., and De La Paz, S. (2016). Effects of critical discussions on middle school students’ written historical arguments. *J. Educ. Psychol.* 108:43. doi: 10.1037/edu0000043
- Yang, Z. J., and Kahlor, L. A. (2013). What, me worry? The role of affect in information seeking and avoidance. *Sci. Commun.* 35, 189–212. doi: 10.1177/1075547012441873
- Zeidler, D. L. (2014). “Socioscientific issues as curriculum emphasis: theory, research and practice,” in *Handbook of Research on Science Education*, ed. N. G. Lederman (New York, NY: Routledge), 697–726
- Zimmermann, M., and Jucks, R. (2018). How experts’ use of medical technical jargon in different types of online health forums affects perceived information credibility: randomized experiment with laypersons. *J. Med. Internet Res.* 20:e30. doi: 10.2196/jmir.8346

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Hendriks, Mayweg-Paus, Felton, Iordanou, Jucks and Zimmermann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Assessing University Students' Critical Online Reasoning Ability: A Conceptual and Assessment Framework With Preliminary Evidence

Dimitri Molerov<sup>1\*</sup>, Olga Zlatkin-Troitschanskaia<sup>2</sup>, Marie-Theres Nagel<sup>2</sup>, Sebastian Brückner<sup>2</sup>, Susanne Schmidt<sup>2</sup> and Richard J. Shavelson<sup>3</sup>

<sup>1</sup> Department of Research Methods in Education, Humboldt University of Berlin, Berlin, Germany, <sup>2</sup> Department of Business and Economics Education, Johannes Gutenberg University, Mainz, Germany, <sup>3</sup> Stanford Graduate School of Education, Stanford University, Palo Alto, CA, United States

## OPEN ACCESS

### Edited by:

Douglas F. Kauffman,  
Medical University of the  
Americas – Nevis, United States

### Reviewed by:

Henk Huijser,  
Queensland University of  
Technology, Australia  
Ronny Scherer,  
University of Oslo, Norway

### \*Correspondence:

Dimitri Molerov  
molerov@hu-berlin.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Education

**Received:** 30 June 2020

**Accepted:** 13 November 2020

**Published:** 15 December 2020

### Citation:

Molerov D, Zlatkin-Troitschanskaia O,  
Nagel M-T, Brückner S, Schmidt S  
and Shavelson RJ (2020) Assessing  
University Students' Critical Online  
Reasoning Ability: A Conceptual and  
Assessment Framework With  
Preliminary Evidence.  
Front. Educ. 5:577843.  
doi: 10.3389/feduc.2020.577843

Critical evaluation skills when using online information are considered important in many research and education frameworks; critical thinking and information literacy are cited as key twenty-first century skills for students. Higher education may play a special role in promoting students' skills in critically evaluating (online) sources. Today, higher education students are more likely to use the Internet instead of offline sources such as textbooks when studying for exams. However, far from being a value-neutral, curated learning environment, the Internet poses various challenges, including a large amount of incomplete, contradictory, erroneous, and biased information. With low barriers to online publication, the responsibility to access, select, process, and use suitable relevant and trustworthy information rests with the (self-directed) learner. Despite the central importance of critically evaluating online information, its assessment in higher education is still an emerging field. In this paper, we present a newly developed theoretical-conceptual framework for Critical Online Reasoning (COR), situated in relation to prior approaches ("information problem-solving," "multiple-source comprehension," "web credibility," "informal argumentation," "critical thinking"), along with an evidence-centered assessment framework and its preliminary validation. In 2016, the Stanford History Education Group developed and validated the assessment of *Civic Online Reasoning* for the United States. At the college level, this assessment holistically measures students' web searches and evaluation of online information using open Internet searches and real websites. Our initial adaptation and validation indicated a need to further develop the construct and assessment framework for evaluating higher education students in Germany across disciplines over their course of studies. Based on our literature review and prior analyses, we classified COR abilities into three uniquely combined facets: (i) online information acquisition, (ii) critical information evaluation, and (iii) reasoning based on evidence, argumentation, and synthesis. We modeled COR ability from a behavior, content, process, and development perspective, specifying scoring rubrics in an evidence-centered design. Preliminary validation results from expert

interviews and content analysis indicated that the assessment covers typical online media and challenges for higher education students in Germany and contains cues to tap modeled COR abilities. We close with a discussion of ongoing research and potentials for future development.

**Keywords:** critical online reasoning assessment, critical thinking, web credibility, higher education, information problem solving using the Internet, multiple-source use, test validation, performance assessment

## INTRODUCTION

### Relevance and Research Background

Today, higher education students use the Internet to access information and sources for learning much more frequently than offline sources such as textbooks (Gasser et al., 2012; Maurer et al., 2020). However, there have been warnings about the harmful effects of online media use on students' learning (Maurer et al., 2018), with misinformation and the acquisition of (domain-specific) misconceptions and erroneous knowledge being prominent examples (Bayer et al., 2019; Center for Humane Technology, 2019). While Internet users are generally concerned about their ability to distinguish warranted, fact-based knowledge from misinformation<sup>1</sup> (Newman et al., 2019), research on web credibility suggests that Internet users pay little attention to cues indicating erroneous information and a lack of trustworthiness; similar findings were determined across a variety of online information environments and learner groups (Fogg et al., 2003; Metzger and Flanagin, 2013, 2015).

For learning in higher education, the Internet may have both a positive and a negative impact (Maurer et al., 2018, 2020). Positive affordances for collaboration, organization, aggregation, presentation, and the ubiquitous accessibility of information have been discussed in research on online and multimedia learning (Mayer, 2009). However, problems such as addictive gratification mechanisms, filter bubbles and algorithm-amplified polarization, political and commercial targeting based on online behavior profiles, censorship, and misinformation (Bayer et al., 2019; Center for Humane Technology, 2019) have recently been critically discussed as well. The potential of online applications and social media for purposes of persuasion has been known for some time (Fogg, 2003), though the impact of online information on knowledge acquisition is still under-researched to date.

As recent research indicates, the multitude of information and sources available online may lead to information overload (Batista and Marques, 2017; Hahnel et al., 2019). Lower barriers to publication and the lack of requirements for quality assurance, fewer gatekeepers, and faster distribution result in a highly diverse online media landscape and varying information quality (Shao et al., 2017). Students are confronted with quality shortcomings such as incomplete, contradictory, or erroneous information when obtaining and integrating new information

from multiple online sources (List and Alexander, 2017; Braasch et al., 2018). Hence, whenever Internet users are acquiring knowledge based on online information or performing online search queries in a way that can be framed as solving an information problem (Brand-Gruwel et al., 2005), they are faced with the challenge of finding, selecting, accessing, and using suitable information. In addition, online learners need to avoid distractions (e.g., advertisements, clickbait) and misinformation as well as evaluate the information they choose with regard to possible biases and specific narrative framing of information (Walton, 2017; Banerjee et al., 2020). To successfully distinguish between trustworthy and untrustworthy online information, students need to judge its relevance to their inquiry and, in particular, evaluate its credibility (Flanagin et al., 2010; Goldman and Brand-Gruwel, 2018). The ability to find suitable information online, distinguish trustworthy from untrustworthy information, and reason based on this information is examined under the term of "critical online reasoning." These abilities are crucial for (self-)regulated (unsupervised) acquisition of warranted (domain-specific) knowledge based on online information.<sup>2</sup> In this context, current studies are focusing on the development of (domain-specific) misconceptions and the acquisition of erroneous knowledge over the course of higher education studies, specifically among students who report that they predominantly use Internet sources when studying (Maurer et al., 2018, 2020).

### University Students' Critical Online Reasoning Assessment (CORA): Study Context

To acquire reliable and warranted (domain-specific) knowledge, students need to access, evaluate, select, and ultimately reason based on relevant and trustworthy information from online sources. At the same time, they need to recognize erroneous or (intentionally) misleading information and possible corresponding bias, for instance, due to underlying framing or unwarranted perspectives, to avoid being misled and acquiring erroneous knowledge. To properly handle online sources featuring incorrect, incomplete, and contradictory information, students need to recognize patterns in the information indicating its trustworthiness or lack thereof (cues for credibility or misinformation) based on self-selected criteria such as perceived

<sup>1</sup>In this study, we focused on misinformation. Misinformation may result from (often unintentional) error, lacking quality assurance, and lacking truth commitment, while disinformation may be spread purposefully due to vested (e.g., business-related, political, ideological, and potentially hidden) interests of stakeholders (Metzger, 2007; Karlova and Fisher, 2013).

<sup>2</sup>We focused on inquiry-based learning using the Internet, information problem solving, and integration of information from multiple sources (Zhang and Duke, 2008; List and Alexander, 2017) in the context of university studies, although the critical evaluation of information when acquiring knowledge while using the Internet for other purposes, such as for entertainment, is important as well.



expertise or communicative intentions to acquire reliable, warranted (domain-specific) knowledge using the Internet.

Students' critical evaluation skills when dealing with online information are considered important in many research frameworks in a multitude of disciplines that address the online learning-and-teaching environment (Section Theoretical and Conceptual Framework; **Table 1**). Like critical thinking and information literacy, they are considered to be among the key twenty-first century skills, and are considered key skills for "Education in the Digital World" (National Research Council, 2012; KMK, 2016). Skills related to the critical-reflective use of online information are more important than ever, which becomes evident especially with regard to the internet-savvy younger generations (Wineburg et al., 2018). Higher education can play a special role in promoting students' critical thinking skills and their skills in evaluating (online) sources (Moore, 2013) due to the evidence-based, research-focused orientation of most academic disciplines (Pellegrino, 2017). For instance, graduate students were found to have advanced critical thinking skills, which has been attributed to the fact that they wrote a bachelor thesis as part of their undergraduate studies (Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019).

Despite being of central importance for studying using the Internet, the assessment of students' skills related to critical online reasoning (COR) is an emerging field with conceptual and theoretical frameworks building on a large number of prior research strands (Section Theoretical and Conceptual Framework; **Table 1**). For instance, computer skills, digital and information literacy, and critical thinking approaches have described and examined (bundles of) related facets. To our knowledge, there is no conceptual and assessment framework to date that describes and operationalizes COR as an interrelated triad of its key facets (i) information acquisition in the online environment, (ii) critical information evaluation, and (iii) reasoning using evidence, argumentation, and synthesis.

In this context, pioneering work has been done by Wineburg et al. (2018) from the Stanford History Education Group (SHEG), who developed an assessment for measuring *Civic Online Reasoning* at the middle school, high school, and college level. At the college level, this holistic assessment of how students evaluate online information and sources comprises short evaluation prompts, real websites, and an open Internet search (Wineburg and McGrew, 2016; Wineburg et al., 2016a,b). The assessment was validated in a nationwide study in the U.S. (Wineburg et al., 2018), which indicated substantial deficits in these skills among higher education students.

Based on this U.S. research, we adapted the assessment framework for higher education in Germany. The preliminary validation of the U.S. assessment for Germany indicated that an adaption and validation in terms of the recommendations by the international Test Adaptation Guidelines [TAGs, International Test Commission (ITC), 2017] was not possible. It became evident that, in addition to the practical difficulties of adapting the U.S. assessment web stimuli for assessing the critical evaluation of online information for learning in the German higher education context, expert interviews (Section Content Analysis: CORA Task Components as Coverage of the Construct)

indicated that due to the differences in terms of historical and socio-cultural traditions between the two countries, in German higher education, the concept of "civic education" is less prominent than "academic education" (for a comparison of the concept of education/ "Bildung" in Germany and in the U.S., see (Beck, 2020); for a model of critical thinking, see Oser and Biedermann, 2020). Moreover, experts noted that students learn from information from a variety of sources not necessarily related to civic issues (e.g., commercial websites), in addition to scientific publications and textbooks, and it remains unclear how new knowledge based on these multiple sources is integrated, which requires further differentiation and specification.

Based on the results of this preliminary validation, we modified the theoretical framework by expanding our focus beyond civic reasoning to include further purposes of online information acquisition, and situated the construct in relation to a number of theories, models, and adjacent fields, focusing on the research traditions of *critical thinking* (Facione, 1990), which are more applicable to Germany (than civic reasoning), as well as in relation to additional relevant constructs such as "web credibility," "multiple source comprehension," "multiple-source use," and "information problem-solving" using the Internet (Metzger, 2007; Braasch et al., 2018; Goldman and Brand-Gruwel, 2018). Based on a combination of converging aspects from these research strands, we developed a new conceptual framework to describe and operationalize the abovementioned triad of key facets underlying the resulting skill of *Critical Online Reasoning (COR)*: (i) *online information acquisition*, (ii) *critical information evaluation*, and (iii) *reasoning using evidence, argumentation, and synthesis*.

## Research Objectives and Questions

The first objective of this paper is to present this newly developed conceptual and assessment framework, and to locate this conceptualization and operationalization approach in the context of prior and current research while critically reflecting on its scope and limitations. The methodological framework is based on an evidence-centered assessment design (ECD) (Mislevy, 2017). According to ECD, alignment of a *student model*, a *task model*, and an *interpretive model* are needed to design assessments with validity in mind. The *student model* covers the abilities that students are to develop and exhibit (RQ1); the *task model* details how abilities are tapped by assessment tasks (RQ2); and the *interpretive model* describes the way in which scores are considered to relate to student abilities (RQ3). The following research questions (RQ) are examined in this context.

RQ1: *What student abilities and mental processes does the CORA cover? How can the COR ability be described and operationalized in terms of its construct definition?*

RQ2: *What kinds of situations (task prompts), with which psychological stimuli (i.e., test definition), are required to validly measure students' abilities and mental processes in accordance with the construct definition?*

As a second objective of this paper, we focus on the preliminary validation of the COR assessment (hereinafter referred to as CORA). The validation framework for

**TABLE 1** | Theoretical and conceptual background of COR.

<b>Critical online reasoning (COR)</b>		
<b>Main assessment frameworks</b>		<b>Studies (selection)</b>
Civic online reasoning assessment (CORA) (assessment in U.S., employing real websites and live open web search as stimuli)		Wineburg et al., 2016a; Wineburg and McGrew, 2017; Wineburg et al., 2018; Breakstone et al., 2019; McGrew et al., 2019
Performance Assessment of Learning in higher education (PAL) (criterion-sampled performance tasks)		Shavelson et al., 2018, 2019; Zlatkin-Troitschanskaia et al., 2019
Positive Learning in the Age of Information (PLATO) framework		Zlatkin-Troitschanskaia et al., 2018, 2020b
<b>Related research strands for main COR facets</b>		<b>Studies (selection)</b>
Overarching	Multiple-source comprehension and use (MSC)	for an overview, see Rouet, 2006; Lawless et al., 2012; List and Alexander, 2017; Braasch et al., 2018
	Information-problem solving using the Internet (IPS-I)	Brand-Gruwel et al., 2009; Walraven et al., 2009; Goldman and Brand-Gruwel, 2018
	Internet reading strategies/online search behavior and self-reported search strategies	Salmerón et al., 2005; Zhang and Duke, 2008; Zhang et al., 2011; Pernice, 2017
	"Sourcing" in MSC/use of source cues about author and meta-data, for credibility evaluation; proactive, repeated, and task-related sourcing	Braten et al., 2018; Hahnel et al., 2019
(I) Online Information Acquisition (OIA)	Higher education students' use of online information	Head and Eisenberg, 2009; Samson, 2010; Maurer et al., 2018
	Students' information needs and problems using online information and databases	Walraven et al., 2008; Catalano, 2013; Sanders et al., 2015; Kohnen and Mertens, 2019
	Heuristic and systematic information seeking strategies	Chen and Chaiken, 1999; De Neys, 2006; Toplak et al., 2007; Evans and Stanovich, 2013; Gronchi and Giovannelli, 2018
	Information foraging theory	Pirolli and Card, 1999; Juvina and van Oostendorp, 2008
	Interactive Information Retrieval/programming and designing search engines, static and dynamic websites, databases, etc. with a focus on user's interaction with them	for an overview, see (Xie, 2008)
(II) Critical Information Evaluation (CIE)	Disinformation and misinformation classification and current examples (EU)/recent misinformation and disinformation phenomena	(Karlova and Fisher, 2013; Ciampaglia, 2018); for a review of disinformation as a threat to democracy, see also (Bayer et al., 2019; Flore et al., 2019)
	Media bias and propaganda/Studies on strategy and system-level media effects promoting spread of misinformation and disinformation, including the documentation of deconstructions of media, framing, interaction measures, message components, power exertion, and manipulation techniques in advertisement, propaganda, and journalism, as well as implications for (civic) education	Herman and Chomsky, 2002; Paul and Elder, 2008; Daniels, 2009; Walton, 2017
	Hierarchy of Influences on a media message/media-sociological approach highlighting agents and practices influencing media messages at different levels of power, correspondingly, different scope and spread of misinformation	Reese and Shoemaker, 2016
	General Web credibility models	Prominence Interpretation Theory/cue identification and interpretation for operator, design, and content; gullibility and incredulity error; deceiver credibility
		Two-step judgment/immediate surface judgment and subsequent message judgment

(Continued)

**TABLE 1** | Continued

<b>Critical online reasoning (COR)</b>		<b>Studies (selection)</b>
	<p>Dual processing model/3 phases: ability and motivation at exposure influence propensity and depth of evaluation; follow-up studies on relations and features</p> <p>Unifying model/ "construct" phase—subjective and context-dependent criterialization of credibility; heuristics and interaction</p> <p>MAIN model—modality, agency, interactivity, navigability/affordances of technology itself as cues for credibility</p> <p>Information trust/3 "s" model (surface, source, semantics), influence by users' domain knowledge, topic knowledge, and information skills</p> <p>New Web Credibility model/juxtaposes P-I website dimensions (operator, content, design) with credibility attributions (expertise and trustworthiness); overview</p> <p>Content Credibility Corpus/corpus collection of websites and topics for credibility evaluation</p>	<p>Metzger, 2007; Winter et al., 2016; Flanagan et al., 2018; Krämer et al., 2018</p> <p>Hilligoss and Rieh, 2008</p> <p>Sundar, 2008</p> <p>Lucassen and Schraagen, 2011, 2013</p> <p>Choi, 2015</p> <p>Kakol et al., 2017; Wierzbicki, 2018</p>
	Web credibility aspects (selection)	<p>appearance (Akamine et al., 2008); web 1.0 to 2.0 (Tanaka, 2009; Tanaka et al., 2010); web experience (Jozsa et al., 2012); conflicting topics (Salmerón et al., 2013), fear appeals (Dunbar et al., 2014), relations with trust in press (Go et al., 2016), message sidedness (Flanagan et al., 2018), source credibility in political communication (Flanagan and Metzger, 2014)</p>
	Further key evaluation criteria of online information, see also web credibility studies	<p>text relevance (McCrudden et al., 2011); accessibility/comprehensibility (Snow, 2002; Coiro, 2003); usefulness (Goldman et al., 2013)</p>
(III) Reasoning with Evidence, Argumentation and Synthesis (REAS)	Integrated model/review of CT constructs for higher education and the online environment, integrated model, ambiguity experience as activation	Jahn, 2012; Jahn and Kenner, 2018
	Logic-based CT/CT pioneering approach	Ennis, 1985
	Education- and psychology-based CT/Delphi study on CT components	Facione, 1990
	Development of reflective CT/6-stage theory based on systematicity of reflection/meta-cognitive monitoring	Elder and Paul, 2010
	Psychology-based CT	Halpern, 2014
	Scientific reasoning and argumentation/studies focusing on students' reasoning and argumentation patterns based on scientific evidence, models, and principles	Fischer et al., 2014; Fischer, 2018
	Rational thinking/rational thinking construct and operationalization	Stanovich et al., 2016
	Valid (informal) argumentation/fundamental components of argumentation, construction, and analysis of valid argumentation patterns (schemes of (un)warranted reasoning, critical questions for knowledge elicitation)	Walton, 2006; Walton et al., 2008
	Adequate heuristics/fast and frugal heuristics for ecological rationality	Goldstein and Gigerenzer, 2002
	Suboptimal heuristics and biases	Kahneman et al., 1982
	Reasoning fallacies	Van Eemeren, 2013

(Continued)

TABLE 1 | Continued

Critical online reasoning (COR)		Studies (selection)
(IV) Metacognitive Activation (MCA)—Overarching metacognitive and regulative component, affective and attitudinal aspects	Definition of bias	Walton, 2006
	Metacognitive processes and regulation/ review of meta-cognitive processes to support information search	Blummer and Kenton, 2015
	Affective reactions and uncertainty/6-phase Information Search Process with varying certainty and affective response	Kuhlthau, 1993
	Context-based initial activation	Jahn, 2012
	Activation via discrepancy detection/ambiguity experience/subfacet of Discrepancy-Induced Source Comprehension (D-ISC) model for mid-task activation	Jahn, 2012; Braasch and Bråten, 2017
	(attitudinal) Critical thinking dispositions	openness to experience etc. Facione (1990) intellectual virtues (Paul and Elder, 2005)
<b>Related research strands for broader COR activity</b>		
Literacies	Digital literacy/media literacy, information literacy, and computer literacy	Koltay, 2011; Bulger et al., 2014; Murray and Pérez, 2014; Sparks et al., 2016
	Information literacy	American Library Association, 2000; Kingsley et al., 2011; Taylor and Dalal, 2014; Sanders et al., 2015; Maurer et al., 2017; Podgornik et al., 2017; McMullin, 2018; Walton et al., 2020
	ICT information and communication technology literacy Media literacy	Zylka et al., 2015 Damico and Panos, 2018; Powers, 2019; Threadgill and Price, 2019
Further Relevant Assessment Frameworks (selection)	Multiple-source comprehension assessment	Lawless et al., 2012
Instructional Approaches (selection)	Assessment of argument evaluation in scientific texts	Münchow et al., 2019
	Critical Thinking assessment in higher education	Liu et al., 2014
	Civic Online Reasoning instructional intervention	McGrew et al., 2019
	Critical source evaluation for improved search	Leeder and Shah, 2016
	Bad News game/multilingual browser game on use of major media disinformation strategies, based on 'inoculation' approach	Rozenbeek and van der Linden, 2019
	Review of CT interventions	Abrami et al., 2008
	Fostering CT using digital media/review of instructional designs in HE	Jahn, 2012; Jahn and Kenner, 2018

CORA is based on approaches by Messick (1989) and Kane (2012). A qualitative evaluation of the CORA yielded preliminary validity evidence based on a content analysis of the CORA tasks, and interviews with experts in media science, linguistics, and test development (Section Content Analysis: CORA Task Components as Coverage of the Construct). Based on the results of content validation studies conducted according to the Standards for Pedagogical and Psychological Testing (AERA et al., 2014; hereinafter referred to as AERA Standards), the following RQ was investigated:

*RQ3: To what extent does the preliminary evidence support the validity claim that CORA measures the participants'*

*personal construct-relevant abilities in the sense of the defined construct definition?*

In Section Theoretical and Conceptual Framework, we first present the theoretical and conceptual COR framework, also in terms of related research approaches. In Section Assessment Framework of Critical Online Reasoning, we describe the U.S. assessment of civic online reasoning and present our work toward adapting and further developing this approach into an expanded assessment framework and scoring scheme for measuring COR in German higher education. In Section Preliminary Validation, we report on initial results from the preliminary validation studies. In Section Research Perspectives, we close with implications for refining CORA tasks and rubrics



and give an outlook on ongoing further validation studies and analyses using CORA in large-scale assessments.

## THEORETICAL AND CONCEPTUAL FRAMEWORK

In this section, we outline the working construct definition for Critical Online Reasoning (COR) as a basis for the CORA framework. We explain the theoretical components and key considerations used to derive this COR construct definition from related prior approaches and frameworks. COR is modeled from a *process, content, domain, and development* perspective. For brevity, we only describe the key facets and central components and list the most relevant references categorized by (sub)facets in **Figure 1**.

### Construct Definition of Critical Online Reasoning

The working construct definition of COR (*RQ1*) describes the personal abilities of searching, selecting, accessing, processing, and using online information to solve a given problem or build knowledge while critically distinguishing trustworthy from untrustworthy information and reasoning argumentatively based on trustworthy and relevant information from the online environment.

This construct definition focuses on a combination of three overlapping facets: (i) *Online Information Acquisition (OIA)* abilities (for inquiry-based learning and information problem-solving), (ii) *Critical Information Evaluation (CIE)* abilities to analyze online information particularly in terms of its credibility and trustworthiness, and (iii) abilities to use the information for *Reasoning based on Evidence, Argumentation, and Synthesis (REAS)*, weighting (contradictory) arguments and (covert) perspectives, while accounting for possible misinformation and biases. In addition, we assume that the activation of these COR facets requires metacognitive skills, described in the *Metacognitive Activation (MCA)* (**Figure 1**).

### Theoretical Components of COR Process Perspective

*Online Information Acquisition (OIA)* focuses on the searching and accessing of online information, for example by using general and specialized search engines and databases, specifying search queries, opening specific websites. Beyond these more technical aspects, COR focuses in particular on searching for specific platform entries and passages and terms on a website in as far as they contribute to an (*efficient*) *accessing of relevant and trustworthy information and avoidance of untrustworthy information* (Braten et al., 2018; the Information Search Process model, Kuhlthau et al., 2008).

*Critical Information Evaluation (CIE)* is crucial for self-directed, cross-sectional learning based on online information. This facet focuses on students' *selection of information sources and evaluation of information and sources* based on website features or specific cues (e.g., text, graphics, audio-visuals). Following comprehension-oriented reception and processing,

CIE is used to differentiate and select high- instead of low-quality information (relative to one's subjective standards and interpretation of task requirements). A cue can be any meaningful pattern in the online environment interpreted as an indicator of (trustworthy or untrustworthy) online media or communicative means. Examples of cues may be a URL, title or keyword on the search engine results page, a layout or design element, media properties, an article title, information about author, publisher or founder, publication date, certain phrasings, legal or technical information. Trustworthiness "evaluations" typically include targeted verification behavior, which results in a (defeasible) "judgment" about a web medium or piece of information, which may be based on an initial heuristic appraisal without further (re-)evaluation. However, CIE as "evaluation" can require a more systematic analytical, criteria-based judgment process for students, possibly using multiple searches to establish reliable and warranted knowledge (for an overview on related multiple document comprehension frameworks, see Braten et al., 2018; e.g., the Discrepancy-Induced Source Comprehension (D-ISC) model, Braasch and Bråten, 2017).

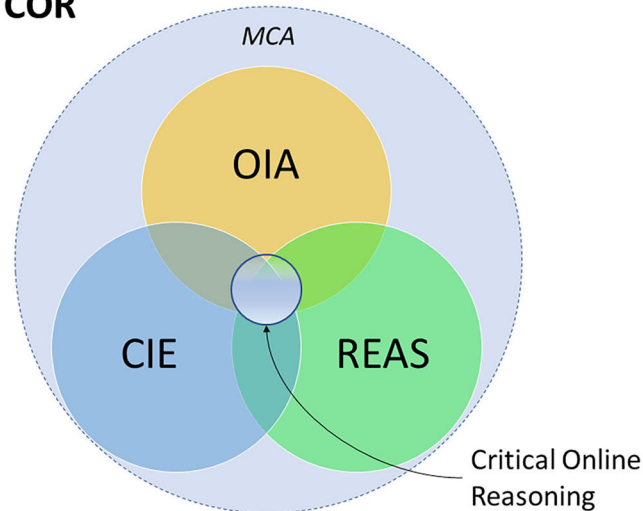
*Reasoning with Evidence, Argumentation, and Synthesis (REAS)* is probably the most important facet of COR, which distinguishes this construct from "literacy" constructs (e.g., digital, information or media literacy). This facet focuses on uniting the initially appraised information, weighting it against further indications and perspectives, and using it as evidence to construct a convincing argument that accounts for uncertainty (Walton, 2006). Argumentation is a well-suited discourse format for deliberating whether to accept a proposition (e.g., to trust or distrust). Evidence-based argumentation imposes certain quality standards for a well-founded judgment (e.g., rationality) and requires minimal components of a claim, reasons, evidence (and data) and conventional inferential connections between them (e.g., Argumentation Schemes, Walton, 2006; Walton et al., 2008; Fischer et al., 2014; Fischer, 2018).

These three main facets, OIA, CIE, and REAS, are primarily considered cognitive abilities. Each of them can also take on a metacognitive quality within the COR process, for example as reasoners (internally) comment on their ongoing search, evaluation or argument construction (e.g., "I would not trust this website"), or (self-)reflect on previously acquired knowledge to identify incorrectness or inconsistencies (e.g., "This sentence here contradicts that other source/what I know about the subject"). The latter reflection can become epistemic if it turns to the method of information acquisition and reasoning itself (e.g., "How did I end up believing this scam?").

These main facets are accompanied by an overarching, self-regulative, metacognitive COR component that activates deliberate COR behavior and coordinates (transitions) between the COR facets in the progression of COR activity, particularly for activating a critical evaluation and deciding when to terminate it, in relation to other events (e.g., during a learning experience, social communication)<sup>3</sup>. Self-regulation can be

<sup>3</sup>Regulation of COR may be performed deliberately using meta-cognition, or in response to a cognitive process outcome, habitual behavior, processing of environmental cues, affective or motivational state.

## Main Facets of COR



**FIGURE 1 |** The COR construct with its main facets: MCA, metacognitive activation; OIA, online information acquisition; CIE, critical information evaluation; REAS, reasoning with evidence, argumentation and synthesis.

applied to monitor and maintain focus (noticing unfocused processing, returning to task) and handle environmental signals (identifying and minimizing distracting information features) (Blummer and Kenton, 2015). As reasoners may have affective responses (Kuhlthau, 1993) to their task progress and to specific information (particularly on controversial topics), affective self-regulation can not only support them in staying on task and keeping an open mind, but they can use it meta-cognitively for COR to gain an insight into unconsciously processed information (e.g., identifying and coping with triggered avoidance reactions or anxiety induced by ambiguity or manipulation attempts) and can critically reflect on triggers in the source cues.

Thus, *Metacognitive Activation (MCA)* is assumed to be an ability required to activate COR in relevant contexts. (Epistemic) metacognition can be characterized by gradations of self-awareness regarding information acquisition, evaluation and reasoning processes, which may activate a “vigilance state” in students and lead to certain (subconscious) reactions (and a habitual affective response, e.g., anxiety, excitement), or can also be interpreted as an indicator of a potential problem with processed information (“am I being lied to/at risk after misjudging the information?”) at the metacognitive level (on uncertainty and emotions when searching for information, see Kuhlthau, 1993; on ambiguity experience as the first stage in a general critical reasoning process, see Jahn and Kenner, 2018), which may lead to the activation of an evaluative COR process.

The main facets of COR and the overarching metacognitive self-regulative component are understood to determine *COR performance* (and are the focus of the CORA, Section Test Definition and Operationalization of COR: Design and Characteristics of CORA Tasks). The main COR facets are

assumed to rely on “secondary” sub-facets that provide support in cases where *related specific problems* occur, including self-regulation for minimizing distractions and on-task focus, as well as diverse knowledge sub-facets.

Knowledge sub-facets may include, for OIA, knowledge of resources and techniques for credibility verification; for CIE, knowledge of credibility indicators and potentially misleading contexts and framings, manipulative genres and communication strategies; for REAS, knowledge of reasoning standards as well as fallacies, heuristics, and perceptual, reasoning and memory biases as well as of epistemic limitations for trustworthiness assertions. The list is non-exhaustive, and the knowledge and skills are problem-dependent (e.g., checking for media bias will yield conclusive results only if there is in fact a bias in the stimulus material); they can be expected to impact COR in related cases. Hence, controlling for corresponding stimuli encompassed in the task is recommended.

Attitudinal dispositions for critical reasoning and thinking, such as open-mindedness, fairness, and intellectual autonomy (Facione, 1990; Paul and Elder, 2005) are equally likely candidates for COR influences. These secondary facets are not examined in the current conceptualization.

### Content Perspective

For acquiring information online in a warranted way, students need to successfully identify and use trustworthy sources and information and avoid untrustworthy ones. In contrast, unsuccessful performance is marked by trusting untrustworthy information, a gullibility error, or refusing to accept trustworthy sources, an incredulity error (Tseng and Fogg, 1999). To decide which information to trust and use, students need to judge information in regard to several criteria, including at least the

following: *usefulness*, *accessibility*, *relevance*, and *trustworthiness*. Information may be judged as useful if it advances the inquiry, for instance by supporting the construction of an argument; usefulness may also be understood as a holistic appraisal based on all other criteria. Lack of *accessibility* (or comprehensibility) limits students to the parts of the information landscape that they can confidently access and process (e.g., students may ignore a search result in a foreign language or leave a website with a paywall, but also abandon a text they deem too difficult to locate or understand in the given task time). In an open information environment, successfully judging *relevance* as relatedness or specificity to the topic of inquiry and *trustworthiness* or quality of information enables students to select and spend more time on high-quality sources and avoid untrustworthy sources. Assuming students will attempt to ignore information they judge as untrustworthy, any decision in this regard affects their available information pool for reasoning and learning.

The judgment of trustworthiness as an (inter-)subjective judgment of the objectively verifiable quality of an online media product against an evidential or epistemic standard is central to COR. In more descriptively oriented “web credibility” research, a credibility judgment is understood as a subjective attribution of trust to an online media product; trustworthiness in COR is closely related, but presupposes that the judgment can be based on valid or invalid reasoning (acceptable or unacceptable based on a normative standard) and hence can be evaluated as a skill. Trustworthiness in COR can be considered a warranted credibility judgment. Consequently, COR enables students to distinguish trustworthy from untrustworthy information and, more specifically, various sub-types based on assumed expertise and communicative intent, for example: accidental misinformation due to error, open or hidden bias, deliberate disinformation, and (non-epistemic) “bullshitting.” A more fine-grained judgment is assumed to afford higher certainty, a more precise information selection, and more adequate response to an information problem. To successfully infer the type of information, reasoners may evaluate cues from at least three major strands of evidence about an online medium, including cues on content, logic, and evidence; cues on design, surface structure, and other representational factors; and cues on author, source, funding, and other media production and publication-related factors. Reasoners may evaluate these themselves (using their own judgment), trust the judgment of experts (external judgment), or a combination of the two; when accepting external judgment, instead of the information itself, reasoners need to judge at least their chosen expert's topic-related expertise and truth-oriented intent.

### Domain-Specificity and Generality

Based on the CORA framework, COR is modeled for generic critical online reasoning (GEN-COR) on tasks and websites that do not require specialized domain knowledge and are suited for young adults after secondary education.

The construct can be specified for study domains (DOM-COR), for instance by defining domain standards of evidence for distinguishing trustworthy from untrustworthy information and typical domain problems regarding the judgment of online information.

### Development Perspective

Different gradations can be derived based on task difficulty, complexity, time, and aspired specificity of reasoning (Sections Test Definition and Operationalization of COR: Design and Characteristics of CORA Tasks and Scoring Rubrics). COR ability levels were distinguished to fit the main construct facets depending on students' performance in (sub-)tasks tapping OIA, CIE, and REAS (see rubrics in Section Scoring Rubrics; Table 1).

## ASSESSMENT FRAMEWORK OF CRITICAL ONLINE REASONING

### Civic Online Reasoning

Wineburg and McGrew (2016) developed an assessment to measure *civic online reasoning*, which they defined as students' skills in interpreting online news sources and social media posts. The assessment includes real, multimodal websites as information sources (and distractors) as well as open web searches. The construct of civic online reasoning was developed from the construct of news media literacy (Wineburg et al., 2016a). It was conceptualized as a key sub-component of analytic thinking while using online media. The assessment aims to measure whether students are able to competently navigate information online and to distinguish reliable, trustworthy sources and information from biased and manipulative information (Wineburg et al., 2016a).

The students' skills required to solve the tasks were assessed under realistic conditions for learning using the Internet, i.e., while students performed website evaluations and self-directed open web searches (Wineburg and McGrew, 2017). The computer-based assessment presents students with short tasks containing links to websites with, for instance news articles or social media text and video posts, which students are asked to evaluate. The task prompts require the test-takers to evaluate the credibility of information, and to justify their decision, also citing web sources as evidence. The topics focus on various political and social issues of most US-centric civic interest, typically with conflicting constellations of sources.

Using this assessment, the SHEG surveyed a sample of 7,804 higher education students across the U.S. (Wineburg et al., 2016a), and compared the students' performance to that of history professors and professional fact checkers. Based on the findings, the search engine results pages designed and implemented an intervention to improve students' civic online reasoning in higher education (Wineburg and McGrew, 2016; McGrew et al., 2019).

## Critical Online Reasoning Assessment (CORA)

In our project<sup>4</sup>, the initial goal was to adapt this instrument to assess the *civic online reasoning* of students in higher education in Germany and to explore the possibility of using this assessment in cross-national comparisons. The assessment of civic online reasoning features realistic judgment and decision-making scenarios with strong socio-cultural roots, which may engage and tap both the (meta)cognition and the emotional responses of test-takers, as well as their critical evaluation skills. While cultural specificity may present advantages in a within-country assessment, these can become idiosyncratic challenges in cross-national adaptations (e.g., Arffman, 2007; Solano-Flores et al., 2009). Even though we followed the state-of-the-art TAGs by the International Test Commission (ITC) (2017) and the best-practice approach of (Double-)Translation, Reconciliation, Adjudication, Pretesting, and Documentation (TRAPD, Harkness, 2003) in assessment adaptation research (as recommended in the TAGs), after the initial adaptation process (Molerov et al., 2019), both the (construct) definition of *civic online reasoning* and the adapted assessment of civic online reasoning showed limitations when applied to the context of learning based on online information in German academic education. The translation team faced several major practical challenges while adapting the real website stimuli, and the results were less favorably evaluated by adaptation experts. This was a key finding from the adaptation attempts and preliminary construct validation by means of curricular analyses and interviews with experts for German higher education. Both analyses indicated the significant differences in terms of historical and socio-cultural traditions between the higher education systems in the two countries (for details, Zlatkin-Troitschanskaia et al., 2018b). Regarding construct limitations, curricular analysis indicated differences in the relevance of “civic education” within German higher education, highlighting problems for the (longitudinal and cross-disciplinary) assessment of generic abilities in learning based on online information. Expert interviews conducted in the context of adaptation attempts and the preliminary validation of the U.S. conceptual and assessment framework of “civic online reasoning” (for details, Molerov et al., 2019) indicated that the concept of “civic education” is related to a specific research strand of political education and is less important in German higher education than “academic education,” which is more strongly related to research traditions focusing on critical thinking (for a comparison of the concept of education in Germany and in the U.S., see Beck, 2020; Oser and Biedermann, 2020).

Based on this preliminary validation of the U.S. assessment in Germany, we modified the conceptual framework (Section Theoretical Components of COR) to accommodate for the close relationship between COR and generic critical thinking, multiple-source comprehension, scientific reasoning and

informal argumentation approaches (Walton, 2006; Fischer et al., 2014, 2018; Goldman and Brand-Gruwel, 2018; Jahn and Kenner, 2018), and expanded the U.S. assessment framework to cover all online sources that students use for learning. We developed the scoring rubrics accordingly to validly measure the *critical online reasoning* (COR) ability of higher education students of all degree programs in Germany in accordance with our construct definition (Section Construct Definition of Critical Online Reasoning). Thus, new CORA tasks with new scenarios were created to cover the (German) online media landscape used for learning and topics including culturally relevant issues and problems. The assessment framework was expanded to comprise tasks stimulating web searches, the critical evaluation of online information, and students' use of this information in reasoning based on evidence, argumentation and synthesis *to obtain warranted knowledge and solve the given information problems, and to develop coherent and conclusive arguments for their decision* (e.g., draft a short essay or evaluative short report). We also developed and validated the scoring scheme to rate the students' responses to the CORA tasks (Section Scoring Rubrics).

## Test Definition and Operationalization of COR: Design and Characteristics of CORA Tasks

The German CORA project developed a holistic, performance assessment that uses criterion-sampled situations to tap students' real-world decision-making and judgment skills. The tasks/situations merit critical evaluation. Students may encounter such tasks when studying and working in academic and professional domains, as well as in their public and private lives (Davey et al., 2015; Shavelson et al., 2018, 2019). CORA comprises six tasks of 10 min each. CORA is characterized by the use of realistic tasks in a natural online environment (for an example, see Figure 2). As tasks are carried out on the Internet, students have an unlimited pool of information from which to search and select suitable sources to verify or refute a claim, while judging and documenting the evidence. Five CORA tasks contain links to websites that may have been published with (covert) commercial or ideological intent, and may, for instance aim to sell products or to convince their audience of a particular point of view by offering low-quality information. The characteristics of the low-quality information offered on websites linked in the CORA tasks included, for instance a selection of information while (intentionally) omitting other perspectives, incorrect or imprecise information, irrelevant and distracting information, and biased framing. The tasks feature snippets of information in online media, such as websites, twitter messages, YouTube videos, put forward by political, financial, religious, media or other groups, some cloaked with covert agendas, others more transparent.

A specific characteristic of the CORA tasks is that only the stimuli and distractors included in the task prompt and the websites linked in the tasks can be manipulated and controlled for by the test developers. Since the task prompt asks the students to evaluate the credibility and trustworthiness of the linked website through a free web search, realistic distractors include,

<sup>4</sup>The German CORA project is part of the cross-university PLATO research program, which examines higher education students' Internet-supported learning for the acquisition of warranted knowledge from various disciplinary perspectives (for an overview, see Zlatkin-Troitschanskaia et al., 2018a; Zlatkin-Troitschanskaia, 2020).



**TASK 1: Vegan protein sources**

Please complete this task within 10 minutes.

Visit the following website:

[URL]

Is this a reliable source of information about vegan protein sources?

---

Note: You can use any information on this website and you can freely search the Internet.

Justify your answer with evidence from the Internet sources used and include the corresponding URLs.

**FIGURE 2** | Example CORA Task prompt (German website).

for instance vividly presented information, a large amount of highly detailed information, (unreferenced) technical, numerical, statistical and graphical data, and alleged (e.g., scientific or political) authority. Depending on the search terms used and the research behavior of the participants, they are confronted with different stimuli and distractors in a free web search, i.e., stimuli and distractors may likely vary significantly from person to person. Thus, while we can control the quality of the websites linked in the CORA tasks, the quality of all other websites that students are confronted with during their Internet research depends solely on their search behavior and can be controlled in the assessment only to a limited extent.

### Stimuli and Distractors of the Linked Websites

Low-quality information on the linked websites can be caused by a lack of expertise of the author(s), belief-related bias, or accidental errors when drawing inferences or citing from other sources. Moreover, the linked sources offer contradictory information or inconsistencies between multiple online texts, which learners need to resolve in the process of acquiring consistent knowledge. In our example (**Figure 2**), the provided link leads to a website that offers information about vegan protein sources. At first glance, the website seems to provide accurate and scientifically sound information about vegan nutrition and protein sources, but upon closer inspection, the information turns out to be biased in favor of vegan protein sources. The article is shaped by a commercial interest, since specific products are advertised. This bias can be noticed by reading the content of the website carefully and critically. The existence of an online shop is another indication of a commercial interest motivating the article. In contrast, the references to scientific studies give a false sense of reliability.

As the construct definition of COR states (Section Construct Definition of Critical Online Reasoning), if students wonder about the trustworthiness of certain online information during the inquiry, this should be a sufficient initial stimulus to activate their COR abilities. Thus, we explicitly include the stimulus at

the beginning of an inquiry task prompt of all CORA tasks. The in-task cues can tap these activation routes even if the students did not respond to the initial prompt at the beginning of the task (**Figure 2**). In the example, the participants are also asked whether the website is reliable to stimulate the COR process and a web search.

Following the ECD (Mislevy, 2017), we describe the *task model* and the *student model* of the CORA in more detail.

### Task Model

Task difficulty in terms of the *cognitive requirements* of the construct dimensions of COR varies through the task properties and the prompt (i.e., difficulty of deciding on a specific solution by considering pros/cons or both). For instance, in the dimension of OIA, task difficulty varies in terms of whether students are required to evaluate a website and related online sources or only a claim and related online sources. The quality of the websites found in the free web searches is likely to significantly vary between test participants, which is not explicitly controlled for in the task and in the scoring of the task performance. This information is only examined in additional process analyses using the recorded log files (Section Analyses of Response Processes and Longitudinal Studies).

In the easy CORA tasks, the web authors were aware that they may be biased and alerted their audience to this fact, for instance by stating their stance directly or by acknowledging their affiliation to a certain position or perspective—the students then had to take these statements into account in their evaluation. In the difficult CORA tasks, the web authors actively tried to conceal the manipulative or biased nature of their published content—and the students had to recognize the techniques these authors employed. In addition, they had to identify the severity of this manipulation and to autonomously decide which information was untrustworthy and should therefore not be taken into consideration. This untrustworthy information can comprise a single word or paragraph, an entire document, all content by a specific author or organization, or even entire platforms (e.g., if

their publication guidelines, practices, and filters allow for low-quality information) or entire geographical areas (e.g., due to biased national discourse).

For each CORA task, we developed a rubric scheme that describes the aforementioned specific features of the websites linked in the task, for instance in terms of credibility and trustworthiness of the information they contained (for details, see Section Scoring Rubrics). To develop the psychological stimuli encountered in CORA tasks (in accordance with the construct definition; Section Construct Definition of Critical Online Reasoning), we based our approach on a specific classification of misinformation by Karlova and Fisher (2013) and on classifications of evaluative criteria of information quality (e.g., topicality, accuracy, trustworthiness, completeness, precision, objectivity) by Arazy and Kopak (2011), Rieh (2010, 2014), and Paul and Elder (2005).

Cues indicating trustworthiness or lack thereof were systematized in evidence strands according to the Information Trust model (Lucassen and Schraagen, 2011, 2013). The model distinguishes evidence on author, content, and presentation, which are aligned with classical routes of persuasion in rhetoric; each requires a different evaluation process. We expand this model by a distinction of personal evaluation vs. trust in a secondary source of information (Table 1).

The *task difficulty level* was gauged in particular by the scope and extent of misinformation based on an adaptation of the Hierarchy of Influences model by Shoemaker and Reese (2014), which assesses agents in the media production process and their relative power to shape the media message—and hence introduce error or manipulation, which need to be judged by students to discern the limits of warranted trust (e.g., at the bottom end are obvious deceptions and errors by the author such as SPAM emails or simple transcription mistakes in a paragraph, while at the top end are high-level secret service operations or a society-wide cultural misconception).

Task difficulty in terms of required argumentative reasoning in CORA was varied in three ways: (1) Scaffolding was added to the task prompts by asking students only for part of the argument (e.g., only pro side, con side, or only specific sub-criteria) to reduce the necessary reasoning steps. (2) The stimuli websites were selected by controlling for (i) scope and (ii) order of bias or misinformation, and for how difficult it is to detect it. *Scope* refers to the comprehensiveness of biases or misinformation based on the adapted Hierarchy of Influences model by Shoemaker and Reese (2014). The *order* is the level of meta-cognition that needs to be assumed in relation to a bias or misinformation. (3) The composition of sources that can be consulted for information (i.e., number of supporting and opposing, or high-quality and low-quality sources) can again be modified only in a closed Internet-like environment (Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019), but it can hardly be controlled for on the open Internet.

The natural online environment used in this assessment constitutes a crucial aspect of the CORA task difficulty (that is also related to the reliable scoring of task performance; Section Scoring Rubrics). In a closed information environment with a finite number of sources, a comprehensive evaluation of all

sources is possible. On the Internet, an indefinitely large number of sources are available. Hence, when solving the CORA tasks, students also need to constantly decide whether to continue examining a selected source to extract more information (and how deeply to process this information, e.g., reading vs. scanning) or whether to attempt to find a more suitable source, and a sample of search hits on a search engine results page, and whether they should use different search terms or even switch to a more specialized search engine or a specific database that might yield more useful information. This aspect is related to the *student model* and the primary aim of students in the context of inquiry-based learning based on online information, which is to gather information to “fill” their knowledge gaps while carrying out a task. Learning in an online environment requires students' initial (and later updated) understanding of the problem in relation to a specific generic or domain-specific task, and recognition of the types of information that are needed to solve a given problem, and then carrying out the steps to locate, access, use, and reason based on information, and finally formulate an evidence-based solution to the problem.

## Student Model

The expected response processes while solving the CORA tasks can be described with a focus on their *basic phases* based on the abovementioned Information Problem-Solving using the Internet (IPS-I) model (Brand-Gruwel et al., 2009): (1) *Defining the information problem*, (2) *Searching information*, (3) *Scanning information*, (4) *Processing Information*, (5) *Organizing and presenting information*. These phases are quite common in many other models and categorizations of information search and also media and digital literacy (e.g., Eisenberg and Berkowitz, 1990; Fisher et al., 2005). For a multi-source information problem, we expect that the processes will be iterated for each new source. An additional meta-cognitive *Regulation* component guides orientation, steering, and evaluation, and can be active throughout interacting at each phase (Brand-Gruwel et al., 2009). Required judgments of information trustworthiness can be situated in the meta-cognitive component of evaluation. Within the evaluative process, trustworthiness judgments might be juxtaposed with judgments of accessibility, relevance and/or usefulness at several points *in addition to the ongoing collection of information for the inquiry*. Based on these categorizations, we developed a fine-grained description of the (sub)processes the students are expected to perform while solving the CORA tasks (Table 2).

In the following, we describe the *student model* with regard to the four main COR facets in more detail.

In CORA, the test takers are required to produce an argumentative conclusive written response based on the consulted and critically evaluated online sources. In line with the older IPS model (Brand-Gruwel et al., 2005), we have added a reflective metacognitive review as an expected process, which may occur at any moment but possibly upon response verification, to highlight that COR may be activated even after an iteration of the IPS-I process or after the whole CORA task has been completed without critical consideration.

**TABLE 2 |** Possible and necessary processes contributing to quality criterion judgments, with web content considered, attributed to IPS-I phase (on evaluative review, see CORA-MCA and IPS).

IPS-I phase	Judgments			
	Accessibility	Relevance	Trustworthiness	Usefulness (task coverage)
Define information problem		o projection of possible relevant information		
Search information	x ignoring inaccessible sources (language, no link) (unless useful based on "scent")	o specifying relevant search terms x judging SERP results for task relevance	o considering URL for trustworthiness x specific fact-checking search	
Scan information	o managing accessibility on site or abandoning site (paid service, media activation; perceptual aides)	x overview and attempt to locate relevant information x title and topics match to search purpose o in-site search	o considering credibility cues	x amount of information appraisal
Process information	o difficult language (skim over, research, or abandon source)	o continuous evaluation during reading (on-task monitoring)	o evaluating argument o evaluating identified cues	o post-processing appraisal of coverage (and further steps)
Organize and present information		x selecting information		o on-task monitoring
Evaluative review (anytime)	o reflection on missed sources	o verification of relation to task	o doubt may re-activate COR	o post-task evaluation

x = required/necessarily tapped; o = possible/can manifest.

Judgments of usefulness, accessibility, relevance, and trustworthiness of online information can be attributed to the COR facet CIE that is represented as a (meta-)cognitive evaluating component in the IPS-I model. A judgment may require a more elaborate evaluation based on additional information searches. Hence, we assume that a spontaneous trustworthiness judgment can occur at any stage in the IPS-I model. Additionally, a more deliberate, likely criterion-based, reflective evaluation of information, for instance in terms of its trustworthiness, can be performed as a specific (scheduled) sub-stage—if the student is aware of the need to evaluate the information.

The sections in the IPS-I process for evaluations of trustworthiness and other judgments also indicate that they can be interwoven with comprehension and reasoning activities and with each other (Section Construct Definition of Critical Online Reasoning, **Figure 1**). However, they are also likely to be distributed across several stages and differ in the content of partial evaluations and possible inferences drawn. The more detailed view of judgments and evaluations by search phase indicate that several judgments are likely to occur per phase and judgments of accessibility, relevance, trustworthiness, and usefulness are *differentially important across phases* and touch upon *different sub-questions per phase*. For instance, trustworthiness evaluations can be both fast, if an exclusion criterion is found, or gradual over one or several stages, including the collection of multiple cues. We assume this to be the case for information and sources that students evaluate as part of the CORA task. For other additional sources found during web searches, it is likely the student will evaluate trustworthiness just once and with little effort, i.e., heuristically, if they know

they can go back to searching a more trustworthy source faster. (i.e., it is not the student's intention to find and determine every untrustworthy source on the Internet, but find one that is not untrustworthy and meets their needs). We therefore expect that the CORA tasks tap a judgment of trustworthiness, including either a systematic criterion-based evaluation (to the extent to which the test-taker is aware of criteria for trustworthy or untrustworthy information) and/or a vigilant recognition of the specific information features that may help the participants identify bias and misinformation.

In this context, the Information Search Process model (Kuhlthau, 1993) links behavior, cognition, and affective responses, with cognition being characterized by gradations of *self-awareness* regarding information search process (**Figure 1**). Here, again, we assume multifold interrelations with the metacognitive facet of COR. Therefore, we expect that recognizing a cue in the linked information in the CORA task (i.e., stimuli), indicating possible bias or misinformation, may activate a "vigilance state" in the students and lead to certain (subconscious) reactions (and a habitual affective response, e.g., anxiety, excitement) or can also be interpreted as an indicator of a potential problem ("am I being lied to?/at risk after misjudging the information?") at the metacognitive level, which may lead to the activation of the facet of COR, i.e., (meta)cognition for critical reasoning activation (on the role of uncertainty and emotions when searching for information, see Kuhlthau, 1993). In this regard, we consider the ambiguity experience as the initial stage in a general critical reasoning and evaluation process, i.e., a cognitive appraisal marked by uncertainty about the validity of one's interpretation of the current situation that leads to a need for more clarity (or to avoid the problem-solving situation, e.g.,

in case of a low self-efficacy), which may prompt an expected response behavior during the CORA tasks, i.e., critical reflection and evaluation.

In terms of the *task model*, this ambiguity is tapped by the CORA task description and the prompt, which explicitly asks students to judge the trustworthiness of a given website or claim. Thus, the task prompt is the initial stimulus for students to activate their trustworthiness evaluation since the question of whether or not information is trustworthy is explicitly given by the task prompt; the second are the cues offered by the stimulus materials embedded in the CORA tasks; the third is the reminder in the response field of the CORA task to formulate a short statement to the task questions and to list the consulted online sources (Section Scoring Rubrics). The CORA task prompts explicitly require students to formulate a response, justify it with reasons and arguments, and back these up by citing URLs of sources used to reach their decision. Thus, students' responses comprise the fundamental components of argumentative reasoning (Section Theoretical Components of COR). In CORA, we framed the trustworthiness evaluation through an *argumentative model*, and modeled (possible) stances on a (trustworthiness) issue and their supporting reasons and evidence (cues). Alternatively, students might not reason deeply about it, but apply cognitive heuristics (Kahneman et al., 1982; Metzger and Flanagin, 2013). However, given that it is explicitly prompted in the task, we expect students to apply argumentative reasoning and to be able to identify cognitive heuristics (e.g., authority biases) within their argumentation.

In this context, one aspect is particularly important in terms of the *interpretative model* (Section Scoring Rubrics). Assuming that cognitive biases (e.g., confirmation bias), and motivated reasoning can be tapped by controversial topics as presented in the CORA tasks, an opposing stance toward a given topic (i.e., skeptical) affords more stimuli to be critical and motivates the student to find evidence of misinformation. This is why a balanced selection of various topics was established in CORA. We assume students' initial personal stance on the task and its topic will depend on a number of influences, controlled for in CORA (e.g., prior domain knowledge, attitude toward the task topic). This aspect is crucial since students may pass different credibility judgments and follow diverse reasoning approaches depending on their initial stance (Kahneman et al., 1982; Flanagin et al., 2018). At a later, longitudinal research stage (e.g., in the context of formative assessments; Section Analyses of Response Processes and Longitudinal Studies), attitude-dependent tasks can be administered to assess COR levels among students for topics they explicitly support or oppose. Not solving the task in a way that accounts for both perspectives would therefore yield a lower CORA test score (Section Scoring Rubrics). This in turn would strengthen the high ecological validity of CORA.

Whichever stance the students choose, they will not be awarded points unless they provide warrant through reasons and arguments, and back them up with evidence from the evaluated website and further consulted online sources (Section Scoring Rubrics). Thus, an evaluation supported by reason and evidence (such as a link to an authoritative website), judged by raters as acceptable against a generic or domain-specific quality standard,

is used to infer the extent to which students' have critically reasoned with and about online information. The call to justify is explicitly prompted in the task ("provide a justification") and the backing with evidence is required in a separate field, asking for the URLs of consulted further websites. Providing citations is a common form of evidence in academic writing, and copying a URL does not require an elaborate evidential standard. The reasons and arguments students cite in their written responses are scored for plausibility and validity based on a few rules (e.g., "trusting only the source's own claims about itself is not sufficient reason"). The indicated URLs are also evaluated in terms of their trustworthiness (Nagel et al., 2020). We assume that students with advanced COR abilities cite only the best sources they found and used to back up their argument. Conversely, indicating many relevant and trustworthy sources as well as irrelevant and untrustworthy ones was considered an indicator of reasoning that was not fully sufficient (see scoring rubrics in Section Scoring Rubrics).

According to the fundamentals of argumentation, the main claim, reasons, and backing (e.g., evidence) are the basic elements of a reasonable argument (Toulmin, 2003; Walton, 2006). Hence, we considered indications of these, which are also explicitly prompted in the CORA task, in a somewhat aligned manner in the students' responses as evidence that students performed argumentative reasoning. Some argumentation frameworks include further basic components, such as rebuttal and undercut as types of opposing reasons or inclusion of consequences (Toulmin, 2003). These components can be included in further CORA tasks (Section Refining and Expanding CORA), but were not required for the short online evaluation tasks. Moreover, in terms of metacognitive evaluation, students are expected to engage the evaluative critical reflection, i.e., "self-reflective review" of their task solution after formulating their response to the task.

## Scoring Rubrics

According to the task and the student models, CORA tasks measure whether students employ critical evaluation of trustworthiness and critically reason based on arguments from the online information they used. Based on our prior research on performance assessments of learning (e.g., for the international Performance Assessment of Learning (iPAL) project, see Shavelson et al., 2019) and the developed scoring approach, we created and applied new scoring rubrics focusing on the main facets of COR and on fine-grained differentiations of scoring subcategories in accordance with our construct definition (Section Construct Definition of Critical Online Reasoning; for an excerpt of the facet "weighting arguments," see Table 3).

Each task is scored with a maximum of 2 points, with to 0.5 points awarded if the response mentions a major bias or credibility cue, for instance noticing a (covert) advertising purpose, and if its implications for the interpretation of information are identified. Up to 0.5 points are awarded if the students support their claim (no matter which stance) with one or two valid reasons that are weighted in relation to each other, and a maximum of 0.5 points if students refer to one or two credible external sources (that are aligned with their overall



**TABLE 3** | Excerpt of the COR scoring scheme; REAS facet, sub-facet “weighing reasons.”

COR facet	Sub-facet	Description	Subscore “weighing reasons” <sup>a</sup>	1	2	3	4	5
				Not fulfilled	Mostly not fulfilled	Partially fulfilled	Mostly fulfilled	Completely fulfilled
Reasoning using Evidence, Argumentation, and Synthesis	Weighing reasons	<ul style="list-style-type: none"> <li>Balanced judgement: pros and cons/diverse perspectives</li> <li>Conclusive argument, comprehensible justification</li> </ul>	0.5 points total 0.1 points per gradation	0.1 tasks not fulfilled at all	0.2	0.3	0.4	0.5 tasks completely fulfilled

Note: completely fulfilled (full score), mostly (not)/partially fulfilled (few, half or most, but not all aspects covered), not fulfilled (zero points).

<sup>a</sup> The complete scoring scheme describes all major facets with scored sub-facets. This excerpt shows the REAS facet with the sub-facet “weighing reasons”.

argumentation). Furthermore, students can achieve 0.5 points if their response is coherent, clearly related to the task prompt, and covers all sub-parts.

In contrast to a simple trustworthiness judgment, which could be performed without further reflection using heuristics, the underlying analytical reasoning requirements of the tasks are more demanding. It is also possible for participants to take the evidence for their criticism of a website from the website itself as long as the argument is warranted and conclusive. Consequently, the scoring rubrics also consider to what extent the students recognize the specific characteristics for or against the trustworthiness of certain websites, cues, and strands of evidence, and whether they consider them in their reasoning and decision-making processes. A student may identify manipulative techniques “X” and “Y” being used by the linked website, which make it untrustworthy, and cite them from the website. In this case, students can receive points for correctly judging the website as unreliable and for identifying a bias, even if they have not accessed external websites. In a follow-up study, in addition to this holistic score per task, further sub-scores can be awarded at different levels of granularity in accordance with the COR construct definition (Section Development of Scoring Modular Rubrics).

Regarding *information trust* strands of evidence, before scoring this aspect of students' responses, we evaluate the stimuli in the CORA tasks in terms of type, number, and location of cues for/against the credibility of a website (Section Test Definition and Operationalization of COR: Design and Characteristics of CORA). In addition to evaluating the stimuli individually, we mark their valence and importance for main argumentative claims (e.g., supporting or contradicting the trustworthiness of the linked website). Given the large number of possible cues, we make some systematic limitations: the collection of cues is mainly restricted to the stimulus materials to be evaluated by all participants. These cues are listed and scored depending on how frequently they are mentioned in the students' argumentative responses (i.e., focus on cues that students selected). In terms of possible verification of plausibility of reasons, we distinguish first-order reasons (e.g., “the website has an imprint”), which may lead to a successful judgment in certain cases and guard against some deceptions if only cues regarding credibility are used, to second-order reasons (e.g., “any website can have an

imprint nowadays, but the indicated organization cannot be found online”).

Further, the cues were systematized following three strands of evidence in accordance with the 3/S Information Trust model (Lucassen and Schraagen, 2011) and Prominence Interpretation Theory (Tseng and Fogg, 1999), including *surface/design*, *semantics/content*, and *source/operator*. Each of these strands can make a specific contribution to an argument about whether to trust information or not. Moreover, they address different reasoning approaches from “aesthetic” appraisal and consideration of mediated presentation, to content and argumentative appraisal as well as to consideration of authorship reputation, intent and expertise (and other cues of the production/publication process).

In addition to the described strands of evidence, the model was expanded by distinguishing a primary- and a secondary sources perspective for each strand. Usually, both perspectives will be used to some extent for an evaluation of trustworthiness, i.e., when verifying a cue oneself, evidence standards (standards related to the information itself) are used than when relying on other persons' judgments (here, one rather uses standards related to the probability of successful judgment of the other person). For example, when judging trustworthiness of the author, a student may complete their own research on relevant aspects from a variety of biographic sources or they may follow a journalist's understanding of this author. Verifying every aspect oneself marks a fully autonomous learner, though we acknowledge that this may not be feasible for all aspects in the short test-taking time. For each task, strands containing important cues were listed. Moreover, major distractors supporting a competing assumption were marked.

The rating was carried out by at least two trained scorers per task. For the overall CORA test score, i.e., the average scores of two or three raters for each participant and for each task, a sufficient interrater agreement was determined, with Cohen's kappa >0.80 ( $p = 0.000$ ).

## PRELIMINARY VALIDATION

The validation of the CORA was integrated with the ECD and follows the AERA Standards (Section Research Objectives

and Questions). Starting from the holistic nature of the CORA (see section Task Model), the construct specification, and the modular extensions of the scoring in this paper (see *Interpretative Model*), we present preliminary validity evidence related to the content of the construct. After the COR construct specification and the assessment design, the newly developed CORA tasks underwent content analyses and were submitted to expert evaluation during interviews. The aims were to examine the coverage of the theoretically derived COR construct facets by the holistic tasks and to obtain expert judgment regarding the suitability of the content and requirements for higher education in Germany. Below, we outline the methodology (Sections Content Analysis: CORA Task Components as Coverage of the Construct and Expert Interviews) and discuss the results for both analyses (Section Findings From the Expert Interviews and Content Analysis).

## Content Analysis: CORA Task Components as Coverage of the Construct

A qualitative content analysis (Mayring, 2014) of the CORA tasks was carried out by the CORA research team members who participated in the construct specification but not the selection of task stimuli. Task prompts and the encompassed stimuli were examined to determine the presence or absence of features that would allow test-takers to draw inferences and generate responses worth partial or full credit according to the scoring rubric (Section Scoring Rubrics). The six higher education CORA tasks that resulted from the design process (Section Test Definition and Operationalization of COR: Design and Characteristics of CORA) were coded according the following features and underlying (theoretical) frameworks:

- (1) As part of the meta-cognitive facet, activation of COR was coded to gather evidence on whether the tasks tap students' overall COR ability, i.e., whether they convey a need for critical evaluation and argumentative reasoning and at which point: at the beginning, middle or end of the task. We coded for activation by prompt or by context, by specific cues that would highlight the need for COR during task processing, and for end-of-task activation by required (metacognitive) review steps or invited by a contradictory or uncertain preliminary conclusion. The expectation was that at least some tasks would have a cue for COR activation at the beginning of the task, whereas others may only have a mid-task activation to tap students' ability to identify situations when it is needed to activate their COR.

Moreover, the aspect of problem definition (in the sense of the IPS-I model, **Table 2**) was examined. We coded whether the task was embedded in a broader activity context to support judgment based on purpose and increase ecological validity (e.g., judging information trustworthiness *for use in a term paper*); in a pretest during task design, students had claimed to apply more or less rigorous evidence standards depending on purpose. We also coded whether the task goal was clearly stated in the prompt and whether solution criteria were given or if they needed to be inferred.

Other MCA subfacets regarding regulation, affective response, or attitudinal aspects were not coded due to the difficulty of assigning them to specific task features (in the online assessment); these could be elicited more efficiently in a future coglabs study (Section Analyses of Response Processes and Longitudinal Studies).

- (2) The OIA and CIE facets were assumed to be organized in order of the phases of the IPS-I process model to highlight similarities and differences among the CORA tasks, while specific features were coded based on other additional models and research foci (Section Scoring Rubrics). The phases of source selection and initial scanning of a website were listed under one facet (OIA or CIE), but are expected to be hybrid search and evaluation activities (to be further examined in coglabs; Section Analyses of Response Processes and Longitudinal Studies).

Among the search-related aspects (OIA), we coded the necessity to use different search interfaces during the process (e.g., a search engine, in-site search) to obtain reliable information. We assumed that basic search skills, but not use of advanced search operators or special databases would be required. Websites that were inaccessible and media that would not play or were too long and not searchable were excluded during the pretest. Hence, suitable information was expected to be fairly easy to locate and access (except on specific search tasks) by performing an external search.

Regarding information source selection, we generally coded the sources students had to evaluate to obtain suitable information, i.e., the given website, additional websites, and linked sources (e.g., a background article to a tweet), and/or websites which students selected themselves. We expected requirements to vary across tasks.

- (3) The facet of CIE united the IPS-I phases of scanning a website and in-depth information processing. For global website appraisal and orientation, we coded to what extent it was necessary to judge the overall layout and design (or if one could ignore the context and start reading/searching immediately), to what extent students needed to get an overview first, for instance to find a suitable paragraph in time by scanning sub-headings, and determine if they had to attend to any specific cues rather than reading the main text. We expected that some websites might have obvious design cues and others might not (e.g., a popular social network could be interpreted as an obvious cue for lower credibility); some websites were expected to be more complex or longer and require initial orientation; however, we expected students to find relevant information on the given landing page and standard sub-pages (e.g., publisher and author listed in the legal notice or "about" page); we expected the task solutions to not be based solely on the identification of a single cue.

Regarding information processing, we generally assumed the required reading comprehension to be a given among higher education students and focused on evidence evaluation, classifying available cues based on the 3'S' Information Trust model (Lucassen and Schraagen, 2011) into cues in the design, content, or source, as well as (jointly

for all three) secondary external sources indicating cue evaluations (Deferring judgment to external sources would also require an evaluation of these sources' expertise and intent). For example, if a website had aggressive popup advertisement, this would be coded as a cue in the design that might indicate lower trustworthiness. We expected that not all tasks would have cues for (un)trustworthiness in all strands, but at least in one strand of evidence. Moreover, different strands of evidence would be tapped across tasks so that no single subset of evaluation skills or strategy (e.g., only using logical critique or looking up the author's reputation) would be universally successful.

- (4) Based on major components of reasoning (Walton, 2006) with evidence, argumentation and synthesizing (REAS), we coded to what extent students needed to cite sources of evidence (expected), to what extent they had to provide reasons why they trusted the information (on some tasks), or arguments against its trustworthiness (expected), to what extent they needed to make an overall evaluative judgment (expected), and to what extent they had to synthesize and weight possibly contradicting information and arguments (expected), to what extent stimulus materials contained a prominent bias, mismatched heuristic, or fallacy to be avoided (expected for most tasks), and if there was a clear-cut solution vs. an undecidable outcome so they had to account for uncertainty (only on few tasks).

In regard to presentation of results (another IPS-I phase), we coded to what extent the quality of the structure and phrasing of students' responses contributed to their score. As we focused on the quality of argumentative links and information nodes rather than their rhetorical arrangement, we expected response structures and phrasing to not matter beyond the general effort of presenting a coherent and conclusively argued response.

- (5) In addition, given that domain- and topic-dependent prior knowledge (and attitudes) might influence participants' searches, evaluation, and reasoning, we collected some descriptive information on the task topics: We labeled the origin of the misinformation as an indicator of how widespread and hard to identify a deception might be (e.g., from a single author's error on a page to a newspaper editorial board's agenda-setting policy to a culturally normalized conviction), as suggested by the Hierarchy of Influences model (Shoemaker and Reese, 2014). We coded the share of supporting and opposing (in terms of the task solution: conducive or distracting) search results for the key terms in the prompt and website title (as an indicator of controversy and how easy it was to find additional online information). We labeled the broader task context in terms of societal sphere (commerce, science, history etc.), the kind of misinformation genre, specific biases, heuristics, and fallacies presented, and the type of online medium. The overall expectation was that CORA tasks would present one or two challenging aspects but not be overly difficult given the short testing time (e.g., no national scandal to be uncovered), and would be varied in their genre and contexts. Results are summarized in **Table 4**.

## Expert Interviews

Semi-structured expert interviews (Schnell et al., 2011) provided a second source of evidence on content representativeness. In semi-structured interviews with experts, we presented examples of CORA tasks and asked experts to comment on *their suitability for higher education in Germany*. The interviewed experts were leading academics in their field and included two of the U.S. developers of the civic online reasoning assessment, four experts in computer-based performance assessments in higher education, and six scholars from the fields of media studies (who focus on online source evaluation or media literacy), linguistics, and cultural studies. After considering the task stimuli, prompt, and rubrics (sent to them in advance), the experts were given the opportunity to ask for clarifications and were then asked to share their first impressions of the assessment before responding to more specific questions regarding the tasks and features. The discussed topics are shown in **Table 5**.

The questions were asked in view of the German context and tasks specifically, since the media landscape and typical challenges with online information, including deception strategies, can be country-specific. Experts' responses were interpreted in light of their disciplinary backgrounds and convergence or divergence between experts.

## Findings From the Expert Interviews and Content Analysis

In the following, we present a summary of the main findings from the expert interviews and content analysis.

### Overall Experts Evaluations

Overall, with regard to the suitability and validity of the CORA tasks for higher education in Germany, most experts agreed and confirmed the content and ecological validity of the CORA tasks and recommended further expansions. For instance: "The task is clear, the instruction is also clear, and it seems obvious that they need to formulate a response."

One expert, after pondering how to translate and adapt the U.S. tasks, and worrying about cultural suitability, considered the CORA tasks and commented: "These [German] tasks are really a hundred times better for Germany."

### Coverage of COR Facets

One question critically discussed with experts addresses the domain-specificity of the CORA tasks. Here, the experts confirmed that the six tasks cover generic COR ability. For instance: "No domain-specific knowledge is required. It's a good selection for the news/science context."

One concern that was raised by most experts regards the suitability of the testing time to assess all facets of COR, and in particular the REAS facet. However, experts also agreed that students may not dedicate more time to the task when evaluating an information source in a real setting. As one expert notes: "There are 10 min to conduct a search. One may doubt if people would commit as much time in everyday life, unless they really took the time to carry out a more detailed search." At the same time, the natural online environment of the assessment was praised in terms of the high ecological validity of CORA by all

**TABLE 4 |** Content analysis of the CORA tasks as coverage of the major facets of the construct.

COR facet	1	2	3	4	5
<b>A. Metacognitive Activation (MCA)</b>					
<i>COR activation</i>					
Initial activation by prompt	++	++	++	++	++
Initial activation by context	0	0 (+)	2nd	0 (+)	0
Mid-task activation by cue identification	+	opt	2nd	+	2nd
End-task activation by synthesis outcome	0	opt	0	opt	opt
End-task activation by review process	0	0	0	0	0
<i>Problem definition</i>	see IPS-I (Brand-Gruwel et al., 2009)				
Clear purpose of activity	0	0	0	0	0
Clear task goal	++	(+)	++	++	(+)
Determining criteria	+	+	+	+	+
<b>1. Online Information Acquisition (OIA)</b>					
<i>Search</i>	see IPS-I, Interactive Information Retrieval (Xie, 2008)				
Search engine use	opt	+	+	opt	+
Data base use	0	opt	0	0	0
Defining query terms	+	+	+	+	+
Spec. search (e.g., operators)	0	0	0	0	0
In-site search	0	+	0	0	+
<i>Information source selection</i>					
Evaluating given source	+	0	+	+	+
Evaluating linked sources	0	0	opt	0	opt
Evaluating self-selected sources	opt	+	+	+	+
<b>2. Critical Information Evaluation (CIE)</b>					
<i>Scanning (global site appraisal)</i>	see IPS-I				
Of design	+	0	+	+	+
Of structure	+	0	+	0	+
Of spec. features	+	0	opt	0	0
<i>Information processing; evaluation of strands of evidence</i>	see IPS-I; Information Trust (Lucassen and Schraagen, 2011); Prominence Interpretation Theory (Tseng and Fogg, 1999)				
Cues in design	++	0	0	+	0
Cues in content	0	+	+	+	+
Cues on author/publisher	+	0	+	0	+
Cues reported on in external sources (alternatively: expertise and intent)	Opt	++	+	+/-	+
<b>3. Reasoning using Evidence, Argumentation, and Synthesis (REAS)</b> see Fundamentals of Critical Argumentation (Walton, 2006)					
Citing external sources of evidence	opt	+	opt	opt	+
Generating supporting reasons	opt	opt	opt	+	+
Generating opposing reasons	+	+	+	+	+
Making a holistic evaluative judgment	+	+	+	0 (scaff)	+
By synthesizing and weighing information	0	+	+	0	+

(Continued)



**TABLE 4 |** Continued

COR facet	1	2	3	4	5
Avoiding biases, heuristics, fallacies	+	+	+	+	+
Adjusting for uncertainty	0	0	opt	opt	+
Structuring and presenting results			see IPS-I		
Organizing structure	0 (scaff)	0 (scaff)	0 (scaff)	0 (scaff)	0 (scaff)
Formulating response	+	+	+	+	+
<b>B. Descriptive Features Hierarchy of Influences (Shoemaker and Reese, 2014)</b>					
Origin and diffusion of misinformation	organization	personal	professional association	news editor	think tank
Share of supporting/distracting external sources	+ / 0	+ / - -	+ / -	0 / -	+ / -
Social context	DE: commerce (1), politics (2), society/ethics (2), history (0), science (0),... US: commerce (0), politics (3), society/ethics (3), history (1), science (0),...				
Misinformation genres	corporate educational texts, hidden advocacy, tendentious media commentary, social media rumor,...				
Specific biases, heuristics, fallacies	commercial bias (limited information selection, overstatement of pros, understatement of cons), oversimplification of opposing stance, ideological bias (religion, economic policy), unsupported one-sided prediction, baseless ridicule,...				
Media type	news article, website, tweet, Facebook post, and news video				

Features per individual task; descriptive features aggregated across tasks for context, genre, biases, media.

++ = cue is obvious from task (e.g., explicit prompt).

+ = cue is present but needs to be identified and used in inference; i.e., ability likely tapped.

opt = cue is available for use, but not required for optimal solution; i.e., supporting sub-facet likely tapped.

0 = cue not available or offers no information; i.e., applying ability on problem yields no result.

0 (scaff) = requirement scaffolded; i.e., partial solution is given.

- = cues misleading; i.e., distractors present, critical evaluation (selection) likely tapped.

-- = most cues misleading; i.e., distractors present, critical evaluation (selection) and search skills likely tapped.

2nd = cue present, but can be identified only after successful inference, e.g., after deception or bias detection, i.e., critical evaluation and argumentation (weighing of alternative explanations) likely tapped.

**TABLE 5 |** Evaluation questions for experts (selection).

Suitability for higher education in Germany	To what extent are the COR tasks suited to assess students' ability to critically use online information and reason based on it?
Coverage of facets	Which (sub)facets of COR might not be assessed by the tasks?
Representativeness of media	Are the media in the stimuli representative of the information environment higher education students typically encounter online?
Representativeness of misinformation types	Are the kinds of biased information representative of the types/genres of misinformation and biases that students should be able to recognize? Which ones might be missing?
Difficulty and source use	How do you judge the difficulty of tasks?/What university level are they suited for? What aspects might contribute to the difficulty? Do students need to evaluate other sources that are significantly more or less difficult to evaluate?
Potential for differential item functioning (DIF)	Would you expect any group of students to perform better on the tasks (e.g., depending on gender, age, study domain)?/Does the assessment disadvantage any group? Does the assessment have a potential bias?
Item design	Which aspects are particularly important to select or construct realistic tasks with adequate difficulty (to tap COR abilities)?

experts: "The mode of administration as given here is important, since it enables assessing internet search behavior."

The new rating scheme with the subscores and evaluation categories was positively evaluated by the experts, although they stressed the high complexity of the scoring rubrics. For instance: "It is also good that you have different degrees, not only "right or wrong." Of course, this places high demands on coders, but with training, it is doable."

## Representativeness of Media

Most experts positively evaluated the representativeness of the chosen media, i.e., media that students frequently use online. However, one expert criticized that "scientific and journalistic media were indeed covered, but the selection could include more reputable media as well as some media more on the lower quality end of the spectrum. The ones here are well chosen; one cannot immediately tell if they are fabricated or not." Another expert

proposed: "These are common media sources. However, you may include even more social media, and not only evaluate news by institutions and organizations, but also by individual users or from the "alternative" news outlets. Influencers on Instagram who present products are another option."

### Representativeness of Misinformation Types

In terms of the presented misinformation, the overall judgment by the experts was positive. For instance: "Item topics are nicely varied; tasks are not too simple, so one does not get bored; and I could not decide right away, I had to click on the [background source] and take a look. Even as a media-competent person I had to examine it to make a judgment." Another expert stated: "I could not solve the items without checking. I had heard nothing about these cases. With unknown issues, ideology also plays a smaller role." In terms of potential biases and DIFs, the experts did not express any concerns. For instance: "I do not think that, given equal competence, it would be easier for students with typically liberal or left-wing attitudes to solve the tasks. The selection of topics in the tasks covers some stances typically accepted in the left and green camp, some typically accepted by the conservative camp.... It is a good mix." In addition, one expert recommended expanding the item pool by a clearly untrustworthy website and one clearly trustworthy website, so that lack of trustworthiness would not be predictable on post-tests. Another expert proposed: "Some other frequently shared information of low trustworthiness can include memes, misattributed or completely wrong quotes, or quotes taken out of context."

### Difficulty and Source Use

At the same time, however, it was questioned whether the task prompts might be too difficult for beginning or undergraduate students. For instance, "Even as a frequent evaluator, I was not always skeptical of the given information." In this context, the appropriateness of the limited testing time was once again questioned. Only one expert was of the opinion that the tapped skills are mastered early on during the course of studies: "What you assess here is what we call study of sources. [...] We teach this the first year in our degree course, and from then on, students should know it, and it is basically part of practice from then on." In this regard, some experts recommend splitting the task into parts that focus on particular facets of COR. For instance, "You could ask for an ad hoc judgment, and have additional tasks [for more detailed search]." Another expert proposed: "If students do not find suitable sources, they may get stuck. Perhaps, it is worth including a separate task format or hints."

Another aspect addressed by most experts concerns participants' prior knowledge, beliefs and critical stances, which may significantly influence their CORA test performance. In this context, one expert stated: "Whether people evaluate sources can also depend on their motivation to put in the time for checking them. Hence, need for cognition could be an influence, people's proclivity to get to the bottom of things and not avoid complexity." Similarly, another expert commented: "People may also carry out a detailed search just

to confirm their worldview or to form an opinion. This can occur despite existing search skills (but they would still be ideologically stuck). Hence, motivation to be open to other positions is key, and it then matters how much time I'm willing to invest in a search." In this context, most experts stressed the need to control for participants' prior knowledge and attitudes. For instance: "Political orientation can be used as a control variable if completely anonymized, for instance asking where they would position themselves on a 1-to-10 left-to-right-wing scale (on a voluntary basis) appears less invasive." Another expert proposed: "You may also want to specify whether it is the successful judgment of a first impression or openness to changing one's opinion. In that case, personality traits would be controlled for. So, a different option would be to assess who changes their mind when they come across new leads."

### Suggestions for the Further Development of CORA

The interviewees did not recommend the exclusion of any tasks. In few cases, the experts recommended removing certain task features. However, the experts provided a number of recommendations in terms of refining the CORA. For instance: "The role of content shared by friends could be expanded, where it is unclear if it has been checked or not... User comments can be read and might influence more passive users... So to increase difficulty, you could add social credibility cues. It would be an even more realistic setting, but you need to see how additional information would influence difficulty." This suggestion is in line with credibility research that highlights the huge role that social persuasion by peers plays in today's social media (Fogg, 2003). Although cues exist in few of the tasks, social persuasion and learning was purposely left for future CORA expansions (Section Refining and Expanding CORA).

### Overview Content Analysis

As task prompts shared a similar structure and wording with only differing topics and source links, the evaluative and argumentative requirements were assumed to be similar as well. The closer content analysis, however, revealed two distinct types of tasks: (1) "website evaluation tasks," tapping particularly CIE, but less OIA if students did not search beyond the presented website; and (2) "fact-checking tasks" that only presented a claim, but no linked website as a stimulus, and therefore forced an Internet search. Fact-checking emphasized OIA more in comparison to CIE since students were not bound to evaluate one particular website; if they were uncertain about a source, they could abandon it to find a better alternative. In this way, the task types afforded use of all three facets but, respectively prioritized one in particular; consequently, a third format emphasizing REAS to complement the other two would be a further development step.

The task response sheet provides students with a clear structure, with the sections overall trustworthiness judgment, warrant (sometimes with separate pro and con sections), and URLs. The scoring rubrics did not contain any specific language requirements. Nonetheless, the students had to fill in the response

sheet sections in a coherent way and formulate a conclusive statement to be awarded points. While the strands of evidence varied systematically, content-related aspects and difficulty were not systematically varied across tasks. At the current stage, given the large number of available topics and types of biases, it is still a small, to-be-expanded task pool (Section Refining and Expanding CORA).

Regarding the individual COR facets, the content analysis showed the following findings:

### Metacognitive Activation (MCA)

In terms of the activation of COR, all task prompts offered clear instructions to evaluate the trustworthiness of the sources at the beginning of the task. Mid-task activation depended on the presence of specific cues. All tasks contained at least one explicit initial and one implicit mid-task cue that might alert students to the need to use their COR. End-of-task activation, for instance a prompt to explicitly review and reflect, was not employed. Moreover, there were no tasks with implicit mid-task or end-of-task activation only, which is a characteristic of deception in online information in real life (i.e., there rarely are prior warnings that a website might contain *misinformation*, compared to automated warnings and filters for, e.g., malware detection). The primary aim of CORA is to measure performance during Internet searches, critical evaluation, and argumentative reasoning; it would be hardly possible to assess these facets if students missed the activation cue.

In regard to the aspect of “problem definition,” while problems were clearly stated in the task prompts, the students need to determine the evaluation criteria for trustworthiness and untrustworthiness themselves. Some experts critically noted that students may be unsure about the required evidence standards. It remains an open question whether deriving criteria for one's trustworthiness judgment should be part of the COR ability. This aspect has been scaffolded in some think-aloud studies, though we are not aware of scaffolding in other assessments. In think-aloud studies, the evaluation of criteria has been separated into different steps; for instance consecutive filtering of sources based first on relevance, then trustworthiness, then usefulness (Walraven et al., 2009; Goldman et al., 2013).

### Online Information Acquisition (OIA)

Regarding expected search skills, content analyses indicated that students can find a suitable source, and in one task a complete website review, even without specific search terms apart from the titles as long as they searched for external sources at all. Only for the fact-checking task did we find an expected larger share of relevant distracting SERP results. For selection of sources for reading, features also varied as expected. Even though some stimuli are quite short (e.g., a tweet), not all students may open the linked background article with more information. However, as this is clearly included as the main piece of evidence backing up the claim in the stimulus, students' attention to the link as a cue and to the background article can be considered a legitimate part of the tapped COR ability. An examination of the SERPs for major keywords showed significant variation in terms of the available information on the first SERP page and across tasks

(see section Descriptive Features), which usually included some supporting, but also multiple irrelevant or misleading sources on the first page. Thus, the tasks appear to tap students' skills in SERP evaluation, as desired; for instance, students need to actively decide which websites to focus on.

### Critical Information Evaluation (CIE)

As expected, some websites contain too much text to process in the limited time and require students to search for or skim the content. Most webpages contained more text on the landing page than fit on a single screen, and had common sub-pages, such as the “legal notice” or “about” section. For their own orientation and for a fast trustworthiness judgment, students need to gain a comprehensive overview of the websites first to be able to deliberately focus on specific sections. Some tasks also required students to recognize and understand cues outside the main text (e.g., an organization logo at the top). This indicates that simply starting to read the text might be an unsuccessful strategy on these tasks and would take too much time.

In terms of strands of evidence, cues were well distributed across tasks, in fact more regularly than expected. There were usually at least two strands of evidence with relevant cues available, so students could take different routes through the task. The linked background webpages usually contained cues that need to be understood and evaluated. Suitable information was also available in (purposefully selected) external sources to help students solve the tasks and, for instance verify the reputation of an unknown author. While tasks could be solved using just one of the available strands of evidence (e.g., only cues on author), combining two or more converging strands could potentially afford higher confidence in task response and possibly minimize effects of interpretation errors. This supports the intended interpretation of task performance.

### Reasoning Based on Evidence, Argumentation, and Synthesis (REAS)

In terms of the argumentative component of COR, students needed to make a judgment in all tasks, mostly by weighting the pros and cons, although some tasks also scaffolded these, asking for both pros and cons separately rather than a final integrated decision. These requirements can be varied more systematically based on empirical evidence regarding task difficulty. All tasks required students to find *disconfirming* evidence or arguments, which supports the interpretation that “critical” reasoning skills are tapped, and some tasks required students to find both confirming and disconfirming evidence or arguments. This could place students who rely only on their confirmation bias at a disadvantage, as intended. However, one expert called for the inclusion of clearly trustworthy or untrustworthy websites to better discriminate performance at the lower skill range and prevent re-testing effects (i.e., students assuming that all websites in the assessment are untrustworthy). While citing external sources is required on all tasks and is often beneficial to building an evidence-based argument, it implies a certain trade-off, as evaluating these external sources takes time and requires a higher cognitive effort. A REAS-focused task format might juxtapose several pre-selected sources with potentially

contradictory information that would need to be argumentatively weighted and synthesized. Such tasks have been developed in the iPAL project (Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019) and MSC research (for an overview, see Braasch et al., 2018).

### Descriptive Features

Task topics were varied, as expected, albeit not all societal spheres were equally covered—experts did not judge this aspect as particularly important. Sources of misinformation, including websites by associations and individuals, small enterprises, and editorial teams, were at the lower to medium level in the Hierarchy of Influences. These sources of misinformation were still mostly identifiable as individual entities within a pluralistic information environment. The CORA tasks did not focus on entities at higher levels of influence, such as media corporations or government agencies. Thus, there were no high-level “scandals” involved (as often referenced in conspiracy-related misinformation). This may imply that the highest levels of COR related to analyzing societal and funding contexts, as typically required from investigative journalists, are not tapped in the CORA tasks. This is reasonable given the time limit and the lack of content-related knowledge requirements. However, as the experts noted, the selection of topics, contexts, and genres covered in the tasks could be varied more systematically (Section Refining and Expanding CORA).

### Summary

The preliminary validity evidence from content analysis and expert interviews indicated some important implications for the CORA. Overall, both the content analysis and the expert interviews indicated that it taps higher education students' abilities to search, access, critically evaluate, use, and reason based on online information in the German context, with a slightly stronger focus on the evaluation components. The preliminary evidence supports our validity claim that the CORA measures the participants' personal construct-relevant abilities in the sense of the defined construct definition (RQ3). Moreover, the expert interviews indicated that the CORA tasks cover a significant portion of the online media landscape relevant for higher education students in Germany as well as typical problems and genres of websites and online texts that call for COR skills in the German higher education context.

## RESEARCH PERSPECTIVES

### Refining and Expanding CORA

The CORA tasks allow for a variety of more detailed (sub)scores, for instance as adaptive feedback based on the navigation logs. Two crucial dimensions that are underrepresented in the scoring so far are the metacognitive activation of the COR abilities in relevant contexts and situations as well as the reviewing of own task-related knowledge and beliefs. This important aspect of COR also aligns with the activation of epistemic metacognition, and can offset own cognitive heuristics (e.g., confirmation bias). COR activation is currently triggered by the task prompt and tapped by the CORA tasks. Sub-tasks

could be developed to assess COR activation in a more focused manner, for instance by using a format that assesses context-dependent choice of action (e.g., decision to evaluate a website or not) as tapped by situational judgment tests (Weekley and Ployhart, 2013). There is also potential for task prompts to include an explicit purpose of the activity indicating subsequent use of the evaluated information. However, as Goldman and Brand-Gruwel (2018) stress, future research might need to focus more intensively on psychological stimuli embedded in the tasks and their complex interrelation with the task solvers' response processes that these stimuli might activate (e.g., rereading, thinking critically).

Metacognitive reviewing of (prior) knowledge and beliefs, accepts the high probability that students do not withstand every manipulation attempt and have likely already acquired prior knowledge based on misinformation, which can only be transformed if it is reconsidered in light of the new knowledge. If an inconsistency between new (warranted) knowledge and prior (misinformed) knowledge occurs, it can only be resolved in an epistemically justified way if the prior misinformed knowledge is altered—the opposite might lead to further misconceptions or motivated reasoning. This can be linked to the epistemic virtue of open-mindedness and implicates that negative experiences and failures can provide unique insights for learners and can be transformed into in-depth knowledge in the future (Oser, 2018), but only if reviewed and successfully reinterpreted by the learner. Hence, conducting an open-minded metacognitive review of prior knowledge and beliefs forms a key component of COR, activated by prompts in CORA tasks.

As another direction for further research, in addition to the generic COR assessment, domain-specific CORA tasks have been developed based on the iPAL assessment framework for specific domains (e.g., economics; Zlatkin-Troitschanskaia et al., 2019, 2020a). Since learning environments and the media used by learners within disciplines change with increasing speed due to digitalization and university students' increasing use of information available on the Internet for their domain learning, we will particularly focus on information gathering and knowledge building from mass and social media when further expanding the assessment of domain-specific COR.

### Development of Scoring Modular Rubrics

A sub-score can be awarded per each single/individual aspect in a facet of COR; that is, for activation of COR, the phases during which a trustworthiness judgment is performed (or not), and additional evaluation process (e.g., a fact-checking search) are initiated (or not) (Section Scoring Rubrics). For the critical evaluation facet of COR, scoring can be extended depending on the strands of evidence used, based on the information trust model (Lucassen and Schraagen, 2011, 2013). Collecting evidence from the three strands—(i) on the author, (ii) design/text surface, and (iii) the content—a more reliable evaluation and reasoning than evidence from only one would be awarded a higher score. Similarly consulting external sources and other's judgments of the same aspect would be awarded a higher score than considering only one. The ratio of self-examined vs. externally consulted vs.



not considered strands of evidence can serve as an indicator of (topic-dependent) intellectual autonomy (Paul and Elder, 2005).

Identifying the (possibly hidden) purpose of a website (e.g., sales, political opinion-forming) is a primary phase in the task-solving process. This also includes recognition and understanding of advertisements and other surface features (e.g., authorship). If these behavior- and process-related facets are included in the scoring categories, a time-sequential diagnosis of the quality of online reasoning is possible. These sub-scores can be further used as a basis for the development of adaptive feedback for teachers and students, which indicates when a student is more or less successful in systematically solving a task (or, e.g., they were spending too much time on searching or on one website).

For the argumentation facet of COR, the score can be further differentiated based on use of each argument sub-component: i.e., central claim, reasons, evidence—and implications for task requiring a recommendation (Walton, 2006; Fischer, 2018). A pool of supporting and attacking reasons can be collected from students' responses, weighted in their contribution to task, and used to score subsequent responses (e.g., depending on whether students' used the most weighted reasons, pros and cons, and only claimed an evidence-orientation, or cited evidence, verified evidence, generated own evidence).

A subscore can be awarded on the level of comprehension and reasoning of single text units. This requires a classification of cues at the text surface as an indication of trustworthy or untrustworthy sources and information. At the moment, this can be efficiently analyzed only for websites given as stimuli in the CORA tasks. The quality of additional websites used by the students can only be estimated based on their URLs. Analyzing the quality of all websites the students accessed while solving the CORA tasks would require comprehensive media-specific and content-qualitative analyses as well as in-depth linguistic and computer linguistic analyses (e.g., text mining). In addition, process data, for instance eye-tracking or navigation logs, can be used to support an on-task detection of cues the student has been exposed to (navigation). Similarly, in the REAS facet, single inferences and conclusions presented in the text, indicating author biases, fallacies, and heuristics can be classified and scored depending on whether students' repeat them uncritically in their responses, avoid them, or qualify them. Against the background of familiarity and a critical approach to the topic given in a task, successful students should not copy statements so much as express their own argument and opinion.

Based on prior research identifying different navigation and reasoning profiles (List and Alexander, 2017), respondents could be classified into specific COR "learner profiles" based on their (sub)scores on facets of the rubric (e.g., using cluster analysis). Based on students' initial stance toward a task topic (for, against, neutral), (self-estimated) prior task- and topic-related knowledge (expert, novice), and topic interest (interested or not), students can be distinguished into distinct initial profiles, for instance "novice in favor" or "expert neutral," which may impact students' information search and reasoning approach while solving the CORA tasks: "Novices" may need to form

an initial stance and identify trustworthy references or experts whose judgment they trust, while "experts" may draw on their knowledge of trustworthy sources, or prior reasoning on the topic, but are challenged to not fall for confirmation bias and need to test their position (self-critically) against opposing views. "Novices" may also adopt a naive strategy of no initial evaluation of online information, but fallibilism over time, for instance compensating low evaluation skills with sophisticated epistemic beliefs and thus being open-minded to change their beliefs based on new evidence (Paul and Elder, 2005). Taking into account a longitudinal learning perspective (Section Analyses of Response Processes and Longitudinal Studies), online reasoning can later also include a meta-cognitive facet of less well-known yet important properties that influence students' learning and mental functioning (e.g., built-in gratification mechanisms and resulting media preferences).

Scoring for formative purposes in educational practice can then focus on certain features in students' response processes (Section Analyses of Response Processes and Longitudinal Studies) depending on initial "learner profile" (e.g., presence of pros and cons in responses of "topic experts" vs. "novices"), whereby the profiles may vary depending on the topic tapped by the CORA task. CORA tasks can be retested across several measurement points over a course of students' studies (Section Analyses of Response Processes and Longitudinal Studies). Here, knowledge tasks (e.g., selected-response items) on key pieces of information and misinformation in CORA task stimuli can be used to control for (prior) knowledge or retesting effects, which would be especially important for domain-specific CORA tasks. A pre-post design can indicate domain learning over time, i.e., when a student accepted misinformation on the pretest but did not accept it on the posttest (e.g., indicating misconceptions or conceptual change). The formative assessments can inform teachers and students how they can improve their search and evaluation behavior and domain-learning using the Internet.

## Analyses of Response Processes and Longitudinal Studies

Given the open-information environment and holistic nature of the performance assessment, a number of more detailed analyses of the information environment and students' navigation thereof is being conducted. We aim to connect the assessment design and outcomes to the complex *Information Landscape (IL)* that the individual student encounters online, and examine how it influences the response process and test result (Nagel et al., 2020). Using logged CORA performance data, the students' browsing activity can be examined to describe which sources they accessed, how much time they spent, what judgments they made, and which cues they considered during which phases (Schmidt et al., 2020).

According to the ECD (Mislevy and Haertel, 2006), response processes indicate which cognitions are generated by a confrontation of a subject (student) with a task. The analysis of the response processes can refer to various indicators that arise during the processing of the CORA tasks (e.g., as described

in **Table 2** in Section Test Definition and Operationalization of COR: Design and Characteristics of CORA, with a focus on quality judgments by IPS-I phases). The log files or think-aloud data can give an indication of the expected (meta)cognitive processes that are elicited during the response processes (Zumbo and Hubley, 2017), for instance on the occurrence of different mental processes, students' attention to particular aspects, and their distribution across the task solving phases to determine whether the theoretically assumed (construct-related) comprehension and reasoning processes were indeed performed by respondents.

In a longitudinal analysis perspective, we aim to investigate the relationship between the students' COR ability and their acquisition of reliable warranted vs. erroneous knowledge over the course of their studies in higher education. Using repeated CORA measurements (i.e., formative assessments), aspects of knowledge development and memory (incl. retesting) effects over the course of study can be analyzed, providing an important basis for instructional interventions in educational practice.

## CONCLUSION

The holistic task format allows for modular extensions of sub-scores, provided abilities are tapped, which can be deployed efficiently in subsequent in-depth validation studies. As Goldman and Brand-Gruwel (2018) conclude for sourcing, which equally applies to trustworthiness evaluation and reasoning based on online information more generally: "We also need a more nuanced approach to the purpose and value of sourcing processes; identifying the perspective of a particular source is not the "end goal." Perspective is not so much about trustworthiness of sources as it is about how perspective informs what learners make of the information with respect to forming interpretations, making decisions, and proposing solutions."

We agree and add that, beyond specific text-types dedicated to arguing about trustworthiness (e.g., research papers, legal opinions), trustworthiness evaluation mainly serves to filter out untrustworthy information. That is, hardly any additional information is added that helps students resolve an information problem, and instead available evidence for reasoning that turns out to be untrustworthy is even detracted from an argument. This can appear demotivating to the novices, unless it supports the achievement of a higher-order goal, such as maintaining a high quality standard. In general, learning based on erroneous knowledge may result in either unverified adoption or incorrectly understood or recognized information that can lead to persistent misconceptions and knowledge inconsistencies, which can become evident in later use of this erroneous knowledge.

With the present COR conceptualization and its assessment framework combining information acquisition, trustworthiness evaluation, and argumentative reasoning, we contribute to a better understanding of how trustworthiness judgments are functionally embedded in the broader information acquisition and online reasoning process, and open up perspectives for long-term studies in this emerging research field.

At the same time, this study is only a starting point for longer-term research on critical reasoning at the higher education level within the specific context of the online information environment, which also marks its limitations. Future research would need to determine the relations with critical thinking skills assessed in other contexts as well as with the other cited, partially overlapping assessments (e.g., iPAL performance assessments) that served as a basis and inspiration in the development of the COR assessment.

## DATA AVAILABILITY STATEMENT

The original contributions generated for the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author.

## ETHICS STATEMENT

Ethical review and approval for the study on human participants was not required in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by participants. Participant statements were anonymized, published in small excerpts only, and checked to not reveal identifiable information.

## AUTHOR CONTRIBUTIONS

DM co-developed the assessment, conducted the analyses, and co-wrote the manuscript. OZ-T co-developed the assessment, supervised the analyses, and co-wrote the manuscript. M-TN co-developed the rating scheme and was involved in preparing, reviewing, and revising the manuscript. SB was involved in preparing, reviewing, and revising the manuscript. SS co-developed the assessment and was involved in its validation. RS was involved in the assessment validation and in reviewing and revising the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was part of the PLATO program, funded by the Rhine-Main Universities fund. Open access publication was supported by the German Research Foundation (DFG) and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

## ACKNOWLEDGMENTS

We would like to thank all members of the Stanford History Education Group as well as all experts who supported this study. We would like to thank the two reviewers and the journal editor who provided helpful critical and constructive comments on the manuscript. We would like to thank the *Frontiers in Education* editorial staff for quality assurance.

## REFERENCES

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., et al. (2008). Instructional interventions affecting critical thinking skills and dispositions: a stage 1 meta-analysis. *Rev. Educ. Res.* 78, 1102–1134. doi: 10.3102/0034654308326084
- AERA, APA, and NCME. (2014). *Standards of Educational and Psychological Testing*. Washington, DC: AERA, APA, and NCME American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.
- Akamine, S., Kato, Y., Inui, K., and Kurohashi, S. (2008). "Using appearance information for web information credibility analysis," in *2nd International Symposium on Universal Communication, 2008: ISUC 2008, 15–16 December, 2008, Osaka, Japan* (Piscataway, NJ: IEEE), 363–365. doi: 10.1109/ISUC.2008.80
- American Library Association (2000). *Information Literacy Competency Standards for Higher Education*. Available online at: <http://www.ala.org/acrl/standards/informationliteracycompetency> (accessed June 24, 2020). doi: 10.5860/crln.61.3.207
- Arazy, O., and Kopak, R. (2011). On the measurability of information quality. *J. Am. Soc. Inf. Sci. Technol.* 62, 89–99. doi: 10.1002/asi.21447
- Arffman, I. (2007). *The problem of equivalence in translating texts in international reading literacy studies: a text analytic study of three English and Finnish texts used in the PISA 2000 reading test* (dissertation). University of Jyväskylä, Jyväskylä, Finland.
- Banerjee, M., Zlatkin-Troitschanskaia, O., and Roeper, J. (2020). Narratives and their impact on students' information seeking and critical online reasoning in higher education economics and medicine. *Front. Educ.* 5:625. doi: 10.3389/educ.2020.570625
- Batista, J. C. L., and Marques, R. P. F. (2017). *Information and Communication Overload in the Digital Age*. Hershey, PA: IGI Global. doi: 10.4018/978-1-5225-2061-0.ch001
- Bayer, J., Bitukova, N., Bárd, P., Szakács, J., Alemanno, A., and Uszkiewicz, E. (2019). *Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and Its Member States*. Directorate General for Internal Policies of the Union, Policy Department for Citizens' Rights and Constitutional Affairs. Available online at: [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/608864/IPOL\\_STU\(2019\)608864\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/608864/IPOL_STU(2019)608864_EN.pdf) (accessed June 24, 2020).
- Beck, K. (2020). "On the relationship between "Education" and "Critical Thinking,"" in *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)*, ed O. Zlatkin-Troitschanskaia (Cham: Springer International Publishing), 73–87.
- Blummer, B., and Kenton, J. M. (2015). *Improving Student Information Search: A Metacognitive Approach*. Amsterdam: Elsevier. doi: 10.1533/9781780634623.23
- Braasch, J. L. G., and Bråten, I. (2017). The discrepancy-induced source comprehension (D-ISC) model: basic assumptions and preliminary evidence. *Educ. Psychol.* 52, 167–181. doi: 10.1080/00461520.2017.1323219
- Braasch, J. L. G., Bråten, I., and McCrudden, M. T. (2018). *Handbook of Multiple Source Use*. New York, NY: Routledge Taylor and Francis Group. doi: 10.4324/9781315627496
- Brand-Gruwel, S., Wopereis, I., and Vermetten, Y. (2005). Information problem solving by experts and novices: analysis of a complex cognitive skill. *Comput. Hum. Behav.* 21, 487–508. doi: 10.1016/j.chb.2004.10.005
- Brand-Gruwel, S., Wopereis, I., and Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Comput. Educ.* 53, 1207–17. doi: 10.1016/j.compedu.2009.06.004
- Braten, I., Stadtler, M., and Salmeron, L. (2018). "The role of sourcing in discourse comprehension," in *Handbook of Discourse Processes*, eds M. F. Schober, D. N. Rapp, and M. A. Britt (New York, NJ: Taylor and Francis), 141–166. doi: 10.4324/9781315687384-10
- Breakstone, J., Smith, M., Wineburg, S., Rapaport, A., Carle, J., Garland, M., et al. (2019). *Students' Civic Online Reasoning: A National Portrait*. Stanford History Education Group and Gibson Consulting. Available online at: <https://purl.stanford.edu/gf151tb4868> (accessed June 25, 2020).
- Bulger, M. E., Mayer, R. E., and Metzger, M. J. (2014). Knowledge and processes that predict proficiency in digital literacy. *Reading Writing* 27, 1567–1583. doi: 10.1007/s11145-014-9507-2
- Catalano, A. (2013). Patterns of graduate students' information seeking behavior: a meta-synthesis of the literature. *J. Doc.* 69, 243–274. doi: 10.1108/00220411311300066
- Center for Humane Technology (2019). *Ledger of Harms*. Available online at: <https://ledger.humanetech.com/> (accessed October 17, 2019).
- Chen, S., and Chaiken, S. (1999). "The heuristic-systematic model in its broader context," in *Dual-Process Theories in Social Psychology*, eds S. Chaiken and Y. Trope (New York, NY: Guilford Press), 73–96.
- Choi, W. (2015). *A new framework of web credibility assessment and an exploratory study of older adults' information behavior on the web* (dissertation). Florida State University, Tallahassee, FL, United States.
- Ciampaglia, G. L. (2018). "The digital misinformation pipeline," in *Positive Learning in the Age of Information*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Wiesbaden: Springer), 413–421. doi: 10.1007/978-3-658-19567-0\_25
- Coiro, J. (2003). Exploring literacy on the internet: reading comprehension on the internet: expanding our understanding of reading comprehension to encompass new literacies. *Reading Teach.* 56, 458–464.
- Damico, J. S., and Panos, A. (2018). Civic media literacy as 21st century source work: future social studies teachers examine web sources about climate change. *J. Soc. Stud. Res.* 42, 345–359. doi: 10.1016/j.jssr.2017.10.001
- Daniels, J. (2009). Cloaked websites: propaganda, cyber-racism and epistemology in the digital era. *N. Media Soc.* 11, 659–683. doi: 10.1177/1461444809105345
- Davey, T., Ferrara, S., Holland, P. W., Shavelson, R., Webb, N. M., and Wise, L. L. (2015). *Psychometric Considerations for the Next Generation of Performance Assessment: Report of the Center for K-12 Assessment and Performance Management at ETS*. Available online at: [https://www.ets.org/Media/Research/pdf/psychometric\\_considerations\\_white\\_paper.pdf](https://www.ets.org/Media/Research/pdf/psychometric_considerations_white_paper.pdf) (accessed July 22, 2018).
- De Neys, W. D. (2006). Dual processing in reasoning: two systems but one reasoner. *Psychol. Sci.* 17, 428–433. doi: 10.1111/j.1467-9280.2006.01723.x
- Dunbar, N. E., Connelly, S., Jensen, M. L., Adame, B. J., Rozzell, B., Griffith, J. A., et al. (2014). Fear appeals, message processing cues, and credibility in the websites of violent, ideological, and nonideological groups. *J. Comput. Mediated Commun.* 19, 871–889. doi: 10.1111/jcc4.12083
- Eisenberg, M. B., and Berkowitz, R. E. (1990). *Information Problem-Solving: The Big Six Skills Approach to Library and Information Skills Instruction*. Norwood, NJ: Ablex.
- Elder, L., and Paul, R. (2010). *Critical Thinking Development: A Stage Theory: With Implications for Instruction*.
- Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educ. Leadersh.* 43, 44–48.
- Evans, J. S. B., and Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685
- Facione, P. A. (1990). Critical thinking: a statement of expert consensus for purposes of educational assessment and instruction: executive summary. *The Delphi Report* (accessed June 25, 2020).
- Fischer, F. (2018). *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*. New York, NY: Routledge.
- Fischer, F., Chinn, C., Engelmann, K., and Osborne, J. (2018). *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge, 1st Edn*. London: Routledge. doi: 10.4324/9780203731826-1
- Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Front. Learn. Res.* 2, 28–45. doi: 10.14786/flr.v2i2.96
- Fisher, K. E., Erdelez, S., and McKechnie, L. (2005). *Theories of Information Behavior, ASIST Monograph Series*. Medford, NJ: Information Today.
- Flanagin, A. J., and Metzger, M. J. (2014). "Digital media and perceptions of source credibility in political communication," in *The Oxford Handbook of Political Communication*, eds K. Kenski, and K. Hall (Oxford: Oxford University Press), 417–436. doi: 10.1093/oxfordhb/9780199793471.013.65
- Flanagin, A. J., Metzger, M. J., and Hartsell, E. (2010). *Kids and Credibility: An Empirical Examination of Youth, Digital Media Use, and Information Credibility*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/8778.001.0001



- Flanagin, A. J., Winter, S., and Metzger, M. J. (2018). Making sense of credibility in complex information environments: the role of message sidedness, information source, and thinking styles in credibility evaluation online. *Inf. Commun. Soc.* 23, 1038–1059. doi: 10.1080/1369118X.2018.1547411
- Flore, M., Balahur, A., Podavini, A., and Verile, M. (2019). *Understanding Citizens' Vulnerability to Disinformation and Data-driven Propaganda*. Luxembourg: Publications Office of the European Union. doi: 10.2760/919835
- Fogg, B. J. (2002). *Stanford Guidelines for Web Credibility. A Research Summary From the Stanford Persuasive Technology Lab*. Available online at: [www.webcredibility.org/guidelines](http://www.webcredibility.org/guidelines) (accessed June 24, 2020).
- Fogg, B. J. (2003). *Persuasive Technology: Using Computers to Change What We Think and Do*. Amsterdam; Boston: Morgan Kaufmann. Available online at: <http://www.loc.gov/catdir/description/els031/2002110617.html> (accessed June 24, 2020).
- Fogg, B. J., Marshall, J., Kameda, T., Solomon, J., Rangnekar, A., and Boyd, J. (2001a). "Web credibility research," in *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, ed M. Tremaine (New York, NY: ACM), 295. doi: 10.1145/634067.634242
- Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., et al. (2001b). "What makes web sites credible? A report on a large quantitative study," in *Proceedings of CHI '01: The SIGCHI Conference on Human Factors in Computing Systems*, eds J. Jacko, and A. Sears (New York, NY: ACM Press), 61–68. doi: 10.1145/365024.365037
- Fogg, B. J., Marshall, J., Osipovich, A., Varma, C., Laraki, O., Fang, N., et al. (2000). "Elements that affect web credibility: early results from a self-report study," in *Chi '00 Extended Abstracts on Human Factors in Computing Systems*, ed M. Tremaine (New York, NY: ACM), 287–288. doi: 10.1145/633292.633460
- Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., and Tauber, E. R. (2003). "How do users evaluate the credibility of web sites?," in *Proceedings of the 2003 Conference on Designing for User Experiences*, ed J. Arnowitz (New York, NY: ACM), 1–15. doi: 10.1145/997078.997097
- Gasser, U., Cortesi, S., Malik, M., and Lee, A. (2012). *Youth and Digital Media: From Credibility to Information Quality*. Cambridge, MA: The Berkman Center for Internet and Society. doi: 10.2139/ssrn.2005272
- George, J. F., Giordano, G., and Tilley, P. A. (2016). Website credibility and deceiver credibility: expanding prominence-interpretation theory. *Comput. Hum. Behav.* 54, 83–93. doi: 10.1016/j.chb.2015.07.065
- George, J. F., Tilley, P., and Giordano, G. (2014). Sender credibility and deception detection. *Comput. Hum. Behav.* 35, 1–11. doi: 10.1016/j.chb.2014.02.027
- Go, E., You, K. H., Jung, E., and Shim, H. (2016). Why do we use different types of websites and assign them different levels of credibility? Structural relations among users' motives, types of websites, information credibility, and trust in the press. *Comput. Hum. Behav.* 54, 231–239. doi: 10.1016/j.chb.2015.07.046
- Goldman, S., Lawless, K., Pellegrino, J., Manning, F., Braasch, J., and Gomez, K. (2013). "A technology for assessing multiple source comprehension: an essential skill of the 21st century," in *Technology-Based Assessments for 21st Century Skills: Theoretical and Practical Implications From Modern Research*, eds M. C. Mayrath, J. Clarke-Midura, and D. H. Robinson (Charlotte, NC: Information Age Publishing), 171–207.
- Goldman, S. R., and Brand-Gruwel, S. (2018). "Learning from multiple sources in a digital society," in *International Handbook of the Learning Sciences*, eds F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, and P. Reimann (London: Routledge), 86–95. doi: 10.4324/9781315617572-9
- Goldstein, D. G., and Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychol. Rev.* 109, 75–90. doi: 10.1037/0033-295X.109.1.75
- Gronchi, G., and Giovannelli, F. (2018). Dual process theory of thought and default mode network: a possible neural foundation of fast thinking. *Front. Psychol.* 9:1237. doi: 10.3389/fpsyg.2018.01237
- Hahnel, C., Kroehne, U., Goldhammer, F., Schoor, C., Mahlow, N., and Artelt, C. (2019). Validating process variables of sourcing in an assessment of multiple document comprehension. *Br. J. Educ. Psychol.* 89, 524–537. doi: 10.1111/bjep.12278
- Halpern, D. F. (2014). *Thought and Knowledge: An Introduction to Critical Thinking, 5th Edn.* New York, NY: Psychology Press. doi: 10.4324/9781315885278
- Harkness, J. A. (2003). "Questionnaire translation," in *Cross-Cultural Survey Methods*, eds J. A. Harkness, F. van de Vijver, and P. P. Mohler (Hoboken, NJ: John Wiley and Sons), 35–56.
- Head, A., and Eisenberg, M. B. (2009). Project information literacy progress report: "lessons learned": how college students seek information in the digital age. *SSRN Electron. J.* doi: 10.2139/ssrn.2281478
- Herman, E. S., and Chomsky, N. (2002). *Manufacturing Consent: The Political Economy of the Mass Media*. New York, NY: Pantheon Books.
- Hilligoss, B., and Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: construct, heuristics, and interaction in context. *Inf. Process. Manag.* 44, 1467–1484. doi: 10.1016/j.ipm.2007.10.001
- International Test Commission (ITC) (2017). *The ITC Guidelines for Translating and Adapting Tests, 2nd Edn.* Available online at: [www.intestcom.org](http://www.intestcom.org) (accessed June 24, 2020).
- Jahn, D. (2012). *Kritisches Denken fördern können: Entwicklung eines didaktischen Designs zur Qualifizierung pädagogischer Professionals [Fostering critical thinking: developing a didactic design for qualification of pedagogical professionals]* (dissertation), Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany.
- Jahn, D., and Kenner, A. (2018). "Critical thinking in higher education: how to foster it using digital media," in *The Digital Turn in Higher Education*, eds D. Kergel, B. Heidkamp, P. K. Telléus, T. Rachwal, and S. Nowakowski (Wiesbaden: Springer), 81–109. doi: 10.1007/978-3-658-19925-8\_7
- Jozsa, E., Komlodi, A., Ahmad, R., and Hercegi, K. (2012). "Trust and credibility on the web: the relationship of web experience levels and user judgments," in *IEEE 3rd international conference on cognitive Infocommunications (CogInfoCom)* (Piscataway, NJ: IEEE), 605–610. doi: 10.1109/CogInfoCom.2012.6422051
- Juvina, I., and van Oostendorp, H. (2008). Modeling semantic and structural knowledge in web navigation. *Discourse Process.* 45, 346–364. doi: 10.1080/01638530802145205
- Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511809477
- Kakol, M., Nielek, R., and Wierzbicki, A. (2017). Understanding and predicting web content credibility using the content credibility corpus. *Inf. Process. Manag.* 53, 1043–1061. doi: 10.1016/j.ipm.2017.04.003
- Kane, M. (2012). Validating score interpretations and uses. *Lang. Test.* 29, 3–17. doi: 10.1177/0265532211417210
- Karova, N. A., and Fisher, K. E. (2013). A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Inf. Res.* 18:573.
- Kingsley, K., Galbraith, G. M., Herring, M., Stowers, E., Stewart, T., and Kingsley, K. V. (2011). Why not just google it? An assessment of information literacy skills in a biomedical science curriculum. *BMC Med. Educ.* 11:1. doi: 10.1186/1472-6920-11-17
- KMK (2016). *Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in Germany. Bildung in der digitalen Welt Strategie der Kultusministerkonferenz. [Education in the digital world. KMK strategy paper]*. Retrieved from [https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2018/Digitalstrategie\\_2017\\_mit\\_Weiterbildung.pdf](https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2018/Digitalstrategie_2017_mit_Weiterbildung.pdf)
- Kohnen, A. M., and Mertens, G. E. (2019). I'm always kind of double-checking: exploring the information-seeking identities of expert generalists. *Reading Res. Q.* 54, 279–297. doi: 10.1002/rrq.245
- Koltay, T. (2011). The media and the literacies: media literacy, information literacy, digital literacy. *Media Cult. Soc.* 33, 211–221. doi: 10.1177/0163443710393382
- Krämer, N. C., Preko, N., Flanagin, A., Winter, S., and Metzger, M. (2018). "What do people attend to when searching for information on the web," in *ICPS, Proceedings of the Technology, Mind, and Society Conference, Washington, DC* (New York, NY: The Association for Computing Machinery). doi: 10.1145/3183654.3183682
- Kuhlthau, C. C. (1993). A principle of uncertainty for information seeking. *J. Doc.* 49, 339–355. doi: 10.1108/eb026918
- Kuhlthau, C. C., Heinström, J., and Todd, R. J. (2008). The 'information search process' revisited: is the model still useful. *Inf. Res.* 13, 13–14.
- Lawless, K. A., Goldman, S. R., Gomez, K., Manning, F., and Braasch, J. (2012). "Assessing multiple source comprehension through evidence-centered design," in *Reaching an Understanding: Innovations in How We View Reading*



- Assessment, eds J. P. Sabatini, T. O'Reilly, and E. Albro (Lanham, MD: Rowman and Littlefield Education), 3–17.
- Leeder, C., and Shah, C. (2016). Practicing critical evaluation of online sources improves student search behavior. *J. Acad. Libr.* 42, 459–468. doi: 10.1016/j.acalib.2016.04.001
- List, A., and Alexander, P. A. (2017). Analyzing and integrating models of multiple text comprehension. *Educ. Psychol.* 52, 143–147. doi: 10.1080/00461520.2017.1328309
- Liu, O. L., Frankel, L., and Crotts Roohs, K. (2014). *Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment*. Princeton, NJ: ETS. doi: 10.1002/ets2.12009
- Lucassen, T., and Schraagen, J. M. (2011). Factual accuracy and trust in information: the role of expertise. *J. Am. Soc. Inf. Sci. Technol.* 62, 1232–1242. doi: 10.1002/asi.21545
- Lucassen, T., and Schraagen, J. M. (2013). The influence of source cues and topic familiarity on credibility evaluation. *Comput. Hum. Behav.* 29, 1387–1392. doi: 10.1016/j.chb.2013.01.036
- Maurer, A., Schloegl, C., and Dreisiebner, S. (2017). Comparing information literacy of student beginners among different branches of study. *Libellarium* 9:2. doi: 10.15291/libellarium.v9i2.280
- Maurer, M., Quiring, O., and Schemer, C. (2018). “Media effects on positive and negative learning,” in *Positive Learning in the Age of Information*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Wiesbaden: Springer), 197–208. doi: 10.1007/978-3-658-19567-0\_11
- Maurer, M., Schemer, C., Zlatkin-Troitschanskaia, O., and Jitomirski, J. (2020). “Positive and negative media effects on university students' learning: preliminary findings and a research program,” in *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*, ed O. Zlatkin-Troitschanskaia (Cham: Springer International Publishing), 109–119. doi: 10.1007/978-3-030-26578-6\_8
- Mayer, R. E. (2009). *Multimedia Learning, 2nd Edn.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511811678
- Mayring, P. (2014). *Qualitative Content Analysis. Theoretical Foundation, Basic Procedures and Software Solution*. Retrieved from <https://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173>
- McCrudden, M. T., Magliano, J. P., and Schraw, G. J. (2011). *Text Relevance and Learning From Text*. Charlotte, NC: Information Age Pub. doi: 10.1007/978-1-4419-1428-6\_354
- McGrew, S., Smith, M., Breakstone, J., Ortega, T., and Wineburg, S. (2019). Improving university students' web savvy: an intervention study. *Br. J. Educ. Psychol.* 89, 485–500. doi: 10.1111/bjep.12279
- McMullin, S. L. (2018). *The correlation between information literacy and critical thinking of college students: an exploratory study* (dissertation thesis). University of North Texas, Denton, TX, United States (ProQuest LLC).
- Messick, S. (1989). “Validity,” in *Educational Measurement*, 3rd Edn., ed R. L. Linn (New York, NY: American Council on education and Macmillan), 13–104.
- Metzger, M. J. (2007). Making sense of credibility on the web: models for evaluating online information and recommendations for future research. *J. Am. Soc. Inf. Sci. Technol.* 58, 2078–2091. doi: 10.1002/asi.20672
- Metzger, M. J., and Flanagin, A. (2015). “Psychological approaches to credibility assessment online” in *The Handbook of the Psychology of Communication Technology*, ed S. S. Sundar (Chichester; Malden, MA: Wiley Blackwell), 445–466. doi: 10.1002/9781118426456.ch20
- Metzger, M. J., and Flanagin, A. J. (2013). Credibility and trust of information in online environments: the use of cognitive heuristics. *J. Pragmatics* 59, 210–220. doi: 10.1016/j.pragma.2013.07.012
- Mislevy, R. J. (2017). *Socio-Cognitive Foundations of Educational Measurement*. London: Routledge. doi: 10.4324/9781315871691
- Mislevy, R. J., and Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educ. Meas.* 25, 6–20. doi: 10.1111/j.1745-3992.2006.00075.x
- Molerov, D., Zlatkin-Troitschanskaia, O., and Schmidt, S. (2019). “Adapting the civic online reasoning assessment cross-nationally using an explicit functional equivalence approach” in *Annual Meeting of the American Educational Research Association (Toronto)*.
- Moore, T. (2013). Critical thinking: seven definitions in search of a concept. *Stud. Higher Educ.* 38, 506–522. doi: 10.1080/03075079.2011.586995
- Münchow, H., Richter, T., von der Mühlen, S., and Schmid, S. (2019). The ability to evaluate arguments in scientific texts: measurement, cognitive processes, nomological network, and relevance for academic success at the university. *Br. J. Educ. Psychol.* 89, 501–523. doi: 10.1111/bjep.12298
- Murray, M. C., and Pérez, J. (2014). Unraveling the digital literacy paradox: how higher education fails at the fourth literacy. *Issues Inf. Sci. Inf. Technol.* 11, 189–210. doi: 10.28945/1982
- Nagel, M.-T., Schäfer, S., Zlatkin-Troitschanskaia, O., Schemer, C., Maurer, M., Molerov, D., et al. (2020). How do university students' web search behavior, website characteristics, and the interaction of both influence students' critical online reasoning? *Front. Educ.* 5:1. doi: 10.3389/feduc.2020.565062
- National Research Council (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: National Academies Press.
- Newman, N., Fletcher, R., Kalogeropoulos, A., and Nielsen, R. K. (2019). *Reuters Institute Digital News Report 2019. Reuters Institut for the Study of Journalism*. Available online at: [https://reutersinstitute.politics.ox.ac.uk/sites/default/files/inline-files/DNR\\_2019\\_FINAL.pdf](https://reutersinstitute.politics.ox.ac.uk/sites/default/files/inline-files/DNR_2019_FINAL.pdf) (accessed January 1, 2020).
- Oser, F. K. (2018). “Positive learning through negative learning - the wonderful burden of PLATO,” in *Positive Learning in the Age of Information: A Blessing or a Curse?* (Wiesbaden: Springer VS), 363–372.
- Oser, F. K., and Biedermann, H. (2020). “A three-level model for critical thinking: critical alertness, critical reflection, and critical analysis,” in *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*, ed O. Zlatkin-Troitschanskaia (Cham: Springer International Publishing), 89–106.
- Paul, R., and Elder, L. (2005). *A Guide for Educators to Critical Thinking Competency Standards, Principles, Performance Indicators, and Outcomes with a Critical Thinking Master Rubric*. Available online at: [www.criticalthinking.org](http://www.criticalthinking.org) (accessed June 24, 2020).
- Paul, R., and Elder, L. (2008). *The Thinker's Guide for Conscientious Citizens on How to Detect Media Bias and Propaganda in National and World News: In National and World News, 4th Edn.* Dillon Beach, CA: The Foundation for Critical Thinking.
- Pellegrino, J. W. (2017). “Teaching, learning and assessing 21st century skills,” in *Educational Research and Innovation. Pedagogical Knowledge and the Changing Nature of the Teaching Profession*, ed S. Guerriero (Paris: OECD Publishing), 223–251. doi: 10.1787/9789264270695-12-en
- Pernice, K. (2017). *F-Shaped Pattern of Reading on the Web: Misunderstood, but Still Relevant (Even on Mobile)*. *World Leaders in Research-Based User Experience*. Available online at: <https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/> (accessed June 24, 2020).
- Pirolli, P., and Card, S. (1999). Information foraging. *Psychol. Rev.* 106, 643–675. doi: 10.1037/0033-295X.106.4.643
- Podgornik, B. B., Dolničar, D., and Glažar, S. A. (2017). Does the information literacy of university students depend on their scientific literacy? *Eurasia J. Math. Sci. Technol. Educ.* 13, 3869–3891. doi: 10.12973/eurasia.2017.00762a
- Powers, E. M. (2019). How students access, filter and evaluate digital news: choices that shape what they consume and the implications for news literacy education. *J. Lit. Technol.* 20:3.
- Reese, S. D., and Shoemaker, P. J. (2016). A media sociology for the networked public sphere: the hierarchy of influences model. *Mass Commun. Soc.* 19, 389–410. doi: 10.1080/15205436.2016.1174268
- Rieh, S. Y. (2010). “Credibility and cognitive authority of information” in *Encyclopedia of Library and Information Sciences*, 1, 1337–1344.
- Rieh, S. Y. (2014). Credibility assessment of online information in context. *J. Inf. Sci. Theory Pract.* 2, 6–17. doi: 10.1633/JISTaP.2014.2.3.1
- Roozenbeek, J., and van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Commun.* 5:133. doi: 10.1057/s41599-019-0279-9
- Rouet, J. F. (2006). *The Skills of Document Use: From Text Comprehension to Web-Based Learning*. Mahwah, NJ: Erlbaum. Available online at: <http://www.loc.gov/catdir/enhancements/fy0625/2005052083-d.html> (accessed June 24, 2020). doi: 10.4324/9780203820094
- Salmerón, L., Cañas, J. J., Kintsch, W., and Fajardo, I. (2005). Reading strategies and hypertext comprehension. *Discourse Process.* 40, 171–191. doi: 10.1207/s15326950dp4003\_1

- Salmerón, L., Kammerer, Y., and García-Carrión, P. (2013). Searching the web for conflicting topics: page and user factors. *Comput. Hum. Behav.* 29, 2161–2171. doi: 10.1016/j.chb.2013.04.034
- Samson, S. (2010). Information literacy learning outcomes and student success. *J. Acad. Libr.* 36, 202–210. doi: 10.1016/j.jacalib.2010.03.002
- Sanders, L., Kurbanoglu, S., Boustany, J., Dogan, G., and Becker, P. (2015). Information behaviors and information literacy skills of LIS students: an international perspective. *J. Educ. Libr. Inf. Sci. Online* 56, 80–99. doi: 10.12783/issn.2328-2967/56/S1/9
- Schmidt, S., Zlatkin-Troitschanskaia, O., Roeper, J., Klose, V., Weber, M., Bültmann, A.-K., et al. (2020). Undergraduate students' critical online reasoning - process mining analysis. *Front. Psychol.* (in press). doi: 10.3389/fpsyg.2020.576273
- Schnell, R., Hill, P. B., and Esser, E. (2011). *Methoden der empirischen Sozialforschung [Methods of Empirical Social Research]*, 9th Edn. Munich: Oldenburg.
- Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., and Menczer, F. (2017). *The Spread of Fake News by Social Bots*. Available online at: <https://arxiv.org/abs/1707.07592> (accessed June 24, 2020).
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Mariño, J. P. (2019). Assessment of university students' critical thinking: next generation performance assessment. *Int. J. Test.* 19, 337–362. doi: 10.1080/15305058.2018.1543309
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., and Mariño, J. (2018). "International performance assessment of learning in higher education (iPAL): research and development," in *Assessment of Learning Outcomes in Higher Education – Cross-National Comparisons and Perspectives*, eds O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, and C. Lautenbach (Wiesbaden: Springer), 193–214. doi: 10.1007/978-3-319-74338-7\_10
- Shoemaker, P. J., and Reese, S. D. (2014). *Mediating the Message in the 21st Century: A Media Sociology Perspective*, 3rd Edn. New York, NY: Routledge/Taylor and Francis Group. doi: 10.4324/9780203930434
- Snow, C. E. (2002). *Reading for Understanding: Toward an RandD Program in Reading Comprehension*. Santa Monica CA: Rand.
- Solano-Flores, G., Backhoff, E., and Contreras-Niño, L. Á. (2009). Theory of test translation error. *Int. J. Test.* 9, 78–91. doi: 10.1080/15305050902880835
- Sparks, J. R., Katz, I. R., and Beile, P. M. (2016). Assessing digital information literacy in higher education: a review of existing frameworks and assessments with recommendations for next-generation assessment. *ETS Res. Rep. Ser.* 2016, 1–33. doi: 10.1002/ets2.12118
- Stanovich, K. E., West, R., and Toplak, M. E. (2016). *The Rationality Quotient: Toward a Test of Rational Thinking*. Cambridge, MA: The MIT Press. doi: 10.7551/mitpress/9780262034845.001.0001
- Sundar, S. S. (2008). "The MAIN model: a heuristic approach to understanding technology effects on credibility," in *Digital Media, Youth, and Credibility*, eds M. J. Metzger, and A. J. Flanagin (Cambridge: MIT Press), 73–100.
- Tanaka, K. (2009). "Web search and information credibility analysis: bridging the gap between web1.0 and web2.0," in *ICUIMC 2009: Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication* (Suwon), 39–44.
- Tanaka, K., Kawai, Y., Zhang, J., Nakajima, S., Inagaki, Y., Ohshima, H., et al. (2010). "Evaluating credibility of web information," in *Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication - ICUIMC '10*, eds W. Kim, D. Won, K.-H. You, and S.-W. Lee (New York, NY: ACM Press), 1–10. doi: 10.1145/2108616.2108645
- Taylor, A., and Dalal, H. A. (2014). Information literacy standards and the world wide web: results from a student survey on evaluation of Internet information sources. *Inf. Res.* 19:4.
- Threadgill, E. J., and Price, L. R. (2019). Assessing online viewing practices among college students. *J. Media Lit. Educ.* 11, 37–55. doi: 10.23860/JMLE-2019-11-2-3
- Toplak, M. E., Liu, E., MacPherson, R., Toneatto, T., and Stanovich, K. E. (2007). The reasoning skills and thinking dispositions of problem gamblers: a dual process taxonomy. *J. Behav. Decis. Mak.* 20, 103–124. doi: 10.1002/bdm.544
- Toulmin, S. (2003). *The Uses of Argument, Updated Edn*. Cambridge, NY: Cambridge University Press. doi: 10.1017/CBO9780511840005
- Tseng, S., and Fogg, B. J. (1999). Credibility and computing technology. *Commun. ACM* 42, 39–44. doi: 10.1145/301353.301402
- Van Eemeren, F. H. (2013). Fallacies as derailments of argumentative discourse: acceptance based on understanding and critical assessment. *J. Pragmatics* 59, 141–152. doi: 10.1016/j.pragma.2013.06.006
- Walraven, A., Brand-Gruwel, S., and Boshuizen, H. P. A. (2008). Information-problem solving: a review of problems students encounter and instructional solutions. *Comput. Hum. Behav.* 24, 623–648. doi: 10.1016/j.chb.2007.01.030
- Walraven, A., Brand-Gruwel, S., and Boshuizen, H. P. A. (2009). How students evaluate information and sources when searching the world wide web for information. *Comput. Educ.* 52, 234–246. doi: 10.1016/j.compedu.2008.08.003
- Walton, D. (2006). *Fundamentals of Critical Argumentation. Critical Reasoning and Argumentation*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511807039
- Walton, D. (2017). Value-based argumentation in mass audience persuasion dialogues. *COGENCY* 9, 139–159.
- Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511802034
- Walton, G., Barker, J., Pointon, M., Turner, M., and Wilkinson, A. (2020). "Information literacy and the societal imperative of information discernment," in *Informed Societies: Why Information Literacy Matters for Citizenship, Participation and Democracy*, ed S. Goldstein (London: Facet Publishing), 149. doi: 10.29085/9781783303922.010
- Wathen, C. N., and Burkell, J. (2002). Believe it or not: factors influencing credibility on the web. *J. Am. Soc. Inf. Sci. Technol.* 53, 134–144. doi: 10.1002/asi.10016
- Weekley, J. A., and Ployhart, R. E. (2013). *Situational Judgment Tests: Theory, Measurement, and Application*. Mahwah: Erlbaum.
- Wierzbicki, A. (2018). *Web Content Credibility*. New York, NY: Springer Berlin Heidelberg. doi: 10.1007/978-3-319-77794-8
- Wineburg, S., Breakstone, J., McGrew, S., and Ortega, T. (2018). "Why google can't save us. The challenges of our post-gutenberg moment," in *Positive Learning in the Age of Information*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Wiesbaden: Springer), 221–228. doi: 10.1007/978-3-658-19567-0\_13
- Wineburg, S., and McGrew, S. (2016). Why students can't google their way to the truth: fact-checkers and students approach websites differently. *Educ. Week* 36, 22–28.
- Wineburg, S., and McGrew, S. (2017). *Lateral Reading: Reading Less and Learning More When Evaluating Digital Information (Working Paper)*. Available online at: <https://ssrn.com/abstract=3048994> (accessed July 22, 2018). doi: 10.2139/ssrn.3048994
- Wineburg, S., McGrew, S., Breakstone, J., and Ortega, T. (2016a). Evaluating information: the cornerstone of civic online reasoning. *Stanford Digital Repository*.
- Wineburg, S., McGrew, S., Breakstone, J., and Ortega, T. (2016b). *Evaluating Information: The Cornerstone of Civic Online Reasoning: Executive summary*. Stanford History Education Group. Available online at: <http://purl.stanford.edu/fv751yt5934> (accessed June 24, 2020).
- Winter, S., Metzger, M. J., and Flanagin, A. J. (2016). Selective use of news cues: a multiple-motive perspective on information selection in social media environments. *J. Commun.* 66, 669–693. doi: 10.1111/jcom.12241
- Xie, I. (2008). *Interactive Information Retrieval in Digital Environments*. Hershey: IGI Global. doi: 10.4018/978-1-59904-240-4
- Zhang, S., and Duke, N. K. (2008). Strategies for internet reading with different reading purposes: a descriptive study of twelve good internet readers. *J. Lit. Res.* 40, 128–162. doi: 10.1080/10862960802070491
- Zhang, S., Duke, N. K., and Jiménez, L. M. (2011). The WWWDOT approach to improving students' critical evaluation of websites. *Reading Teach.* 65, 150–158. doi: 10.1002/TRTR.01016
- Zlatkin-Troitschanskaia, O. (2020). *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)*. Cham: Springer International Publishing. doi: 10.1007/978-3-030-26578-6
- Zlatkin-Troitschanskaia, O., Beck, K., Fischer, J., Braunheim, D., Schmidt, S., and Shavelson, R. J. (2020a). The role of students' beliefs when critically reasoning from multiple contradictory sources of information in performance assessments. *Front. Psychol.* 11:2192. doi: 10.3389/fpsyg.2020.02192
- Zlatkin-Troitschanskaia, O., Brückner, S., Molero, D., and Bisang, W. (2020b). "What can we learn from theoretical considerations and empirical evidence on learning in higher education? Implications for an interdisciplinary research framework," in *Frontiers and Advances in Positive Learning in the Age*

- of *InformaTiOn (PLATO)*, ed. O. Zlatkin-Troitschanskaia (Cham: Springer International Publishing), 287–309.
- Zlatkin-Troitschanskaia, O., Dengel, A., and Wittum, G. (2018a). *Positive Learning in the Age of Information: A Blessing or a Curse?* Wiesbaden: Springer VS. doi: 10.1007/978-3-658-19567-0
- Zlatkin-Troitschanskaia, O., Schmidt, S., Molerov, D., Shavelson, R. J., and Berliner, D. (2018). “Conceptual fundamentals for a theoretical and empirical framework of positive learning,” in *Positive Learning in the Age of Information: A Blessing or a Curse?*, eds O. Zlatkin-Troitschanskaia, A. Dengel, and G. Wittum (Wiesbaden: Springer VS.), 29–50.
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., and Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *Br. J. Educ. Psychol.* 89, 468–484. doi: 10.1111/bjep.12286
- Zlatkin-Troitschanskaia, O., Toepper, M., Molerov, D., Buske, R., Brückner, S., Pant, H. A., et al. (2018b). “Adapting and validating the collegiate learning assessment to measure generic academic skills of students in Germany: implications for international assessment studies in higher education,” in *Assessment of Learning Outcomes in Higher Education*, eds O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, and C. Kuhn (Cham: Springer International Publishing), 245–266. doi: 10.1007/978-3-319-74338-7\_12
- Zumbo, B. D., and Hubley, A. M. (2017). Understanding and Investigating Response. *Processes in Validation Research*. Cham: Springer, 69. doi: 10.1007/978-3-319-56129-5
- Zylka, J., Christoph, G., Kröhne, U., Hartig, J., and Goldhammer, F. (2015). Moving beyond cognitive elements of ICT literacy. First evidence on the structure of ICT engagement. *Comput. Hum. Behav.* 53, 149–160. doi: 10.1016/j.chb.2015.07.008

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Molerov, Zlatkin-Troitschanskaia, Nagel, Brückner, Schmidt and Shavelson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Evaluation of Online Information in University Students: Development and Scaling of the Screening Instrument EVON

Carolin Hahnel<sup>1,2\*</sup>, Beate Eichmann<sup>1,2</sup> and Frank Goldhammer<sup>1,2</sup>

<sup>1</sup> DIPF | Leibniz Institute for Research and Information in Education, Frankfurt, Germany, <sup>2</sup> Centre for International Student Assessment (ZIB), Frankfurt, Germany

## OPEN ACCESS

### Edited by:

Patricia A. Alexander,  
University of Maryland, United States

### Reviewed by:

Tom Rosman,  
Leibniz Institute for Psychology  
Information and Documentation  
(ZPID), Germany  
Martin Senkbeil,  
University of Kiel, Germany

### \*Correspondence:

Carolin Hahnel  
hahnel@dipf.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Psychology

**Received:** 14 May 2020

**Accepted:** 16 November 2020

**Published:** 16 December 2020

### Citation:

Hahnel C, Eichmann B and  
Goldhammer F (2020) Evaluation  
of Online Information in University  
Students: Development and Scaling  
of the Screening Instrument EVON.  
*Front. Psychol.* 11:562128.  
doi: 10.3389/fpsyg.2020.562128

As Internet sources provide information of varying quality, it is an indispensable prerequisite skill to evaluate the relevance and credibility of online information. Based on the assumption that competent individuals can use different properties of information to assess its relevance and credibility, we developed the EVON (evaluation of online information), an interactive computer-based test for university students. The developed instrument consists of eight items that assess the skill to evaluate online information in six languages. Within a simulated search engine environment, students are requested to select the most relevant and credible link for a respective task. To evaluate the developed instrument, we conducted two studies: (1) a pre-study for quality assurance and observing the response process (cognitive interviews of  $n = 8$  students) and (2) a main study aimed at investigating the psychometric properties of the EVON and its relation to other variables ( $n = 152$  students). The results of the pre-study provided first evidence for a theoretically sound test construction with regard to students' item processing behavior. The results of the main study showed acceptable psychometric outcomes for a standardized screening instrument with a small number of items. The item design criteria affected the item difficulty as intended, and students' choice to visit a website had an impact on their task success. Furthermore, the probability of task success was positively predicted by general cognitive performance and reading skill. Although the results uncovered a few weaknesses (e.g., a lack of difficult items), and the efforts of validating the interpretation of EVON outcomes still need to be continued, the overall results speak in favor of a successful test construction and provide first indication that the EVON assesses students' skill in evaluating online information in search engine environments.

**Keywords:** evaluating online information, link selection, information relevance and credibility, university students, test development and validation

## INTRODUCTION

Information literacy and related competencies have become essential in the digital era, as they refer to skills and knowledge that students need in order to act effectively, confidently, and successfully in dynamic and interconnected information environments. However, there is an urgent need to improve students' information literacy beyond simply making necessary tools and resources



available. For example, according to the international large-scale assessment ICILS (Fraillon et al., 2020), only a small percentage of the participating school students were able to critically evaluate and use information when searching online (see also Breakstone et al., 2019). University students, who are expected to possess a certain level of competence (Association of College and Research Libraries, 2000), are no exception to this phenomenon. Studies indicate difficulties in identifying information and information sources that are reliable and trustworthy (e.g., Walraven et al., 2008; Maurer et al., 2017), but there are efforts to support students in developing their information literacy (e.g., Peter et al., 2017; McGrew et al., 2019). One recent European example is the multilingual Massive Open Online Course (MOOC) of the Erasmus+ project Information Literacy Online (ILO; Mandl et al., 2018)<sup>1</sup>. This MOOC provides students with open learning materials, quizzes, and achievement tests for self-assessment purposes. The EVON is one of those achievement tests, with the aim of giving students a first impression of their performance in evaluating the relevance and credibility of online information from search engine results—a central component skill of information literacy. In this article, we report on the test development and first efforts to validate the interpretation of its test score (i.e., construct interpretation).

## EVALUATING AND SELECTING ONLINE INFORMATION

Processing, evaluating, and deciding on the use of information during a web search is a complex phenomenon. Accordingly, there are various interdisciplinary approaches in research, often focusing on selected aspects. In this section, we give a short introduction to different conceptualizations, theories, and related empirical observations. We start with a formal description of the web search process and elaborate on when evaluations are triggered in this process, what purposes they serve and why their depth will vary depending on the context (see section “Web Search as a Decision-Making Problem”). We then go into detail about how individuals determine the relevance of information for a particular task (see section “Determination of Relevance”) and how they make credibility assessments of information and information sources (see section “Determination of Credibility”). We conclude the introduction with a short overview of previous assessment approaches that capture how individuals assess the relevance and credibility of information (see section “Assessment Approaches”).

### Web Search as a Decision-Making Problem

Search engines usually provide web users with large amounts of information that can relate to a topic of interest in many different ways (see e.g., Bendersky et al., 2012). In procedural descriptions of the web search process, such as the IPS model (information problem solving; Brand-Gruwel et al., 2005, 2009), it is distinguished that web search requires individuals to

(1) identify their information needs; (2) specify their search strategy and select links on a search engine result page (SERP) based on initial judgments; (3) scan the information on the websites visited to get an idea of whether it could be useful; (4) deeply process the information identified as useful in the previous step and integrate it with previously found information and prior knowledge; and (5) compare and integrate all collected information to form some kind of response. Steps (2) to (4) require web users to evaluate information in order to decide which information object should be selected from multiple alternatives and considered as part of a response.

The assessment of relevance and credibility is considered an iterative process in which a person makes a series of judgments about the available information (Hilligoss and Rieh, 2008). The scientific literature mainly distinguishes between two types of judgment, which serve different purposes: Predictive judgments are made before accessing the object of evaluation (e.g., a website); evaluative judgments are made when confronted with the object of evaluation (Rieh and Danielson, 2007). Predictive judgments are used to anticipate the value of information for a task and to decide whether or not to follow a SERP link or consider a particular website. A web user's perception of the value of information (“information scent”) is obtained by cues in the immediate task environment (“proximal cues”; e.g., Sundar et al., 2007). Such cues can manifest themselves in many ways, for example, semantically (e.g., keywords from the search query; Rouet et al., 2011) or by describing structural, message-related and sponsor-based features of information (e.g., website layout, topicality, or source reputation; Metzger and Flanagin, 2013). They are often only examined for the first few entries of a SERP, which indicates an implicit trust in the optimization of search engine algorithms (e.g., Pan et al., 2007; Kiili et al., 2008; Walraven et al., 2008; Kammerer and Gerjets, 2014). Failure to find “valuable” information is more likely to prompt web users to modify their search query rather than to continue examining other SERP entries (e.g., Huang and Efthimiadis, 2009; Hollink et al., 2012). Accordingly, predictive judgments represent in some way a “bouncer” in deciding whether information should be processed at all, with the accessibility and interpretability of cues being crucial to this decision. This also means that web users may omit important information or turn to less suitable information if their predictive judgments are inadequate (see Kiili et al., 2008). Evaluative judgments, in contrast, serve to determine whether and how identified information is suitable for solving the information problem. If individuals come to the conclusion that the information is of value for providing a sufficient outcome (Pirulli and Card, 1999), they will process this information in further detail and integrate it as part of fulfilling their search task. If not, the website is likely to be discarded (e.g., Salmerón et al., 2017).

The depth and level of detail of evaluations made will depend on the way in which web users process the identified cues. Dual-processing theories (e.g., Wirth et al., 2007) distinguish between systematic processing, which involves a relatively analytical, thorough, and comprehensive examination of information versus heuristic processing, which is fast and automatic and

<sup>1</sup><https://informationliteracy.eu/en>

does not consume too much processing resources (e.g., time and attention). They suggest that online information is not fully processed, with the result that individuals use cognitive “shortcuts” based on the cues considered (Gigerenzer and Gaissmaier, 2011). Similar predictions are made based on information foraging theory (Pirulli and Card, 1999) that postulates that web users search in a way to maximize their gain of valuable information while keeping their effort as low as possible. Depending on the context, however, heuristics can be inadequate, leading to erroneous assessments (e.g., Rouet et al., 2011; Metzger and Flanagin, 2013).

Web users will primarily select information based on its relevance to the task at hand (see Rouet, 2006; Kiili et al., 2008), although a concurrent critical evaluation of source characteristics of information is indispensable as it can help individuals to avoid misinformation and overcome misconceptions (overview in Braasch and Graesser, 2020). A source of information might be recognized as credible, but is unlikely to be considered further if it does not provide any indication of relevance. Accordingly, relevance assessments traditionally are important criteria for assessing the credibility of information (Rieh and Danielson, 2007). Nevertheless, in order to understand the mechanisms of individuals' assessment of relevance and credibility, it is useful to consider both aspects in their own right.

## Determination of Relevance

Relevance concerns the extent to which information matches the needs given the specifications of a task (McCrudden et al., 2005). Accordingly, the degree to which information segments are evaluated as relevant will mainly depend on a web user's search goal. To determine relevance, web users will rely on the use of surface cues and deep semantic cues that require decoding and comprehension. They can benefit from both types, although an overreliance on superficial cues can result in neglecting important aspects. There is evidence that adolescents show increasing skill in recognizing deep semantic cues over time (Rouet et al., 2011). Compared to older students, early secondary school students tended to rely more on surface cues (e.g., keywords that are written in upper cases), indicating that younger students experience more difficulties in balancing the use of surface and deep cues when selecting website titles (ibid.). Keil and Kominsky (2013) came to a similar conclusion studying how 11-year-olds to over 18-year-old students increasingly include discipline-related cues in their evaluation of search results. The recognition of deep conceptual relationships between a search task and a search result that are not entirely obvious (i.e., due to the absence of lexical similarity on the surface) increased over high school years and received a level in grade 10 that was comparable to adult-like performance. Although an overreliance on surface cues seems to decrease over time, it remains crucial that web users do not falsely determine relevance from an uncritical use of surface cues.

Besides prior knowledge (e.g., Hölscher and Strube, 2000), other important factors that influence how web users determine the relevance of information clearly concern information processing skills (or conditional skills in the IPS framework; Brand-Gruwel et al., 2009), such as reading. Reading skills

support web users in identifying and locating relevant information, for example, by enabling them to extract main ideas from text (Hahnel et al., 2016). Highly skilled readers also seem to be in a better position to identify deep semantic cues and make use of them to efficiently discard irrelevant information (Hahnel et al., 2018). However, this does not necessarily mean that skilled readers are also skilled searchers. Salmerón et al. (2017) found that if skilled readers fell for irrelevant sections of a digital text, they were at a greater disadvantage than less skilled readers, indicating that skilled readers do not automatically recognize deep semantic cues correctly or sufficiently process them.

## Determination of Credibility

Traditional “gatekeepers” such as editors, reviewers, and publishers are often not available to ensure the integrity of online information (Flanagin and Metzger, 2007; Rieh and Danielson, 2007). Accordingly, the recognition of credibility aspects of information has become increasingly necessary, in particular when information is presented in a way that resembles editorial content but is paid for by an advertiser (sponsored content as part of native advertising; see Amazeen and Muddiman, 2018). This is a difficult task for students, even when the advertisements are explicitly marked (Wineburg et al., 2018). Students rarely spontaneously evaluate credibility aspects of information obtained (for an overview, see Bråten et al., 2018), and although they tend to select information from seemingly credible sources, students lower their evaluation standards if they do not have access to better information sources (Kiili et al., 2008).

According to feature or checklist approaches (Flanagin and Metzger, 2007; Metzger, 2007; see also Chinn and Rinehart, 2016; van Zyl et al., 2020), web users' perception of credibility will depend on their judgments referring to structural (e.g., design features and website complexity), message-based (e.g., accuracy and writing style), and sponsor-based features (e.g., personal experience with the sponsor). The weight given to each feature may vary depending on the genre of website or other circumstances (e.g., websites from news organizations are generally rated more credible than personal websites; Flanagin and Metzger, 2007). It is noteworthy that we distinguish between semantic cues and structural, message-based and sponsor-based features, although there is a strong conceptual overlap in the properties addressed. This is done with the purpose of distinguishing whether a cue or feature is primarily used to determine relevance or credibility. For example, recognizing the intention of a text will inform both the assessment of relevance and credibility, but might be evaluated with an emphasis either on whether the content can contribute to solving the information problem or whether the text has secondary motives.

The recognition and use of specific features are assumed to trigger heuristics to aid the assessment of credibility (Metzger and Flanagin, 2013). Accordingly, participants, interviewed in focus groups, showed to employ a wide variety of cognitive heuristics, which Metzger et al. (2010) classified as rooted in social confirmation (e.g., reputation heuristics, such as the

rule of thumb that URLs of .org domains are credible) or rooted in expectancies within specific contexts (e.g., persuasive intent heuristics, such as the presence of advertisements as negative credibility indicators). Although such heuristics are often helpful, they can still lead to biased assessments, for example, when information is dismissed as not credible only because of discrepancies with one's own beliefs or those of peers and vice versa (see Braasch and Graesser, 2020).

Checklist approaches imply that information credibility is determined by whether or not the information and its source show certain characteristics. It should be noted that Chinn and Rinehart (2016) argue that such characteristics are only valid if they actually correspond to the use of reliable epistemic processes to produce knowledge claims. That means, for example, that a news website should be considered credible not because it is operated by a news agency, but because its journalists produce knowledge claims that are accurate and plausible in their argumentation, which rely on processes of thorough search, evaluation, and synthesis of evidence to produce them. Recent considerations support this view arguing that core components of critical thinking (e.g., evaluating whether a claim is validated by examining the argument surrounding it) can enrich checklist approaches and should be considered to foster students' credibility assessment (van Zyl et al., 2020; see also Stadler and Bromme, 2014, on strategies to reconcile conflicts about competing scientific claims). Nevertheless, provided that they are closely related to such epistemic processes, structural, message-based, and sponsor-based features are useful markers that present web users with comparatively simple and straightforward ways to assess the credibility of information.

## Assessment Approaches

Many instruments claim to assess information literacy, which emphasizes the importance of this construct in research and society. In an attempt to structure the field, Walsh (2009) reviewed 91 scientific articles, summarizing several approaches to assess information literacy. He identified in total nine different methodologies (e.g., essays, observations, portfolios, "self-assessments" in the sense of self-report). Most prominently were multiple-choice questionnaires and quizzes, but Walsh remarks that the respective studies have often not been thorough in their efforts to investigate the reliability and validity aspects of their instrument (see also Rosman et al., 2016, for a discussion of different test formats).

Recent approaches are increasingly focusing not only on declarative knowledge aspects of students' information literacy, but also on procedural knowledge and actual behavior. We briefly highlight some instruments of information literacy that we think have a convincing approach. For example, Lechner et al. (2014) suggested a taxonomy to create information search tasks that request students to find a scientific article about a subject. After each task, the students are asked several questions about their task processing, which serves as the basis of scoring students' procedure. Rosman et al. (2016) proposed a less resource-consuming vignette-based approach. They constructed a test of 28 situational judgment tasks that provided students

with a scenario description and several possible procedures to solve the scenario and requested them to rate each procedure according to its usefulness. Also worth mentioning is the serious game of Steinrück et al. (2020). They measured information literacy by classifying the in-game behavior of individuals playing a crisis situation manager game. However, their validation approach strongly relied on a self-report, not an independent performance measure.

Especially students' evaluation of information from search engines is often examined based on their performance in open search tasks of varying complexity (e.g., fact-finding vs. research-oriented tasks, closed-ended vs. open-ended tasks; e.g., Wirth et al., 2007; Kiili et al., 2008; Brand-Gruwel et al., 2009; Bilal and Gwizdka, 2018; Pardi et al., 2020). The assessment, scoring, and evaluation of performance are usually recorded by an additional tracking application, such as screen recording or a proxy server that retrieves search engine data in the background. Although such task setups can provide substantial information about the evaluation skills of individuals, they are often not standardized or lack controlled and comparable conditions. Therefore, a number of researchers have moved toward the development of search tasks in mock environments. That means they have created search engine results and/or websites that were identical for all participants or groups of participants to ensure comparability (e.g., Rouet et al., 2011; Keil and Kominsky, 2013; Metzger and Flanagin, 2013; Kammerer and Gerjets, 2014). Such simulation-based approaches are also often used to assess constructs that are closely related to information literacy, such as individuals' skills in dealing with information and communication technologies (e.g., ICILS, Fraillon et al., 2019; for an overview see Siddiq et al., 2016), problem-solving in technology-rich environments (e.g., Goldhammer et al., 2020), digital reading (OECD, 2011), or skills in online research and comprehension (ORCA; e.g., Leu et al., 2014).

A simulation-based approach was also implemented by Keßel (2017) to test the evaluation skill of adolescents (see also Hahnel et al., 2018). She developed 24 items that simulated search results and Internet forums in which students were requested to identify and select the most credible entry for the respective search task. Eight of these items presented students with a page of search results (SERP) related to topics on health, crafts, sports, and education. The items were interactive, as students are allowed to access a website through the links, providing them with detailed information. A correct answer was defined by the search result (i.e., the target) with the highest number of features that identified it as credible. The items varied according to the attractiveness of non-target search results (low vs. high attractiveness) and the congruence of features indicating the credibility of the source underlying the search results (congruence vs. incongruence). Keßel defined these criteria based on the number of features that indicate the credibility of the SERP results (attractiveness) and based on whether the information of a SERP result and its corresponding website signal a similar degree of credibility (congruence). Inspired by her instrument, we developed the EVON (evaluation of online information) to assess the evaluation skill of students in higher education.

## FRAMEWORK AND TEST DEVELOPMENT

Based on the theoretical background, we define the skill to evaluate online information as the cognitive skill to recognize and make use of semantic cues and structural, message-based, and sponsor-based features in order to evaluate the relevance and credibility of information in search engine environments (after Keßel, 2017). We assume that students who engage in web search first scan a SERP and generate a series of predictive judgments to preselect websites for close examination (Rieh and Danielson, 2007; Brand-Gruwel et al., 2009). When a website is accessed, we assume that students make evaluative judgments to determine the extent to which the website contributes to the completion of their search task. If a decision has to be made between several positively evaluated alternatives, the identified relevance and credibility aspects are compared and weighed against each other. Accordingly, a student competent in evaluating online information is able to select websites suited for a specific task based on informed conclusions about the relevance and credibility of information. A test that claims to assess how students evaluate online information should therefore take this process into account and provide students with opportunities to judge different features of links and websites of varying relevance and credibility. In the following, we describe the development of the interactive computer-based instrument EVON, which aims to provide students who wish to improve their information literacy (Mandl et al., 2018) with a screening of their evaluation skills.

### Guidelines for Item Design

The EVON was designed to request students to select the most relevant and credible link in a simulated search engine environment for a respective task. Accordingly, we have adopted the basic task structure of Keßel's (2017) items simulating a SERP and websites. However, we have decided to emphasize the role of relevance assessment because it is likely that information in web search contexts will not be further processed if it is not found to be related to a task at hand. Although checklist approaches consider relevance as part of the credibility assessment (especially with regard to message-based features), we intended to acknowledge in particular situations where websites can be credible but may be not relevant and vice versa.

The new items were designed to present a target that is the optimal solution in terms of both relevance and credibility of information. Competing non-targets were characterized by flaws and shortcomings compared to the target. In the revision process, we made sure that the provided cues and features were consistent with the expected epistemic processes (e.g., if a website was authored by an expert, the knowledge claim would be accurate; see Chinn and Rinehart, 2016). **Table 1** summarizes the combinations of the two main design criteria, attractiveness and congruence. However, we have broadened the definition of Keßel's (2017) design criteria to explicitly consider relevance aspects and implications for the expected item solution process. For each of the four resulting types, two tasks were developed that presented either three or five information sources on a SERP.

**TABLE 1 |** Guidelines for item design.

Item type	Guiding characteristics	Description	Expectation for the solution process
1	Low attractiveness of non-targets. Congruence between link and website	The target link already stands out from the non-target links in terms of features signaling relevance and credibility	Navigation is not necessary, as predictive judgments are sufficient, but can consolidate a decision
2	High attractiveness of non-targets. Congruence between links and websites	The target differs only marginally from non-targets in features signaling its relevance and credibility	Individuals need to judge and consider several aspects of information from both link and website to identify the best option
3	High attractiveness of non-targets. Incongruence between target link and website	The target link differs only marginally from non-target links in features signaling its relevance and credibility, but its website stands out compared to non-targets	Individuals can identify the target as the best option by inspecting its website
4	High attractiveness of non-targets. Incongruence between non-target links and websites	The target link differs only marginally from non-target links in features signaling its relevance and credibility, but the non-target websites violate the expectations generated by their links	Individuals can exclude non-targets by inspecting their websites

The attractiveness criterion addresses the extent to which non-target SERP links display cues that affect their perceived information value. Non-target links of low attractiveness are only superficially related with a search task (item type 1; e.g., when searching for a solution to an email attachment problem, the results not only present a link addressing the problem, but also a link about dangerous attachments in phishing emails). As in these conditions students can potentially identify the target based on predictive judgments, these tasks are supposed to be the easiest tasks. In contrast, highly attractive non-target links signal an information value similar to the target link, which means that predictive judgments cannot be used exclusively to identify the best source of information (type 2; e.g., when searching for information about diving equipment for beginners, the results present a link about basic equipment and links about special equipment). Accordingly, a high non-target attractiveness is expected to increase the item difficulty.

The congruence criterion addresses the extent to which SERP links can raise expectations that may be violated by the information on the website. Because of the extended scope compared to the definition of Keßel (2017), we considered that with regard to authentic web search situations, this criterion is only meaningful for non-targets that are as attractive as the target (i.e., the condition of high attractiveness). With respect to the incongruence condition, the most significant change that we made was to indicate the object and the direction of incongruence. That means we distinguished between situations in which the target link (type 3) or the non-target links (type 4) violate the expectations formed by predictive judgments. In type 3 items, the SERP presents a list of moderately useful-looking links (e.g., when searching for remedies against a cold, the SERP



lists websites from a news agency, a pharmaceutical journal, or a discussion forum), with the target being clearly identifiable as suitable by the information on its website. In type 4 items, all links on the SERP indicate to provide useful information, but when visiting the non-targets, it becomes evident that their websites are less appropriate (e.g., they indicate primary commercial intentions or address a different audience). As students may need to reconsider their initial assessment of relevance and credibility after new (incongruent) information is discovered, the tasks of the incongruent conditions are supposed to be difficult, but visits to websites can facilitate the evaluation, as more information becomes available to make an informed decision.

An overview of all developed items is presented in **Table 2** (with detailed information about the respective item type in **Table 1**). An example item is displayed in **Figure 1**. The item “Recovering from a cold” instructs students to search for useful and trustworthy information to treat a common cold. This item belongs to item type 3 (i.e., high attractiveness of non-targets, incongruence between target link and website). According to the high attractiveness condition, the search results on the SERP were created to appear equally suited to solve the underlying information problem (“get a grip on a cold quickly,” “get rid of your unwanted cold,” “What should I do to get well quickly,” etc.). A SERP of low attractiveness would require non-target search results to be only superficially related to the search task (e.g., with regard to the word “cold,” a website could refer to chronic obstructive lung diseases or a rock band). According to the target-incongruence criterion, the target website is supposed to stand out in terms of relevance and credibility. In case of the example item, the target link (“Pharmaceutical newspaper”) suggests that the website is directed to a professional audience, but when inspecting the website (and eventually comparing it to the other websites), it becomes clear that its information is suitable to solve the search task, information about the author and publisher is clearly stated, and it can be expected that the author and publisher have authority in the respective field. For comparison, in the case of congruence, the link would actually lead to a website with highly specific pharmaceutical information.

## The Developed Test

The developed items of the EVON cover different topics that were chosen in consultation with representatives of the target population to ensure that the topics are relevant and authentic

(**Table 2**). Nevertheless, we aimed at constructing the test in a way that students had as little advantage as possible due to their prior knowledge. Accordingly, the contents are fictitious, with existing websites having served as loose templates. Mainly due to copyrights, we have also refrained from using real brand and organization names.

The EVON is a power test in which students are asked to perform at their best (see Klehe and Anderson, 2007). We aimed for a setting that was as authentic and unobtrusive as possible, but the purpose of the assessment is not masked in any way. The item instructions explicitly request students to select a link for a respective task with regard to relevance and credibility aspects (“[...] select the website with the most useful and trustworthy information [...]”). Students’ performance is scored dichotomously to whether they selected the target or a non-target. During the test-taking process, mouse-click data with timestamps are collected in log files. An interactive tutorial introduces students to the environment and all available functionalities. We recommend a total test time of 18 min to complete the EVON assessment. The EVON was implemented with the software CBA ItemBuilder<sup>2</sup> and is available in six different languages (German, English, Spanish, Catalan, Croatian, and Slovenian). The corresponding author can be requested for test uses and modifications.

## Examining the Intended Test Score Interpretation

Based on students’ information selection, the EVON claims to assess their skill to evaluate the relevance and credibility of online information in search engine environments. A first step to support this claim was taken with the theory-based design of the interactive and authentic task environment. To further ensure the quality of the assessment and to validate the intended interpretation of the EVON score, we conducted a pre-study during the phase of item development and a main study after the EVON item set was finalized. The overarching goal of these studies was to collect validity evidence from different sources that provide information on the perception of item content, response processes, the internal structure of the EVON, and the nomological network of its score, allowing to evaluate arguments for and against the intended interpretation of the EVON score (see American Educational Research Association et al., 2014).

With regard to the test construction, we investigated whether the items are suitable to elicit and observe information selection based on students’ assessment of relevance and credibility (pre-study). After finalizing the test development, we investigated the internal structure of the EVON and effects of the item design criteria on the item difficulty by means of a larger student sample (main study). To investigate evidence referring to the nomological network of the EVON score, the network of relations to construct-related variables was also examined (main study). We focused on the relationship of the EVON score

<sup>2</sup>The CBA ItemBuilder is an authoring tool to create dynamic and interactive assessment and learning environments. It is free of charge and can be requested from the Centre for Technology-Based Assessment at DIPF (ib-support@dipf.de). [https://tba.dipf.de/en/infrastructure/software-development/cba-itembuilder/cba-itembuilder-1?set\\_language=en](https://tba.dipf.de/en/infrastructure/software-development/cba-itembuilder/cba-itembuilder-1?set_language=en)

**TABLE 2 |** EVON item overview.

Item	Description	Item type	No. links
1	Restoring the charging capacity of a laptop battery	1	3
2	Recovering from a cold	3	5
3	Writing a scientific paper	4	5
4	Repairing a broken bicycle chain	3	3
5	Finding out about basic equipment required for diving	2	5
6	Preparing for a stress-free examination period	4	3
7	Resolving the blocking of an email attachment	1	5
8	Financing a semester abroad	2	3

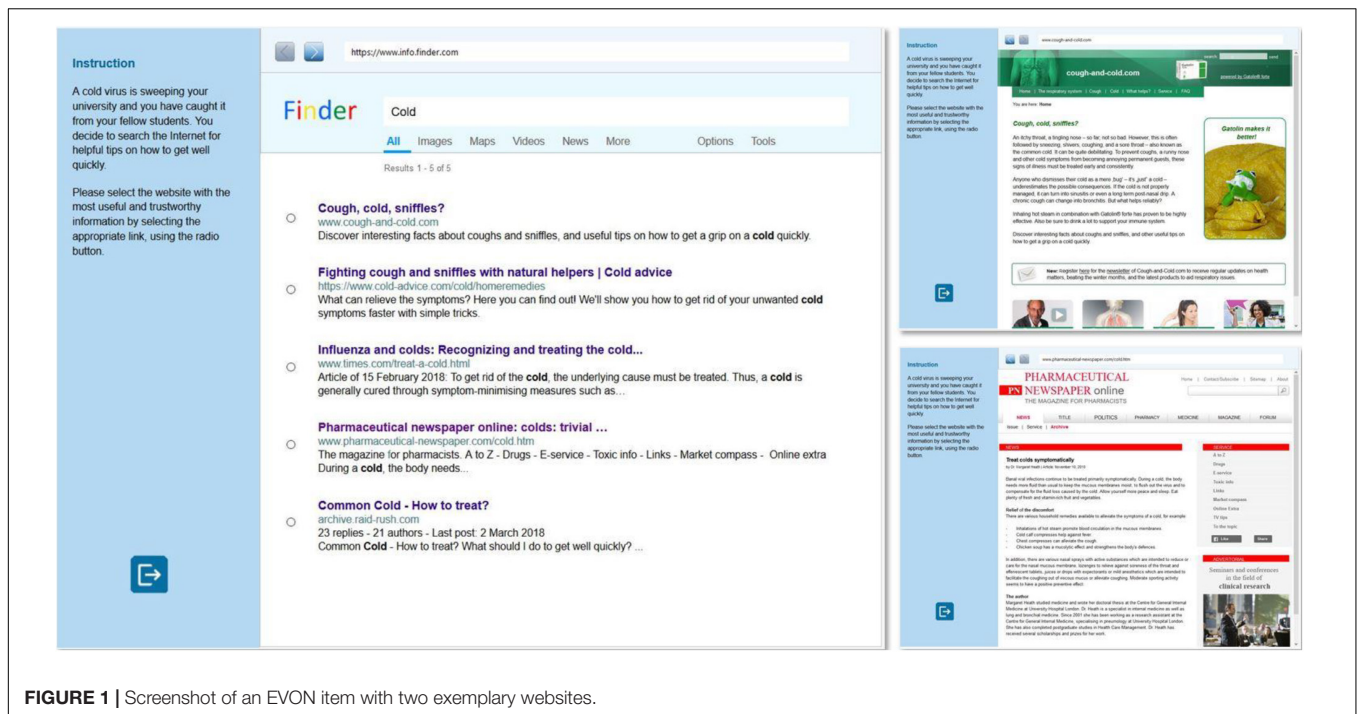


FIGURE 1 | Screenshot of an EVON item with two exemplary websites.

to students' general cognitive performance and basic reading skills, taking into account their self-reported prior knowledge of the EVON topics.

## PRE-STUDY

### Aim of the Study

Cognitive interviews were carried out to observe the course of students' processing of the constructed items. The objectives were twofold: First, the study served to ensure the comprehensibility of item content and the usability of the test environment. Second, it was investigated whether the presented semantic, structural, message-based, and sponsor-based cues were identified and used to assess the relevance and credibility of information. Adjustments were made in response to participants' feedback on incomprehensibility and misconceptions (e.g., clarifying instructions, modifying link and website information to provide more or less relevance and credibility related cues).

### Method

We collected the data of eight students (five females;  $\text{mean}_{\text{age}} = 25.6$  years; seven enrolled in a master's program). The test sessions were organized individually and lasted for 1.5–2 h, depending on the participants' speed. An interviewer welcomed and instructed the participants and monitored the session. After giving their written and informed consent, the participants were instructed to think aloud while working on the German version of the EVON. Camtasia Studio 6 was used to synchronously record participants' voice and processing behavior (via screen capture). To familiarize participants with the think-aloud procedure, each session started with a warm-up

task. If the participants stopped verbalizing their thoughts, the interviewer reminded them to keep talking (see van Someren et al., 1994). During the assessment, the interviewer took notes about a participant's behavior (e.g., which link attracted the participant's attention first, which link was ignored, or which websites were clicked but left quickly). After completing the EVON, the interviewer asked the participants questions about the appropriateness of the tutorial, the clarity of content and instructions, the authenticity of the simulated web environment, and any specificities identified during the session (e.g., why was a particular link ignored). The interviewer also asked the participants for an assessment of their prior knowledge of the EVON content, as well as demographic information (age, gender, study program, and semester). Afterward, the test session was completed, and participants could choose to receive course credit or a monetary compensation for their participation. The resulting screen-capture videos with the participants' verbalized thoughts and their answers during the interview were transcribed. The transcripts and the interviewer's notes were analyzed to determine if the items were processed as intended.

### Results and Discussion

The simulated web environment was generally perceived as authentic and natural, with only two remarks indicating astonishment that someone was looking for remedies for a common cold (remark by Charlotte<sup>3</sup>) or that only three results were returned from the search engine (Fiona). Overall, the responses and comments of participants suggested that they processed the EVON items as intended. During the processing of the EVON items, they commented on specific semantic,

<sup>3</sup>The names of all participants are fictitious.

structural, message-based and sponsor-based properties of the EVON stimuli indicating that they recognized and interpreted these cues to determine the relevance and credibility of links and websites. In addition, they explicitly reported on their use of these cues during the interviews. Below, we illustrate our findings with selected interview snippets from the item “Recovering from a cold” (Figure 1).

Examples that indicate the use of surface and deep semantic cues for assessing information relevance are presented in Table 3. The participants demonstrated to integrate surface cues in their judgments by mentioning keywords in the SERP link (Alexander and Emily) or scanning the website (David and Fiona). Alexander's and Emily's comments are examples of predictive judgments that are generated to decide whether to visit or dismiss a particular website. In contrast, David's and Fiona's comments rather reflect evaluations to decide whether a website is worth reading thoroughly. David's comment even incorporates the use of a message-based feature that backs up his decision with an initial credibility judgment of the website (“if a doctor even writes that”). The examples for the use of deep semantic cues suggest that the participants reflected deeply on how the encountered information contributes to solve the associated search task by evaluating it in light of their personal experiences and world knowledge (Alexander and Bianca) or in terms of whether the provided information meets the requirements of the search task (David and Henry).

Table 4 shows examples of how the participants referred to structural, message-based, and sponsor-based features of websites to infer on information credibility. Structural features

**TABLE 3 |** Indications for the use of semantic cues to determine information relevance.

Cue	Example quotes
Surface	<p>Alexander: [Inspects link 1] “So... Cough, cold and sniffles [mumbles]. Okay, that sounds pretty good.”</p> <p>David: [Scans website 4] “Hmm, treat symptomatically. OK, I'll have a look. OK, it's probably a newspaper... so if a doctor [...] even writes that, then I would have a closer look.”</p> <p>Emily: [Inspects link 2] “[reads ‘Fight coughs and colds with natural helpers... relieve symptoms’] Well that would be something, it's all about getting me healthy again quickly.”</p> <p>Fiona: [Scans website 3] “Help with cold... bacterial infections... antibiotics may be necessary... allergic rhinitis... allergies... active ingredient... [pause]. Okay doesn't quite seem to be it.”</p>
Deep	<p>Alexander: [Website 2] “There are even some... Exactly, there are also recipes with which I can make myself something to drink or eat. And I know that with ginger, lemon juice, honey, yeah that should probably help.”</p> <p>Bianca: [Website 2] “Is it important to drink a lot [...] I find that good, because the doctor, when I am sick, always tells me: Drink, drink, drink. So this is also what the doctor advises me.”</p> <p>David: [Website 3] “OK I don't have a... yeah I don't have an allergy. Well, that's not very helpful.”</p> <p>Henry: [After having visited websites 1 and 2] “So here the [link 1] was just an introduction, so I couldn't really know if that helps what is offered there. Here [link 2] was at least directly something visible.”</p>

**TABLE 4 |** Indications for the use of features to determine information credibility.

Feature	Example quotes
Structural	<p>Bianca: [Website 1] “That looks like a commercial to me with that medicine up there. [...] I find it funny with the frog on the side [pause], but ‘Gatolin makes it better’... well that uh puts me off.”</p> <p>Charlotte: [Website 2] “Well, the second page looks a bit trashy from a layout point of view, so not so reliable.”</p> <p>Emily: [Website 4] “Yes, I find that quite... well, somehow it's not so vivid, because there's such a small font and all, but hmm. Well, [it is] relatively clear, actually explains what you can do anyway, but don't find the page so likeable actually.”</p>
Message-based	<p>Alexander: [Website 4] “Uh especially in the field of medicine there are many who simply tell something that doesn't have to be true. [...] it is good to know that at least a doctor wrote it and not just anyone.”</p> <p>Bianca: [Website 3] “The source, ‘Internal differential diagnosis,’ OK. But the source is pretty old, from 1999!”</p> <p>[Inspects link 4] “Pharmaceutical newspaper, hmm magazine for pharmacists, okay I think that is... if it's for pharmacists, it will probably be too complicated for me.”</p> <p>Fiona: [Website 2] “Ginger tea usually always works well, says my mum.”</p>
Sponsor-based	<p>Alexander: [Website 4] “So I would say, since really, uhm, the publisher is named, and ah it's a serious publisher, I would prefer this source.”</p> <p>Bianca: [Scans website 1] “Ok, with the expert interview I automatically think that this site works with experts and therefore is qualitative.”</p> <p>Giselle: [Visited only website 4, retrospective interview] “That sounded trustworthy. Not because the others weren't any good, [...] I just had no reason to keep on searching [...] it was published in a newspaper and, I don't know, sounded better than [link 5].”</p>

mentioned referred to the presence of pictures (Bianca), the general layout of websites (Charlotte), or typesetting (Emily). Multiple features are sometimes blended and get integrated or weighted against each other, as the comments of Bianca and Emily demonstrate. In their comments, they refer to both structural features (“that medicine up there,” “such a small font”) as well as message-based (“relatively clear, actually explains”) and sponsor-based features (“looks like a commercial”). The comments classified as referring to message-based features show that the participants took different aspects into account when judging the message that a website intends to convey. They elaborated on the author's background (Alexander), evaluated information in terms of its currency and the comprehensibility of information provided (Bianca), or considered whether information was legitimated by trusted authorities (Fiona; also Bianca's comment in Table 3). With regard to sponsor-based cues, it might have been suspected that students would rather base their judgments primarily on structural and message-based features due to the lack of real brand and organization names (Flanagin and Metzger, 2007). However, sponsor-based cues were identified and taken into account, as shown by references to publishing organizations (Alexander and Giselle) or recognized expertise (Bianca).



The overall impression gained from the participants' comments is that they made use of several cues to infer both the relevance and credibility of the information provided and that they combined different heuristic strategies to process the EVON items, which is consistent with the assumptions of the test construction (e.g., Rouet, 2006; Brand-Gruwel et al., 2009; Metzger and Flanagan, 2013). In this respect, the results of the cognitive interviews provide first empirical evidence based on the item contents and the response processes observed, supporting the intended interpretation of the EVON score.

It should be noted that, in terms of performance, the participants showed high rates of correct responses (success rates per item between 50 and 88%). Accordingly, the test was rather easy. However, this might be due to the setup of cognitive interviews. As participants were asked to verbalize their thoughts and comment on the material as part of improving the items, they might have adopted a higher desired level of understanding the provided information and engaged in strategic rather than automatic processes of reading (standards of coherence; van den Broek et al., 2011). Accordingly, they might have reflected upon the links and websites more thoroughly than they would have done otherwise.

## MAIN STUDY

### Aim of the Study and Hypotheses

With the overarching objective of validating the interpretation of the EVON score, an online assessment was conducted to investigate the psychometric properties of the EVON and to test hypotheses relating to the design of its items and nomological network. With regard to the psychometric properties, it was expected that the EVON items contribute to the assessment of a unidimensional skill that is part of the broader construct of (online) information literacy. Support for the assumption of unidimensionality would allow for the differentiation of different skill levels in evaluating online information.

With regard to the item design (Table 1), we expected to find differences in item difficulty related to the item type and to whether or not students visited target or non-target websites. In general, items where non-targets signal a low value of information (type 1) were supposed to be the easiest items, whereas items where the target link differs only marginally from non-target links in features signaling its relevance and credibility (type 2 to 4) should be more difficult (H1.1). Visiting a target's website (i.e., target navigation) should facilitate solving the item correctly, as the target website is designed to provide information that marks the website as the best choice in terms of relevance and credibility (H1.2). On the contrary, there can be several reasons for visiting a non-target website (i.e., non-target navigation), from ensuring to not miss anything to just drawing inadequate inferences from the SERP information. We expected to see an overall negative effect of non-target navigation on the probability of task success, as it might indicate the result of inappropriate judgment (H1.3), but also a differential effect of non-target navigation in type 4 items (i.e., the incongruent condition where the website information fails the

link information). As non-targets in these items were designed to look highly attractive, but disappoint when visited, non-target navigation should actually support students in discarding the attractive alternative (H1.4).

With regard to the nomological network of the EVON, we investigated the relations of students' EVON performance with other variables. A test that claims to represent a skill to evaluate written information should mandatorily be associated with indicators of cognitive information processing. To examine this aspect, we investigated the relationship of the EVON with students' graduation grades (German "Abiturnote") as indicator of general cognitive performance and sentence-level comprehension as indicator of reading skill. German graduation grades are an aggregate of subject-specific grades assessed by several teachers over a couple of years. Accordingly, they do not reflect specific domain knowledge and are discussed as indicators of general cognitive abilities (e.g., Sorge et al., 2016). They also show a high predictive value for academic success (Trapmann et al., 2007). Note that lower numerical values of German grades indicate better performance. Reading skill is necessary to decode and understand written information. Unsurprisingly, reading skills on word, sentence, and text levels were shown to predict school students' evaluation of online information (Hahnel et al., 2018). Therefore, we expected that the probability to solve an EVON item correctly increases by better (lower) graduation grades (H2.1) and higher reading skill (H2.2).

When investigating web search behavior, prior knowledge usually needs to be taken into account, as it supports web users in interpreting and evaluating semantic and message-related cues and contributes to both the assessment of relevance and credibility (e.g., Hölscher and Strube, 2000; Lucassen et al., 2013). Despite the importance of prior knowledge, however, we did not explicitly expect to find any effect of prior knowledge of the EVON topics on performance. Topic-specific knowledge might facilitate item processing, but due to the item design, it was not necessary to solve the items correctly. Nevertheless, we regarded prior knowledge as an important covariate.

### Method Sample

A convenience sample of 173 students was recruited on the campus of a German university. Because of technical issues (e.g., server connection problems) or commitment (e.g., withdrawal from test), 21 cases were excluded, resulting in a final sample of 152 students (66.2% female) aged from 18 to 37 years (mean = 23.2, *SD* = 3.4). The participants were enrolled in different programs (54.7% bachelor, 14.0% master, 31.3% teacher training and others) from the humanities and social sciences, natural sciences, engineering sciences, economics, and medicine (semesters 1–19, mean = 6.9, *SD* = 3.7). Participants' final school grades ranged from 1 ("very good") to 4 ("sufficient"; mean = 2.3, *SD* = 0.7).

### Procedure

The study was hosted on a server within our institute, on which the data of the participants were also collected and stored. Participants were recruited by posters on the campus, social



media, and direct contact. Most students took an individual test session with a test administrator and received a small gift for participation (e.g., a candy or a ballpoint pen). To increase the reach of our recruitment, we also offered participants to conduct the test independently online; 15 students made use of this offer and received an invitation email with a link. Participation was voluntary and anonymous. After giving their informed consent, the participants were asked to complete a questionnaire assessing demographic variables and their educational background. Afterwards, the participants were asked to work on a speeded test assessing reading skill at sentence level as well as on the tutorial and the eight items of the EVON. Finally, the participants were requested to state how familiar they were with the topics of the EVON items. A test session took about half an hour.

## Measures

### *Evaluating online information*

Students' performance on the EVON items was assessed in terms of dichotomous item scores (0 = incorrect, 1 = correct). The data showed 2.14% missing values in total (including omitted responses and not-reached items). Because of this small amount, missing values were treated as if the respective item had not been administered (Pohl et al., 2014). In addition to the item scores and based on students' log files, it was assessed whether or not the students visited the target website (0 = no visit, 1 = at least one visit) or one or more of the non-target websites (0 = no visit, 1 = at least one visit). Across all cases (152 students  $\times$  8 items), the target was visited in 52.6% and the non-targets in 57.8% of cases.

### *Topic-specific knowledge*

After the EVON assessment, the participants were asked to indicate how familiar they were with the topics in the EVON items. For each topic, they were requested to rate their previous knowledge and experience, responding on a 5-point Likert scale (1 = "don't know what it is," 2 = "heard of it," 3 = "little prior knowledge," 4 = "solid prior knowledge," 5 = "excellent prior knowledge"). Across items, students reported little prior knowledge on average (mean = 3.13,  $SD$  = 0.49, min = 1.88, max = 4.38).

### *Reading skill*

Reading skill was assessed by a sentence verification task that measures the ability to read accurately and quickly (i.e., automatized basic reading processes of lexical access and semantic and syntactic integration of propositions at sentence level; see Johnson et al., 2011; Zimmermann et al., 2014). The test consisted of 58 items that the participants were asked to evaluate as "true" or "false" as quickly and accurately as possible by pressing a respective button ( $\alpha$  = 0.97; e.g., "Sugar is sweet," "A cactus is a little furry animal"; Richter et al., 2012). The test has a total time limit of 80 s. The item contents draw upon common knowledge and are easy to understand (i.e., without uncommon words, complex syntactic structures, or specific knowledge requirements). The stimuli were half true and half false and varied in their semantic abstractness, the number of propositions (one to three propositions), and the sentence length (16–61 characters). The participants processed

between 12 and all 58 sentences (mean = 41.1,  $SD$  = 11.9). The reading score was calculated as the number of correct responses minus the number of incorrect ones (mean = 39.9,  $SD$  = 12.1, min = 8, max = 58).

## Data Analysis

For investigating the EVON assessment, a Rasch model was fitted on students' item scores (Embretson and Reise, 2000). Relative frequencies of correct scores and descriptive point-biserial correlations of the item scores with the sum of scores were inspected. The fit of the Rasch model was examined by inspecting values of item infit and outfit (thresholds between 0.7 and 1.3; Wright and Linacre, 1994) and visual inspection of item characteristic curves and observed non-parametric response functions with respect to non-monotony and unexpected asymptotes. For testing the assumptions of local independence and unidimensionality, we examined Q3 statistics (cutoff: |value| > 0.2; Chen and Thissen, 1997) and conducted modified parallel analyses (Drasgow and Lissak, 1983).

For hypothesis testing, a series of generalized linear mixed models (GLMMs) was carried out (De Boeck et al., 2011). In these models, the probability of successfully solving an EVON item is predicted by fixed and random effects with regard to the hierarchical data structure of item responses nested in persons. Fixed effects are constant across observed units (e.g., students and items), while random effects vary across units. We specified a baseline model including a fixed intercept and random intercepts for students and items.

For examining the effects of item design and navigation behavior (H1.1–H1.4), the baseline model was extended to include fixed effects of the item types (model M1), of target navigation and non-target navigation (M2), and of both the item types and the navigation variables and an additional interaction of item types and non-target navigation (M3). The item type, target and non-target navigation were categorical variables with the reference categories of "type 1 (low attractiveness, congruent)" and "no navigation".

For examining the nomological network of the EVON, the baseline model was extended by students' graduation grades and reading skill (H2.1 and H2.2). Topic-specific prior knowledge was included as a person-by-item covariate. The continuous predictors were  $z$ -standardized before entered to the regression models. Accordingly, the regression coefficients represent the predicted change of the probability of task success when a predictor increases by one standard deviation in a logit metric.

The analyses were carried out in R 3.5.3 (R Core Team, 2019) with the R packages TAM (Robitzsch et al., 2019; for IRT modeling) and lme4 (Bates et al., 2015; for estimating GLMMs). The tests were one-tailed, with a type I error probability of 5%.

## Results and Discussion

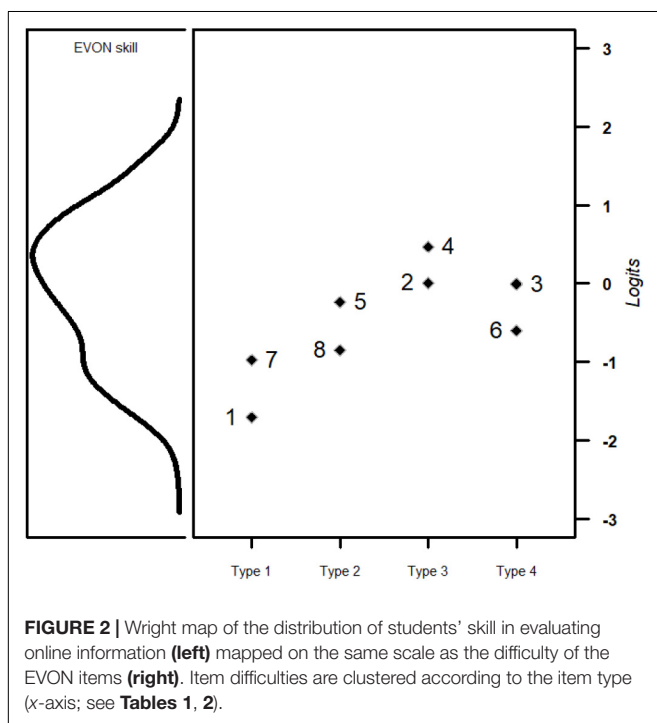
### Scaling

Fitting a Rasch model, the estimated *expected a posteriori* (EAP) scores showed an EAP reliability of 0.62 (range of EAP scores = -1.99 to 1.46, variance = 1.14). Like in the pre-study, the items revealed relatively high rates of correct responses (Table 5). Figure 2 illustrates the estimated ability distribution of students

**TABLE 5 |** Results of item analyses.

Item	% Correct	$r_{pb}$	Missing	Difficulty	Infit	Outfit
1	62.2	0.54	4	−0.60	0.87	0.81
2	50.3	0.48	1	0.00	0.91	0.88
3	68.7	0.32	5	−0.97	1.02	1.05
4	40.9	0.35	3	0.47	1.03	1.04
5	80.5	0.21	3	−1.71	1.07	1.21
6	66.7	0.11	5	−0.85	1.18	1.34
7	55.0	0.44	3	−0.24	0.92	0.90
8	50.0	0.33	2	0.01	1.00	1.00

$r_{pb}$  is the point-biserial correlation of the item with the total sum score (excl. item).



**FIGURE 2 |** Wright map of the distribution of students' skill in evaluating online information (left) mapped on the same scale as the difficulty of the EVON items (right). Item difficulties are clustered according to the item type (x-axis; see Tables 1, 2).

simultaneously with the item difficulty parameters, underlining this lack of difficult items and indicating difficulty differences that seem to correspond with the item types.

The visual inspection of item characteristic curves and the non-parametric response functions showed no severe model violations and even indicated an overfit for some items (i.e., a tendency to underestimate the probability of success for highly skilled students and to overestimate it for low-skilled students; see the **Supplementary Material**). Inspecting the infit and outfit values, item 6 revealed an outfit value beyond the threshold, indicating that it describes students of high or low skill poorly. Its point-biserial correlation with the sum score for all items was also rather low, but positive.

Supporting the assumption of local independence, the mean of all Q3 item pair statistics was slightly negative (−0.08). Only in four cases (14.3%), all involved item 6, a value above the cutoff was shown. The result of the modified parallel analysis was significant, indicating a violation of the unidimensionality

assumption (second eigenvalue observed = 1.01, second eigenvalues averaged across 100 Monte Carlo samples = 0.75,  $p = 0.040$ ). Without item 6, though, the result was opposite (second eigenvalue observed = 0.65, second eigenvalue sampled = 0.63,  $p = 0.401$ ). Although the identified deviations of item 6 are not statistically negligible, they were still relatively small. Therefore and with respect to the construct representation, we decided to keep the item.

### Analysis of the Item Type and Navigation Behavior

The GLMM baseline model showed an intercept of 0.48 ( $SE = 0.24$ ), indicating that students' probability to correctly solve an average EVON item was 61.7% (SD random person intercepts = 1.02; SD random item intercepts = 0.61). As also indicated in **Figure 2** and in line with H1.1, the differentiation according to item types showed that students were most likely to correctly solve type 1 items and least likely to solve the other item types (**Table 6**, model M1). When the logit metric was transformed back into probabilities, the probability of correctly solving an average type 1 item was about 78.9%, which was reduced in items of type 2 (63.1%), type 3 (44.1%), and type 4 (57.3%).

With regard to navigation, the results of model M2 in **Table 6** show that both target and non-target navigation significantly affected task success in an average EVON item, which is in line with H1.2 and H1.3. When students visited the target website, they were very likely to solve an average EVON item correctly ( $b = 2.69$ ). In contrast, keeping the level of target navigation constant, non-target navigation was on average detrimental for students' task success ( $b = -0.58$ ). The tetrachoric correlation between target and non-target navigation was 0.86, indicating a general tendency to navigate or to not inspect the websites at all. The probability of task success without having navigated at all was 36.4% (intercept of M2), which is descriptively larger than the

**TABLE 6 |** Results of the GLMMs examining the effect of item type and navigation on the probability of successfully solving an EVON item.

Predictor	M1	M2	M3
	Est. (SE)	Est. (SE)	Est. (SE)
Intercept	1.32 (0.24)***	−0.55 (0.28)*	0.37 (0.24)
Type 2	−0.78 (0.30)**		−0.45 (0.33)
Type 3	−1.55 (0.30)***		−1.67 (0.36)***
Type 4	−1.02 (0.30)***		−1.62 (0.35)***
Target navigation		2.69 (0.03)***	2.86 (0.25)***
Non-target navigation		−0.58 (0.03)**	−0.20 (0.33)
Non-target navigation × type 2			−1.49 (0.44)***
Non-target navigation × type 3			−0.61 (0.45)
Non-target navigation × type 4			0.21 (0.46)
SD random item intercepts	0.23	0.71	0.18
SD random person intercepts	1.02	0.43	0.44

\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

probability of guessing correctly on average in items with three or five response alternatives (26.7%).

Finally, the last model, M3 in **Table 6**, revealed—as predicted in H1.4—a differential positive effect of non-target navigation in item type 4 ( $b = 0.21$ ), which, however, was not significant. The high standard error suggests that it might be a comparatively small effect that we cannot find as the item types are represented by only two items. Unexpectedly, there was a negative effect of non-target navigation in type 2 items, which means that the negative effect of non-target navigation was especially pronounced in these items.

### Analysis of Relations to Other Variables

Before predicting students' task success, we determined the correlations of the estimated EVON score with students' graduation grades, reading skill, and the sum score of topic-specific knowledge ratings over all items. They showed that the EVON score significantly relates to better (lower) graduation grades [ $r(145) = -0.24$ ,  $p = 0.004$ ] and higher reading skill [ $r(150) = 0.25$ ,  $p = 0.002$ ]. Surprisingly, it was also negatively related to the overall sum score of students' prior knowledge [ $r(147) = -0.24$ ,  $p = 0.003$ ], indicating that students who self-report a broad knowledge about all EVON topics would be less critical of search results.

The GLMM, investigating the effects of these variables on the probability of task success, explained a total of 12.90% of interindividual variation (SD random person intercepts = 0.96; SD random item intercepts = 0.63; intercept:  $b = 0.45$ ,  $SE = 0.25$ ,  $p = 0.068$ ). In line with the hypotheses H2.1 and H2.2, students who were more likely to correctly solve an EVON item also showed significantly better (lower) graduation grades ( $b = -0.24$ ,  $SE = 0.11$ ,  $p = 0.033$ ) and higher reading scores ( $b = 0.30$ ,  $SE = 0.11$ ,  $p = 0.007$ ).

## OVERALL DISCUSSION

With the aim of giving university students a first impression of their performance in evaluating online information, we developed a simulation-based achievement test for a MOOC that addresses the development of information literacy. In the present study, we reported on the development of the resulting instrument, the EVON. The test development and design of the interactive task environment followed a theory-based approach and distinguished four types of situations in which the use of certain heuristics is more or less suitable for making informed judgments about the appropriateness of information in search engine environments. Accordingly, the EVON claims to assess students' skill to evaluate the relevance and credibility of such online information. In order to preliminarily validate this interpretation, we have analyzed several aspects concerning the response process, the internal structure of the instrument, and its relation to third variables.

With regard to the underlying response process, the pre-study showed that students identify and reflect on different aspects of the information provided based on semantic, structural, message-based, and sponsor-based cues. The resulting assessment of

information relevance and credibility formed the basis for their selection of a link and its website. The results of the main study supported this assumption by showing different effects for different situations (item types). If supporting cues were identified early in the evaluation process and used appropriately, students were indeed able to make adequate predictive judgments beyond guessing based on the SERP information alone, as the average probability of task success without visiting a website suggests (36.4%). If the students' decisions were enriched by evaluative assessments of website content, their chance of correctly solving the tasks increased, which is suggested by the positive effect of target navigation. In contrast, as indicated by the negative effect of non-target navigation, if their predictive judgments were inadequate, students may have turned their attention to less appropriate information and remained with it, perhaps because processing effort has already been made. This is also suggested by the unexpected but not implausible observation of the pronounced negative effect of non-target navigation in item type 2. If a website fails to meet web users' expectations built up by predictive judgments, web users will find this source less trustworthy (Metzger and Flanagin, 2013). However, inadequate predictive judgments might be confirmed by the non-target website information in type 2, as it was not incongruent. The findings rather suggest that predictive judgments, once made, may already be quite robust. The positive effect of non-target navigation in item type 4 would have been in line with the empirical observation that web users rate websites as less trustworthy when their initial expectations are disappointed. However, as pointed out, it was not significant, potentially for reasons of the limited item set.

Insights into the internal test structure showed that the EVON sufficiently fitted a Rasch model, with the implication that it assesses a unidimensional construct. Although the results indicated minor difficulties with the psychometric properties of one item, as well as a lack of difficult items, these shortcomings can be overcome by adapting and refining the test on the solid foundation of the present test. To develop more difficult items, it might be worthwhile to create items that keep certain information features constant across links on the SERP (e.g., all website authors show the same level of expertise), thereby reducing the value that students can already gain from predictive judgments. For use in individual diagnostics, the development of further items is generally necessary, as this improves the reliability of the instrument and reduces the imprecision of the measurement. In summary, however, given the small number of items, the present psychometric results can be interpreted as acceptable for a standardized screening tool.

The investigation of evidence referring to the EVON's relations to other construct-related variables showed weak but, as expected, significant relationships to cognitive performance measures such as graduation grades and reading skill. This indicates that the EVON reflects the cognitive performance of a person to some extent and adds to the empirical evidence on the relationship between reading and the evaluation of online information (Hahnel et al., 2018). Future research might extend investigations of the nomological network of

the EVON score, especially with regard to motivational and personality-related aspects beyond cognitive variables. Studies on the use of digital media indicate that different online reading activities or specific motives underlying the use of digital media (e.g., information seeking vs. hedonic or social interaction purposes) are associated with mental processes of recognizing and interpreting web information (e.g., Lee and Wu, 2013; Senkbeil and Ihme, 2017; Senkbeil, 2018). Accordingly, it can be expected that the motivation of web users to process information has an impact on when and how they rely on certain heuristics affecting their credibility assessment of information (Metzger, 2007; Metzger and Flanagin, 2013).

Despite the overall promising findings supporting the test score interpretation of the EVON, the present attempt at validation can only be regarded as preliminary. Accordingly, there are a number of limitations that cannot be resolved by our study, but that also stimulate further research based on our findings. First, further validity evidence needs to be investigated, for example, on students' EVON performance together with other measures of their information literacy or evaluation skill. Demonstrating positive relationships between the EVON and such skills would provide other strong validity arguments. A promising candidate for providing detailed insight into processes assessed by the EVON are, for instance, facets of source evaluation, such as the identification of source features, the evaluation of author credentials and the actual use of source information (e.g., Potocki et al., 2020). Positive relationships should emerge between students' EVON performance and these facets of source evaluation skills, as the EVON claims to assess students' assessment of credibility based on the identification and critical evaluation of source information. In this regard, it is noteworthy that the EVON might not reflect "typical behavior" of students dealing with online information (see Klehe and Anderson, 2007). As students should perform at their best (power test), we explicitly requested them to select a useful and trustworthy link. Without such an instruction, students might have paid less attention to information credibility. Although our results do not speak against interpreting the EVON score in terms of "typical behavior," our validation arguments are weak in this respect. To validate such an interpretation, for example, an experiment would be needed in which one group works on the current EVON test and another group on the EVON test without the instruction amendment on trustworthiness.

Second, we scored the students' answers dichotomously, but this does not mean that a more nuanced coding would not be possible. In particular, we see two directions for improvement, which could also be combined. On the one hand, enriched information could be obtained from alternative response formats, for example, by asking students directly about their perception of why a website appears to be more or less credible or by asking them to rate the relevance and credibility of each link. This option could easily be added to the EVON (e.g., in the form of a separate test part). On the other hand, the stimulus material could be further developed to the extent that it allows partially correct or even multiple correct response options. A partial

credit coding might acknowledge responses that demonstrate moderate assessment skills but still show a lack of thoroughness, rigor, or critical thinking. The challenge would then be to construct such websites that would distinguish between moderate and high levels of competence. Given our psychometric results, which show a lack of difficult items, and recent proposals to consider aspects of critical thinking research (van Zyl et al., 2020), the checklist approach may have limited potential to meet this challenge. However, a more promising attempt might be to develop items that require students to identify and evaluate knowledge claims of websites and evidence that speaks for or against these claims. With respect to both directions, our article shows that EVON provides a solid basis for pursuing such developments.

Third, we only investigated the German version of the EVON. The test is available in five other languages. Although the other language versions do not automatically restrict the applicability of our findings, they should be subject to empirical testing for establishing measurement invariance between the different versions. Measurement invariance ensures that a test measures the same latent construct across several groups. Accordingly, it is an important prerequisite for comparability. Therefore and with respect to restrictions due to our small-scale convenience sample, further research is needed to investigate the generalizability of our findings.

Finally, the EVON was conceptualized as a screening instrument. Accordingly, the ILO MOOC currently uses the EVON as a warm-up test for a lesson on the subject of information evaluation, without further consequences for the course. However, there are possible other uses for which the EVON might be suitable after further adaptation. The EVON might be extended and adapted to serve as a preintervention–postintervention measure to investigate the effectiveness of interventions, such as technology-assisted trainings of evaluating information (for overviews, see Bråten et al., 2018; Braasch and Graesser, 2020). Based on the comprehensive item content and the process data collected during an EVON assessment, it might be even worthwhile to implement a feedback component that provides students not only with their EVON test score or raw item responses, but also information on why a selected alternative might have been suboptimal or how students approached the EVON tasks for purposes of self-reflection. For sure, the usefulness of such feedback for learners would need to be investigated. Yet, if it is found to improve students' evaluation skill, the EVON has the potential to provide elaborate feedback to learners for improving a critical aspect of their information literacy.

In summary, with the EVON, we constructed a complex interactive assessment with an authentic task environment. We observed supporting evidence that its items elicited students to make use of different information features and employed various heuristics for assessing the relevance and credibility of information. Although our findings also uncovered a few weaknesses, and the efforts of validating the interpretation of EVON outcomes still need to be continued, the overall results speak in favor of a successful test construction and provide first



indications that the EVON assesses students' skill in evaluating online information in search engine environments.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

CH: concept, coordination and item development, study design, data collection and preparation, analysis, and writing. BE: item development, data collection and preparation, analysis, and writing. FG: preparation of the manuscript. All authors contributed to the article and approved the submitted version.

## REFERENCES

- Amazeen, M. A., and Muddiman, A. R. (2018). Saving media or trading on trust? *Digit. Journal.* 6, 176–195. doi: 10.1080/21670811.2017.1293488
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Association of College and Research Libraries, (2000). *Information Literacy Competency Standards for Higher Education*. Association of College & Research Libraries. Available online at: <http://www.ala.org/ala/acrl/acrlstandards/informationliteracycompetency.htm> (accessed November 9, 2020).
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bendersky, M., Metzler, D., and Croft, W. B. (2012). "Effective query formulation with multiple information sources," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining – WSDM '12*, New York, NY, 443–452. doi: 10.1145/2124295.2124349
- Bilal, D., and Gwizdka, J. (2018). Children's query types and reformulations in Google search. *Inf. Process. Manage.* 54, 1022–1041. doi: 10.1016/j.ipm.2018.06.008
- Braasch, J. L. G., and Graesser, A. C. (2020). "Avoiding and overcoming misinformation on the internet," in *Critical Thinking in Psychology*, 2nd Edn, eds R. J. Sternberg, and D. F. Halpern (Cambridge: Cambridge University Press), 125–151. doi: 10.1017/9781108684354.007
- Brand-Gruwel, S., Wopereis, I., and Vermetten, Y. (2005). Information problem solving by experts and novices: analysis of a complex cognitive skill. *Comput. Human Behav.* 21, 487–508. doi: 10.1016/j.chb.2004.10.005
- Brand-Gruwel, S., Wopereis, I., and Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Comput. Educ.* 53, 1207–1217. doi: 10.1016/j.compedu.2009.06.004
- Bråten, I., Stadler, M., and Salmerón, L. (2018). "The role of sourcing in discourse comprehension," in *Routledge Handbooks in linguistics. the Routledge Handbook of Discourse Processes*, eds M. F. Schober, D. N. Rapp, and M. A. Britt (Abingdon: Taylor & Francis).

## FUNDING

This research was funded by the Erasmus+ project ILO and the Centre for International Student Assessment (ZIB), Germany.

## ACKNOWLEDGMENTS

We want to thank our student assistants — Carolin Riedel, Christina Röper, Jan Krause, Christina Weers, Mirjana Malešević, Nina Riemenschneider, Isabel Schramm, and Tobias Christoffel — who were invaluable for content development, the implementation and testing of the computer-based instrument as well as several parts in the process of preparing, administering and analyzing the studies reported. We also thank Johannes Naumann and Tobias Richter for providing the items of the sentence verification task for assessing sentence-level reading skill.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.562128/full#supplementary-material>

- Breakstone, J., Smith, M., Wineburg, S., Rapaport, A., Carle, J., Garland, M., et al. (2019). *Students' Civic Online Reasoning: A National Portrait*. Stanford, CA: The Stanford History Education Group.
- Chen, W.-H., and Thissen, D. (1997). local dependence indexes for item Pairs using item response theory. *J. Educ. Behav. Stat.* 22:265. doi: 10.2307/1165285
- Chinn, C. A., and Rinehart, R. W. (2016). Commentary: advances in research on sourcing—source credibility and reliable processes for producing knowledge claims. *Read. Writ.* 29, 1701–1717. doi: 10.1007/s11145-016-9675-3
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *J. Stat. Softw.* 39, 1–28.
- Drasgow, F., and Lissak, R. I. (1983). Modified parallel analysis: a procedure for examining the latent dimensionality of dichotomously scored item responses. *J. Appl. Psychol.* 68, 363–373. doi: 10.1037/0021-9010.68.3.363
- Embretson, S., and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: L. Erlbaum Associates.
- Flanagin, A. J., and Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media Soc.* 9, 319–342. doi: 10.1177/1461444807075015
- Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., and Friedman, T. (2019). *IEA International Computer and Information Literacy Study 2018 Assessment Framework*. Amsterdam: Springer International Publishing. doi: 10.1007/978-3-030-19389-8
- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., and Duckworth, D. (2020). *Preparing for Life in a Digital World: IEA International Computer and Information Literacy Study 2018 International Report*. Amsterdam: Springer International Publishing. doi: 10.1007/978-3-030-38781-5
- Gigerenzer, G., and Gaissmaier, W. (2011). Heuristic decision making. *Ann. Rev. Psychol.* 62, 451–482. doi: 10.1146/annurev-psych-120709-145346
- Goldhammer, F., Hahnel, C., and Kroehne, U. (2020). "Analysing log file data from PIAAC," in *Large-Scale Cognitive Assessment: Analyzing PIAAC Data*, eds D. B. Maehler, and B. Rammstedt (Cham: Springer), 239–269. doi: 10.1007/978-3-030-47515-4\_10

- Hahnel, C., Goldhammer, F., Kröhne, U., and Naumann, J. (2018). The role of reading skills in the evaluation of online information gathered from search engine environments. *Comput. Human Behav.* 78, 223–234. doi: 10.1016/j.chb.2017.10.004
- Hahnel, C., Goldhammer, F., Naumann, J., and Kröhne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text. *Comput. Human Behav.* 55, 486–500. doi: 10.1016/j.chb.2015.09.042
- Hilligoss, B., and Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: construct, heuristics, and interaction in context. *Inf. Process. Manag.* 44, 1467–1484. doi: 10.1016/j.ipm.2007.10.001
- Hollink, V., He, J., and de Vries, A. (2012). “Explaining query modifications,” in *Advances in Information Retrieval. ECIR 2012. Lecture Notes in Computer Science*, Vol. 7224, eds R. Baeza-Yates, A. P. de Vries, H. Zaragoza, B. B. Cambazoglu, V. Murdock, R. Lempel, et al. (Berlin: Springer), 1–12. doi: 10.1007/978-3-642-28997-2\_1
- Hölscher, C., and Strube, G. (2000). Web search behavior of Internet experts and newbies. *Comput. Netw.* 33, 337–346. doi: 10.1016/S1389-1286(00)00031-1
- Huang, J., and Efthimiadis, E. N. (2009). “Analyzing and evaluating query reformulation strategies in web search logs,” in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, New York, NY, 77–86. doi: 10.1145/1645953.1645966
- Johnson, E. S., Pool, J. L., and Carter, D. R. (2011). Validity evidence for the test of silent reading efficiency and comprehension (TOSREC). *Assess. Eff. Interv.* 37, 50–57. doi: 10.1177/1534508411395556
- Kammerer, Y., and Gerjets, P. (2014). The role of search result position and source trustworthiness in the selection of web search results when using a list or a grid interface. *Int. J. Hum. Comput. Interact.* 30, 177–191. doi: 10.1080/10447318.2013.846790
- Keil, F. C., and Kominsky, J. F. (2013). Missing links in middle school: developing use of disciplinary relatedness in evaluating internet search results. *PLoS One* 8:e67777. doi: 10.1371/journal.pone.0067777
- Keßel, Y. (2017). *Development of Interactive Performance Measures for Two Components of ICT Literacy: Successfully Accessing and Evaluating Information*. Ph.D. dissertation, Johann Wolfgang Goethe-Universität, Frankfurt.
- Kiili, C., Laurinen, L., and Marttunen, M. (2008). Students evaluating internet sources: from versatile evaluators to uncritical readers. *J. Educ. Comput. Res.* 39, 75–95. doi: 10.2190/EC.39.1.e
- Klehe, U.-C., and Anderson, N. (2007). Working hard and working smart: motivation and ability during typical and maximum performance. *J. Appl. Psychol.* 92, 978–992. doi: 10.1037/0021-9010.92.4.978
- Lee, Y.-H., and Wu, J.-Y. (2013). The indirect effects of online social entertainment and information seeking activities on reading literacy. *Comput. Educ.* 67, 168–177. doi: 10.1016/j.compedu.2013.03.001
- Leichner, N., Peter, J., Mayer, A.-K., and Krampen, G. (2014). Assessing information literacy programmes using information search tasks. *J. Inf. Lit.* 8, 3–20.
- Leu, D. J., Forzani, E., Rhoads, C., Maykel, C., Kennedy, C., and Timbrell, N. (2014). The new literacies of online research and comprehension: rethinking the reading achievement gap. *Read. Res. Q.* 50, 37–59. doi: 10.1002/rrq.85
- Lucassen, T., Muilwijk, R., Noordzij, M. L., and Schraagen, J. M. (2013). Topic familiarity and information skills in online credibility evaluation. *J. Am. Soc. Inf. Sci. Technol.* 64, 254–264. doi: 10.1002/asi.22743
- Mandl, T., Dreisiebner, S., Libbrecht, P., and Boté, J.-J. (2018). “Challenges for international and multilingual MOOCs: experiences with the information literacy online (ILO) learning service,” in *Proceedings of the International Symposium on the Future of Education in Information Science (FEIS)*, Pisa, Italy (Osijek: University of Osijek).
- Maurer, A., Schloegl, C., and Dreisiebner, S. (2017). Comparing information literacy of student beginners among different branches of study. *Libellarium* 9, 309–319. doi: 10.15291/libellarium.v9i2.280
- McCrudden, M. T., Schraw, G., and Kambe, G. (2005). The effect of relevance instructions on reading time and learning. *J. Educ. Psychol.* 97, 88–102. doi: 10.1037/0022-0663.97.1.88
- McGrew, S., Smith, M., Breakstone, J., Ortega, T., and Wineburg, S. (2019). Improving university students' web savvy: an intervention study. *Br. J. Educ. Psychol.* 89, 485–500. doi: 10.1111/bjep.12279
- Metzger, M. J. (2007). Making sense of credibility on the web: models for evaluating online information and recommendations for future research. *J. Am. Soc. Inf. Sci. Technol.* 58, 2078–2091. doi: 10.1002/asi.20672
- Metzger, M. J., and Flanagin, A. J. (2013). Credibility and trust of information in online environments: the use of cognitive heuristics. *J. Pragmat.* 59, 210–220. doi: 10.1016/j.pragma.2013.07.012
- Metzger, M. J., Flanagin, A. J., and Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *J. Commun.* 60, 413–439. doi: 10.1111/j.1460-2466.2010.01488.x
- OECD (2011). *PISA 2009 Results: Students On Line*. Paris: OECD Publishing.
- Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., and Granka, L. (2007). In Google we trust: users' decisions on rank, position, and relevance. *J. Comput. Mediat. Commun.* 12, 801–823. doi: 10.1111/j.1083-6101.2007.00351.x
- Pardi, G., von Hoyer, J., Holtz, P., and Kammerer, Y. (2020). “The role of cognitive abilities and time spent on texts and videos in a multimodal searching as learning task,” in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, New York, NY, 378–382. doi: 10.1145/3343413.3378001
- Peter, J., Leichner, N., Mayer, A.-K., and Krampen, G. (2017). Making information literacy instruction more efficient by providing individual feedback. *Stud. High. Educ.* 42, 1110–1125. doi: 10.1080/03075079.2015.1079607
- Pirolli, P., and Card, S. (1999). Information foraging. *Psychol. Rev.* 106, 643–675. doi: 10.1037/0033-295X.106.4.643
- Pohl, S., Gräfe, L., and Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: evaluating approaches accounting for missing responses in item response theory models. *Educ. Psychol. Meas.* 74, 423–452. doi: 10.1177/0013164413504926
- Potocki, A., de Pereyra, G., Ros, C., Macedo-Rouet, M., Stadler, M., Salmerón, L., et al. (2020). The development of source evaluation skills during adolescence: exploring different levels of source processing and their relationships (El desarrollo de las habilidades de evaluación de las fuentes durante la adolescencia: una exploración de los distintos niveles de procesamiento de las fuentes y sus relaciones). *J. Study Educ. Dev.* 43, 19–59. doi: 10.1080/02103702.2019.1690848
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Available online at: <https://www.R-project.org/> (accessed November 9, 2020).
- Richter, T., Isberner, M.-B., Naumann, J., and Kutzner, Y. (2012). Prozessbezogene Diagnostik von Lesefähigkeiten bei Grundschulkindern. *Z. Pädagog. Psychol.* 26, 313–331. doi: 10.1024/1010-0652/a000079
- Rieh, S. Y., and Danielson, D. R. (2007). Credibility: a multidisciplinary framework. *Ann. Rev. Inf. Sci. Technol.* 4, 307–364. doi: 10.1002/aris.2007.1440410114
- Robitzsch, A., Kiefer, T., and Wu, M. (2019). *TAM: Test Analysis Modules*. Available online at: <https://CRAN.R-project.org/package=TAM> (accessed November 9, 2020).
- Rosman, T., Mayer, A.-K., and Krampen, G. (2016). Measuring psychology students' information-seeking skills in a situational judgment test format: construction and validation of the PIKE-P test. *Eur. J. Psychol. Assess.* 32, 220–229. doi: 10.1027/1015-5759/a000239
- Rouet, J.-F. (2006). *The Skills of Document Use: From Text Comprehension to Web-based Learning*. Mahwah, NJ: Erlbaum.
- Rouet, J.-F., Ros, C., Goumi, A., Macedo-Rouet, M., and Dinot, J. (2011). The influence of surface and deep cues on primary and secondary school students' assessment of relevance in Web menus. *Learn. Instr.* 21, 205–219. doi: 10.1016/j.learninstruc.2010.02.007
- Salmerón, L., Naumann, J., García, V., and Fajardo, I. (2017). Scanning and deep processing of information in hypertext: an eye tracking and cued retrospective think-aloud study. *J. Comput. Assist. Learn.* 33, 222–233. doi: 10.1111/jcal.12152
- Senkbeil, M. (2018). Development and validation of the ICT motivation scale for young adolescents. Results of the international school assessment study ICILS 2013 in Germany. *Learn. Individ. Differ.* 67, 167–176. doi: 10.1016/j.lindif.2018.08.007
- Senkbeil, M., and Ihme, J. M. (2017). Motivational factors predicting ICT literacy: first evidence on the structure of an ICT motivation inventory. *Comput. Educ.* 108, 145–158. doi: 10.1016/j.compedu.2017.02.003
- Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I., and Scherer, R. (2016). Taking a future perspective by learning from the past – a systematic review

- of assessment instruments that aim to measure primary and secondary school students' ICT literacy. *Educ. Res. Rev.* 19, 58–84. doi: 10.1016/j.edurev.2016.05.002
- Sorge, S., Petersen, S., and Neumann, K. (2016). Die bedeutung der studierfähigkeit für den studienfolg im 1. Semester in physik [The importance of the ability to study for the success in the 1st semester in physics]. *Z. D. Naturwiss.* 22, 165–180. doi: 10.1007/s40573-016-0048-x
- Stadtler, M., and Bromme, R. (2014). "The content–source integration model: a taxonomic description of how readers comprehend conflicting scientific information," in *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences*, eds D. N. Rapp, and J. Braasch (Cambridge, MA: MIT Press), 379–402.
- Steinrück, J., Veldkamp, B. P., and de Jong, T. (2020). Information literacy skills assessment in digital crisis management training for the safety domain: developing an unobtrusive method. *Front. Educ.* 5:140. doi: 10.3389/educ.2020.00140
- Sundar, S. S., Knobloch-Westerwick, S., and Hastall, M. R. (2007). News cues: information scent and cognitive heuristics. *J. Am. Soc. Inf. Sci. Technol.* 58, 366–378. doi: 10.1002/asi.20511
- Trapmann, S., Hell, B., Weigand, S., and Schuler, H. (2007). Die validität von schulnoten zur vorhersage des studienersfolgs—eine metaanalyse [the validity of school grades for academic achievement – a meta-analysis]. *Z. Pädagog. Psychol.* 21, 11–27. doi: 10.1024/1010-0652.21.1.11
- van den Broek, P., Bohn-Gettler, C., Kendeou, P., Carlson, S., and White, M. J. (2011). "When a reader meets a text: the role of standards of coherence in reading comprehension," in *Text Relevance and Learning from Text*, eds M. McCrudden, J. Magliano, and G. Schraw (Charlotte, NC: Information Age Publishing), 123–140.
- van Someren, M. W., Barnard, Y. F., and Sandberg, J. A. C. (1994). *The Think Aloud Method—A Practical Guide to Modelling Cognitive Processes*, Vol. 31. Cambridge, MA: Academic Press.
- van Zyl, A., Turpin, M., and Matthee, M. (2020). "How can critical thinking be used to assess the credibility of online information?," in *Responsible Design, Implementation and Use of Information and Communication Technology I3E 2020*. Lecture Notes in Computer Science, Vol. 12067, eds M. Hattingh, M. Matthee, H. Smuts, I. Pappas, Y. K. Dwivedi, and M. Mäntymäki (Cham: Springer International Publishing), 199–210. doi: 10.1007/978-3-030-45002-1\_17
- Walraven, A., Brand-Gruwel, S., and Boshuizen, H. P. A. (2008). Information-problem solving: a review of problems students encounter and instructional solutions. *Comput. Human Behav.* 24, 623–648. doi: 10.1016/j.chb.2007.01.030
- Walsh, A. (2009). Information literacy assessment: where do we start? *J. Librariansh. Inf. Sci.* 41, 19–28. doi: 10.1177/0961000608099896
- Wineburg, S., Breakstone, J., McGrew, S., and Ortega, T. (2018). "Why google can't save us," in *Positive Learning in the Age of Information*, eds O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Wiesbaden: Springer), 221–228. doi: 10.1007/978-3-658-19567-0\_13
- Wirth, W., Böcking, T., Karnowski, V., and von Pape, T. (2007). Heuristic and systematic use of search engines. *J. Comput. Mediat. Commun.* 12, 778–800. doi: 10.1111/j.1083-6101.2007.00350.x
- Wright, B. D., and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Meas. Trans.* 8:370.
- Zimmermann, S., Artelt, C., and Weinert, S. (2014). *The Assessment of Reading Speed in Adults and First-Year Students*. Leibniz Institute for Educational Trajectories (LIfBi). Available online at: [https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC5/3-0-0/com\\_rs\\_SC5\\_SC6.pdf](https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC5/3-0-0/com_rs_SC5_SC6.pdf) (accessed November 9, 2020).

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Hahnel, Eichmann and Goldhammer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Computational Linguistic Assessment of Textbooks and Online Texts by Means of Threshold Concepts in Economics

Andy Lücking<sup>1,2\*</sup>, Sebastian Brückner<sup>3</sup>, Giuseppe Abrami<sup>1</sup>, Tolga Uslu<sup>1</sup> and Alexander Mehler<sup>1</sup>

<sup>1</sup> Text Technology Lab, Faculty of Computer Science and Mathematics, Institute of Computer Science, Goethe University Frankfurt, Frankfurt, Germany, <sup>2</sup> Laboratoire de Linguistique Formelle, Laboratory of Excellence "Empirical Foundations of Linguistics", Université de Paris, Paris, France, <sup>3</sup> Department of Business and Economics Education, Johannes Gutenberg University Mainz, Mainz, Germany

## OPEN ACCESS

### Edited by:

Patricia A. Alexander,  
University of Maryland, United States

### Reviewed by:

Teresa Pozo-Rico,  
University of Alicante, Spain  
Anisha Singh,  
University of Maryland, College Park,  
United States

### \*Correspondence:

Andy Lücking  
luecking@em.uni-frankfurt.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Education

Received: 30 June 2020

Accepted: 11 December 2020

Published: 19 January 2021

### Citation:

Lücking A, Brückner S, Abrami G,  
Uslu T and Mehler A (2021)  
Computational Linguistic Assessment  
of Textbooks and Online Texts by  
Means of Threshold Concepts in  
Economics. *Front. Educ.* 5:578475.  
doi: 10.3389/feduc.2020.578475

The ongoing digitalization of educational resources and the use of the internet lead to a steady increase of potentially available learning media. However, many of the media which are used for educational purposes have not been designed specifically for teaching and learning. Usually, linguistic criteria of readability and comprehensibility as well as content-related criteria are used independently to assess and compare the quality of educational media. This also holds true for educational media used in economics. This article aims to improve the analysis of textual learning media used in economic education by drawing on threshold concepts. Threshold concepts are key terms in knowledge acquisition within a domain. From a linguistic perspective, however, threshold concepts are instances of specialized vocabularies, exhibiting particular linguistic features. In three kinds of (German) resources, namely in textbooks, in newspapers, and on Wikipedia, we investigate the distributive profiles of 63 threshold concepts identified in economics education (which have been collected from threshold concept research). We looked at the threshold concepts' frequency distribution, their compound distribution, and their network structure within the three kinds of resources. The two main findings of our analysis show that firstly, the three kinds of resources can indeed be distinguished in terms of their threshold concepts' profiles. Secondly, Wikipedia definitely shows stronger associative connections between economic threshold concepts than the other sources. We discuss the findings in relation to adequate media use for teaching and learning—not only in economic education.

**Keywords:** threshold concepts, corpus study, wikipedia, newspaper, specialized vocabulary, network model, economics, textbooks

## 1. INTRODUCTION

In recent years, research on how to facilitate teaching, curriculum development, and the diagnostic of competences acquired during higher education studies has intensified significantly in many disciplines, not only in Germany but also worldwide (Nicola-Richmond et al., 2018; Zlatkin-Troitschanskaia et al., 2018). As shown in various instructional models (e.g., the *offer-use model*



by Helmke, 2009), the quality of learning media is of crucial importance for the learning success of students. A central challenge for higher education lecturers from all disciplines is to select high-quality learning media for their teaching, for which learning media research provides corresponding findings. The investigation of the quality of learning media can be investigated on the basis of a variety of criteria. Expert ratings are often used for the evaluation of learning media using criteria, such as *Accuracy*, *Clarity*, *Comprehensiveness*, *Consistency*, *Grammar*, *Readability*, *Modularity*, and *Cultural Relevance* (Fischer et al., 2017). An important quality criterion and benchmark besides didactic, pictorial, and further media-structural characteristics is that the learning media address central concepts of a subject area and their interconnectedness, because the extent to which a digital medium supports learning success depends largely on the quality of the content presented in it (Devetak and Vogrinc, 2013).

In economic education in higher education in particular, a large amount of different media sources are frequently used because economic phenomena are the subject of everyday encounters and historical events (Simkins, 1999; Davies and Mangan, 2007; Meier, 2008; Hoyt and McGoldrick, 2012; Schuhen and Kunde, 2016). Traditionally, of course, the major learning resources are textbooks (Jadin and Zöserl, 2009; Maurer et al., 2019; Dalimunte and Pramoolsook, 2020), whose didactic purposes include, among others, the introduction of technical vocabulary. Currently, many textbooks are available to students as *Open Educational Resources* (OER), but the predominant use of textbooks as a learning resource has emerged over the years. This development has been attributed to the professional quality and the connection to lectures and courses (Devetak and Vogrinc, 2013; Fischer et al., 2017; Maurer et al., 2019; Dalimunte and Pramoolsook, 2020). Furthermore, in economics education, textbooks are central for teaching and learning in formal teaching-learning environments (Leet and Lopus, 2003; Richardson, 2004; Tinkler and Woods, 2013). In connection with the increasing digitization of university teaching, digital learning platforms, forums, and online encyclopedias are increasingly used by students as a complementary source of learning alongside textbooks because of their easy and often free access (Brooks, 2016; Kilgour et al., 2019; Maurer et al., 2019). According to several studies (Knight and Pryke, 2012; Steffens et al., 2017; Jihinke and Di Lauro, 2020), Wikipedia is one the topmost used internet services. Online encyclopedias, such as Wikipedia, are often used to quickly access summaries and definitions or as a first encounter with subject-specific concepts (Jadin and Zöserl, 2009; Lim, 2009; Knight and Pryke, 2012; Maurer et al., 2019; Jihinke and Di Lauro, 2020). Sources for learning-related purposes mainly are used by learners who explore the core contents and concepts of their respective fields (Knight and Pryke, 2012; Steffens et al., 2017; Maurer et al., 2019). In college-level economic education, an increase of the usage of digital learning tools in formal education has been acknowledged for many years (Simkins, 1999). Besides textbooks (Hu and Gao, 2019), Wikipedia is considered for initial orientation and for dealing with economic content, too (Meier, 2008; Haab et al., 2012; Freire and Li, 2016).

However, economic education is a differentiated field of study, the content may gradually change, for example, in the light of current news or changes in legislation. For current events, newspapers offer a way to stay up-to-date on the economic situation in businesses and countries (Croushore, 2012). In addition, many newspapers are easier to understand, especially for novice learners (Dalimunte and Pramoolsook, 2020), and are therefore sometimes read as frequently as online economic blogs (Haab et al., 2012). For a long time, newspapers have been one of the main resources used in economics education. Especially lecturers in introductory courses often use the variable prior knowledge of students regarding current issues in business and economics to encourage more active engagement with the subject. As current research suggests, students frequently come into contact with economic content in their everyday life by reading newspapers (Hoyt and McGoldrick, 2012). Especially in Germany, unlike in other industrial nations, business or economics has not yet been established as a school subject in Germany (Schuhen and Kunde, 2016). The majority of first-year students at German universities usually have previous knowledge that was acquired in an informal<sup>1</sup> context (cf. Schumann et al., 2010). The first-year students' knowledge of economics often comes from various media that are not directly related to a learning-intended purpose (e.g., online magazines, news magazines, videos) (Maurer et al., 2019), social interactions on financial topics (e.g., as a consumer in a supermarket or buying a mobile phone) (Davies and Mangan, 2007; Schuhen and Kunde, 2016), or other behavior of economic relevance (e.g., retirement planning). Consequently, students may also use textbooks, Wikipedia and newspapers as central learning media in economic education. In order to ensure that learning media with the highest possible quality of content are used in a way that is appropriate for the target group, lecturers are therefore inevitably faced with the question of which media to select for a given topic or concept to be taught. However, a comparative analysis of digital learning media in economics education with regard to concrete professional concepts is still pending. Since the core of these teaching-learning media in economics is always central focal content (Leet and Lopus, 2003), we will compare these media using domain specific economic concepts. In economics education the so-called *thresholds concepts* are a current and frequently discussed approach that seeks to identify the most important concepts for learning economics (Meyer and Land, 2006; Davies and Mangan, 2007). Therefore, the three media types can be compared using the linguistic features of these concepts. In order to provide teachers with a certain information basis for the selection of media based on the comparison of threshold concepts in learning media, we will address the following research question in this paper: To what extent do textbooks, Wikipedia, and newspapers used (by students) for learning about economic concepts differ in terms of

<sup>1</sup>Informal learning can—taking into account the variety of definitions—essentially be understood as learning *en passant*; i.e., learning that takes place implicitly when carrying out other activities (e.g., learn about costs when reading a newspaper article), is usually not consciously controlled by the learner (Neuweg, 2000; Hofhues, 2016).

the structure and linguistic characteristics of threshold concepts that are important for learning?

In section 2, threshold concepts are introduced in more detail and discussed in relation to domain-specificity, conceptual change, and specialized vocabularies. A theoretical linguistic perspective is outlined in section 3. In section 3, a theoretical linguistic perspective is outlined that shows how learning how learning can be construed in terms of a dynamic update semantics and how linked *mental files* represent relations between threshold concept terms in texts. Some terminological and conceptual distinctions that arise in this context are drawn in section 4. Section 5 then introduces a computational linguistic approach for deriving networks of linked threshold concepts on a large scale. The method is applied to three types of (online) resources, namely newspaper articles, textbooks, and Wikipedia article. The results are finally discussed in section 6.

## 2. THRESHOLD CONCEPTS APPROACH AND CONCEPTUAL CHANGE

An approach using threshold concepts rather than simplified content categories (Kricks et al., 2013) has been introduced into didactic discussions that focuses on highest potentials for developing a professional disciplinary understanding for both novice and experienced learners (Meyer and Land, 2013). The authors describe threshold concepts as “akin to a portal, opening up a new and previously inaccessible way of thinking about something” (Meyer and Land, 2006, p. 3). Due to their special character within a discipline, they thus represent a threshold that needs to be crossed and that fundamentally changes the learner’s understanding of the discipline. Concepts can thus describe principles and rules, objects, theories, modeling methods on an abstract level, which contribute to the development of a comprehensive understanding of the learner within an individual discipline (Sender, 2017).

Often the threshold concepts approach refers to learning in the sense of conceptual change (Davies and Mangan, 2007): it is assumed that knowledge gain is not just an accumulative process of mere addition of new knowledge, but that the learner’s existing knowledge structures are (possibly fundamentally) transformed (Davies and Mangan, 2007). If the learner develops a new understanding of a concept, the conceptual change can be very sudden and unexpected, namely when the learner experiences the new concept as expanding his or her previous field of imagination. This initial change of concepts can be demonstrated didactically by a change of perspective for the learner, e.g., by looking at a purchase decision from the roles of buyer and supplier and thus better understanding the formation of prices (Sender, 2017, p. 56). This illustrates a short-term event in the learning process. If the learner is able to adapt and transfer his new concept to other contexts and examples, or if he experiences the limits of his newly developed conceptions, the knowledge structures are gradually changed and consolidated, so that a threshold concept also has a long-term effect (Sender, 2017). Thus, the more short- and long-term support the understanding of a concept has, the more irreversible the understanding is

(Cousin, 2008). Accordingly, *irreversibility* is one characteristic of threshold concepts, alongside *transformativity*, *integrativity*, *limitedness*, and *difficulty* (Meyer and Land, 2005, 2006). The constant transformation and application of the acquired knowledge to a variety of known phenomena promotes the intertwining of knowledge. Integrativity leads to the fact that different knowledge structures, which previously could not be put into context for the learner, are increasingly brought into a semantic relation. Threshold concepts are also limited, since the new conceptual spaces created by linking content-related ideas simultaneously create new boundaries that distinguish the discipline from other academic disciplines (Meyer and Land, 2005).

### 2.1. Threshold Concepts in Business and Economics

A large number of studies focus on the identification of threshold concepts (Sender, 2017; Brückner and Zlatkin-Troitschanskaia, 2018; Hatt, 2018; Lamb et al., 2019; van Mourik and Wilkin, 2019; Ivan Montiel and Antolin-Lopez, 2020). *Opportunity costs* was the initial threshold concept that has been identified for the discipline of economics (Meyer and Shanahan, 2003) and has since been taken up in several studies (Shanahan et al., 2006; Davies and Mangan, 2007). The critical discourse and empirical examination as to which concepts can be considered threshold concepts and which are important for the curriculum but not mandatory is ongoing and has since been discussed in a number of papers (Davies and Mangan, 2007; Lucas and Mladenovic, 2009; Ivan Montiel and Antolin-Lopez, 2020). Over the years, in addition to opportunity costs, a large number of concepts have been proposed and empirically tested in economics, e.g., on depreciation (Lucas and Mladenovic, 2009), elasticity (Reimann and Jackson, 2006), information asymmetry (Hoadley et al., 2015), and many more, on the basis of multiple research methods, e.g., using interviews with teachers or learners, videographies, curriculum analyses or standardized tests. For example, in a Delphi study, Hatt (2018) use interviews with entrepreneurs to investigate which concepts they regard as threshold concepts. Ivan Montiel and Antolin-Lopez (2020) conduct a literature analysis and develops 33 threshold concepts for corporate sustainable management. Davies and Mangan (2007) identify threshold concepts in economics on the basis of literature analysis and Hoadley et al. (2015) use expert interviews to find out whether or not a pre-selected sample of threshold concepts actually consists of threshold concepts. Some studies also examine facets of conceptual change on this basis. For example, Sender (2017) analyzes how affective and cognitive states develop in liminal phases of understanding when confronted with threshold concepts in economics courses. Brückner and Zlatkin-Troitschanskaia (2018) examine how confident students are in their ability to assess their solution behavior in tests when the complexity of threshold concepts increases. A number of studies also describe that the relationships established between the threshold concepts by the learner are of great importance for generating a deeper understanding (Davies

and Mangan, 2007; Vidal et al., 2015; Ivan Montiel and Antolin-Lopez, 2020). A central area of research also lies in various types of conceptual change. Davies and Mangan (2007) distinguish three forms, i.e., the *basic*, *discipline*, and *procedural* form of conceptual change. This three-part categorization has been taken up frequently, especially in recent years, by integrating further concepts from the economic sciences and further developing existing concept attributions (Lucas and Mladenovic, 2009; Kricks et al., 2013; Hoadley et al., 2015; Sender, 2017; Brückner and Zlatkin-Troitschanskaia, 2018; van Mourik and Wilkin, 2019). A basic conceptual change is defined as “Understanding of everyday experience transformed through integration of personal experience with ideas from discipline” (Davies and Mangan, 2007, p. 715). This is a conceptual change, which is fundamental and which a learner experiences as soon as he develops a first disciplinary understanding, e.g., of the concept of *cost*. Concepts documented along the basic threshold are accessible to most learners, as they are confronted with their everyday life (e.g., in their behavior as consumers) (Davies and Mangan, 2007). At the level of the disciplinary threshold, the learner succeeds in developing and linking conceptual understandings based on a theoretically elaborated perspective, which is hardly accessible from everyday life. This concerns concepts that are mainly accessible within the economic sciences (e.g., the concept of opportunity costs, hedging; depreciation; see Davies and Mangan, 2007; Lucas and Mladenovic, 2009; Hoadley et al., 2015). Some of the concepts require that a first encounter with a subject has already taken place and that the learner has a basic level of knowledge (Davies and Mangan, 2007), for example, the concept of costs should be understood before the opportunity cost principle is understood. The procedural threshold comprises concepts that are deeply integrated in the subject structures and require an understanding of modeling in economics. These are abstract modeling methods, procedures or argumentations that are used to analyze economic phenomena, but also to further develop economic theories (e.g., comparative statics, intertemporality; Davies and Mangan, 2007; Sender, 2017; Brückner and Zlatkin-Troitschanskaia, 2018). However, it can be seen that the studies mainly focus on learning processes and learning success as well as on personal prerequisites. According to the *offer-use-model* (Helmke and Schrader, 2008), it is important to investigate whether learning media also offer learners the possibility to go through this conceptual change and to connect concepts with each other. It is therefore important to investigate to what extent the threshold concepts are represented in the learning media used by the learners. The frequency of occurrence in learning media and the cross-linking of threshold concepts (Davies and Mangan, 2007) are thus a central aspect of the investigation of the potential of learning media.

Due to their fundamental character for the genesis of a disciplinary economic understanding, threshold concepts are often a central content in textbooks and are sometimes referred to as “building blocks” (Davies and Mangan, 2007, p. 724). A number of studies also start in their investigations in textbooks, often analyzing the variable views and differences in their understanding by learners (Lucas and Mladenovic, 2009). Less frequently, linguistic characteristics and representations of

threshold concepts are considered, although these have been shown to be of great importance for learning and understanding processes (Mayer, 2005). For example, Shanahan et al. (2006, p. 105) explicate: “Many first-year economics students report, that they find ‘economic jargon’ the most difficult barrier to their understanding. For economists ‘learning the language’ is one of the necessary elements to ‘think like an economist’.”

## 2.2. Threshold Concepts and Specialized Vocabularies

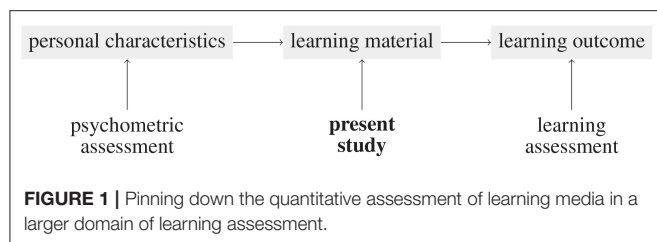
Since threshold concepts in business and economics are addressed by words it comes as no surprise that there is a connection to investigations from linguistics, in particular in studies of a certain kind of a manner of speaking (a socio-, functo-, or technolect) known as *specialized languages*, or the “language of science.” A specialized language is more than just a specialized vocabulary since it involves grammatical aspects as well (Crystal, 1997, p. 384)—however, the vocabulary is the most salient part of a scientific sociolect and threshold concepts are no exception to this impression. Accordingly, there is a branch of linguistics specialized on specialized languages (see Roelcke, 2010 for an introduction), in particular in lexicography (Hoffmann et al., 1998). Interestingly, lexicographic work on specialized vocabularies distinguishes three classes of scientific expressions: “technical terms, semi-technical terms, and general vocabulary frequently used in a specialized domain” (Motos, 2011, p. 9, quoted from Nagy, 2014, p. 267). Obviously, there is a coincidence with the 3-fold distinction of threshold concepts into basic, discipline, and procedural, which could be worth pursuing. The present study, however, investigates textual features with regard to threshold concepts, based on linguistic considerations concerning specialized languages.

## 3. THEORETICAL LINGUISTICS PERSPECTIVE: THRESHOLD CONCEPTS IN DISCOURSE REPRESENTATION STRUCTURES

Three general factors from the complex network of factors that influence learning introduced in sections 1 and 2 can be extracted: personal characteristics, learning material, and learning outcome (cf. **Figure 1**). There are statistical assessments for both the personal characteristics and the learning outcome (Lodico et al., 2006). However, assessments regarding the learning material (e.g., quantification of texts) are rare. The aim of the present study is to develop a methodological proposal in this respect. Threshold concepts seem to be particularly suited for obtaining a reference frame that is needed for frequentist analyses and comparisons. Threshold concepts are especially suited for this task since the corresponding word forms are easily identifiable in texts (see section 4 on words and concepts) and they are related to conceptual change (cf. section 2).

Let us illustrate this with a very simple example, namely *Kosten* “cost.” The everyday sense of *cost* is derived from buying events. This is encoded in natural language grammar where the *lexical frame* (Fillmore et al., 2012) for the noun *cost* has four core

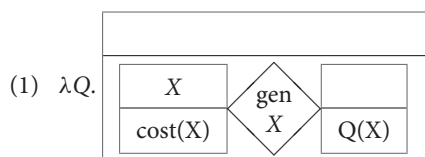




elements: *asset*, *goods*, *intended\_event*, and *payer*<sup>2</sup>. Accordingly, we can think of the psycholinguistic, everyday concept of *cost* as a *sensorimotor simulator* of such buying events (Barsalou, 1999). Now this does not square easily with the economic sense of *cost*. The associated German Wikipedia page, for instance, starts as follows [translated by AL]<sup>3</sup>:

Costs are the negative consequences of the use of production factors with an impact on profits. The exact definitions differ depending on the subject area. In the economic sense of cost accounting, costs are usually understood to be the consumption of production factors valued in monetary units.

The economic definition of *cost* is at most indirectly related to buying events (each “production factor” has eventual to be paid in the everyday sense, though, hence providing evidence that *cost* is to be classified as a *basic* threshold concept). The subject noun *costs* which starts the Wikipedia article compiles a new *mental file* (Heim, 2002; Murez and Recanati, 2016) or *discourse referent* (Karttunen, 1969; Kamp and Reyle, 1993) which becomes the information structural *topic* (Cohen and Erteschik-Shir, 2002). Since *costs* is a bare plural noun, it introduces a plurality (represented by capital *X*) and receives a generic interpretation (Link, 1983; Krifka, 2003). Using the graphical discourse representation format of Asher (1993) and Kamp and Reyle (1993), the semantic representation of *costs* at this point is as follows:

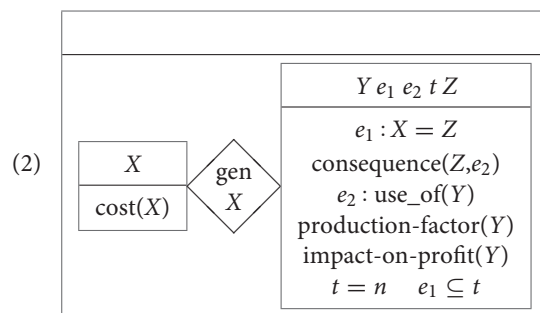


The file or discourse universe is in the following populated with the predication, regimented by a syntax-driven construction algorithm (cf. subsection 4.1.3). Plural *be*, “are,” triggers the parsing hypothesis (Demberg et al., 2013) that *are* is a copula

<sup>2</sup><https://framenet2.icsi.berkeley.edu/fnReports/data/lu/lu9191.xml?mode=lexentry> (accessed October 29, 2020).

<sup>3</sup><https://de.wikipedia.org/wiki/Kosten> (accessed October 29, 2020). “Kosten sind die negativen Konsequenzen einer erfolgswirksamen Nutzung von Produktionsfaktoren. Die genauen Definitionen unterscheiden sich je nach Fachgebiet. Im betriebswirtschaftlichen Sinn der Kostenrechnung wird darunter meist der in Geldeinheiten bewertete Verbrauch an Produktionsfaktoren verstanden.”

which initiates a predication on its subject (this parsing prediction turns out to be correct). The copula is interpreted in terms of the identity function (Russell, 1919) and introduces the corresponding condition (the predicate variable *Q* provides the interface for composition). Having processed the remainder of the sentence in this fashion, the semantic representation given in (2) is obtained (note that the deverbal noun *use* receives an eventive interpretation, as does the identity relation; processing present tense introduces the condition that the main event  $e_1$  holds at the indexical time point  $n$ , “now”).



Part of the predicative content in (2) is the information that costs follow from production factors. Since this sounds different from what the learner knows from his or her everyday language competence, the new mental file *costs* is not merged with the eponymous pre-theoretic one (though both remain related at least due to phonological identity). Furthermore, if the learner already has a (rich) mental file for the noun compound *production factors* (*Y*), integration of both files will happen at this point. This integration obviously depends on the learner’s prior knowledge<sup>4</sup>.

The Wikipedia article continues with mentioning opportunity costs alongside costs. This mention again compiles a mental file. Since *costs* and *opportunity costs* share a great deal of surface form (namely the head noun *costs*) they will be connected, but their precise connection at this point is still unspecified (given that the learner has no prior knowledge in this regard). Thus, already after a few sentences, two threshold concepts will be initialized and connected—in terms of operations on a knowledge base: the knowledge base is expanded (by introducing mental files) and denser connected (by a *Consequence* relation) (Chi and Ohlsson, 2005, p. 376 f.).

Textbooks often refrain from an initial definition of costs in favor of a distinction of different *types* of costs (namely elaborating on the production factors mentioned in the above-given quotation)<sup>5</sup>. Instead they list examples for costs, such as delivery costs, holding costs, production costs, retooling costs, etc. Accordingly, the mental file is populated with sub-types of costs. These sub-types are

<sup>4</sup>Additionally the learner may associate, for instance, personal experiences with any of the mental files, but this is not part of text meaning, see subsection 4.1.2.

<sup>5</sup>This enumerative way is taken, for instance, in Weber et al. (2014), Mumm (2015), and Blum (2017).



connected to the file's header by means of *Elaboration* relations, which involve a *part-of* condition by default (Asher and Lascarides, 2003, p. 160). Hence, the knowledge base is expanded and this expansion receives a more fine-grained representation (Chi and Ohlsson, 2005, p. 376, p. 382).

Since “opportunity costs” follows the same compound structure as the other just mentioned sub-types of costs, the initial hypothesis is to add it *via Elaboration* to *costs*'s mental file, too. However, in this case linguistic structure is deceptive: while all sub-type of costs are related to the everyday *buying* concept, opportunity costs are not. Hence, they eventually have to be compiled in a file of their own.

This sketch of a linguistic analysis shows that different texts present what can be assumed to be the same topic in different ways. These different ways can be made precise in terms of a dynamic update semantics (a closely related, cognitive model of text meaning has been developed by Asher, 1993), which then can be used as a model of learning (Lücking, 2019)<sup>6</sup>. Semantic updates are equivalent to changes in a knowledge base, which characterizes (declarative) learning. We have seen three types of changes or updates. In general the following types of changes can be distinguished (see Chi and Ohlsson, 2005 for details): *larger size*, *denser connectedness*, *increased consistency*, *finer grain of representation*, *greater complexity*, *higher level of abstraction*, and *shifted vantage point*. Acquiring threshold concepts from the *discipline* category essentially involves *denser connectedness* changes, where acquiring those of category *procedural* rest on a *higher level of abstraction* or even a *shifted vantage point*. Now there is no large-scale implementation of construction algorithms leading to semantic representations as studied in theoretical linguistics, nor is there a construction algorithm for further operations on mental files. For that reason, current computational linguistics employs shallow processing methods that aim at approximating such representations (cf. subsection 4.1.3). An approach to applying computational linguistics methods to texts in order to derive networks of threshold concept expressions is developed in section 5.

Linguistic semantics (and pragmatics, for that matter) studies the *normative* dimension of meaning: the interpretation of words and sentences of a language that any speaker should get if he or she is a speaker of that language. This does not guarantee that the speaker *actually* or *de facto* gets the normative interpretation; nor does it follow that the normative interpretation exhausts the speaker's understanding. So let us first elaborate on this and related issues to avoid any possibility of confusion.

<sup>6</sup>It should be emphasized that we just informally sketched how to derive conceptual discourse *representations*, which are just the first step of semantic interpretation. The second step consists in interpreting these representations in models. The propositional meaning of a discourse representation is the set of input-output assignments that provide a successful embedding in a model: its *context change potential*. Construing a learner as a model (knowledge base), it is suggestive to define conceptual change in terms of context change potential change.

## 4. THRESHOLD CONCEPTS: MENTAL, REFERENTIAL, AND DIFFERENTIAL MEANING

As outlined in section 2, threshold concepts from the disciplines of business and economics can be approached from various perspectives: they are defined as specialized terms, they are building blocks of students' learning development and they are expressed by words. Each of these perspectives corresponds to different scientific (sub-)disciplines (namely business and economics, learning psychology and education, and linguistics and lexicography, in that order; for a related view see Lenci, 2008). But how are they related?

### 4.1. Different Concepts of “Threshold Concepts”

According to a widely accepted sign-based conception, a word is a couple of a *form* (hereafter also called *expression*) and a *meaning*. The form side can be a token, an inflected morpho-syntactic expression of a type (lemma), or it can be the lemma itself. With respect to the meaning side, any scholar dealing with meaning faces a dilemma: she has to use meaningful words in order to describe the meaning of words (cf. Neurath, 1932). In order to avoid vicious circles, a distinction between *metalanguage* (the language used to describe meanings) and *object language* (the language whose meanings are described) is to be adhered to (cf. subsection 4.1.1). The basic idea is that the metalanguage provides an interpreted descriptive framework according to which meanings (of the object language) can be specified. In fact, there are (good) reasons to assume that such an approach cannot be circumvented—the irreducibility of language principle (cf. either Wittgenstein, 1984 for a usage-based view or Hjelmslev, 1961 for a structuralist view of this argument).

Now one can think that the meanings of words *are* concepts. However, the concept a speaker associates with a word includes private episodes. Such private episodes do not belong to the shared (i.e., *normative*) lexical meanings of words. Accordingly, we also distinguish between the (idealized) lexical meaning of a threshold concept expression and (a student's) concept of it (subsection 4.1.2).

But one can just look up the meaning of a word in a dictionary, can't one? Although there is a kernel of truth in it, dictionaries completely avail themselves on the meanings of the object language of the dictionary; in other words, dictionaries contain *paraphrases* of meanings (subsection 4.1.3).

#### 4.1.1. Lexical Meanings

The term *meaning* applies to various relations, as pointed out by means of the examples (3a–c) by Murphy (2010, p. 30):

- (3) a. *Happiness* means “the state of being happy.”
- b. Happiness means never having to frown.
- c. *Glädje* means *happiness* in Swedish.
- d. By *happiness* Peter means *ecstasy*.

In (3) only the first example (3a) involves lexical meaning. In (3b) a consequence relation is expressed and in (3c) a translation relation. (3d) finally is a about speaker meaning (Linsky, 1971). Speaker meaning is usually conceived as pragmatic while lexical meaning is semantic (“Speaker’s Reference and Semantic Reference,” re-published in Kripke, 2011).

Besides lexical meaning there is *compositional meaning* (which for instance accounts for the ambiguity within a simple sentence, such as *every dog chased a cat*, which as a relational (a single cat is chased) and a dependent (there are as many cats as dogs, that is, a plural interpretation of the singular noun phrase *a cat*) reading; see e.g., Zeevat, 2018).

Lexical meaning has to be distinguished into *sense* and *denotation* (this distinction goes back to Frege’s, 1892)<sup>7</sup>. The denotation relation gives rise to the phenomenon that natural language expressions *are about* something in the first place. The denotation of a word is the set of things (potentially) “picked out” by that word. In lexical semantics, senses are directly represented in terms of semantic components (see Jackendoff, 1983, 1991, 2002; Pustejovsky, 1995; Wierzbicka, 1996). We know, however, of no lexical semantic analysis of threshold concept. Thus, describing the meaning of threshold concept expressions in terms of a (existing or specifically developed) metalanguage and their interactions w.r.t. to compositionality and inference could be a desideratum for further studies.

So far, meanings have been ascribed to both words and thoughts. The tension is resolved when considering that senses are *types*, that is, abstract properties which have a normative (and therefore also coordinative) dimension (this issue will be briefly taken up in subsubsection 4.1.2). These sense types are tokened in thoughts of individuals. Accordingly, in cognitive sciences concepts are construed as “temporary constructions in working memory” (Barsalou, 1993, p. 34). Each speaker instantiating a lexical sense instantiates his or her *perspective* or *understanding* of the lexical sense, or indexed concept.

#### 4.1.2. Indexed Concepts

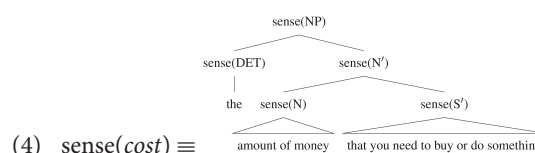
A *concept* is a psychological entity, namely a mental representation and therefore a property of an individual. A concept in the sense of the threshold concept approach (cf. section 2) integrates a disciplinary perspective—an normative description of an economic fact or a principle identified by experts—with the individual perspective—the individual mental representations that the learner associates with a fact—within learning, the individual perspective matches the disciplinary one (Sender, 2017). This means that (i) concepts are not directly observable (they can be evinced by learning assessments or (neuro-)psychological testing, however); (ii) concepts are charged with individual-specific content (which partly accounts for individual-specific understanding); (iii) that concepts are the place where learning takes place.

Now speakers have knowledge about the meaning of lexical items; that is, part of speakers’ lexicalized concepts is *their understanding* of the sense of an expression—this is also one of

the hallmarks of Cognitive Grammar (Langacker, 2013, p. 29)<sup>8</sup>. Hence, the senses identified and modeled in lexical semantics are idealizations; these senses are only realized in meaning-making minds<sup>9</sup>. Thus, when we talk about *the* meaning or *the* concept of an expression, we rely on an idealization, namely the assumption that we share meanings and have a common understanding. Of course, this issue has not gone unnoticed. In fact, there are several genealogical reasons that prevent a “conceptual solipsism.” These include: *coordination* (Lewis, 1969; meanings get coordinated between communities of language users *via* situation of language use), and *evolution* (Millikan, 1984; meanings have a historic yet normative force acquired as biological functions in evolutionary processes). Following a semiotic variant of the principle of methodological individualism (Keller, 1995), socially accepted concepts have to be explained in terms of individual concepts (further examples are known from social ontologies; Searle, 2006). Following the advice of Klein and Kracht (2014, p. 304), namely “the more we talk to each other, the easier it gets, and the more we can come to understand each other,” natural language dialog is the best way for securing mutual understanding. Such an approach is actually pursued in learning studies, where, e.g., classroom interactions are observed. In particular non-verbal behavior of the learners provide evidence on their conceptualizations (Cook and Goldin-Meadow, 2006), in line with the dictum that, for instance, manual gestures are “postcards from the mind” (de Ruiter, 2007).

#### 4.1.3. Dictionary Concepts

While lexical semantics is a useful tool for linguistic analyses of word meanings (cf. subsubsection 4.1.1), it is less useful for everyday use and computational applications. After all, when one wants to know what a word means, one looks it up in a dictionary. According to the British English Online Dictionary<sup>10</sup>, the meaning of *cost* is “the amount of money that you need to buy or do something.” In contrast to lexical semantics, a dictionary describes object language terms in terms of object language terms<sup>11</sup>. The sketch of meanings from subsubsection 4.1.1 suffices in order to make more precise what claim a dictionary entry makes.



<sup>8</sup>Despite claims that concepts and meanings are complementary contents (e.g., Barsalou et al., 1993). Note further that according to Cognitive Grammar “meanings are in the minds of the speakers who produce and understand the expressions” (Langacker, 2013, p. 27). Obviously this claim can only be made because Cognitive Grammar lacks a notion of denotation, leaving it with the identity problem of conceptual content.

<sup>9</sup>There are positions that postulate an objective existence of senses, though—Frege’s (1892) “third realm” is a classic example.

<sup>10</sup><https://dictionary.cambridge.org/> (accessed May 14, 2020).

<sup>11</sup>Murphy (2010, p. 34) is very explicit: “Such paraphrases, also called glosses, are indicated in single quotation marks. One must keep in mind, however, that these glosses are not themselves the meanings of the words (as they are represented in our minds)—they are descriptions of the meanings of the words.”

<sup>7</sup>This pair of kinds of meanings are often translated as *sense* and *reference*. However, since most semanticists would agree that reference is a pragmatic notion (Searle, 1969; Roberts, 2019), we reserve it for that purpose.

The lexical meaning of *cost* is the sense of the syntactic parse (compositional meaning) of the gloss. The reader learns the meaning of *cost*, if he or she knows *sense(NP)*. Furthermore, in order to derive *sense(NP)* not only the lexical meanings but also the compositional meanings have to be computed (cf. also section 3). In order to avoid this, a further simplification can be made by abstracting away from compositional meanings. Now the lexical meaning of *cost* is related (but not equivalent any more) to the lexical meanings of the content words from the gloss, as in (5)

- (5) *sense(cost)* is related to *sense(amount)*, *sense(money)*, *sense(need)*, *sense(buy)*, *sense(do)*, and *sense(something)*

Interestingly, for the dictionary user (5) is nearly as helpful as (4). Most notably, however, dictionary concepts give rise to a notion of *context* of a learning media (cf. Braun et al., 2014): the context in (5) is just the collection of expressions of the dictionary gloss. But in general a context can be any stretch of text from a few words to entire corpora or online resources. Given a context of expressions (dictionary entry, corpus, ...), the expressions are transferred into a claim about their senses, as is made precise in (5). What happens here is that a statement about meanings is given in purely relational manner in terms of the object language—just like in a dictionary paraphrase. That is, (5) exemplifies the scheme of a *differential* rather than *referential* approach to word meaning (Sahlgren, 2008)<sup>12</sup>. Ultimately based on word frequency measures within text corpora, the relation of an expression can also be assigned different strengths by means of vector-valued word representations (Spärck Jones, 1972; Mikolov et al., 2013; Levy et al., 2015)—reflecting their respective “importance.” Now dictionary concepts have a further property which is useful for present purposes: for any two non-identical contexts  $c_1$  and  $c_2$ , the dictionary concept of a random expression will differ with respect to  $c_1$  and  $c_2$ . In other words, dictionary concepts are *text-bound*, and text-boundedness is a prerequisite for comparing different resources in the first place. From a learning perspective, an interpreter of a dictionary entry has to entertain an indexed concept for each of its elements—amounting to the transient nature of threshold concepts and the mental linkage emphasized in subsection 2.1.

#### 4.1.4. Concept Expressions and the “Law of Denotation”

(Lexical) semantics discovered a couple of principles and generalizations. The most important one for current purposes is what Murphy (2010, p. 36) calls the *Law of Denotation* (LoD): the “bigger” a word’s sense (i.e., the more conditions that it places on what counts as a referent for that word), the smaller its extension will be. There are several phenomena to which this principle applies. For instance, the hypernym–hyponym relation fulfills the law of denotation, as does compounding. A broader term like *dog* has less lexical meaning components than a narrower term like

*dachshund*<sup>13</sup>. Since the modifying noun of a nominal compound adds its meaning in some way or other to the head noun, the law of denotation is trivially fulfilled.

Since every expression is bound up with a sense<sup>14</sup>, larger constituents are necessarily accumulative (in fact, compositional). Now assuming expressions, sentences or discourses to be coherent (a notion on which see Asher and Lascarides, 2003, p. 21, and various other places and Ginzburg, 2012, p. 208), this gives rise to the simple but useful generalization: *the more expressions, the more elaborate the combined sense* (where “combined” is intended to cover both compositional derivation as well as accumulation).

The relation between senses and denotations is regimented by LoD. It applies likewise to words, phrases and sentences. The more fine-grained the senses of these constituents, the more detailed are their denotations. The connection to sciences and the language of sciences is obvious: (natural) sciences aim at precise descriptions of the world. That is, scientific languages are about very detailed denotations. In order to achieve this level of detail, guided by LoD, the expressions of the specialized vocabularies need to have elaborate senses, which, by dint of compositional meanings, get even more specific in phrases and sentences. Since natural languages are devices of ontology construction, as has been pointed out by some versions of semantics (e.g., Barwise and Perry, 1983), it is also possible to “postulate new denotations,” so to speak, as has famously been done in the history of physics several times, for instance. LoD and *making things precise* has repercussions to linguistic expressions. Against this backdrop, we discuss observable features of expressions of threshold concepts in the following.

## 4.2. Linguistic Features

Following the guideline that threshold concepts are instances of specialized vocabularies, we expect their expressions to exhibit features which will be described in more detail in the following: (1) compounding potential, (2) large nominal groups, and (3) web of threshold expressions.

1. **Compounding potential.** Of how many compounds is an expression a part? The compounding potential is a long-known feature of specialized vocabulary where specialized languages are characterized by a large number of compounds (Widdowson, 1974). It has also been highlighted by business and economics studies on threshold concepts (e.g., Meyer and Land, 2006). A large number of (compound) nouns is also confirmed in the textbook study by Hu and Gao (2019). In light of the above-mentioned specificity demand of languages of science, this feature is expected. But why are compounds semantically specific and distinguish themselves from *prima vista* synonymous syntactic realizations? Most

<sup>12</sup>This line of thought is rooted in structuralism (de Saussure, 1916; Hjelmslev, 1961).

<sup>13</sup>In this case one must of course know that *dachshund* is a hyponym of the hypernym *dog*. According to dictionary approaches, such knowledge is part of the speaker’s mental lexicon, according to conceptual semantics it is computed based on semantic componential representations.

<sup>14</sup>This is less clear, however, for syncategorematic expressions, such as conjuncts. However, since they do not remove any sense components, they do no harm to the generalization.



nominal compounds [that are compounds whose head is a noun while the modifying component may be an adjective (*green tea*), a verb (*swimming pool*), or a further noun (*football*)] are determinative, meaning that the modifying expression determines the head noun. For instance, a football is not just a ball, but a ball meant to be moved along by one's feet. But there are more interesting properties of compounds. Most importantly, a compound induces a *kind reading* (Bücking, 2010). Given this feature, we expect compounding (as a form of name-giving) to be coupled to the dynamic ontological modifications within the sciences, as is evinced by findings for specialized vocabulary (Widdowson, 1974).

If we conceive the kind-reading of compounds in relation to LoD and the specificity demands of scientific languages, a few trends can be derived:

- (a) For all compounds that share the same threshold concept expression head it holds that the more modifying constituents the compound has, the more specific it is. This follows trivially from sense accumulation. For instance, both *Grenzkosten* “terminal cost” and *Marginalkosten* “marginal cost” are more specific than *Kosten* “cost.”
- (b) The inverse formulation of the previous item is that the more specific a given threshold concept head is, the less compounds it will show. Note that this is a recursive notion: (more) complex compounds may consist of (less) complex heads.
- (c) Going from expressions to the use of these expressions in sentences and texts it is very likely that the more compounds a sentence or text contains, the more specific the sentence or text is (see also the following linguistic feature, “large nominal groups”).

These trends can directly be read off the concept expressions.

2. **Large nominal groups.** Related to the compounding potential is the elaborateness of the whole nominal group of which a concept expression (compound or not) is a part. Expressions of specialized vocabularies tend to occur in elaborate environments (Stevens, 1977). Contexts of elaborateness are constructed by adjectives and relative clauses (mainly restrictive ones). Obviously, nominal groups are more specific according to LoD. This feature is a further linguistic feature of threshold concept expressions to look for.
3. **“Web of threshold expressions.”** Based on postulations of threshold concept research from subsection 2.1 and the linguistic perspective sketched in section 3, concept expressions are to be expected to be related to each other, i.e., forming a “web” of threshold expressions (Davies and Mangan, 2007). Thus, in terms of subsection 4.1.4 we can make the claim more precise in saying that the web of threshold concepts is a context of weighted expressions where the context consists exclusively of threshold concepts. Now the different contexts under consideration (textbooks, newspaper, Wikipedia) trivially give rise to different dictionary concepts. However, since the different contexts are an independent variable, differences can point at meaningful differences in the independent variable (i.e., contexts). Further support

for this claim comes from qualitative investigations of specialized vocabularies, where the context is accredited to be most important feature of special terms (Vaňková, 2018). From that we can derive the expectation that the web of threshold concepts is “stronger woven” in formal than in informal contexts.

## 5. METHODS

### 5.1. Guiding Questions

From subsection 2.2 we take the assumption that resources from formal learning environments are more specific than resources from informal learning environments, since formal environments are characterized by special vocabularies, among others. What tends to be more specific in its use, however, will also form more specific associations with similarly used units: threshold concepts should therefore be more strongly associated with each other if they tend to be used together in specific and equally rare contexts. In this way of thinking, specificity and associative strength seem to be two related concepts that help to compare the use of threshold concepts in different corpora. Thus, it is reasonable to operationalize the above introduced linguistics of threshold concepts by *quantifying* their specificity properties and association relations: the former will be carried out by means of a classical distribution analysis using appropriately quantified specificity values; the latter will be performed by means of a network analysis in which threshold concepts are the nodes whose association relations are interpreted as node connections or links, weighted by the strengths of these associations. In this way we gain access to two types of information: a node-related one (*specificity*) and a link-related one (*association strength*). This enables us to explore both sources of information independently as well as simultaneously using a unified, network-based representation format. However, let us first look at which guiding questions can be formulated either node- or link-related in more general terms<sup>15</sup>:

1. **Q1:** *Do formal corpora show “longer” compounds (i.e., words composed by two or more other words) than informal ones, that is, do formal corpora have more modifying constituents for a given threshold expression head?*
2. **Q2:** *Are there more compounds with threshold concepts (whether as head or not) in formal corpora than in informal ones?*
3. **Q3:** *Are threshold concepts within formal corpora part of larger nominal groups than in informal corpora?*
4. **Q4:** *Does the “web of threshold concepts” derived from formal corpora give rise to a stronger connected threshold concept context than the one derived from informal corpora?*

More formally speaking, the questions Q1–Q3 are all node-related: by operationalizing answers to these questions, we quantify the specificity of threshold concepts in the underlying

<sup>15</sup>Here we focus on threshold concepts within formal and informal learning contexts. For an assessment of the three classes of threshold concepts—basic, discipline, modeling—see the study of Brückner and Lücking (2019).



corpora. Question Q4 is link-related: this question addresses the association strengths in networks of threshold concepts. In any event, according to the current state of our explanations, questions Q1–Q4 are formulated too unspecifically: what does it mean to be connected, for example, and how should this be numerically weighted? In other words, Q1–Q4 cannot yet be tested by means of an exact measurement procedure. To ensure this, we must first translate them into a formal language: in our case this is network theory. This will also mean that we consider variants of selected hypotheses addressing these questions. Ultimately, this approach serves to precisely measure the two core hypotheses about the greater specificity and stronger associativity of threshold concepts in textbooks. To this end, 63 threshold concepts (see **Appendix A**) are compared across several corpora where the textbook corpus consists of the textbooks listed in **Appendix B**. This comparison is based on the measurement procedure described in the next section.

## 5.2. A Two-Part Procedure for Measuring the Use of Threshold Concepts

To tackle the guiding questions Q2 and Q4, we develop a two-part procedure to measure significant differences in the use of threshold concepts. Our first aim is to quantify the difference in the specificity of uses of threshold concepts. In order to operationalize this notion, we start from the following assumptions:

- The more often a threshold concept  $x$  manifests itself as a component in compounds and the higher the frequencies of these compounds in corpus  $C$ , the higher the *degree of specification* of  $x$  and thus its use in  $C$ . We call this sort of specificity *compounding-related specificity* or just *compounding-specificity* of  $x$  in  $C$ . Furthermore, the more frequently the concept occurs in  $C$  as a whole, the higher its polytextuality in the sense of Köhler (1986) (i.e., the higher the number of sentences by which it is semantically specified), the higher its degree of specification. We call this sort of specificity *sentence-related specificity* or just *sentence-specificity*. Finally, the higher the number of threshold concepts with the higher degrees of compounding- or sentence-specificity, the higher the overall specificity of this set of concepts in the underlying corpus.
- The more compounding- or sentence-specific the use of a threshold concept in a corpus, the more detailed and differentiated knowledge can be acquired about this concept by reading texts of this corpus (i.e., the larger the context of the dictionary concept of the threshold concept expression in question).

Starting from these considerations we arrive at the following hypothesis about the difference between formal and informal language corpora (manifesting formal and informal learning contexts) in terms of the compounding- and sentence-specificity with which they manifest threshold concepts:

**H1:** *The use of threshold concepts in formal language corpora is more compounding- or sentence-specific than in informal language corpora.*

Our second aim is to quantify the differences in the associative networks of threshold concepts as induced by corpora of three different genres, i.e., of press communication, encyclopedic communication, and technical communication. From subsection 2.2 we know that newspapers are an example for informal learning contexts, whereas textbooks make up formal contexts. Since to our knowledge there is no *linguistic* judgment of Wikipedia in this respect yet, we remain neutral and will see how Wikipedia compares to formal and informal resources used in the following. For this purpose, we start from the following consideration:

- The greater the differences in the ways threshold concepts are used in two corpora, the more different the associative relations that can be learned as a result of reading homogeneous subsets of texts of these corpora.

By a *homogeneous subset* we mean a set of texts sampled from the same corpus. It should be noted that we do not directly observe the acquisition of semantic associations between threshold concepts. Rather, this acquisition will be estimated by means of word embeddings (Mikolov et al., 2013). The embeddings are compared for the purpose of measuring the semantic associations of the embedded concepts, in the sense of the *Weak Contextual Hypothesis* (WCH) of Miller and Charles (1991): Words that tend to be used in similar contexts are then regarded as semantically similar and correspondingly more strongly associated. That is, if a corpus exhibits such contextual similarities, reading subsets of texts from that corpus makes the acquisition of corresponding syntagmatic or paradigmatic associations, as we assume, more likely. Thus, if the semantic associations of a corpus deviate significantly from those that can be expected, for example, from a thematically similar corpus of textbooks, this may have negative consequences for the acquisition of the concepts concerned. Even if we do not investigate this consequence ourselves, we at least measure the previously mentioned similarity or dissimilarity of association networks. These considerations are a prerequisite for operationalizing the falsification of the following hypothesis about the difference between formal and informal language corpora in terms of the semantic networking of threshold concepts:

**H2:** *Due to their usage contexts in formal language corpora, threshold concepts are more strongly associated than due to their usage in informal language corpora.*

By falsifying the alternative hypotheses of H1 and H2, we obtain evidence that the threshold concepts we are looking at are used significantly differently in the genres under consideration, insofar as their uses correspond to different degrees of specificity (a), while spanning different semantic networks (b). However, what differs in two ways, in that it induces the acquisition of concepts of different specificity (*node-related*) and different associations (*edge-related*), ultimately represents a different learning basis or learning context. From this point of view, it becomes clear that

we understand the structure induced by threshold concepts as a *network of concept nodes and their association relations*, whose “shape” depends on what is said about them in the underlying corpus or how they are specified by means of compounding. More precisely, let  $T = \{a_1, \dots, a_n\}$  be a set of threshold concepts and  $C = \{x_1, \dots, x_m\}$  a text corpus. Then, we denote by

$$C(T) = (V, E, \mu, \nu, \lambda) \quad (1)$$

the *Threshold Concept Network* (TCN) induced by  $C$  over  $T$  where  $E \subseteq V^2$ ,  $\mu: V \rightarrow \mathbb{R}_0^+$  is a function measuring the specificity  $\mu(v)$  of each  $v \in V \subseteq T$  in  $C$ ,  $\nu: E \rightarrow \mathbb{R}$  is a function measuring the semantic association  $\nu(\{v, w\})$  between  $v$  and  $w$  for each  $\{v, w\} \in E$  and  $\lambda: V \rightarrow T$  is an injective vertex labeling function. More specifically,  $\nu(\{v, w\})$  equals the cosine similarity of the embedding vectors computed for  $v$  and  $w$ , respectively, by the operative embedding method that is used to explore  $C$ .

Let  $C_i(T) = (V_i, E_i, \mu_i, \nu_i, \lambda_i)$  and  $C_j(T) = (V_j, E_j, \mu_j, \nu_j, \lambda_j)$  be two TCNs induced by the corpora  $C_i$  and  $C_j$ . For any pair of vertices  $v \in V_i, w \in V_j$ , for which  $\lambda_i(v) = \lambda_j(w)$ , we will write  $\dot{v} = \dot{w}$ . To operationalize the falsification of H1 and H2, we now specify the functions  $\mu$  and  $\nu$  in more detail:

- *On  $\mu$  and H1:* We consider a simple frequency-related definition of  $\mu$ , according to which  $\mu(v)$  corresponds to the number of tokens of the lemma  $v$  in  $C$  plus the number of occurrences of compounds in  $C$  that contain  $v$  as a component (*compounding- + sentence-specificity*). A first variant of  $\mu$ , denoted by  $\mu'$ , considers only the former number (*compounding-specificity*), a second, denoted by  $\mu''$ , only the latter number (*sentence-specificity*). Let  $\mu$  be any of these variants, then we derive the following rank-frequency distribution

$$\mu(V) = ((v_{i_1}, \mu(v_{i_1})), \dots, (v_{i_n}, \mu(v_{i_n})), \mu(v_{i_1}) \geq \dots \geq \mu(v_{i_n}), v_{i_1}, \dots, v_{i_n} \in V \quad (2)$$

for which we compute the exponent  $\alpha$  of the power law that best fits this rank distribution. In this way, we test the skewness of the distribution of the specificities of threshold concepts as induced by  $C$ : the higher the value of  $\alpha$ , the faster the frequency-related transition from high-rank (frequent or highly specified) to low-rank (rare or rarely specified) concepts; note that we always consider small numbers of concepts for the distributions, so the slope cannot be the result of a larger number of rare concepts and especially *hapax legomena*. The alternative to H1 is now considered falsified if the corpus length-normalized rank specificity distribution of formal language corpora is above that of informal language ones, under the condition of a Zipfian, power law-like character of such distributions as normally observed for word frequency distributions (Zipf, 1949; Tuldava, 1998) and also assumed for threshold concepts. Beyond that, we assume that power laws better fit the use of threshold concepts in textbook corpora or in formal language corpora in general than in informal language corpora (e.g., of press communication). Furthermore, we assume that the rank specificity distributions

of formal language corpora differ significantly from those obtained for informal language corpora. Finally, we assume that the rank correlation between the rank specificity distributions of formal and informal language corpora is lower than in cases where the corpora manifest either both formal or informal language—provided that these corpora are all sufficiently similar thematically. If we succeed in falsifying the alternative to H1 in these senses, we get the information that formal language contributes to the development of more specific threshold concepts, the specificity distribution of which follows a Zipfian distribution in a more pronounced and significantly different way compared to corpora of informal language, that the specificity of the concepts in the latter corpora tends to be lower, and that, finally, thematically and formally similar corpora are more similar to each other than corpora of different formality.

- *On  $\nu$  and H2:* The association strength of TCNs in relation to the degree of formality of the underlying corpus will be measured using methods of network theory (Newman, 2010) and especially of the theory of linguistic networks (Mehler et al., 2020a). More specifically, we test H2 by quantifying the densities of TCNs derived from different corpora using the approach of Mehler et al. (2020b). That is, we utilize the notion of  $\alpha$ -cuts, as introduced in the description of fuzzy sets, and apply it to weighted graphs as follows: let  $C(T) = (V, E, \mu, \nu, \lambda)$  be a TCN. Then we define:

$$a(C(T)) = (\alpha_1, \dots, \alpha_l)^T \quad (3)$$

$$\alpha_1 = \min\{s(v(e)) \mid e \in E\} \quad (4)$$

$$\forall k \in \{2, \dots, l\}: \alpha_k = \min\{s(v(e)) \mid s(v(e)) > \alpha_{k-1}\} \quad (5)$$

$$\forall e \in E: s(v(e)) = \frac{\nu(e) - \text{Min}}{\text{Max} - \text{Min}} \in [0, 1] \quad (6)$$

where Max (Min) is the theoretical maximum (minimum) that  $\nu$  can assume. Then we define the  $\alpha$ -cut of  $C(T) = (V, E, \mu, \nu, \lambda)$ , that is, the so-called  *$\alpha$ -cut graph*  $C(T, \alpha) = (V, E|_\alpha, \mu|_\alpha, \nu|_\alpha, \lambda|_\alpha)$  where

$$E|_\alpha = \{e \in E \mid \nu(e) \geq \alpha\} \quad (7)$$

and  $\mu|_\alpha, \nu|_\alpha$  are the restrictions of  $\mu, \nu$  to the vertex set induced by  $E|_\alpha$  and where  $\nu|_\alpha: E|_\alpha \rightarrow [0, 1], \forall e \in E|_\alpha: \nu|_\alpha(e) = s(v(e))$ . This allows us to define the graph series

$$\text{cuts}(a(C(T))) = (C(T, \alpha_1), \dots, C(T, \alpha_l)) \quad (8)$$

Finally, for any graph index  $\iota: \mathbb{G} \rightarrow \mathbb{R}$ , we get a series of index values:

$$\iota(\text{cuts}(a(C(T)))) = (\iota(C(T, \alpha_1)), \dots, \iota(C(T, \alpha_l))) \quad (9)$$

In this paper, we experiment with graph cohesion and graph clustering (Newman, 2010). For each of these indices, we want to know how (1) early, (2) fast, and (3) differently its values for the different series of  $\alpha$ -cut graphs calculated for the targeted corpora are decreasing or increasing. Now Hypothesis H2 is considered falsified if the cohesion of the series of  $\alpha$ -cut

**TABLE 1** | Summary of corpora used in the study.

	No. of articles	No. of token	Period of publication
SZ-Eco	288,792.000,0	85,826,410.000,0	1992–2014
SZ-All	1,707,666.000,0	630,588,082.000,0	1992–2014
WP-Eco	653,397.000,0	265,063,077.000,0	2001–2016
WP-Top-1	37,895.000,0	20,090,166.000,0	2001–2016
WP-Top-3	71,013.000,0	28,145,793.000,0	2001–2016
WP-All	1,760,875.000,0	736,071,291.000,0	2001–2016
ZEIT	184,186.000,0	179,327,441.000,0	1994–2014
TB	14.000,0 books	2,326,374.000,0	2015–2020

See main text for a description. The acronyms are resolved in section 5.3.

graphs calculated for the textbook corpus decreases later than in the case of alpha-cut graphs calculated for non-textbook corpora, and in such a way that the behaviors of these series differ significantly from each other. Further, we expect the same behavior with regard to the corresponding series of graph clustering or transitivity values.

In a nutshell: H1 is considered falsified if the alternative hypotheses to H1 and H2 are falsified. If such a double falsification succeeds, we obtain evidence that formal language corpora support the development of more strongly specified threshold concepts that are at the same time more strongly associated with each other or semantically networked. According to our guiding idea, such an observation is linked to the assumption that reading formal language corpora facilitates the acquisition of threshold concepts according to the associated learning objective.

### 5.3. Data and Pre-processing

We consider corpora from press communication, encyclopedic communication and technical communication (see **Table 1**):

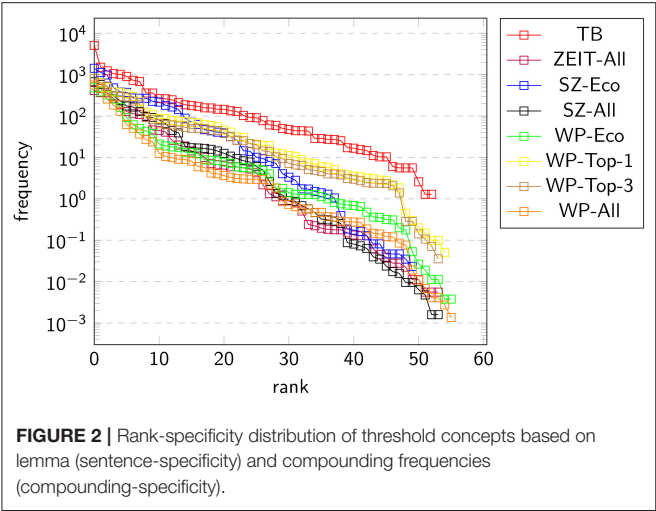
1. *Corpus SZ-Eco*: as an informal language corpus of texts about economics, we process 288,792.000,0 texts from the *Süddeutsche Zeitung* (SZ), one of the largest daily German newspapers, all of which belong to the register *Wirtschaft [economics]*—see **Table 1** for the corpus statistics.
2. *Corpus SZ-All*: SZ-Eco is contrasted with SZ-All, that is, the corpus of all 1,707,666.000,0 articles of SZ published in the years 1992 to 2014 (see **Table 1**). In this way we get access to the usage regularities of threshold concepts in arbitrary press articles of whatever topic.
3. *Corpus WP-Top-1*: as a formal language corpus of texts on economics, we determine the subset of all Wikipedia articles whose top-level topic category corresponds to the *Dewey Decimal Classification* (DDC) *Category 330 (Economics)*. In other words, we DDC-categorize all Wikipedia articles of the German Wikipedia using *text2ddc* (Uslu et al., 2019) and select those articles whose top-level topic category corresponds to DDC category 330. In this way, we obtain a subset of Wikipedia articles that can be very reliably assigned to our target topic of economics: anyone who reads articles of the Wikipedia article network, which is spanned by these articles, navigates, so to speak, in the thematically homogeneous area of economically relevant articles.

4. *Corpus WP-Top-3*: in analogy to WP-Top-1, WP-Top-3 is the set of all German Wikipedia articles where the DDC category 330 is among the first three DDC categories assigned to this article by *text2ddc* with a membership value of at least 10%. Obviously WP-Top-3 contains larger parts of WP-Top-1 (10% threshold) or even this corpus as a whole, but likely also articles whose relation to economics is less confirmed, even if they do not fall below the 10% threshold.
5. *Corpus WP-Eco*: WP-Eco is the corpus of all articles in Wikipedia that are directly or indirectly assigned to the category *Wirtschaft [economics]* from Wikipedia's category system. WP-Eco contains 653,397.000,0 articles and thus about a third of all 1,760,875.000,0 articles of German Wikipedia; WP-Eco also contains articles that are (possibly) only (very) indirectly related to the topic of economics. Whoever reads articles from the corresponding article network navigates, so to speak, in the wider area of economics-related articles, while possibly changing the topic (starting from economics), but in a frame that still has to do with economics.
6. *Corpus WP-All*: the largest corpus we look at includes the 1,760,875.000,0 articles from the German Wikipedia, most of which are not related to economics (see **Table 1**).
7. *Corpus ZEIT*: as a second corpus of informal language of press communication, we process the 184,186.000,0 texts of the German weekly newspaper *Die Zeit* published in the years 1994–2014.
8. *Corpus TB*: Last but not least we analyze a corpus of formal language, that is, a corpus of 14 textbooks all about economics in the narrow sense (see **Appendix B**).

In total, we consider eight corpora, three of which are informal language corpora of press communication (SZ-Eco, SZ-All, ZEIT), three of which mainly comprise texts that are not related to economics (SZ-All, WP-All, ZEIT) and five of which are formal language corpora (WP-All, WP-Eco, WP-Top-1, WP-Top-3, TB). Moreover, one of the informal language corpora (SZ-Eco) and four of the formal language corpora (WP-Eco, WP-Top-1, WP-Top-3, TB) focus more or less on economics. For preprocessing all these corpora, we use *TextImager* (Hemati et al., 2016). That is, the corpora are tokenized, part of speech-tagged and lemmatized. Furthermore, sentences are split and tokens are segmented to identify candidate compounds, their heads and modifiers. Text classification regarding the second level of the DDC is performed by means of *text2ddc* (Uslu et al., 2019). Embeddings are computed for all corpora separately using *word2vec* based on standard settings (i.e., word vector size = 100, window size = 5, with five training iterations) for skip-gram and cbow (see Mehler et al., 2020c for a related procedure). Finally, the embeddings are used to induce TCNs according to section 5.2, which are then processed with *GraphMiner*, a network analysis software under development at TTLab ([www.texttechnologylab.org](http://www.texttechnologylab.org)).

### 5.4. Results

In **Figure 2** we show the rank-specificity distribution of our set of threshold concepts based on the variant  $\mu$  of vertex weights in TCNs. It is remarkable that the specificity values of threshold concepts in textbooks are above all distributions induced by



the comparison corpora. Furthermore, the specificity values for concepts from formal language corpora dedicated to economics, such as WP-Top-1 and WP-Top-3 are also higher. In contrast, specificity values from corpora of more general content (WP-All, SZ-All, ZEIT-All) do not achieve such high levels. In the middle of the spectrum of specificity distributions we observe SZ-Eco and WP-Eco, two corpora of medium size, which deal with economic issues in a larger thematic context. Note that we calculate relative frequencies in order to rule out size effects and scale the distributions (by multiplying with 1,000,000.000,0) in order to enhance readability.

In order to estimate whether the distributions actually differ from each other, we perform pairwise Kolmogorov-Smirnov goodness-of-fit tests. If the  $p$ -values of any such fit is high, then we cannot reject the hypothesis that the distributions of the two samples are the same. In other words: small  $p$ -values indicate a significant difference between two distributions. Results are collected in **Table 2**, where  $p < 0.1$  is highlighted in green (likewise for **Tables 3–8** below): obviously, in most cases the distributions differ from each other. Remarkable exceptions are SZ-Eco in relation to SZ-All (the latter contains the former), WP-Top-1 and WP-Top-3 (also a matter of inclusion) and especially SZ-Eco in relation to WP-All.

The scenario observed in **Figure 2** is also displayed by **Figure 3** (sentence-specificity) and **Figure 4** (compounding-specificity): the specificity distributions are all topped by the distribution for textbooks. In this sense, it can be said that the threshold concepts considered here are most specifically described in the formal language textbook corpus, followed by the two formal language Wikipedia-based corpora WP-Top-1 and WP-Top-3 and least specifically in the informal newspaper corpora SZ-All and ZEIT-All, although in the case of compounding-specificity the situation is not so obvious. A borderline case is WP-Eco, a corpus that consists of Wikipedia articles that are directly or indirectly assigned to the thematic field of economics.

**TABLE 2 |**  $P$ -values of the Kolmogorov-Smirnov goodness-of-fit test applied to the pairwise combinations of the distributions in **Figure 2**.

	TB	SZ-Eco	SZ-All	WP-Eco	WP-Top-1	WP-Top-3	WP-All	Zeit-All
TB	—	—	—	—	—	—	—	—
SZ-Eco	—	—	—	—	—	—	—	—
SZ-All	—	—	—	—	—	—	—	—
WP-Eco	—	—	—	—	—	—	—	—
WP-Top-1	—	—	—	—	—	—	—	—
WP-Top-3	—	—	—	—	—	—	—	—
WP-All	—	—	—	—	—	—	—	—
Zeit-All	—	—	—	—	—	—	—	—

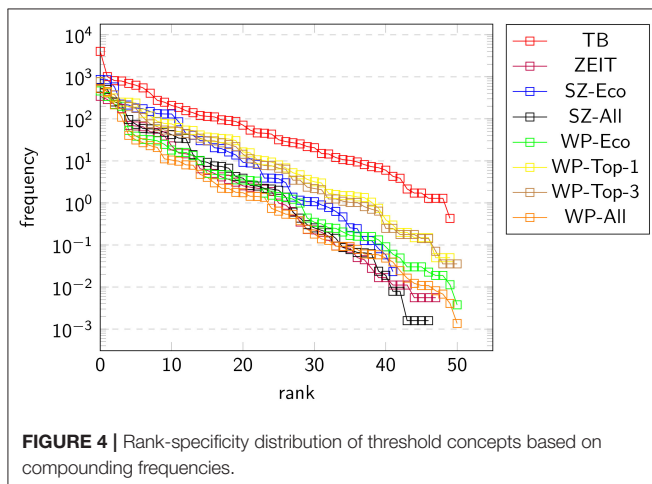
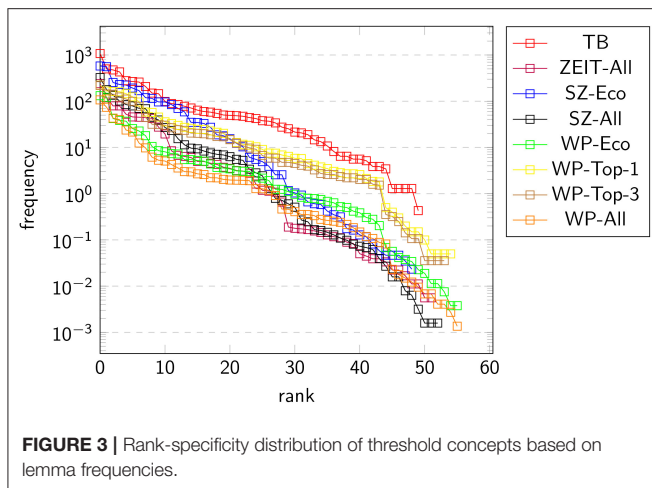


**TABLE 3** | *P*-values of the Kolmogorov-Smirnov goodness-of-fit test applied to the pairwise combinations of the distributions in **Figure 3** (sentence-specificity).

	TB	SZ-Eco	SZ-All	WP-Eco	WP-Top-1	WP-Top-3	WP-All	Zeit-All
TB	—	0.000,3	$1.080,0 \times 10^{-05}$	0.000,5	0.138,4	0.043,4	0.000,1	0.007,1
SZ-Eco	—	—	0.562,7	0.184,0	0.013,2	0.013,5	0.538,8	0.233,0
SZ-All	—	—	—	0.015,0	0.000,2	0.001,5	0.075,0	0.023,2
WP-Eco	—	—	—	—	0.049,1	0.096,5	0.468,1	0.099,5
WP-Top-1	—	—	—	—	—	0.966,6	0.008,1	0.069,5
WP-Top-3	—	—	—	—	—	—	0.005,6	0.047,5
WP-All	—	—	—	—	—	—	—	0.061,9
Zeit-All	—	—	—	—	—	—	—	—

**TABLE 4** | *P*-values of the Kolmogorov-Smirnov goodness-of-fit test applied to the pairwise combinations of the distributions in **Figure 4** (compounding-specificity).

	TB	SZ-Eco	SZ-All	WP-Eco	WP-Top-1	WP-Top-3	WP-All	Zeit-All
TB	—	0.010,8	0.000,7	0.039,3	0.218,4	0.180,3	0.012,2	0.066,4
SZ-Eco	—	—	0.313,6	0.693,7	0.194,1	0.243,6	0.995,6	0.597,8
SZ-All	—	—	—	0.201,5	0.022,5	0.052,7	0.650,5	0.139,3
WP-Eco	—	—	—	—	0.435,4	0.593,0	0.980,7	0.484,3
WP-Top-1	—	—	—	—	—	0.996,2	0.188,1	0.525,9
WP-Top-3	—	—	—	—	—	—	0.307,9	0.598,5
WP-All	—	—	—	—	—	—	—	0.298,5
Zeit-All	—	—	—	—	—	—	—	—



When we look at **Tables 3, 4**, we get the information that while the frequency distributions (sentence-specificity) tend to be distinguishable, the distinguishability of the compounding-specificities is much less: obviously, the frequencies of compounds to which our threshold concepts belong are more independent of the underlying corpus. Moreover, the distributions in **Figures 2–4** tend to be all Zipfian: although a lognormal distribution is also a good fit in 17 (of 24) cases, power law fitting is still a valid option (there is not a single significant  $p$ -value  $< 0.05$  for any  $R < 0$ ; note further that a lognormal distribution is a heavy-tailed distribution, too): the exponent  $\alpha$  ranges from  $\approx 1.3$  to  $\approx 2.8$ , where the minimum  $x$  value of the fit is given as “ $x$ -min” (see **Table 5**)<sup>16</sup>.

From this perspective, we see the alternative of hypothesis H1a, which states that the use of threshold concepts in formal

language corpora is neither more compounding-specific nor more sentence-specific than in informal language corpora, as being falsified.

Next we consider Hypothesis H1b. For this purpose, we compare the series of cohesion values induced by the series of alpha-cut graphs (see above) based on our eight different corpora. We start with exemplifying alpha-cut graphs based on three different corpora using the same set of threshold concepts and cutting for the same  $\alpha = 0.7$ : corpus SZ-All (**Figure 5**), corpus TB (**Figure 6**), and corpus WP-Eco (**Figure 7**). These graphs, which are all based on the same vertex set, illustrate a networking effect that is later confirmed by our analysis of the entire time series of alpha-cut graphs: Wikipedia-based corpora exhibit the densest networking, followed by textbook corpora and newspaper corpora. Threshold concepts associate more strongly and more often in the case of the former compared to the latter. Moreover, in the case of the newspaper corpus, the number of network components is highest (so that the number of isolated nodes is also highest), while in the case of the textbook corpus there is a unique dominant vertex (*costs/Kosten*) in terms of compounding- and sentence-specificity. But what exactly does the network density look like when we look at the entire time series of these alpha-cut graphs? **Figure 8** shows the corresponding distributions starting from the TCNs derived from word embedding similarities based on the skip-gram model of word2vec and thus for syntagmatic associations (starting from the respective seed word to the probable context in the sense of being defined by neighboring words). Very remarkably, all four Wikipedia corpora behave very alike: the cohesion values of the TCN series induced by these corpora decrease at the latest compared to all other corpora and their corresponding TCN series, i.e., they decrease for the comparatively highest  $\alpha$  values. Conversely, the cohesion values of the corresponding TCN series induced by the newspaper corpora (SZ-All, ZEIT-All) decrease the fastest. In the middle of this spectrum we surprisingly observe two series of cohesion values: that for the textbook corpus and that for the economics-related SZ-Eco corpus, though rather in the neighborhood of the Wikipedia corpora than in the one of the newspaper corpora. At this point, we have to ask whether the distributions shown in **Figure 8** are actually different or not. For this purpose we again perform Kolmogorov-Smirnov tests of goodness-of-fit, but now separately for both axes from **Figure 8**. The reason is that neither axis is ordinal scaled, so we first perform a corresponding scaling before we can compare the corresponding feature distributions. As shown in **Table 6**, we get a mixed result: while the alpha-cuts of the individual distributions increase very differently (so that the distributions are mostly clearly distinguishable from each other), this does not apply to the decreases in cohesion values caused by the increasing alpha-cuts: here the distributions are all indistinguishable. For the distributions of the cohesion values this means that they are in fact all almost “identical” and therefore indistinguishable mirrored S-curves when being scaled appropriately.

From this spectrum of distributions, we get the following assessment: in Wikipedia-based corpora, the threshold concepts are most strongly associated with each other—metaphorically

<sup>16</sup>We apply the toolbox of Alstott et al. (2014) according to Clauset et al. (2009): Power laws (first) are compared to lognormal distributions (second): “ $R$  is the log likelihood ratio between the two candidate distributions. This number will be positive if the data is more likely in the first distribution, and negative if the data is more likely in the second distribution. The significance value for that direction is  $p$ .” (Alstott et al., 2014, p. 5).

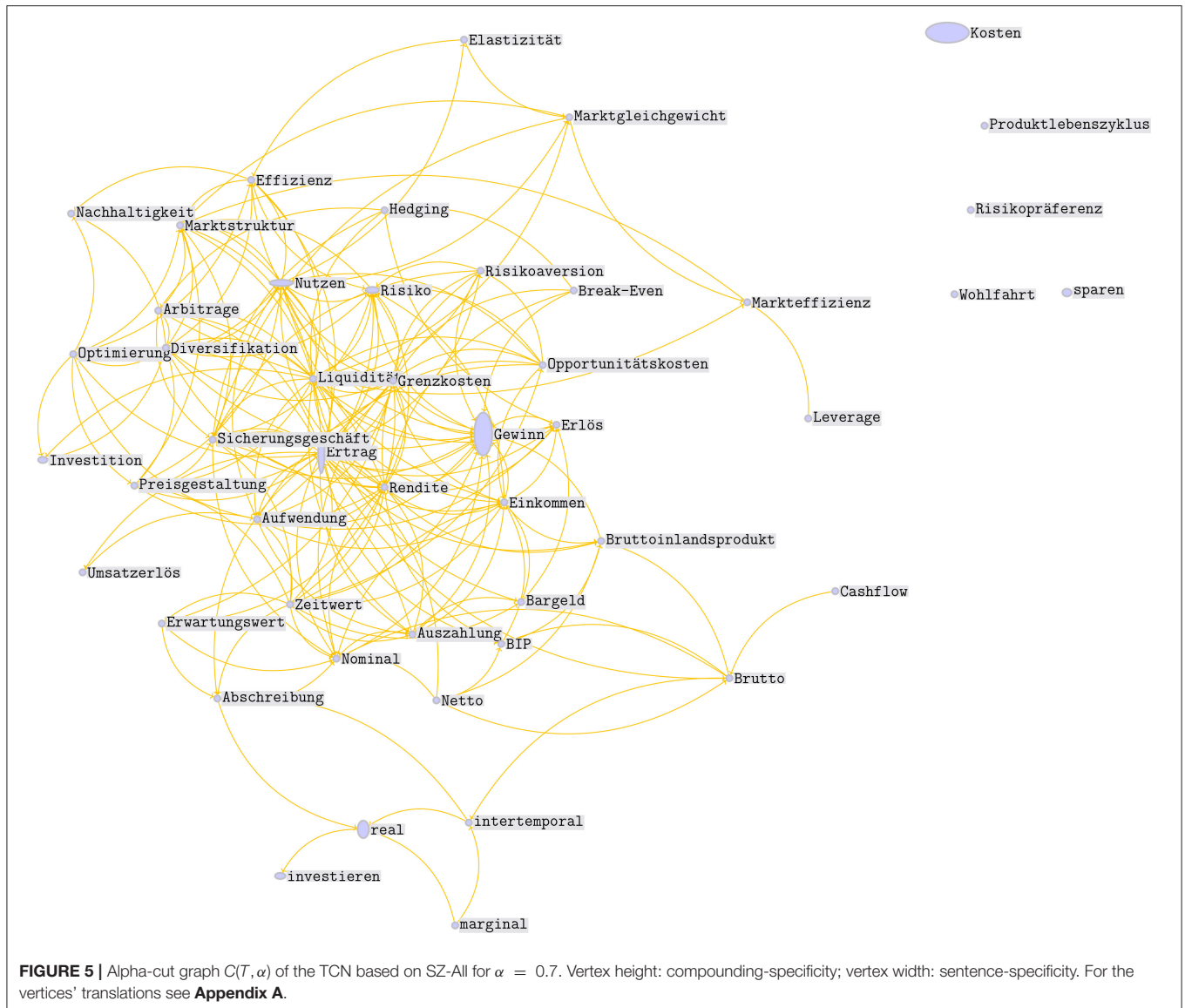
**TABLE 5 |** Power law goodness-of-fit tests for the distributions from **Figures 2–4**.

	Alpha	x-min	R	P
<b>Lemma and compound, Figure 2</b>				
SZ-All	1.711629	4.0	0.004996	0.737173
SZ-Eco	1.565592	2.0	−0.686457	0.305550
TB	2.825375	41.0	0.009203	0.853003
WP-All	1.593086	3.0	0.009178	0.719042
WP-Eco	1.613395	3.0	−0.568429	0.322225
WP-Top-1	1.522315	3.0	−0.002717	0.793323
WP-Top-3	1.485334	2.0	−0.242887	0.866997
Zeit-All	1.395048	1.0	−0.666070	0.403735
<b>Only lemma, Figure 3</b>				
SZ-All	1.735390	5.0	−0.126319	0.294422
SZ-Eco	1.548944	2.0	−0.689713	0.307030
TB	1.336778	1.0	−0.627145	0.792154
WP-All	1.593086	3.0	0.009178	0.719042
WP-Eco	1.483664	2.0	0.000933	0.964148
WP-Top-1	1.433428	2.0	−0.173451	0.273327
WP-Top-3	1.546151	4.0	0.015474	0.301308
Zeit-All	1.395048	1.0	−0.666070	0.403735
<b>Only compound, Figure 4</b>				
SZ-All	1.355914	1.0	−0.666272	0.268328
SZ-Eco	1.559961	3.0	0.028074	0.235693
TB	1.395212	2.0	−0.683652	0.307499
WP-All	1.395048	1.0	−0.666070	0.403735
WP-Eco	1.676775	4.0	−0.580707	0.369533
WP-Top-1	1.545465	3.0	−0.641716	0.320173
WP-Top-3	1.464465	2.0	−0.649692	0.317607
Zeit-All	1.418524	2.0	−0.080603	0.606211

speaking, they form a denser network of particles that are located much closer to each other. For much higher values than for any other corpus, the network cohesion (starting from a completely connected graph) takes a maximum value of 1; and for equally maximum values the cohesion is at least 50, 75%, etc.: the deletion of lower weighted edges in TCNs based on Wikipedia corpora is therefore more likely to lead to more cohesive networks compared to the other TCNs. In view of this finding, the textbook-based TCNs are surprisingly less cohesive. *Based on our cognitive model, this suggests that reading such textbooks makes stronger syntagmatic associations under threshold concepts less likely.* Wikipedia seems to write more densely about these concepts, in a way that makes their associations more probable and also more pronounced. This may be related to the text type of Wikipedia (*encyclopedic communication*) as opposed to textbooks, which may also contain longer motivational, exemplary or elaborating text passages. In any case, however, we see the hypothesis confirmed that formal language corpora make stronger associations between threshold concepts more likely than informal language corpora—this is

indirectly confirmed by the values of **Table 6** regarding the  $x$ -axis (formal language corpora are significantly “shifted” to the right compared to their newspaper-based counterparts, i.e., SZ-All and ZEIT-All). An extreme-value-forming special position of textbooks, however, cannot be confirmed. Moreover, the strengths of the associations of threshold concepts obtained by means of informal texts on topics related to economics (SZ-Eco) can hardly be distinguished from those obtained with the help of textbooks: *from this point of view, we do not see a special role for textbooks compared to quasi informal newspaper articles.* The only exception is Wikipedia—regardless of the topic of economics.

**Figures 9 and 10** essentially confirm the results obtained so far. However, we now observe, for higher  $\alpha$  values, that the cluster values of textbook-based networks become seemingly indistinguishable from those observable for Wikipedia corpora-based networks—the same observation concerns the SZ-Eco-based networks. Textbook-based TCNs are again hardly distinguishable from TCNs derived from informal language newspaper articles about topics related to economics (SZ-Eco).



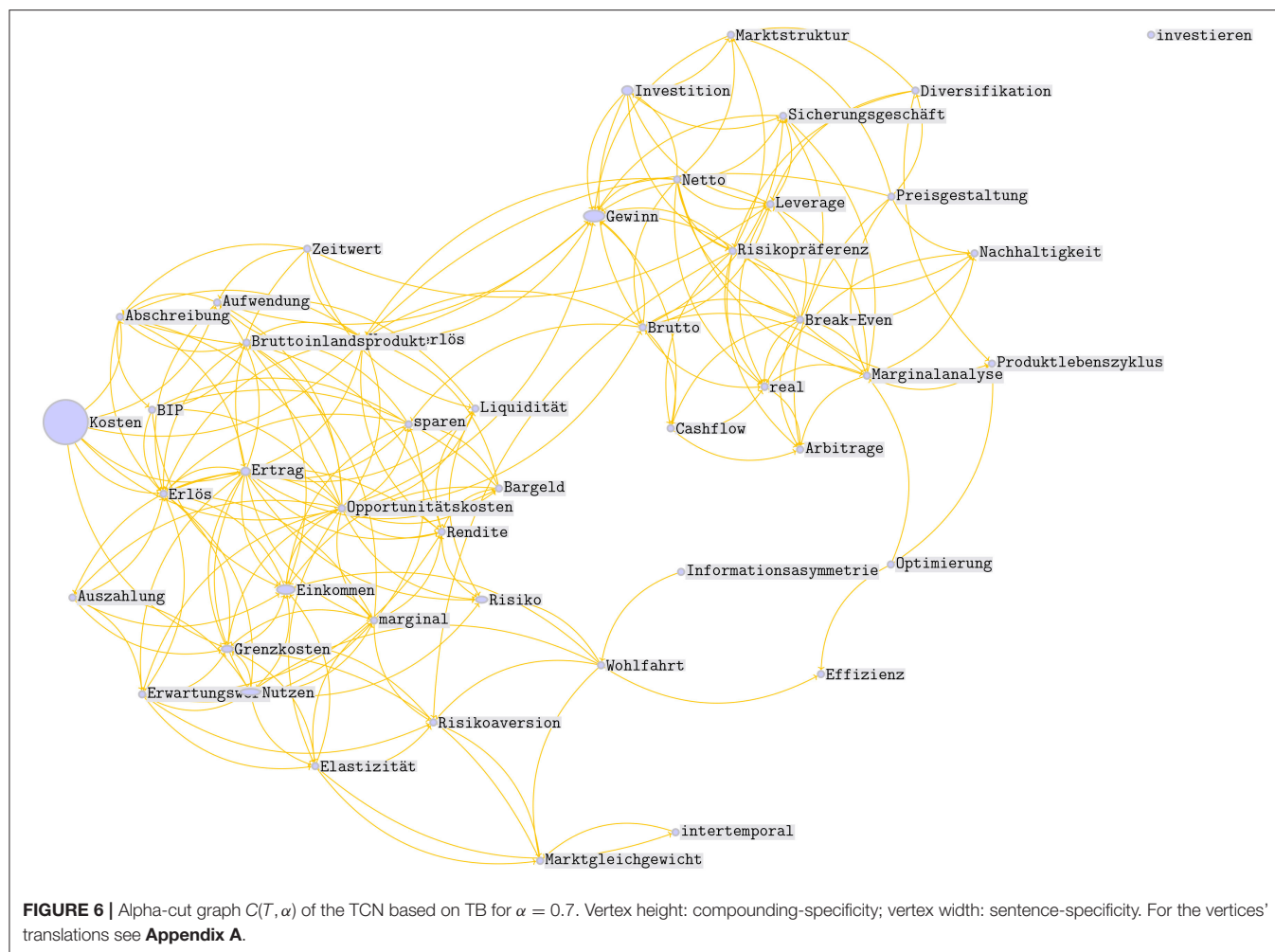
In any case, **Table 7** also shows that all value distributions along the x and y axis are now distinguishable with only three exceptions: the dynamics of clustering is obviously more corpus specific.

Any special role of textbooks almost completely disappears if we consider the cbow model of word2vec (i.e., associations starting from lexical contexts toward target words and thus paradigmatic associations) (see **Figure 8**). In other words, paradigmatic associations of the sort *Bruttoinlandsprodukt/gross domestic product* and *BIP/GDP* seem to be highest from the perspective of Wikipedia-based corpora and higher from the perspective of newspaper corpora than from the perspective of the textbook corpus, while syntagmatic associations of the sort *Gewinn/profit* and *marginal/marginal* are still highest in the case of Wikipedia-based corpora, but are more pronounced from the

perspective of textbooks than from newspapers. **Table 8** leads to an assessment similar to **Table 6**.

Note that in all these cases of cbow (**Figure 9**) and skip-gram-based (**Figure 10**) networks and their underlying embeddings we use standard parameter settings and especially a rate of five iterations: from this point of view, it could be that shorter corpora are more negatively affected by such iterations than longer ones. Scaling their size by increasing the number of iterations can lead to false dissociations of words (as a test of 100 iterations based on the textbook corpus actually suggests). Instead, the sizes of the larger corpora should be reduced to those of the smallest corpora, i.e., the corpus of textbooks—but the corresponding sampling routine and experimentation will be part of future work. In any case, it should be noted that our results are conditioned by the latter assessment. And this means that the alternative of





Hypothesis H1b is only falsified if we compare Wikipedia-based corpora with newspaper corpora. However, in the case of WP-All, we must refrain from a focus on economics-related topics. The inclusion of the textbook corpus in the set of formal language corpora definitely does not allow such a falsification: so either H1b is wrong or our current measuring procedure does not allow yet for falsifying the alternative of H1b.

## 6. DISCUSSION

As evidenced in section 5.4, threshold concepts occur significantly more frequently in formal textbook corpora than in Wikipedia and newspaper corpora, both with respect to the naming variants investigated here and with respect to their frequencies as components of compounds: In line with Hypothesis H1a, the textbook corpora examined had a higher density of compounds and unique sentences than all other corpora investigated here. However, we have also shown that their surrounding networks are not exceptional (in terms of stronger syntagmatic or paradigmatic connections). Regarding the network structure we observed Wikipedia to be exceptional, and this observation is independent of the topic of economics

as it holds for the non-economic corpora, as well. This finding points to a special role of encyclopedic communication as a representative of formal language communication, a role that may have been underestimated in educational sciences until now. However, based on our experiments we must also note that we could not confirm H1b (or falsify its alternative hypothesis).

### 6.1. Limitations and Suggestions for Future Research

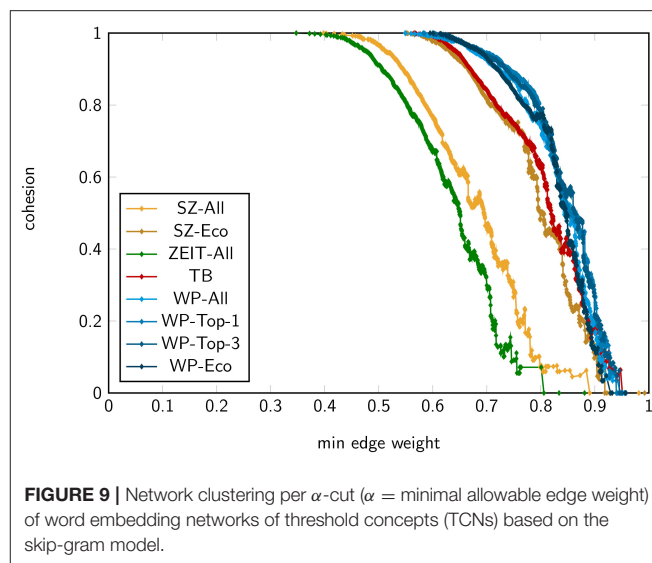
There are several points of departure for improving the procedure we have developed for measuring the usage frequencies of threshold concepts in corpora of formal and informal language:

- We observe that Wikipedia stands out in terms of networking of threshold concepts. Since this observation extends beyond the domain of economics, Wikipedia seems to be characterized by a rather high level of density of specialized language terms in general. As indicated in subsection 5.4, this property is likely due to the encyclopedic genre of Wikipedia, which raises issues



TABLE 6 | *P*-values of the Kolmogorov-Smirnov goodness-of-fit test applied to the pairwise combinations of the *x* and *y* values of the distributions in Figure 8.

	SZ-All	SZ-Eco	TB	WP-All	WP-Eco	WP-Top-1	WP-Top-3	Zeit-All
<b>X-values</b>								
SZ-All	—	$1.332,3 \times 10^{-15}$	$6.661,3 \times 10^{-16}$	$3.774,8 \times 10^{-15}$	$3.774,8 \times 10^{-15}$	$3.254,7 \times 10^{-300}$	$2.164,6 \times 10^{-299}$	$1.443,3 \times 10^{-15}$
SZ-Eco	—	—	$2.133,0 \times 10^{-05}$	$1.554,3 \times 10^{-15}$	$1.554,3 \times 10^{-15}$	$1.332,3 \times 10^{-15}$	$1.332,3 \times 10^{-15}$	$1.110,2 \times 10^{-16}$
TB	—	—	—	$1.554,3 \times 10^{-15}$	$1.554,3 \times 10^{-15}$	$6.661,3 \times 10^{-16}$	$6.661,3 \times 10^{-16}$	$1.554,3 \times 10^{-15}$
WP-All	—	—	—	—	0.304,7	$9.832,6 \times 10^{-07}$	0.000,2	$2.109,4 \times 10^{-15}$
WP-Eco	—	—	—	—	—	$1.718,7 \times 10^{-07}$	$6.944,0 \times 10^{-05}$	$2.109,4 \times 10^{-15}$
WP-Top-1	—	—	—	—	—	—	0.477,5	$2.109,4 \times 10^{-15}$
WP-Top-3	—	—	—	—	—	—	—	$1.443,3 \times 10^{-15}$
Zeit-All	—	—	—	—	—	—	—	$1.443,3 \times 10^{-15}$
<b>y-values</b>								
SZ-All	—	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0
SZ-Eco	—	—	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0
TB	—	—	—	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0
WP-All	—	—	—	—	1.000,0	1.000,0	1.000,0	1.000,0
WP-Eco	—	—	—	—	—	1.000,0	1.000,0	1.000,0
WP-Top-1	—	—	—	—	—	—	1.000,0	1.000,0
WP-Top-3	—	—	—	—	—	—	—	1.000,0
Zeit-All	—	—	—	—	—	—	—	—

FIGURE 9 | Network clustering per  $\alpha$ -cut ( $\alpha$  = minimal allowable edge weight) of word embedding networks of threshold concepts (TCNs) based on the skip-gram model.

- A second extension concerns the detailed consideration of *basic*-, *discipline*-, and *procedural*-level concepts. More specifically, formal language corpora could be divided into subsets of texts depending on their learning level, which are either at the basic, disciplinary or procedural level. In this way, we gain access to contexts of use of threshold concepts that allow us to assign them to one of these levels or to determine linguistic evidence of what was described above as conceptual change, i.e., the transition in the use of a concept between these levels that might indicate a higher dynamics relevant to formal learning contexts.
- A third extension concerns the broadening of the basis of comparison of threshold concepts. That is, instead of just networking them with each other, we could additionally examine how they network with non-threshold concepts or with concepts that belong to one of the three basic, disciplinary or procedural learner levels. In any event, this should again be done in such a way that each of these reference sets is small and selected in advance in order to allow transparent comparisons.

## 6.2. Implications of Learning Media for Learning Assessment

Different resources can be interpreted to make different claims about the relations between threshold concepts. For the 63 threshold concept expressions  $t_1, \dots, t_{63}$  under consideration, this claim can be represented in the form: “sense ( $t_1$ ) is related to sense ( $t_2$ ), sense ( $t_2$ ) is more related to sense( $t_8$ ),” and so on, where the degree of relatedness differs between the corpora (cf. subsection 4.1.3). That is, different resources express a different “take on threshold concepts.” This in turn leads to the question whether the different resources also imply or lead learners to assume a significantly different understanding of threshold concepts, or consequently whether different learning media might be appropriate in different contexts (e.g., depending

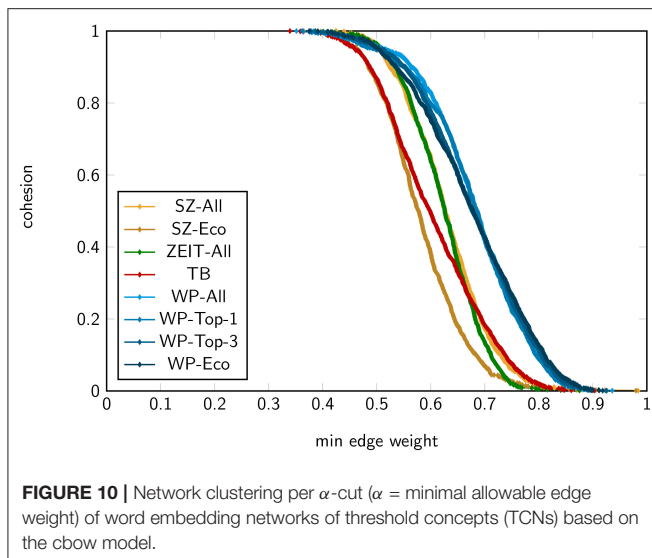
**TABLE 7** | *P*-values of the Kolmogorov-Smirnov goodness-of-fit test applied to the pairwise combinations of the *x* and *y* values of the distributions in **Figure 9**.

	SZ-All	SZ-Eco	TB	WP-All	WP-Eco	WP-Top-1	WP-Top-3	Zeit-All
<b>x-values</b>								
SZ-All	—	$1.332,3 \times 10^{-15}$	$6.661,3 \times 10^{-16}$	$3.774,8 \times 10^{-15}$	$3.774,8 \times 10^{-15}$	$3.254,7 \times 10^{-300}$	$2.164,6 \times 10^{-299}$	$1.443,3 \times 10^{-15}$
SZ-Eco	—	—	$2.133,0 \times 10^{-05}$	$1.554,3 \times 10^{-15}$	$1.554,3 \times 10^{-15}$	$1.332,3 \times 10^{-15}$	$1.332,3 \times 10^{-15}$	$1.110,2 \times 10^{-16}$
TB	—	—	—	$1.554,3 \times 10^{-15}$	$1.554,3 \times 10^{-15}$	$6.661,3 \times 10^{-16}$	$6.661,3 \times 10^{-16}$	$1.554,3 \times 10^{-15}$
WP-All	—	—	—	—	0.304,7	$9.832,6 \times 10^{-07}$	0.000,2	$2.109,4 \times 10^{-15}$
WP-Eco	—	—	—	—	—	$1.718,7 \times 10^{-07}$	$6.944,0 \times 10^{-05}$	$2.109,4 \times 10^{-15}$
WP-Top-1	—	—	—	—	—	—	0.477,5	$1.443,3 \times 10^{-15}$
WP-Top-3	—	—	—	—	—	—	—	$1.443,3 \times 10^{-15}$
Zeit-All	—	—	—	—	—	—	—	—
<b>y-values</b>								
SZ-All	—	1.000,0	$6.661,3 \times 10^{-16}$	$3.774,8 \times 10^{-15}$	$3.774,8 \times 10^{-15}$	$1.086,4 \times 10^{-54}$	$2.445,3 \times 10^{-71}$	$1.443,3 \times 10^{-15}$
SZ-Eco	—	—	$1.221,2 \times 10^{-15}$	$1.554,3 \times 10^{-15}$	$1.554,3 \times 10^{-15}$	$1.332,3 \times 10^{-15}$	$1.332,3 \times 10^{-15}$	$1.110,2 \times 10^{-16}$
TB	—	—	—	$1.176,7 \times 10^{-09}$	$5.162,5 \times 10^{-14}$	$4.872,1 \times 10^{-11}$	$4.872,1 \times 10^{-10}$	$1.767,3 \times 10^{-10}$
WP-All	—	—	—	—	$6.728,1 \times 10^{-05}$	$3.734,6 \times 10^{-05}$	0.012,6	$7.908,8 \times 10^{-12}$
WP-Eco	—	—	—	—	—	$4.078,8 \times 10^{-11}$	$3.815,2 \times 10^{-06}$	$2.109,4 \times 10^{-15}$
WP-Top-1	—	—	—	—	—	—	0.000,9	$1.042,1 \times 10^{-12}$
WP-Top-3	—	—	—	—	—	—	—	$3.153,0 \times 10^{-14}$
Zeit-All	—	—	—	—	—	—	—	—

**TABLE 8** | *P*-values of the Kolmogorov-Smirnov goodness-of-fit test applied to the pairwise combinations of the *x* and *y* values of the distributions in **Figure 10**.

	SZ-All	SZ-Eco	TB	WP-All	WP-Eco	WP-Top-1	WP-Top-3	Zeit-All
<b>x-values</b>								
SZ-All	—	$1.332,3 \times 10^{-15}$	$1.776,4 \times 10^{-15}$	$3.774,8 \times 10^{-15}$	$3.774,8 \times 10^{-15}$	$3.835,6 \times 10^{-41}$	$2.193,2 \times 10^{-33}$	0.199,8
SZ-Eco	—	—	$5.596,8 \times 10^{-11}$	$1.554,3 \times 10^{-15}$	$1.554,3 \times 10^{-15}$	$1.332,3 \times 10^{-15}$	$1.332,3 \times 10^{-15}$	$1.110,2 \times 10^{-16}$
TB	—	—	—	$1.554,3 \times 10^{-15}$	$1.554,3 \times 10^{-15}$	$6.661,3 \times 10^{-16}$	$6.661,3 \times 10^{-16}$	$3.108,6 \times 10^{-15}$
WP-All	—	—	—	—	0.000,6	0.595,3	0.041,9	$2.109,4 \times 10^{-15}$
WP-Eco	—	—	—	—	—	0.001,7	0.471,6	$2.109,4 \times 10^{-15}$
WP-Top-1	—	—	—	—	—	—	0.047,2	$1.443,3 \times 10^{-15}$
WP-Top-3	—	—	—	—	—	—	—	$1.443,3 \times 10^{-15}$
Zeit-All	—	—	—	—	—	—	—	—
<b>y-values</b>								
SZ-All	—	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0
SZ-Eco	—	—	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0
TB	—	—	—	1.000,0	1.000,0	1.000,0	1.000,0	1.000,0
WP-All	—	—	—	—	1.000,0	1.000,0	1.000,0	1.000,0
WP-Eco	—	—	—	—	—	1.000,0	1.000,0	1.000,0
WP-Top-1	—	—	—	—	—	—	1.000,0	1.000,0
WP-Top-3	—	—	—	—	—	—	—	1.000,0
Zeit-All	—	—	—	—	—	—	—	—





on current level of a learner). Such questions obviously pertain to the encompassing methodological structure outlined in **Figure 1**, not just to the learning media. Studies combining educational and computational linguistic methods, as presented here, make it possible to derive assumptions of the effects of different types of texts used for learning on the learning outcome, include individual learner-internal influence factors, as most explicitly formulated in the *offer-use model* (Helmke, 2009). Accordingly, a most straightforward continuation of our approach is to implement an educational assessment of student learning, related to the examined threshold concepts. The computational linguistic assessment very likely has implications for text comprehension (Kintsch, 1988) and domain learning (Alexander, 2018). For instance using educational assessment, we can link findings on threshold concept profiles in texts to findings on learners' understandings of these threshold concepts, as evidenced in their assessment performance (Brückner and Zlatkin-Troitschanskaia, 2018). Such an assessment is necessary since there is no straightforward mapping between dictionary concepts and indexed concepts (a student's private understanding of lexical meanings), as mentioned in sections 3 and 4. Although such studies are future work (but see Mehler and Ramesh, 2019 for a formal learning assessment framework), a few points of departure can already be considered, as we do in the following.

As observed by Rincke (2010), there is a striking similarity of the acquisition of a special language with language acquisition in general (both include terminological and conceptual change). Hence, we might expect to find some empirical evidence relevant for the acquisition of threshold concepts in language acquisition studies. In this regard, Oakhill et al. (2003) show in a study on language development that word reading and text comprehension are dissociated. This implies that text comprehension and word decoding follow different developmental trajectories and can be taught at least to some degree independently. The acquisition of threshold concepts

proceeds at least on these two routes, meaning that developing respective understandings draws on text comprehension as well as on lexical definitions. This line of thought emphasizes the need for a semantic analysis of threshold concepts in business education which, as far as we know, is missing (see section 4.1.3). Furthermore, we may hypothesize that denser networks of threshold concepts pose higher requirements on word decoding, while looser networks pose higher requirements on text comprehension (very likely there is an interaction with text type which is discussed in subsection 6.3 below). Now given that from learning assessments (cf. **Figure 1**) we know that a student cohort has a better developed definitional than applicational competence in dealing with threshold concepts, a deliberate choice of learning media can foster or balance this asymmetry in competence.

Text comprehension is not only based on memory processes but also on constructionist processes (van den Broek et al., 2005). The latter can, for instance, arise due to associations bound up with readers' indexed concepts. This includes personal preferences as well as all sorts of top-down processes. Constructionist aspects of comprehension are bound up with learners' everyday language and prior knowledge and experience. If we liken the acquisition of a specialized language to second language acquisition, this implies that also the first or prior language(s) should be taken into account (cf. Shanahan et al., 2006). Here we meet advice from educational research, namely *to regard everyday language and specialized language as respectively developable in their own right and to address them both in class* ("Alltags- und Fachsprache als je für sich entwicklungsfähig anzusehen und im Unterricht zu thematisieren" [translated by AL]) (Rincke, 2010, p. 235). The everyday language competence can only be tapped in the classroom, if one can classify the text resources according to their language level. On a large scale this can only be done with the help of automatic methods.

The prior knowledge of learners also plays a role in reading hypertexts, such as Wikipedia articles. Interestingly, using hypertexts as a learning resource can be advantageous in particular for informed learners, since the hypertext structure allows them to exert a strategic reading processes (Salmerón et al., 2006). That is, online (hypertext) resources, such as Wikipedia can enrich the learning landscape for formal education (as they already do as a matter of fact by virtue of student selection, cf. subsection 2.2).

This poses the question of reliability of Wikipedia<sup>17</sup>. Wikipedia articles seem to be reliable in general (e.g., Wilkinson and Huberman, 2007). However, with regard to specific topics, such as *respiration in medicine* Wikipedia turned out to be an insufficiently reliable resource for learning (Meier, 2008; Azer, 2015)<sup>18</sup>. Accordingly, a qualitative assessment is needed in order to find evidence on which of these

<sup>17</sup>We are thankful to an anonymous reviewer for emphasizing this issue.

<sup>18</sup>To be fair, since Wikipedia (and related specialized Wikis) is a highly dynamic resource the situation may have changed already since the time of publication of the study. We know of no recent replication, however.

opposing sides Wikipedia's economic articles belong—there is no *a priori* reason to exclude newspapers from such an assessment<sup>19</sup>.

### 6.3. Text Types in and for Learning

Wikipedia, newspapers and textbooks are all examples of different text types. These text types differ with respect to narrative structure, content, target group, and many more properties. In particular the didactic structure of different learning media genres interact with conceptual change, as discussed in the following for each text type used in our study.

1. *Interpretation variant I: Textbooks are optimal.* Assuming that textbooks are optimized for the transfer of specialized information in higher education, the question arises as to the significance of our findings. Section 5 indicates that the network of threshold concepts based on the usage regularities confirmed in Wikipedia is much *denser* than in the case of newspaper corpora or even textbooks. Metaphorically speaking, the encyclopedia-based association network manifests a more densely distributed “matter” of much more closely associated conceptual units. Higher density and stronger associativity also mean a higher degree of confirmation and thus stability, because the underlying associative relations are more strongly confirmed by co-occurrences that can actually be observed in many sentence windows. Stability is here a simple consequence of the fact that a change of such strongly confirmed associations would require a higher amount of textual information contradicting the already confirmed associations by aiming into other directions of associations instead. This amount probably would be equal to the amount of the original textual information, which underlies the stabilized associations. According to this interpretation we may say that *encyclopedic textual information seems to over-confirm the associations of threshold concepts*. Conversely, the network of associations based on newspaper corpora seems to be *under-confirmed* and therefore too unstable: by positioning the same concepts in ever new contexts, their association relations virtually fan out, so that each individual association is far less confirmed. New textual information then does not necessarily confirm what already exists, but rather refers to ever new possibilities of association. In the middle of these two extreme cases we find the association network resulting from textbook corpora. Under the interpretation that this network is optimally organized, we find that textbooks balance the under-confirmation induced by newspaper corpora with the over-confirmation by encyclopedias in terms of a fluent equilibrium: an optimum, so to speak, as a balance of firstness and confirmation according to the notion of pragmatic information (von Weizsäcker, 1974). Textbooks are organized around threshold concepts in such a way that their readers can learn

the targeted concepts with sufficient conceptual density (outcome perspective), but not in such a way that they would not be able to recontextualize them or transfer them between different contexts (process perspective) whereby these recontextualizations do not excessively disturb and consequently do not dissolve the previously confirmed associations of threshold concepts. This is supported by the directional way in which textbooks guide learners through the learning process by providing them with an epistemic structure of the discipline (Dalimunte and Pramoolsook, 2020).

2. *Interpretation variant II: Encyclopedic texts are optimal.* As conjectured in subsection 5.4, the dense network of threshold concepts observed in Wikipedia is probably due the fact that Wikipedia is an encyclopedic resource and as such introduces special threshold concept terms by means of definitions. In this sense, Wikipedia represents the result state of threshold concept knowledge. In contrast, a textbook often *develops* a concept and takes a more process-oriented approach (see, e.g., Dalimunte and Pramoolsook, 2020). The semantic flavors of both approaches have already been observed in the sample sketch in section 3. Such differences, we argue, are finally reflected in different network densities. Teleologically understood, as in domain learning (Alexander et al., 1995), a result is the goal of a process. A process can be conceived as a succession of (intermediate) states (cf., e.g., Fernando, 2011). In section 3 we suggested to connect conceptual change in particular to the update operations *greater level of abstraction* and *shifted vantage point* (Chi and Ohlsson, 2005). That is, the intermediate states are related in terms of semantic update operations. A consequence then is that a successor state is more developed than its predecessor state. Furthermore, the hypothesis is that update operations apply at a larger range to looser linked concept networks than to denser linked ones. To put it another way: denser networks are closer to a result state and make further conceptual changes more unlikely, and rightly so since result state are closer to an optimum. In order to make this line of argumentation and modeling more precise, however, the need for a semantic characterization of mental updates and the differences between different kinds of updates are required.
3. *Interpretation variant III: Newspaper articles are optimal.* Newspapers also offer a wide range of potential for learning despite the low frequency of compound-specific and sentence-specific threshold concepts and a lower semantic density compared to textbooks and Wikipedia. Depending on the curricular goal, e.g., whether the focus is on economic education in the sense of general maturity for social participation or on the professional expertise of an economist, alternative uses may be suitable. A lower density of the threshold web in learning media, as was evident in the present findings, leaves room for a more in-depth examination of individual threshold conceptions by learners and can promote their motivation and understanding. Newspapers are by no means only complementary materials in economics courses. They offer the possibility of an active application

<sup>19</sup>Since textbooks are submitted to a quality control procedure, there is an *a priori* reason to exclude textbooks from a further quality assessment.

of what has been learned in the course due to the potential of the articles' alternative interpretations, current topics, events, and ever new contexts (see the articles in Hoyt and McGoldrick, 2012). The looser density of threshold concepts promotes newspapers to be used as an introductory learning opportunity (Helmke and Schrader, 2008). As Dalimunte and Pramoolsook (2020) note, despite the central structure textbooks provide for teaching, their texts are often more difficult for novices to read, so that newspapers can provide a first access to subject-specific learning in economics. Depending on the objective, e.g., a critical examination of the definition of concepts in newspaper articles, a certain amount of prior knowledge of the learning group is required. Newspapers corpora of *SZ* and *Zeit* in particular often require reading skills and prior subject-specific knowledge, so that they can also be used effectively by lecturers during their classes (McEachern, 2012). Newspapers can also be useful for cooperative forms of learning (McGoldrick et al., 2010), e.g., for jointly comparing and evaluating threshold concepts in different newspaper articles from different corpora. In addition, however, they can not only serve as exemplary texts and information materials, but can also be a central object for the design of lessons. For example, in his conclusion on the analysis of learning media, that are not originally developed for educational purposes, Croushore (2012, p. 636) writes: "[...] instructors of money and banking must be on constant alert for changes in the material. While this may seem difficult, these constant changes actually make the course easy to teach because nearly every day's newspaper provides new course material." Therefore, assuming that newspapers have a lower threshold concepts density, it seems reasonable to expect that more diverse associations are possible for the learner. In other words: Since the network is less stable, teachers have more freedom to design their courses.

4. *Interpretation variant IV: Synthetic view, or mixture model.* The previous bullet points provided reasons that each learning text type can be considered "optimal." But optimal with reference to what? Adopting the view that learning is a process (for a recent affirmation of this (somewhat obvious) view see Dalimunte and Pramoolsook, 2020) that conceptually develops in the triangle between lexical, dictionary and indexed concepts, among others, as outlined in section 4, one also adopts a *dynamic* rather than a *static* perspective. A dynamic perspective allows for a synthetic view on learning media since it conceives learning in its ecological niche. In relation to students' prior knowledge, current interest, curricular goals and teachers' content-related and pedagogical focus each text type can be used for its respective strengths. In the end, thus, a synthetic view amounts to an adjusted and combined approach. However, in order to be of value, it needs to be complemented with an assessment of learning situations in order to gain evidence about the most suitable learning resources for a given learning situation. Since this issue leads to the topic of this special issue, we want to elaborate on it in the subsequent section.

## 6.4. Comparative Media Analysis of Threshold Concept Webs and (Online) Information Processing and Learning

We argued that a dedicated linguistic analysis of textual learning media used in economics education is necessary due to the increasing digitalization of teaching and learning in economics. Digitalization is constantly increasing the range of learning media that can potentially be used by teachers and lecturers (Johinke and Di Lauro, 2020). The aforementioned, more frequent use of Wikipedia by students in economic learning contexts (Freire and Li, 2016) and the increasing digitization of textbooks and distribution as Open Educational Resources (Fischer et al., 2017) are facilitating computer-based and internet-based learning. These multimedia environments afford learning based on multiple representations (Mayer, 2014). In order to support teachers in their decisions for selecting media for teaching-and-learning purposes, it is necessary to apply a content quality criterion to compare media used for learning. For this purpose, the linguistic properties of threshold concepts were compared between several corpora from Wikipedia, the business-related newspaper sections of *Süddeutsche Zeitung* and *Zeit*, and 14 business and economics textbooks. Given the large amounts of learning resources in digital media and the associated comprehensive (corpus) data sets, the computational linguistic approach is advantageous in comparison to already established qualitative content-analytical procedures, in order to provide teachers with general and innovative information on the usefulness of media for learning in a condensed form. Linguistic procedures, which explore the morpho-syntactic structure of the underlying threshold concepts are particularly suitable, since the primary access of novices and beginners to economics is text-based. Text-based introduction to threshold is more commonly used than via diagrams and other visualizations (Tinkler and Woods, 2013). With the exception of studies on readability (Tinkler and Woods, 2013), word frequency counts (Leet and Lopus, 2003) or genre-specific analyses of a few textbooks (Dalimunte and Pramoolsook, 2020), there are no comparisons of digital media of different types and genres in economics education. The present analysis thus provides an important comparison of different media types and implications for their use in digital learning contexts. The use of threshold concept webs for the comparative media analysis of learning sources (Helmke and Schrader, 2008) is often a prerequisite for studies on learning success which rely on mental association patterns, such as those found in the studies by Davies and Mangan (2007), Vidal et al. (2015), and Ivan Montiel and Antolin-Lopez (2020). In this study, we found that the lexical and semantic density of threshold concepts is higher in Wikipedia than in textbooks and newspapers. This analysis of subject-specific concepts goes beyond the density analyses of pronouns found in textbook analyses of foreign language research (e.g., Kong, 2009) and offers a number of implications for the initiation and design of learning processes. On the one hand, threshold concept density can be an advantage for students who want to learn about a content area in a short time (Meier, 2008; Freire and Li, 2016), on the other hand, students need not only basic skills for



researching and evaluating web resources, but, especially with the difficulty of learning new threshold concepts (see section 1), they also need prior subject knowledge (Sender, 2017). Nevertheless, especially in introductory economics courses, Wikipedia can also stimulate creative learning processes, because on the one hand the platform includes references to external literature or alternative perspectives and thus can generate interest in different topics (Meier, 2008). On the other hand, due to the density of concepts and the editability of content, it also offers opportunities for students to critically reflect on content, to review existing articles (Johinke and Di Lauro, 2020) or to check their own misconceptions (Freire and Li, 2016). In turn, a lower density of threshold concepts, as is the case with newspapers in particular, does not imply a lower quality of newspapers for didactic purposes. Threshold concepts are special subject-specific concepts that require a gradual development of expertise over several phases (Davies and Mangan, 2007). This development requires examples, practical and professional applications in which the concepts are didactically embedded. The more variable the context is regularly updating newspapers reporting on changing topics, the more application possibilities are offered to the learner for an in-depth examination of acquired threshold concepts. In addition, the disciplinary and semantic density is not too high, so that even learners with little previous knowledge can approach the threshold concepts and develop initial ideas, which may need to be corrected or refined over time. Furthermore, lower specialization and stronger contextualization as well as a change of media from textbook to newspaper afford didactic advantages and enable learning through multiple-representations, which can be used in a targeted manner, especially in phases of learner activation and topic introducing, to cognitively activate and motivate learners. However, it should be kept in mind that newspapers are subject to daily change. The presented study is fundamental for future research on information and learning processes. It offers a number of links for further research that can be taken into account in conventional educational assessments. For example, it could be investigated how the density of threshold concepts between the different media types affects learning success, whether students with varying levels of prior knowledge benefit more or less from certain media, or which specificity (e.g., compound or sentence specificity) affects the learning process and how. These central linguistic characteristics can in turn help to determine how textbooks could be structured, which language use would support teaching or how closely threshold concepts should be linked to be as conducive to learning from learning media as possible.

## 7. CONCLUSION

The computational linguistic perspective adopted in the present contribution pursues an orientation which, in terms of educational research on threshold concepts, has two special features. On the one hand, it complements content analyses, which are classically used to analyze textbooks, protocols, or other textually and graphically represented materials in order to work out education-related meanings from the materials

(e.g., Krippendorff, 2013). The often tedious and lengthy manual evaluation with only a limited number of documents and the corresponding susceptibility to errors is as a matter of fact limited to a small amount of data. Computational linguistic analyses, to the contrary, can process huge corpora. Secondly, so-called *utilization-of-learning-opportunities* models are used to model the mechanisms of action of teaching-learning arrangements in educational research (e.g., Braun et al., 2014). These models show the interactions between learning-relevant aspects in terms of input-process-output paths. Very often learning outcomes are analyzed in connection with different input factors (e.g., socio-economic status, gender, intelligence, self-assessed use of learning media). Significantly less frequently, however, the learning potentials of the respective learning environments or learning materials are considered independently of a learner's assessment. With the computational linguistic approach presented here, especially the learning media that are used as input into the learning processes are processed on a large scale and thus a description of the learning environment is presented that can be considered in informal as well as formal learning processes. Ultimately learning, the meaning of threshold concept expressions and their use in text resources are embraced within the contour of an emerging research program—encompassing specialized vocabularies, learning and education, and computational linguistics—in terms of mental, referential and differential meanings. The latter two (referential and differential meanings) are used in order to derive hypotheses concerning formal and informal learning contexts with respect to a special class of expressions, *viz.* threshold concepts. A second focus was the development of a computational linguistic model for operationalizing threshold concepts for the analysis of learning resources. In this context, we developed the notion of a Threshold Concept Network (TCN) and quantified it by means of alpha-cuts, taking into account the “web of threshold concepts” (Davies and Mangan, 2007). In this way, we were able to prove an exceptional status of threshold concepts in textbooks, at least at the node level. The main result was that formal and informal resources can indeed be distinguished in terms of their threshold concepts' profiles. Furthermore, Wikipedia turns out to be a first class formal learning resource. Continuing this line of research will include at least the following steps: the methodological considerations discussed in subsection 6 are to be addressed. A lexical semantic analysis of threshold concepts is due. And, most importantly, our findings have to be tied back to education assessments of learners. Furthermore, experimental studies have to be designed that investigate systematically the impact of different resources on learning. Very often experimental studies are developed on assumptions that have not been tested themselves. On the basis of the computational linguistic assessment, however, it is possible to develop more specific questions. Most notably, the threshold concept acquisition of learners can be compared depending on the media to learn (e.g., Wikipedia vs. textbook vs. daily newspaper, and their interaction and complementary uses)—whereby, of course, the corresponding media competencies and information literacy or other (intellectual) characteristics must also be controlled (Vernooij, 2000). The assessments from the study presented here



provide a starting point for such experiments which in turn would round out the emerging research program we sketched.

## DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available due to copyright restrictions, the newspaper and the textbook corpora are not publicly available. Wikipedia can be obtained via Wikipedia dumps. Further queries regarding the material and analysis presented here should be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

AL mainly has written subsection 2.2, and sections 3 and 4, and performed the Kolmogorov-Smirnov goodness-of-fit tests and the power law fitting (Tables 2–8). AM designed the computational linguistic measurement procedure for threshold concepts, implemented the corresponding network analyses, has written almost all parts of section 5 and generated Figures 5–10. Sections 6 and 7 have been jointly written by SB, AM, and AL, with the exception of subsection 6.4, which is mainly due to SB. GA carried out the preprocessing of the corpora and the word embeddings. SB selected the textbook corpus and mainly has written section 1, the preamble of section 2 and subsection 2.1. TU calculated the compound distributions and the threshold

concept networks and produced Table 1. All authors contributed to the article and approved the submitted version.

## FUNDING

This work on this article by AL was partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the program *Investissements d'Avenir* (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris—ANR-18-IDEX-0001. The work of SB and AL on this article is partly supported by the RMU project PLATO. RMU is an initiative of the Rhein-Main Universities Johannes Gutenberg-Universität Mainz (JGU), TU Darmstadt and the Goethe University Frankfurt.

## ACKNOWLEDGMENTS

We were thankful to two reviewers for their helpful comments. They helped to improve the presentation and discussion of our topic a lot. We are also thankful to Jasmin Schlax, Dimitri Molerov, and Andreas Falke for commenting on a near-final version of the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2020.578475/full#supplementary-material>

## REFERENCES

- Alexander, P. A. (2018). "Into the future. A prospective look at the model of domain learning," in *The Model of Domain Learning: Understanding the Development of Expertise, Chapter 10*, eds H. Fives and D. L. Dinsmore (New York, NY: Routledge), 195–214. doi: 10.4324/9781315458014-12
- Alexander, P. A., Jetton, T. L., and Kulikowich, J. A. (1995). Interrelationship of knowledge, interest, and recall: assessing a model of domain learning. *J. Educ. Psychol.* 87, 559–575. doi: 10.1037/0022-0663.87.4.559
- Alstott, J., Bullmore, E., and Plenz, D. (2014). powerlaw: A Python package for analysis of heavy-tailed distributions. *PLoS ONE* 9:e95816. doi: 10.1371/journal.pone.0085777
- Asher, N. (1993). *Reference to Abstract Objects in Discourse. Number 50 in Studies in Linguistics and Philosophy*. Dordrecht: Kluwer Academic Publishers.
- Asher, N., and Lascarides, A. (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.
- Azer, S. A. (2015). Is Wikipedia a reliable learning resource for medical students? Evaluating respiratory topics. *Adv. Physiol. Educ.* 39, 5–14. doi: 10.1152/advan.00110.2014
- Barsalou, L. W. (1993). "Flexibility, structure, and linguistic vagary in concepts: manifestations of a compositional system of perceptual symbols," in *Theories of Memory, Chapter 3*, eds A. F. Collins, M. A. Conway, and P. E. Morris (Hillsdale, NJ: Lawrence Erlbaum Associates), 29–101. doi: 10.4324/9781315782119-3
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behav. Brain Sci.* 22, 577–660. doi: 10.1017/S0140525X99002149
- Barsalou, L. W., Yeh, W., Luka, B. J., Olseth, K. L., Mix, K. S., and Wu, L.-L. (1993). "Concepts and meaning," in *Chicago Linguistics Society 29: Papers from the Parasessions on Conceptual Representations*, eds K. Beals, G. Cooke,
- D. Kathman, K. E. McCullough, S. Kita, and D. Testen (Chicago, IL: Chicago Linguistics Society), 23–61.
- Barwise, J., and Perry, J. (1983). *Situations and Attitudes. The David Hume Series on Philosophy and Cognitive Science Reissues*. Stanford, CA: CSLI Publications.
- Blum, U. (2017). *Grundlagen der Volkswirtschaftslehre*. Berlin; Boston, MA: De Gruyter Oldenbourg.
- Braun, E., Weiß, T., and Seidel, T. (2014). "Lernumwelten in der Hochschule," in *Pädagogische Psychologie: Mit Online-Materialien zum Download*, 6th Edn., eds T. Seidel and A. Krapp (Weinheim: Beltz), 433–454.
- Brooks, D. C. (2016). *ECAR Study of Undergraduate Students and Information Technology*. Available online at: <https://er.educause.edu/~media/files/library/2016/10/ers1605.pdf?la=en>
- Brückner, S., and Lücking, A. (2019). "Computerlinguistische Analyse des Schwellenkonzeptansatzes in der Wirtschaftsdidaktik," in *Talk at the Jahrestagung Sektion Berufs- und Wirtschaftspädagogik* (Graz: Karl-Franzens-University).
- Brückner, S., and Zlatkin-Troitschanskaia, O. (2018). "Threshold concepts for modeling and assessing higher education students' understanding and learning in economics," in *Assessment of Learning Outcomes in Higher Education: Cross-National Comparisons and Perspectives*, eds O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, and C. Kuhn (Wiesbaden: Springer International Publishing), 103–121. doi: 10.1007/978-3-319-74338-7\_6
- Bücking, S. (2010). "German nominal compounds as underspecified names for kinds," in *New Impulses in Word-Formation*, ed S. Olsen (Hamburg: Buske), 253–281.
- Chi, M. T. H., and Ohlsson, S. (2005). "Complex declarative learning," in *The Cambridge Handbook of Thinking and Reasoning, Chapter 16*, eds K. J. Holyoak and R. G. Morrison (Cambridge, NY: Cambridge University Press), 371–399.
- Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.* 51, 661–703. doi: 10.1137/070710111

- Cohen, A., and Erteschik-Shir, N. (2002). Topic, focus, and the interpretation of bare plurals. *Nat. Lang. Semant.* 10, 125–165. doi: 10.1023/A:1016576614139
- Cook, S. W., and Goldin-Meadow, S. (2006). The role of gesture in learning: do children use their hands to change their minds? *J. Cogn. Dev.* 7, 211–232. doi: 10.1207/s15327647jcd0702\_4
- Cousin, G. (2008). “Threshold concepts: old wine in new bottles or a new form of transactional curriculum inquiry,” in *Threshold Concepts Within the Disciplines*, eds R. Land, J. H. Meyer, and J. Smith (Rotterdam: Sense Publishers), 261–272. doi: 10.1163/9789460911477\_020
- Croushore, D. (2012). “Using real-world applications to policy and everyday life to teach money and banking,” in *International Handbook on Teaching and Learning in Economics. Elgar Original Reference*, eds G. M. Hoyt and K. McGoldrick (Cheltenham: Edward Elgar), 628–637. Retrieved from: <http://www.elgaronline.com/view/9781848449688.xml>
- Crystal, D. (1997). *The Cambridge Encyclopedia of Language*, 2nd Edn. Cambridge: Cambridge University Press.
- Dalimunte, A. A., and Pramoolsook, I. (2020). Genres classification and generic structures in the English language textbooks of economics and Islamic economics in an Indonesian university. *Lang. Educ. Acquisit. Res. Netw.* 13, 1–19.
- Davies, P., and Mangan, J. (2007). Threshold concepts and the integration of understanding in economics. *Stud. High. Educ.* 32, 711–726. doi: 10.1080/03075070701685148
- de Ruiter, J. P. (2007). Postcards from the mind: the relationship between speech, imagistic gesture, and thought. *Gesture* 7, 21–38. doi: 10.1075/gest.7.1.03rui
- de Saussure, F. (1916). *Course de linguistique générale*. Lausanne; Paris: Payot.
- Demberg, V., Keller, F., and Koller, A. (2013). Incremental, predictive parsing with psycholinguistically motivated tree-adjointing grammar. *Comput. Linguist.* 39, 1025–1066. doi: 10.1162/COLI\_a\_00160
- Devetak, I., and Vogrinc, J. (2013). “The criteria for evaluating the quality of the science textbooks,” in *Critical Analysis of Science Textbooks*, Vol. 10, ed M. S. Khine (Dordrecht: Springer Netherlands), 3–15. doi: 10.1007/978-94-007-4168-3\_1
- Fernando, T. (2011). Constructing situations and time. *J. Philos. Logic* 40, 371–396. doi: 10.1007/s10992-010-9155-1
- Fillmore, C. J., Lee-Goodman, R. R., and Rhomieux, R. (2012). “The FrameNet construction,” in *Sign-Based Construction Grammar, Number 193 in CSLI Lecture Notes, Chapter 7*, eds H. C. Boas and I. A. Sag (Stanford, CA: CSLI Publications), 309–372.
- Fischer, L., Ernst, D., and Mason, S. (2017). Rating the quality of open textbooks: how reviewer and text characteristics predict ratings. *Int. Rev. Res. Open Distrib. Learn.* 18, 142–154. doi: 10.19173/irrod.v18i4.2985
- Frege, G. (1892). Über Sinn und Bedeutung. *Z. Philos. Philos. Krit.* 100, 25–50.
- Freire, T., and Li, J. (2016). Using Wikipedia to enhance student learning: a case study in economics. *Educ. Inform. Technol.* 21, 1169–1181. doi: 10.1007/s10639-014-9374-0
- Ginzburg, J. (2012). *The Interactive Stance: Meaning for Conversation*. Oxford: Oxford University Press.
- Haab, T. C., Schiff, A., and Whitehead, J. C. (2012). “Economics blogs and economic education,” in *International Handbook on Teaching and Learning in Economics. Elgar Original Reference*, eds G. M. Hoyt and K. McGoldrick (Cheltenham: Edward Elgar), 167–173. Retrieved from: <http://www.elgaronline.com/view/9781848449688.xml>
- Hatt, L. (2018). Threshold concepts in entrepreneurship-the entrepreneurs’ perspective. *Educ. Train.* 60, 155–167. doi: 10.1108/ET-08-2017-0119
- Heim, I. (2002). “File change semantics,” in *Formal Semantics: The Essential Readings, Number 1 in Linguistics: The Essential Readings, Chapter 9*, eds P. Portner and B. H. Partee (Oxford; Malden, MA: Wiley-Blackwell), 223–248.
- Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett-Kallmeyer.
- Helmke, A., and Schrader, F.-W. (2008). Merkmale der Unterrichtsqualität: Potential, Reichweite und Grenzen. *Seminar* 14, 17–47.
- Hemati, W., Uslu, T., and Mehler, A. (2016). “TextImager: a distributed UIMA-based system for NLP,” in *Proceedings of the COLING 2016 System Demonstrations. Federated Conference on Computer Science and Information Systems* (Osaka).
- Hjelmslev, L. (1961). *Prolegomena to a Theory of Language*, 7th Edn. Madison, WI: University of Wisconsin Press.
- Hoadley, S., Tickle, L., Wood, L. N., and Kyng, T. (2015). Threshold concepts in finance: conceptualizing the curriculum. *Int. J. Math. Educ. Sci. Technol.* 46, 824–840. doi: 10.1080/0020739X.2015.1011244
- Hoffmann, L., Kalverkämper, H., Wiegand, H. E., Galinskim, C., and Hullen, W. (Eds.). (1998). *Fachsprachen/Languages for Special Purposes: Ein Internationales Handbuch Zur Fachsprachenforschung und Terminologiewissenschaft*. Berlin; Boston, MA: De Gruyter, Inc.
- Hofhues, S. (2016). “Informelles Lernen mit digitalen Medien in der Hochschule,” in *Handbuch Informelles Lernen*, ed M. Rohs (Wiesbaden: Springer Fachmedien), 529–546. doi: 10.1007/978-3-658-05953-8\_28
- Hoyt, G. M., and McGoldrick, K. (Eds.). (2012). *International Handbook on Teaching and Learning in Economics. Elgar Original Reference* (Cheltenham: Edward Elgar). Retrieved from: <http://www.elgaronline.com/view/9781848449688.xml>
- Hu, C., and Gao, H. (2019). Nouns and nominalizations in economics textbooks. *Lang. Context Text Soc. Semiot. Forum* 1, 288–312. doi: 10.1075/langct.00012.hu
- Ivan Montiel, P. J. G., and Antolin-Lopez, R. (2020). What on earth should managers learn about corporate sustainability? A threshold concept approach. *J. Bus. Ethics* 162, 857–880. doi: 10.1007/s10551-019-04361-y
- Jackendoff, R. (1983). *Semantics and Cognition*. Cambridge, MA: MIT Press.
- Jackendoff, R. (1991). *Semantic Structures. Number 18 in Current Studies in Linguistics*. Cambridge, MA: MIT Press.
- Jackendoff, R. (2002). *Foundations of Language. Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.
- Jadin, T., and Zöserl, E. (2009). Informelles Lernen mit Web-2.0-Medien. *Bildungsforschung* 6, 41–61.
- Johinke, R., and Di Lauro, F. (2020). Wikipedia in higher education: practice what you teach. *Stud. High. Educ.* 45, 947–949. doi: 10.1080/03075079.2020.1763014
- Kamp, H., and Reyle, U. (1993). *From Discourse to Logic*. Dordrecht: Kluwer Academic Publishers.
- Karttunen, L. (1969). “Discourse referents,” in *Proceedings of the 1969 Conference on Computational Linguistics, COLING ’69* (Stockholm), 1–38. doi: 10.3115/990403.990487
- Keller, R. (1995). *Zeichentheorie. Zu einer Theorie semiotischen Wissens*. Tübingen: Francke.
- Kilgour, P., Reynaud, D., Northcote, M., McLoughlin, C., and Gosselin, K. P. (2019). Threshold concepts about online pedagogy for novice online teachers in higher education. *High. Educ. Res. Dev.* 38, 1417–1431. doi: 10.1080/07294360.2018.1450360
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. Rev.* 95, 163–182. doi: 10.1037/0033-295X.95.2.163
- Klein, U., and Kracht, M. (2014). “Notes on disagreement,” in *Approaches to Meaning. Composition, Values, and Interpretation, Number 32 in Current Research in the Semantics/Pragmatics Interface*, eds D. Gutzman, J. Köpping, and C. Meier (Leiden: Brill), 276–305. doi: 10.1163/9789004279377\_013
- Knight, C., and Pryke, S. (2012). Wikipedia and the university, a case study. *Teach. High. Educ.* 17, 649–659. doi: 10.1080/13562517.2012.666734
- Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.
- Kong, K. (2009). A comparison of the linguistic and interactional features of language learning websites and textbooks. *Comput. Assist. Lang. Learn.* 22, 31–55. doi: 10.1080/09588220802613799
- Kricks, K., Mittelstädt, E., and Lening, A. (2013). Schwellenkonzepte und Phänomenografie. Explorative Studie zur Messung von Unterschieden im ökonomischen Verstehen. *Z. ökonom. Bild.* 2, 17–41. doi: 10.7808/zfoeb.1.2.74
- Krifka, M. (2003). “Bare NPs: Kind-referring, indefinites, both, neither?” in *Proceedings of the 13th Conference on Semantics and Linguistic Theory, SALT XIII*, eds R. Young and Y. Zhou (Ithaca, NY: Cornell University), 180–203. doi: 10.3765/salt.v13i0.2880
- Kripke, S. A. (2011). *Philosophical Troubles: Collected Papers*, Vol. 1. New York, NY: Oxford University Press.
- Krippendorff, K. (2013). *Content Analysis: An Introduction to Its Methodology*, 3rd Edn. Thousand Oaks, CA: SAGE Publications, Inc.

- Lamb, P., Hsu, S.-W., and Lemanski, M. (2019). A threshold concept and capability approach to the cross-cultural contextualization of Western management education. *J. Manag. Educ.* 44, 101–120. doi: 10.1177/1052562919851826
- Langacker, R. W. (2013). *Essentials of Cognitive Grammar*. Oxford, NY: Oxford University Press.
- Leet, D. R., and Lopus, J. S. (2003). *A Review of High School Economics Textbooks*. Available online at: <https://ssrn.com/abstract=381760>
- Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Riv. Linguist.* 20, 1–31.
- Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Trans. Assoc. Comput. Linguist.* 3, 211–225. doi: 10.1162/tacL\_a\_00134
- Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.
- Lim, S. (2009). How and why do college students use Wikipedia? *J. Am. Soc. Inform. Sci. Technol.* 60, 2189–2202. doi: 10.1002/asi.21142
- Link, G. (1983). “The logical analysis of plurals and mass terms: a lattice-theoretic approach,” in *Meaning, Use, and Interpretation of Language, Grundlagen der Kommunikation und Kognition/Foundations of Communication and Cognition*, eds R. Bäuerle, C. Schwarze, and A. von Stechow (Berlin: de Gruyter), 302–323. doi: 10.1515/9783110852820.302
- Linsky, L. (1971). “Reference and referents,” in *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*, eds D. D. Steinberg and L. A. Jakobovitz (Cambridge: Cambridge University Press), 76–85.
- Lodico, M. G., Spaulding, D. T., and Voegtli, K. H. (2006). *Methods in Educational Research: From Theory to Practice*. San Francisco, CA: Jossey-Bass.
- Lucas, U., and Mladenovic, R. (2009). The identification of variation in students’ understandings of disciplinary concepts: the application of the SOLO taxonomy within introductory accounting. *High. Educ.* 58, 257–283. doi: 10.1007/s10734-009-9218-9
- Lücking, A. (2019). “Dialogue semantics: from cognitive structures to positive and negative learning,” in *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)*, ed O. Zlatkin-Troitschanskaia (Cham: Springer), 197–205. doi: 10.1007/978-3-030-26578-6\_15
- Maurer, M., Schemer, C., Zlatkin-Troitschanskaia, O., and Jitomirski, J. (2019). “Positive and negative media effects on university students’ learning: preliminary findings and a research program,” in *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)*, ed O. Zlatkin-Troitschanskaia (Cham: Springer), 109–119. doi: 10.1007/978-3-030-26578-6\_8
- Mayer, R. E. (Eds.). (2005). “Introduction to multimedia learning,” in *The Cambridge Handbook of Multimedia Learning* (Cambridge, NY: Cambridge University Press), 1–16. doi: 10.1017/CBO9780511816819.002
- Mayer, R. E. (Eds.). (2014). “Cognitive theory of multimedia learning,” in *The Cambridge Handbook of Multimedia Learning* (Cambridge, NY: Cambridge University Press), 43–71.
- McEachern, W. A. (2012). “Macroeconomic principles are still relevant and still important,” in *International Handbook on Teaching and Learning in Economics. Elgar Original Reference*, eds G. M. Hoyt and K. McGoldrick (Cheltenham: Edward Elgar), 413–422. Retrieved from: <http://www.elgaronline.com/view/9781848449688.xml>
- McGoldrick, K., Rebelein, R., Rhoads, J., and Stockly, S. (2010). “Making cooperative learning effective for economics,” in *Teaching Innovations in Economics*, eds M. Salemi and W. Walstad (Cheltenham; Northampton, MA: Edward Elgar), 65–94. doi: 10.4337/9780857930620.00011
- Mehler, A., Gleim, R., Gaitsch, R., Uslu, T., and Hemati, W. (2020a). From topic networks to distributed cognitive maps: Zipfian topic universes in the area of volunteered geographic information. *Complexity* 4, 1–47. doi: 10.1155/2020/4607025
- Mehler, A., Hemati, W., Welke, P., Konca, M., and Uslu, T. (2020b). Multiple texts as a limiting factor in online learning: quantifying (dis-)similarities of knowledge networks. *Front. Educ.* 5:206. doi: 10.3389/educ.2020.562670
- Mehler, A., Jussen, B., Geelhaar, T., Henlein, A., Abrami, G., Baumartz, D., et al. (2020c). The Frankfurt Latin Lexicon. From morphological expansion and word embeddings to SemioGraphs. *Stud. Saggi Linguist.* 58, 121–155. doi: 10.4454/ssl.v58i1.276
- Mehler, A., and Ramesh, V. (2019). *TextInContext: On the Way to a Framework for Measuring the Context-Sensitive Complexity of Educationally Relevant Texts—A Combined Cognitive and Computational Linguistic Approach*. Cham: Springer International Publishing.
- Meier, S. (2008). Is Wikipedia a credible source for undergraduate economics students? *Major Themes Econ.* 10, 79–105.
- Meyer, J. H., and Land, R. (2005). Threshold concepts and troublesome knowledge (2). Epistemological considerations and a conceptual framework for teaching and learning. *High. Educ.* 49, 373–388. doi: 10.1007/s10734-004-6779-5
- Meyer, J. H., and Land, R. (2013). “Threshold concepts and troublesome knowledge (1): linkages to ways of thinking and practising within the disciplines,” in *Improving Student Learning. Theory and Practice-Ten Years On*, ed C. Rust (Oxford: Oxford Centre for Staff and Learning Development), 412–424.
- Meyer, J. H., and Land, R. (Eds.). (2006). “Threshold concepts and troublesome knowledge: an introduction,” in *Overcoming Barriers to Student Understanding: Threshold Concepts and Troublesome Knowledge* (London: Routledge), 3–18. doi: 10.4324/9780203966273
- Meyer, J. H., and Shanahan, M. (2003). “The troublesome nature of a threshold concept in economics,” in *Paper presented to the 10th Conference of the European Association for Research on Learning and Instruction (EARLI)* (Padova).
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*.
- Miller, G. A., and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Lang. Cogn. Process.* 6, 1–28. doi: 10.1080/01690969108406936
- Millikan, R. G. (1984). *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.
- Motos, R. M. (2011). “The role of interdisciplinary in lexicography and lexicology,” in *New Approaches to Specialized English Lexicology and Lexicography*, ed I. B. Fernández (Newcastle upon Tyne: Cambridge Scholars Publishing), 3–13.
- Mumm, M. (2015). *Kosten- und Leistungsrechnung. Internes Rechnungswesen für Industrie- und Handelsbetriebe, 2nd Edn.* Berlin; Heidelberg: Springer Gabler.
- Murez, M., and Recanati, F. (2016). Mental files: an introduction. *Rev. Philos. Psychol.* 7, 265–281. doi: 10.1007/s13164-016-0314-3
- Murphy, M. L. (2010). *Lexical Meaning. Cambridge Textbooks in Linguistics*. Cambridge, NY: Cambridge University Press.
- Nagy, I. K. (2014). English for special purposes: specialized languages and problems of terminology. *Acta Univ. Sapient. Philol.* 6, 261–273. doi: 10.1515/ausp-2015-0018
- Neurath, O. (1932). Protokollsätze. *Erkenntnis* 3, 204–214. doi: 10.1007/BF01886420
- Neuweg, G. H. (2000). Mehr lernen, als man sagen kann: Konzepte und didaktische Perspektiven impliziten Lernens. *Unterrichtswissenschaft* 28, 197–217.
- Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford: Oxford University Press.
- Nicola-Richmond, K., Pépin, G., Larkin, H., and Taylor, C. (2018). Threshold concepts in higher education: a synthesis of the literature relating to measurement of threshold crossing. *High. Educ. Res. Dev.* 37, 101–114. doi: 10.1080/07294360.2017.1339181
- Oakhill, J., Cain, K., and Bryant, P. (2003). The dissociation of word reading and text comprehension: evidence from component skills. *Lang. Cogn. Process.* 18, 443–468. doi: 10.1080/01690960344000008
- Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.
- Reimann, N., and Jackson, I. (2006). “Threshold concepts in economics: a case study,” in *Overcoming Barriers to Student Understanding. Threshold Concepts and Troublesome Knowledge, Chapter 8*, eds J. H. Meyer and R. Land (Abingdon; New York, NY: Routledge), 115–133.
- Richardson, P. W. (2004). Reading and writing from textbooks in higher education: a case study from economics. *Stud. High. Educ.* 29, 505–521. doi: 10.1080/0307507042000236399
- Rincke, K. (2010). Alltagssprache, Fachsprache und ihre besonderen Bedeutungen für das Lernen. *Z. Didaktik Naturwissensch.* 16, 235–260.
- Roberts, C. (2019). “Contextual influences on reference,” in *The Oxford Handbook of Reference, Chapter 13*, eds B. Abbott and J. Gundel (Oxford: Oxford University Press), 260–282. doi: 10.1093/oxfordhb/9780199687305.013.13
- Roelcke, T. (2010). *Fachsprachen. Number 37 in Grundlagen der Germanistik, 3rd Edn.* Berlin: Schmidt.
- Russell, B. (1919). *Introduction to Mathematical Philosophy, 2nd Edn.* London: George Allen & Unwin.



- Sahlgren, M. (2008). The distributional hypothesis. *Ital. J. Linguist.* 20, 33–54.
- Salmerón, L., Kintsch, W., and Cañas, J. J. (2006). Reading strategies and prior knowledge in learning from hypertext. *Mem. Cogn.* 34, 1157–1171. doi: 10.3758/BF03193262
- Schuhlen, M., and Kunde, F. (2016). “Informelles Lernen und ökonomische Bildung,” in *Handbuch Informelles Lernen*, ed M. Rohs (Wiesbaden: Springer Fachmedien), 455–466. doi: 10.1007/978-3-658-05953-8\_35
- Schumann, S., Eberle, F., Oepke, M., Pflüger, M., Gruber, C., Stamm, P., et al. (2010). *Inhaltsauswahl für den Test zur Erfassung ökonomischen Wissens und Könnens im Projekt “ökonomische Kompetenzen von Maturandinnen und Maturanden (oekoma)”*. Technical report, Universität Zürich, Institut für Gymnasial- und Berufspädagogik.
- Searle, J. R. (1969). *Speech Acts*. Cambridge: Cambridge University Press.
- Searle, J. R. (2006). Social ontology: some basic principles. *Anthropol. Theory* 6, 12–29. doi: 10.1177/1463499606061731
- Sender, T. (2017). *Wirtschaftsdidaktische Lerndiagnostik und Komplexität. Lokalisierung liminaler Unsicherheitsphasen im Hinblick auf Schwellenübergänge. Komplexität, Entrepreneurship und Ökonomische Bildung* (Zugl. dissertation). Springer Gabler, TU Dortmund, Wiesbaden, Germany.
- Shanahan, M. P., Foster, G., and Meyer, J. H. (2006). Operationalising a threshold concept in economics: a pilot study using multiple choice questions on opportunity cost. *Int. Rev. Econ. Educ.* 5, 29–57. doi: 10.1016/S1477-3880(15)30119-5
- Simkins, S. P. (1999). Promoting active-student learning using the world wide web in economics courses. *J. Econ. Educ.* 30, 278–287. doi: 10.1080/00220489909595990
- Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *J. Document.* 28, 11–21. doi: 10.1108/eb026526
- Steffens, Y., Schmitt, I. L., and Amann, S. (2017). *Mediennutzung Studierender: über den Umgang mit Medien in hochschulischen Kontexten-Systematisches Review nationaler und internationaler Studien zur Mediennutzung Studierender*. Köln: Universität zu Köln.
- Strevens, P. (1977). Special-purpose language learning: a perspective. *Lang. Teach. Linguist. Abstr.* 10, 145–163. doi: 10.1017/S0261444800003402
- Tinkler, S., and Woods, J. (2013). The readability of principles of macroeconomics textbooks. *J. Econ. Educ.* 44, 178–191. doi: 10.1080/00220485.2013.770345
- Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Trier: Wissenschaftlicher Verlag.
- Uslu, T., Mehler, A., and Baumartz, D. (2019). “Computing classifier-based embeddings with the help of text2ddc,” in *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing, (CICLing 2019), CICLing 2019* (La Rochelle).
- van den Broek, P., Rapp, D. N., and Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Process.* 39, 299–316. doi: 10.1080/0163853X.2005.9651685
- van Mourik, G., and Wilkin, C. L. (2019). Educational implications and the changing role of accountants: a conceptual approach to accounting education. *J. Vocat. Educ. Train.* 71, 312–335. doi: 10.1080/13636820.2018.1535517
- Vaňková, L. (2018). “Fachsprachen und der Alltag: eine Untersuchung anhand der deutschen Tagespresse,” in *Zentrum und Peripherie: aus sprachwissenschaftlicher Sicht*, eds V. Kotólková and G. Rykalová (Opava: Slezská Univerzita v Opavě), 51–64.
- Vernooij, F. (2000). “Tracking down the knowledge structure of students,” in *Business Education for the Changing Workplace, Number 5 in Educational Innovation in Economics and Business*, eds L. Borghans, W. H. Gijselaers, R. G. Milter, and J. E. Stinson (Dordrecht: Kluwer), 437–450.
- Vidal, N., Smith, R., and Spetic, W. (2015). Designing and teaching business & society courses from a threshold concept approach. *J. Manag. Educ.* 39, 497–530. doi: 10.1177/1052562915574595
- von Weizsäcker, E. U. (1974). “Erstmaligkeit und Bestätigung als Komponenten der pragmatischen Information,” in *Offene Systeme I. Beiträge zur Zeitstruktur von Information, Entropie und Evolution*, ed E. U. von Weizsäcker (Stuttgart: Klett), 82–113.
- Weber, W., Kabst, R., and Baum, M. (2014). *Einführung in die Betriebswirtschaftslehre, 9th Edn*. Wiesbaden: Springer Gabler.
- Widdowson, H. (1974). Literary and scientific uses of English. *English Lang. Teach. J.* 28, 282–292. doi: 10.1093/elt/XXVIII.4.282
- Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford: Oxford University Press.
- Wilkinson, D. M., and Huberman, B. A. (2007). “Cooperation and quality in Wikipedia,” in *Proceedings of the 2007 International Symposium on Wikis, WikiSym '07* (Montreal, QC), 157–164. doi: 10.1145/1296951.1296968
- Wittgenstein, L. (1984). *Tractatus Logico-Philosophicus; Tagebücher 1914–1916; Philosophische Untersuchungen. Number 1 in Werkausgabe*. Frankfurt am Main: Suhrkamp.
- Zeevat, H. (2018). “Interpreting dependent NPs,” in *Proceedings of Cognitive Structures: Linguistic, Philosophical and Psychological Perspectives, CoSt'18* (Düsseldorf).
- Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley.
- Zlatkin-Troitschanskaia, O., Toepper, M., Pant, H. A., Lautenbach, C., and Kuhn, C. (Eds.). (2018). *Assessment of Learning Outcomes in Higher Education: Cross-National Comparisons and Perspectives*. Wiesbaden: Springer International Publishing.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Lücking, Brückner, Abrami, Uslu and Mehler. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Keep Calm in Heated Debates: How People Perceive Different Styles of Discourse in a Scientific Debate

Juliane Tkotz<sup>1,2,3†</sup>, Dorothe Kienhues<sup>4\*†</sup>, Regina Jucks<sup>4</sup> and Rainer Bromme<sup>4</sup>

<sup>1</sup>Department of Clinical Psychology, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany, <sup>2</sup>Department of Addiction Behavior and Addiction Medicine, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany, <sup>3</sup>Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany, <sup>4</sup>Department of Psychology, University of Muenster, Muenster, Germany

## OPEN ACCESS

### Edited by:

Olga Zlatkin-Troitschanskaia,  
Johannes Gutenberg University  
Mainz, Germany

### Reviewed by:

Susan R Goldman,  
University of Illinois at Chicago,  
United States  
Ying-Chih Chen,  
Arizona State University, United States

### \*Correspondence:

Dorothe Kienhues  
kienhues@uni-muenster.de

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Education

**Received:** 14 June 2020

**Accepted:** 21 December 2020

**Published:** 11 February 2021

### Citation:

Tkotz J, Kienhues D, Jucks R and  
Bromme R (2021) Keep Calm in  
Heated Debates: How People Perceive  
Different Styles of Discourse in a  
Scientific Debate.  
Front. Educ. 5:572503.  
doi: 10.3389/feduc.2020.572503

Scientific debates are, in an epistemological sense, argumentative approaches aimed at coming to the most appropriate conclusion. However, as these debates sometimes involve interpersonal rather than content-driven attacks (e.g., an argument between scientific experts might involve personal dislike), the following question arises: How do such communication behaviors affect people's perception of the argument? In an empirical study, we presented prospective teachers ( $N = 222$ ) with a newspaper article about two scientific experts controversially discussing the pros and cons of a fictional vocabulary training program. Using a  $1 \times 2$  between-subject design, the article contained either a neutral or an incivil discourse style. The dependent measures evaluated how participants perceived the experts' trustworthiness and how they viewed the practical relevance of the scientific topic at hand. Results revealed that participants who read the neutral-style discourse perceived the two experts as having more expertise, higher integrity, and higher benevolence than participants who read the incivil-style discourse. However, the groups did not differ in their ratings of how beneficial the scientific findings might be in the classroom. Overall, this study shows that discourse style indeed influences the perceived trustworthiness of experts, in that it might be damaged in heated debates. The study therefore suggests that the scientific community's methodological and social conventions should be addressed in higher education, in this case teacher education, as understanding these conventions is important for substantially evaluating heated scientific debates.

**Keywords:** scientific debate, understanding controversies, epistemic trust, discourse style, scientists' ethos

## INTRODUCTION

In the scientific community, true scientific knowledge is, in conjunction with other practices, determined through discussions and arguments, namely scientific debates. For example, at the beginning of an empirical research project, researchers develop ideas and collect data; after they submitted their results to a journal in written form, their ideas are critically discussed by other scientists (Douglas, 2015). During a formal review process, other experts will reflect on the results and discuss them with the authors, sometimes implicitly via the journal's editor, sometimes directly via rounds of reviews. Further, discussions of research results take place at conferences as well as within social media

(Peters, 2013). A piece of new scientific knowledge is deemed to be true by the scientific community if it survives these discussions (Kitcher, 2001); that is, a group of scientists has formed a consensus. In this sense, scientific debates are an inherent epistemic feature of how knowledge is produced.

Nevertheless, as these debates are enacted by social subjects, scientific debates can also be considered social interactions. Therefore, features of social interactions, such as interpersonal attacks or rude behavior may occur in such debates. In consequence, the controversies arising in scientific debates may be twofold: Beyond the topic-inherent scientific controversy, a scientific debate may also be fueled by interpersonal controversies. Usually, people view the scientific knowledge they deal with as being intimately linked with its social source (Jenkins, 1999), so they might overlook the epistemic reasons for scientific debates. Yet, individuals should be aware that scientific debates (regardless of their civility) are required for achieving scientific truths. This awareness should also entail an understanding that scientists' uncertainty does neither imply unreliability (Kienhues et al., 2020) nor represent an excuse not to act based on the available evidence (Tversky and Shafir, 1992). Such a nuanced understanding of scientific debates is a cornerstone of individuals' scientific literacy, as it encourages people to value scientific evidence as the most rational approach for answering questions in their personal or professional lives.

Improving people's (professional) decision-making by having them consider the best available scientific knowledge is an important goal for higher education, such as in medical education or teacher education. It is particularly crucial for teacher education, because evidence-based teaching and teacher education is not straightforward (Murphy, 2015) and often falls upon deaf ears (Zlatkin-Troitschanskaia, 2016; Alexander, 2018). One reason for this is that people devalue the scientific quality of educational knowledge. Educational research, as a social science, is often perceived to provide rather weak and uncertain knowledge (in comparison to natural sciences, e.g., Hofer, 2000; Lonka et al., 2020), and disciplines contributing to educational research, such as psychology, are perceived "as largely nonscientific and as lacking in scientific rigor" (Lilienfeld, 2012, p. 114). Nevertheless, to ensure that students receive the best education possible, teaching should be evidence based (Bromme et al., 2016; National Research Council, 2001; Slavin, 2002), meaning that teachers have "to identify approaches and practices that work to promote learning and performance" (Alexander, 2018, p. 158). Given this tension, it would be interesting to understand how individuals view controversies about evidence from educational research. Especially prospective teachers, who deal with educational evidence and accompanying debates in their studies, need to understand what scientific evidence is and how it evolves, which involves understanding the epistemic and the interpersonal reasons for scientific debates.

In this study, we aim to investigate how prospective teachers understand scientific debates, especially how their epistemic judgments are influenced by controversies that are intertwined with social interaction.

## Scientific Controversies and Debates in Everyday and Professional Life

Controversies are vital for scientific progress, as they cause evidence to be revisited and mistakes to be uncovered (Paletz et al., 2016). In an epistemological sense, opposing viewpoints represent argumentative approaches toward finding reliable knowledge (Lakatos and Musgrave, 1970). Importantly, to be scientifically literate and to participate in a democratic society, individuals must be able to navigate such controversies where there is not yet scientific consensus (Kolstø, 2001). This entails to understand why controversies between scientists occur and that they are essential for achieving scientific truth. For scientific controversies that are relevant to the public, disagreements among scientists are often publicly accessible. Recent examples of such public disagreements among scientists (and the evolving knowledge that comes along with them) are the discussions about face masks or ibuprofen in the context of the Covid-19 outbreak (Chan and Yuen, 2020; Sodhi and Etmann, 2020).

However, general science education often does not prepare the public to handle scientific controversies productively: It seldom highlights science as argumentative in nature, but instead portrays science as the mere accumulation of undebated facts (Osborne, 2010), and disregards the productivity of moments of uncertainty for science understanding (Chen et al., 2019). Thus, to people who see scientific findings as undebatable factual information, the idea that scientists use consensus to create scientific knowledge might seem underhanded or manipulative. Consequently, attempts to cast doubt on science might fall on fertile ground; people may be vulnerable to post-truth attempts where partisan actors try to attack and devalue science using the very idea that scientific knowledge is created through scientific consensus (McKee and Diethelm, 2010; Oreskes and Conway, 2010). Thus, if individuals are not able to navigate scientific controversies, they may neither value scientific evidence nor act in accordance with it.

## Individuals' Evaluation of Scientific Debates and Their Protagonists

Reasons for disagreements in science can be multifarious and, thus, may not only refer to *epistemic* conflicts (e.g., methodological problems in experiments, new and evolving knowledge) but may also involve *interpersonal* conflicts, especially when scientists are clearly at odds with one another [(case in point, the famous disagreement between Leibniz and Newton (Hall, 1980))]. Such reasons for disagreement may also partly evolve from the communicative goals of scientific debates, which are not always to co-construct consensus but sometimes to convince someone or to win a debate (Leitão, 2000; Fisher et al., 2018).

We have mentioned above that scientific debates can also be considered social interactions, differing in their civility (Rowe, 2015). Incivility is used as an umbrella term including rudeness, aggressiveness, and impoliteness. That is, a scientific debate may involve a kind of personal tone, as scientific experts might dislike each other and be rude to one another. In consequence, someone who is confronted with a scientific debate may not only encounter

opposing views but may also be introduced to interpersonal conflicts that result in ad hominem attacks (Carlson, 2017). Various studies show that such ad hominem attacks can influence individuals' evaluation of scientific debates and their protagonists. For example, the civility of an interaction influences (among other issues) whether a bystander perceives the protagonists as being rational (Popan et al., 2019). Participants watching a video of a scientific debate evaluated a scientist using an aggressive discourse style as less credible, less competent, less sincere, less benevolent, and less likable (König and Jucks, 2019) than a scientist using a neutral discourse style. Further, ad hominem arguments (e.g., questioning a researcher's motives) seem to challenge the perceived credibility of the attacked scientist just as much as arguments targeting the empirical basis of their claims (Barnes et al., 2018; Gierth and Bromme, 2020). That is, incivility in scientific debates can have detrimental effects. Further evidence for such effects comes from research in political science: Participants agreed less with verbally aggressive political speakers and perceived them as less credible than nonaggressive speakers (Nau and Stewart, 2013). In contrast to incivil debates, civil discussions have positive effects, e.g., increasing participants' willingness to vaccinate their future children (Jennings and Russell, 2019). That is, watching an incivil or impolite scientific debate influences how individuals evaluate the content of that debate and its protagonists.

### Practical Relevance of Science

Typically, laypeople engage with conflicting scientific arguments in order to reach a solid answer to a specific question, e.g., when Googling the side effects of a vaccination (Bromme and Goldman, 2014; Brummernhenrich and Jucks, 2019). Individuals want to reach the most reasonable conclusion. On the other hand, one's political orientation and analytical thinking are related to one's agreement with scientific conclusions and factual statements (Lobato and Zimmerman, 2019; Medlin et al., 2019). One strategy people use to reject scientific evidence that does not align with their own beliefs on a topic is to question the perceived potency of the scientific methods that were used to investigate that topic (Munro, 2010). Perceived potency refers to the degree to which science is capable of providing reliable knowledge in response to the problem under consideration. That is, in how far science can really address the problem at hand. Inspired by Munro's findings, we assume that discourse style influences this perceived potency of science.

Regarding educational practice, prospective teachers (teacher students) may evaluate science not only in terms of its general potency but also regarding its relevance for their teaching practice (Zeuch and Souvignier, 2015; Merk et al., 2017). Especially nowadays teachers need to be able to evaluate empirical evidence, and it is of practical relevance to scrutinize whether specific styles of discourse in a scientific debate differently influence the perceived practical importance of an educational science issue. Specifically, it would be important to know whether prospective teachers overlook the potency of certain scientific findings for their forthcoming professional careers when these findings are discussed in an incivil manner.

### Epistemic Trustworthiness

In our society, knowledge is highly specialized and unevenly distributed, and it is almost impossible for laypersons to directly evaluate such specialized knowledge (Kitcher, 1990; Bromme and Goldman, 2014). Therefore, instead of evaluating the scientific evidence itself, individuals often select the best arguments by assessing whether the person providing the information is a reliable and credible source; that is, individuals might evaluate a science communicator's epistemic trustworthiness (Hendriks and Kienhues, 2019). Such judgments focus on three features: an expert's expertise, integrity, and benevolence (Cummings, 2014; Hendriks et al., 2015). *Expertise* refers to the extent that someone is truly knowledgeable and trained in her domain, such as methodological competencies; *integrity* indicates that an expert adheres to the rules and norms of science; and *benevolence* suggests that an expert does not pursue personal benefit or aims but focuses on the interests of others. Various studies have revealed that individuals are capable of nuanced trustworthiness judgments. For example, Jensen (2008) showed that individuals' judgments are sensitive to scientists' disclosures of uncertainty: They showed that messages are perceived as more trustworthy when scientists reported study limitations as opposed to when scientists did not report such limitations. Further, Hendriks et al. (2016) showed that trustworthiness judgments differ depending on whether a scientist self-discloses the limitations of his work or another scientist discloses these limitations. Research by König and Jucks (2019) indicated that an aggressive language (vs. neutral language) style negatively affects trustworthiness judgments. Trustworthiness judgments are crucial, as they lead to informed trust; that is, individuals will not trust blindly.

### Scientists' Ethos

When laypeople observe scientific discourse, they might partly judge it based on their assumptions of how scientists should or should not behave. Such idealized behaviors have been described in the fields of sociology and philosophy of science by Merton (1942) and Mitroff (1974). Merton's norms (1942) refer to the ethos of science and capture views of idealized scientific practice. They, for example, include that scientists are only motivated by the pursuit of knowledge but not by personal gain, and always work objectively (Mitroff, 1974). counter-norms serve as counterpoints to Merton's norms and describe practices that scientists ideally should not do, such as to compete with others for recognition of achievements. These are obviously ideal norms and not descriptions of the actual motives and behaviors of scientists. Nevertheless, as norms they might have constraining effects on scientists, (e.g., such norms differ between scientific faculty and undergraduates as shown by Kardash and Edwards (2012); here, scientific faculty more strongly advocated Merton's norms than did undergraduates). The explication of such norms and counter-norms is also helpful for analyzing empirically how laypersons generally and also university students (in our case, teacher students) think about how scientists should behave.

## Present Study

In the present study, we aimed to investigate the everyday situation that people need to make sense of science-based information they come across in their personal or professional life. We specifically aimed to study the reception of different styles of discourse in a scientific debate on an educational topic. Therefore, we investigated how discourse style affects prospective teachers' perception of the debate and of the scientists involved in the debate as well as how it affects how they view educational science. In a  $1 \times 2$  experimental group design, we presented a newspaper article about two educational experts debating a fictitious computer program for vocabulary training. These experts adopted either a neutral or an incivil discourse style. Note, a neutral discourse style refers to a communication without any elements of attack and aggression. We chose neutral instead of civil for the wording of the (control) group in order to explain that it not necessary to include expressions of mutual personal appreciation or esteem in order to have a civil discourse—at least not within a scientific discourse.

Our hypotheses derive from the distinction between the epistemic and the social sides of scientific discourse outlined above: While scientific discourse can be conceived as an epistemic endeavor to constitute knowledge, it can also be conceived as an interpersonal conflict where scientists are at odds with one another and are fighting because of personal differences between the debaters. We are interested in whether this interpersonal conflict might somehow mask the nevertheless existing fact that discourse is necessary to achieve scientific truth.

In consequence, we first of all hypothesized that an incivil discourse style would influence *conflict explanation* and strengthen participants' assumption that the conflict stemmed from personal differences between the debaters rather than from reason-based aspects (e.g., methodological differences) (H1).

Furthermore, we expected an incivil discourse style to polarize participants' opinions about the debate topic (as it might be perceived as rather opinion-based than reason-based), hence leading to more extreme *opinion ratings* (H2) and higher *confidence ratings* (H3).

Our hypotheses also take into account how discourse style might affect participants' views on educational science. Regarding the *potency of science*, we expected an incivil discourse style to make participants think science is less equipped to answer the question of the debate (H4).

We also assumed that participants who read the incivil discourse style would see less *practical benefit of science*; specifically, we hypothesized that they would find science less useful for their teaching practice (H5).

Concerning the *epistemic trustworthiness* of the scientists involved in the debate, we hypothesized that participants who read the incivil discourse style would place less epistemic trust in the debaters (H6).

Further, regarding participants' assumptions about scientific norms, we expected that an incivil discourse style would lead participants to devalue scientific ethos; that is, we thought participants reading the incivil article would rate *scientists' ethos* as being aligned more strongly with counter-norms than with norms (H7).

## METHODS

### Participants

An a priori power analysis in G\*Power (Faul et al., 2007) for an independent two-tailed *t*-test with  $\alpha = 0.05$  as significance level yielded a minimal sample size of  $N = 210$  in order to detect a medium effect of (Cohen, 1988)  $d = 0.5$  with a power of 0.95. We recruited  $N = 245$  German-speaking teacher students for an online study which was advertised via Facebook groups for teacher students across Germany and in lectures for teacher students at the University of Münster. A short demographic questionnaire collected information about participants' gender, age, the type of school they planned to teach in after university, the subjects they were currently studying, the university at which they were studying, and how many semesters they had studied so far (summed number of bachelor and master semesters). We excluded participants who were not currently studying to become teachers, who did not report at the end of the study that they answered all questions honestly and attentively (Aust et al., 2013), those who completed the study implausibly fast (i.e., 1 *SD* faster than the mean time it took five trained readers to complete a test run of the study) and those who did not focus on the survey page throughout the whole session, leaving  $N = 222$  for final analysis (see **Supplementary Table SA,SB**). After completing the survey, participants had the chance to win one of eleven booksellers' vouchers ( $1 \times 50\text{€}$ ,  $10 \times 15\text{€}$ ). The study was approved by the ethics commission of the University of Münster.

### Materials

The whole study was conducted online via Unipark (Questback GmbH, 2018), and all materials and questionnaires were presented in German.

### Debate Scenario

In both conditions, the newspaper articles featured the same brief information about the program PAVLOV and the debate on it. Subsequently, the article continued to describe the arguments of two educational scientists, (named Dr. Frank Völkel and Dr. Frederick Mische) each taking turns to provide their viewpoint and the evidence for it. In both conditions the content and wording was the same, except in the incivil condition the verbs were exchanged and accompanying adverbs were inserted in order to express that the debaters had an aggressive stance toward each other. For example, the neutral version stated "Völkel replied," while the incivil version read "Völkel retorted aggressively." The words in question were generated based on synonyms and antonyms as provided by a dictionary of the German language (Questback GmbH, 2018). For example, in generating the incivil version of the text, it was important that the words clearly described aggression directed toward the other debater (vs. general, undirected negative emotion). We wanted to clarify that the emotional language one debater expressed was due to the discussion with the other debater and not a result of events unrelated to the panel discussion. In the final version, both texts were of comparable length and both debaters had an equal share of the discussion.



**TABLE 1 |** Overview on hypotheses and instruments used.

Hypothesis	Dependent variable	Number of items
(1) An incivil discourse style strengthens participants' assumption that the conflict stemmed from personal differences between the debaters rather than from reason-based aspects.	Conflict explanation	4
(2) An incivil discourse style leads to more extreme opinion ratings.	Opinion rating	1
(3) An incivil discourse style leads to higher confidence ratings.	Confidence rating	1
(4) An incivil discourse style lets science appear to be less equipped to answer the central question of the debate.	Potency of science (Munro, 2010)	1
(5) Participants who read the incivil discourse style find science less useful for their teaching practice.	Practical benefit of science (Zeuch and Souvignier, 2015)	9
(6) Participants who read the incivil discourse style place less epistemic trust in the debaters.	Epistemic trustworthiness - METI (Hendriks et al., 2015)	14
(7) Participants who read the incivil discourse style rate scientists' ethos as being aligned more strongly with counter-norms than with norms.	Scientists' ethos (Kardash and Edwards, 2012)	8

Participants read at their own pace. At the end of the article, participants were informed that they would not be able to read the article again once beginning the following questionnaires. Both versions of the newspaper article are provided in **Supplementary Material**.

### Task Instructions

Via a random generator implemented in the survey, participants were assigned to one of two conditions: either a neutral debate scenario or an incivil debate scenario. In both conditions, participants were instructed to imagine being a teacher at a school that was deciding whether to use a new vocabulary training program (PAVLOV: Programmed Associative Visualization Learning of Vocabularies). Since a school's choice on media use affects every teacher, the whole staff was involved in the decision. Participants were told that the principal requested that they carefully read a newspaper article on a panel discussion that took place as part of a congress on educational sciences. During the congress, two educational scientists debated the evidence for and against PAVLOV, and the principal was awaiting each participant's opinion on whether the program should be used in classes. Scenario descriptions included a short introduction to the congress at which the debate took place. For both conditions, scenario descriptions read the same, except in the incivil condition participants were informed that they were about to read a heated debate (neutral version: debate).

### Measures

All measures and how they relate to our hypotheses are listed in **Table 1**.

#### Conflict Explanation

After reading the article, participants were asked to provide an explanation for the conflict they just had read about. To induce reasoning about the conflict, we first asked the following open question: In your opinion, what are the reasons for the conflict that emerged in the panel discussion? Participants were instructed to provide their perspective in short sentences. The free responses were not analyzed further. Participants then answered four closed items about their explanation for the

conflict (cf. H1). For each item, they indicated their agreement on a 7-point Likert-type scale (1 = do not agree at all, 7 = fully agree). The first item stated that the debaters referred to different research findings; the second item stated that there was a personal conflict between the debaters; the third item stated that the debaters referred to different effects the program had; the fourth item stated that the debaters focused on different goals when evaluating the program.

#### Opinion About PAVLOV

Participants were asked whether the vocabulary training program should be used at the imaginary school (definitely no - definitely yes, opinion rating, cf. H2). In a second item, participants indicated how much confidence they had in their previously stated opinion (not confident at all - very confident, confidence rating, cf. H3). For both items, they provided their answers using a slider ranging from 1 to 100 (numbers were not shown).

#### Potency of Science

One item adapted from Munro (2010) was used to capture the perceived potency of science (H4): Participants indicated whether they believed that the question about using PAVLOV could ultimately be answered unambiguously with scientific research on a 7-point Likert-type scale (1 = do not agree at all, 7 = fully agree).

#### Practical Benefit of Science

To assess participants' perceived benefit of educational sciences in day-to-day teaching (cf. H5), we asked them to complete the subscale Benefit of Science for Professional Practice as implemented in the questionnaire about scientific thinking of pre-service teachers by Zeuch and Souvignier (2015). For nine statements, participants indicated their agreement on a 7-point Likert-type scale (1 = do not agree at all, 7 = fully agree), e.g., In the classroom, it would be best if teachers rely on their experiences instead of findings from the educational sciences. (Reverse scored; original questionnaire in German). The authors report a Cronbach's  $\alpha$  of 0.76, an item discrimination power range of 0.33–0.52 and a mean score of 4.45 (SD = 0.76). Participants were asked to refer to the field of educational science, so we adapted the original items from

Zeuch and Souvignier (2015) by changing science to educational science.

### Muenster Epistemic Trustworthiness Inventory (METI)

Participants provided their judgment on the debaters' trustworthiness in the Muenster Epistemic Trustworthiness Inventory (cf. H6, Hendriks et al., 2015). They did so by choosing between 14 word-pairs on a 7-point Likert-type scale, presented as semantic differentials (e.g., competent vs. incompetent). Three mean scores were computed for each of the following sub-dimensions: expertise (Six items), integrity (four items), and benevolence (four items). Hendriks et al. (2015) report a Cronbach's  $\alpha$  of 0.91 for expertise, 0.82 for integrity, and 0.90 for benevolence. Participants were instructed to rate both debaters simultaneously, as we were interested in their overall impression of the debate as source of information.

### Scientists' Ethos

Additionally, participants filled in a questionnaire reflecting their perception of scientists' ethos (cf. H7). We translated into German the version used by Kardash and Edwards (2012), which is a slight adaption of the questionnaire proposed by Anderson and Louis (1994). In eight items, participants indicated how much they thought statements about norms and counter-norms described actual scientific practice on a 5-point Likert-type scale (1 = not representable at all, 5 = fully representable), e.g., Scientists are generally motivated by the desire for knowledge and discovery, and not by the possibility of personal gain (norm of disinterestedness). Each norm proposed by Merton (1942) and each counter-norm proposed by Mitroff (1974) was represented by one item, and participants were reminded to refer to the field of educational science. For the original version of the questionnaire, Anderson and Louis (1994) report a moderate reliability of 0.49 for the norm scale and a reliability of 0.64 for the counter-norm scale. This may be due to the fact that the scale consists of several (counter-) norms, each represented by one item. That means that different constructs are reflected within the same scale, which might lower the internal consistency. Unfortunately, no indices are reported by Kardash and Edwards (2012).

### Procedure

Participants gave their informed consent and filled in the demographic questionnaire. They were then introduced to either the neutral or the incivil debate scenario description. Participants read the corresponding newspaper article and then expressed their opinion on PAVLOV. Afterward, they answered the items on conflict explanation, the METI, the questionnaire about the practical benefit of science and the questionnaire on scientists' ethos. In a final item, participants indicated whether they had honestly and attentively answered the items or whether we should discard their data. Lastly, we thanked participants for taking part in the study and debriefed them. If they wished, participants could then follow a link to a separate survey where they could provide their email addresses for the lottery of booksellers' vouchers.

### Data Analysis

We used R (Version 3.6.0; R Core Team, 2018) for all analyses, which were carried out using  $\alpha = 0.05$  as significance level. To assess whether METI subscale correlations significantly differed between experimental conditions, we compared  $z$ -standardized correlation coefficients  $r$  with respect to the sample size, as implemented in the R-package cocor (Version 1.1–3; Diedenhofen and Musch, 2015).

### RESULTS

In the following, the results of our statistical analyses are described according to the order of the hypotheses formulated above.

### Conflict Explanation

Discourse style did not affect the rather objective aspects of conflict explanation (H1): Between conditions, participants did not differ in the degree they thought the debaters referred to different research results, to different effects of PAVLOV or to different goals when using PAVLOV. However, participants reading the incivil panel discussion more strongly assumed the conflict to be personal than those reading a neutral panel discussion (Table 2).

### Opinion About PAVLOV

Overall, participants supported using PAVLOV; that is, their mean rating of the program was greater than 50 ( $M = 62.60$ ,  $SD = 20.89$ ),  $t(221) = 8.99$ ,  $p < 0.001$ . With regard to hypothesis 2, participants reading the neutral debate ( $M = 63.04$ ,  $SD = 20.14$ ) and participants reading the incivil debate ( $M = 62.20$ ,  $SD = 21.65$ ) were equally in favor of using the program,  $t(220) = 0.30$ ,  $p = 0.766$ ,  $d = 0.04$ . Furthermore, regarding hypothesis 3, participants in the neutral condition ( $M = 72.58$ ,  $SD = 21.12$ ) and those in the incivil condition ( $M = 69.48$ ,  $SD = 23.51$ ) expressed equal confidence in their opinions,  $t(219.74) = 1.03$ ,  $p = 0.302$ ,  $d = 0.14$ . Having a strong opinion about PAVLOV was associated with more confidence in it,  $r(220) = 0.40$ ,  $p < 0.001$ .

### Potency of Science

With regard to hypothesis 4, here was no difference between the neutral ( $M = 3.09$ ,  $SD = 1.80$ ) and the incivil condition ( $M = 2.77$ ,

TABLE 2 | Conflict explanation.

	Condition		<i>t</i>	<i>df</i>	<i>p</i>	<i>d</i>
	Neutral	Incivil				
Research results	4.72 (1.46)	4.48 (1.39)	1.26	216.90	0.209	0.17
personal conflict	2.45 (1.50)	4.19 (2.03)	−7.31	209.47	<0.001	−0.97
effects	5.52 (1.29)	5.36 (1.52)	0.88	218.15	0.378	0.12
Goals	4.64 (1.65)	4.93 (1.67)	−1.32	219.17	0.187	−0.18

Mean and standard deviation for the items capturing conflict explanation, each rated on a Likert-type Scale from 1 to 7. *t*-tests between the neutral and the incivil condition for each item are reported

$SD = 1.75$ ) regarding the question of whether science is equipped to resolve the conflict about PAVLOV,  $t(217.66) = 1.37$ ,  $p = 0.171$ ,  $d = 0.18$ .

## Practical Benefit of Science

Regarding hypothesis 5, participants who read the incivil debate ( $M = 4.52$ ,  $SD = 0.92$ ) did not differ from those who read the neutral debate ( $M = 4.43$ ,  $SD = 1.01$ ) in the degree to which they think scientific findings are beneficial in the classroom,  $t(214.26) = -0.72$ ,  $p = 0.470$ ,  $d = -0.10$ . Contrary to Zeuch and Souvignier (2015), participants studying STEM subjects ( $M = 4.48$ ,  $SD = 0.95$ ) did not perceive science to be more beneficial than participants studying non-STEM subjects ( $M = 4.48$ ,  $SD = 0.98$ ),  $t(219.83) = -0.03$ ,  $p = 0.975$ ,  $d = 0.00$ . In our sample, the scale reached a Cronbach's  $\alpha$  of 0.83.

## Epistemic Trustworthiness

Regarding hypothesis 6, participants placed more epistemic trust in the debaters when reading a neutral debate: Compared to participants in the incivil condition ( $M = 4.79$ ,  $SD = 0.99$ ), participants in the neutral condition ( $M = 5.06$ ,  $SD = 1.00$ ) perceived the debaters as having more expertise,  $t(218.49) = 1.99$ ,  $p = 0.047$ ,  $d = 0.27$ . Furthermore, participants reading the neutral debate ( $M = 4.76$ ,  $SD = 1.02$ ) reported higher ratings of debaters' integrity than those reading the incivil debate ( $M = 4.05$ ,  $SD = 1.15$ ),  $t(219.41) = 4.87$ ,  $p < 0.001$ ,  $d = 0.65$ . Additionally, ratings of benevolence were higher in the neutral condition ( $M = 4.77$ ,  $SD = 0.98$ ) than in the incivil condition ( $M = 4.05$ ,  $SD = 0.89$ ),  $t(214.11) = 5.67$ ,  $p < 0.001$ ,  $d = 0.76$ .

In addition, we explored the correlation between the METI subscales and the four conflict explanation items to determine whether the perception of various aspects of a conflict was associated with different degrees of epistemic trust. Those whose explained the conflict by stating that the debaters referred to *different research results* (item 1) also thought them to have more expertise,  $r(220) = 0.14$ ,  $p = 0.039$ . No relation was found with integrity,  $r(220) = 0.07$ ,  $p = 0.321$ , or benevolence,  $r(220) = 0.03$ ,  $p = 0.679$ . Conflict explanations that assumed *personal reasons* (item 2) were most strongly related with epistemic trust; in particular, the more participants perceived the conflict to be personal, the less expertise they assigned to the debaters,  $r(220) = -0.25$ ,  $p < 0.001$ . Similarly, the perception of a personal conflict led to decreased ratings of integrity,  $r(220) = -0.36$ ,  $p < 0.001$ , and benevolence,  $r(220) = -0.41$ ,  $p < 0.001$ . The degree to which participants agreed that the debaters referred to *different goals of PAVLOV* (item 3) did not correlate with any of the METI subscales (expertise:  $r(220) = 0.10$ ,  $p = 0.122$ ; integrity:  $r(220) = -0.00$ ,  $p = 0.946$ ; benevolence:  $r(220) = -0.00$ ,  $p = 0.994$ ). Further, the degree to which participants agreed that the debaters referred to *different effects of PAVLOV* (item 4) was not associated with epistemic trust either (expertise:  $r(220) = 0.01$ ,  $p = 0.863$ ; integrity:  $r(220) = -0.06$ ,  $p = 0.348$ ; benevolence:  $r(220) = -0.05$ ,  $p = 0.475$ ). Internal consistency of the METI subscales was somewhat lower than initially found by Hendriks et al. (2015), with a Cronbach's  $\alpha$  of 0.87 for expertise, 0.83 for integrity, and 0.76 for benevolence.

## Scientists' Ethos

With regard to hypothesis 7, participants more strongly agreed with the statements that described scientists' ethos in terms of counter-norms ( $M = 14.50$ ,  $SD = 2.36$ ) rather than norms ( $M = 12.73$ ,  $SD = 2.29$ ),  $t(221) = 7.41$ ,  $p < 0.001$ ,  $d_z = 0.76$ ; a significant negative relationship existed between participants' agreement with norms and counter-norms,  $r(220) = -0.17$ ,  $p = 0.011$ . Discourse style, however, left the perception of scientists' ethos largely unaffected. The only difference that emerged was that participants who read the neutral debate compared to the incivil one more strongly thought that educational scientists follow the norm of organized skepticism (Table 3). For the set of norms and counter-norms, we found a Cronbach's  $\alpha$  of 0.41 and 0.58, respectively.

## Further Exploratory Analysis: Correlation of METI Subscales

Given the effect that our manipulation of discourse style had on the debaters' epistemic trustworthiness, we further investigated the METI and its subscales. Specifically, we were interested in the correlations between the different subscales as a function of discourse style, that is whether an incivil discourse style increases or decreases the associations between different aspects of epistemic trustworthiness. For this purpose, we compared the z-standardized correlation coefficients of the subscales between the civil and the incivil condition. Descriptively, the correlations between subscales was weaker in the incivil condition. However, only the correlations involving integrity (i.e. integrity and expertise; integrity and benevolence) were significantly different between the two conditions (Table 4).

## DISCUSSION

We examined whether the discourse style of a scientific debate affected participants' perception of the conflict and the assumed potency of science, the perceived practical relevance of science, participants' epistemic trust in the debaters, and their perceived

TABLE 3 | (Counter-)norms.

	Condition		<i>t</i>	<i>df</i>	<i>p</i>	<i>d</i>
	Neutral	Incivil				
Disinterestedness	3.51 (1.02)	3.47 (0.90)	0.34	211.81	0.732	0.05
Organized skepticism	3.20 (0.97)	2.82 (0.88)	3.04	214.51	0.003	0.41
Communitary	2.92 (0.95)	2.79 (1.06)	0.92	219.69	0.358	0.12
Universalism	3.35 (0.91)	3.42 (0.87)	-0.60	216.80	0.550	-0.08
particularism	2.99 (0.94)	3.11 (0.94)	-0.97	219.08	0.333	-0.13
Organized dogmatism	3.85 (0.80)	3.70 (0.91)	1.28	219.32	0.204	0.17
Self-interestedness	3.80 (0.93)	3.98 (0.90)	-1.46	217.70	0.146	-0.20
Solitariness	3.86 (0.72)	3.69 (0.94)	1.54	212.27	0.124	0.21

Mean and standard deviation for the items capturing scientific practice via norms and counter-norms, rated on a Likert-Scale from 1 to 5. *t*-tests between the neutral and the incivil condition for each item are reported.

**TABLE 4 |** Correlations of METI subscales.

	Overall		Neutral		Incivil		Comparison	
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>z</i>	<i>p</i>
Expertize - integrity	0.44	<0.001	0.59	<0.001	0.29	0.002	2.82	0.005
Expertize - benevolence	0.47	<0.001	0.54	<0.001	0.38	<0.001	1.56	0.118
Integrity - benevolence	0.58	<0.001	0.67	<0.001	0.41	<0.001	2.80	0.005

Pearson correlation coefficients between the METI subscales, overall and grouped by experimental condition. Comparison of correlation coefficients between conditions are reported as Fisher's *z*.

scientific ethos. In the following, we will first of all briefly summarize our findings.

With regard to hypothesis 1 on conflict explanation, while participants reading the incivil debate more strongly assumed that debaters' personal differences caused the conflict, their agreement to interpret the debate as a mere epistemic conflict (e.g., that the debaters referred to different research results) was not affected. Hence, discourse style only differently influenced interpersonal conflict explanations but not epistemic conflict explanations and the personal nature of the conflict did not distract participants from the underlying methodological arguments. In contrast to hypotheses 2 and 3, an incivil discourse style did neither lead to more extreme opinion ratings nor to higher confidence in one's opinion. Regarding the perceived potency of science, in contrast to hypothesis 4, the discourse style did not differently influence in how far participants perceived science to be equipped to answer the question of the debate (but see our further analyses below). Turning to hypothesis 5, the discourse style had no effect on participants' willingness to implement findings from educational science into their teaching practice. All in all, participants in our study notably assigned a rather high value to evidence-based teaching practices.

With regard to hypothesis 6 on epistemic trustworthiness, an incivil discourse style led participants to place less epistemic trust in the debaters. Thus, participants rated experts who keep their temper as a more reliable source of knowledge. The effect was most pronounced for the subscales benevolence and integrity, while only a small effect was detectable for scientists' expertise.

Concerning hypothesis 7 on scientists' ethos, we found that the discourse style largely did not affect participants' perception of scientists' ethos. One exception was the norm of organized skepticism, which participants reading the incivil debate thought that educational scientists fulfill to a lesser degree.

In sum, our findings indicate that only questions regarding the perception of scientists, but not regarding the perception of science as such, were differently tackled by the different discourse styles. For a further discussion of these findings, it is especially interesting to see that the findings for hypothesis 6 are in line with the conflict explanation results (H1): An incivil discourse style mainly affected the rather social components of trustworthiness as opposed to the comparably technical component of expertise. Indeed, expertise does not seem to be

a requirement for high ratings of benevolence and integrity: When experts admit flaws, participants perceive them as holding less expertise but ascribe more integrity and benevolence to them (Hendriks et al., 2016). Thus, benevolence and integrity might be the key factors to consider in scientific communication that aims to increase epistemic trustworthiness.

Further, the explanations participants assumed for the conflict were associated with the epistemic trust they placed in the debaters. That is, when participants perceived the conflict to be interpersonal, they also ascribed less epistemic trustworthiness to the debaters. However, when participants perceived the conflict to be caused by debaters referring to different research findings, they also thought the debaters held more expertise. This could mean that when individuals are aware that conflicting evidence is being discussed, they tend to value the experts' methodological skills. An alternative explanation might be that participants who ascribe more expertise to the debaters are more likely to notice the conflicting research results behind the debate.

In our exploratory analysis, we found that the correlations between the METI subscale "integrity" and the other subscales ("expertise" and "benevolence") were reduced for participants who read the incivil debate. One interpretation is that participants in the incivil condition might have developed a more nuanced view of the debaters' epistemic trustworthiness, rating the different components independent of each other. Ratings of integrity and benevolence were more strongly affected by the debaters' incivil behavior than those of expertise. This supports the idea that epistemic trustworthiness consists of at least partly independent components.

Revisiting the results for hypothesis 7 where the discourse style only affected participants' perception of scientific ethos with regard to the norm of organized skepticism, we which to elaborate further why reading the incivil debate caused the impression that educational scientists fulfill organized skepticism to a lesser degree. Indeed, an incivil debate can be seen as a deviation from the behavior described in the norm: Scientists should consider all evidence, even if that means questioning themselves. In a personal conflict, however, it might appear that they are questioning the other person rather than carefully checking their own perspective. For all other norms and counter-norms, no differences emerged: Even though incivility affected epistemic trust in the debaters, participants



did not generalize to the perception of scientific ethos overall. It is reassuring that a single debate was not sufficient to change participants' perspectives on a whole research community. However, there is the possibility that repeated experience with a certain type of discourse style can modify how people view scientific ethos. An alternative explanation for why the other norms and counter-norms were not affected is that they were not clearly reflected in the newspaper article. For example, the article offered no information about the personal motivations of the debaters (typically manipulated via information about debaters' affiliations), which otherwise could have affected the norm of disinterestedness.

## Limitations and Implications

In the following, limitations and implications are outlined focusing on 1) the study design and 2) the setting of teacher education and higher education.

(1) A minimal intervention sparing the content of a debate is sufficient to cause participants perceive a conflict differently. From an applied perspective, this instance can be worrying because a third party (e.g., a journalist) might influence the perception of a topic via the descriptions about what is being said, even though he may literally be quoting the experts' statements. On the other hand, a neutral description of an incivil debate might, in fact, increase the epistemic trust readers place in the experts depicted.

Heated debates in many informal learning settings might impact differently on readers' evaluations than in our experimental setting. Merely changing subtle descriptions in a newspaper article is less multi-faceted than the discourse style in a *real-life debate*. For example, in another media format such as video recordings of a debate (e.g., König and Jucks, 2019), one could additionally alter the tone or volume of the experts' voices. Furthermore, in a real-life incivil debate, it is not only the discourse style that changes, but also argument substance and the way arguments are exchanged are likely to be different. In a conflict, debaters tend to give less consideration to the other's arguments and do not address them in their replies (Fisher et al., 2017). It might be a more realistic manipulation to additionally vary *what* debaters are saying, not just how they say it. In such a design, however, it would not have been possible to isolate the effect of discourse style.

Another limitation is the *topic of debate* and the fact that this heated debate was provided in an area that is not under heated debate in general. Further studies might transfer our experimental manipulation to issues that are under ongoing heated discourse (such as climate change; Hendriks and Jucks, 2020). Especially *value-based evaluations* on the content of information might impact the evaluation of heated debates and their appropriateness (Kienhues et al., 2020).

(2) Teacher students have at least three roles and tasks when interacting with scientific information and heated debates. First, they are users/readers and simple participants and *recipients* of scientific discourses. They directly engage with scientific information, e.g., when

reflecting upon the role of digitization in school. Second, they (prospectively) *teach* and play a pivotal role in conveying how scientific conflicts should be dealt with. Though there is evidence that teachers ignore empirical evidence (like that on waiting time in teacher-pupil interactions; Borko et al., 1990), teachers teach how scientific information should be used. In so far, teacher students form a group with specific interests: they are learners in the setting of higher education and trained to be teachers in their former jobs. However, focusing on this specific group in a empirical study provides some limitations: teacher students are more familiar with the topic of education itself than other students in higher education. Hence, our findings might not generalize to scenarios where laypeople are confronted with scientific information in a less academic setting. Furthermore, future studies might expose teachers to a topic less related to school settings, such as a medical debate, and compare their perception of the conflict with that of other laypeople or experts in the field. In a similar fashion, the impact of a debate in the educational sciences on people without expertise in teaching could be examined. Since results from Kardash and Edwards (2012) and Zeuch and Souvignier (2015) indicate that perception of science is altered by professional or educational experience, including experienced teachers in the sample would provide further insights.

Furthermore, the setting of teacher education might reduce a direct immersion into the topic. Though the study used a heated debate, the role of emotions might be stronger in a setting where readers have direct and strong emotions regarding the topic (e.g., flat earthers). Hence, the findings might be limited to an educational setting like the one used in the experiment. Again, taking the perspective of teacher education, training teacher students how to address heated debates and how to support their learners to tear emotional language apart from scientific correctness is important.

An incivil discourse style can negatively affect the epistemic trust placed in scientific debaters. Yet, epistemic trust in scientists is needed for people to perceive them as a reliable source of knowledge. That means that we should encourage neutral debates, especially when they take place in public. On the other hand, teaching science as debate as part of the curriculum in science education could empower students to see past seemingly personal conflicts. Here, teachers are multipliers of their perspective on science. As such, they need to be able to teach their students how to navigate scientific debates, irrespective of discourse style. Hence, scientific controversies need to be evaluated in light of the scientific progress as such, and they should also be a part of teacher education. At this point, higher education sets the stage for what is needed in society and in science education: The knowledge and insights in how to cope with scientific information and debates. However, teachers should be prepared to confront the paradox of personalized communication, emotional coloring and scientific standards. Hence, they are expected to solve this personally and as part of an educational approach.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation. The dataset supporting the conclusions of this article is available at PsychArchives by the following DOI: <http://dx.doi.org/10.23668/psycharchives.4483>.

## ETHICS STATEMENT

The study was reviewed and approved by the ethics commission of the Department of Psychology, University of Muenster. The participants provided their written informed consent to participate in this study.

## REFERENCES

- Alexander, P. A. (2018). Past as prologue: educational psychology's legacy and progeny. *J. Educ. Psychol.* 110 (2), 147–162. doi:10.1037/edu0000200.
- Anderson, M. S., and Louis, K. S. (1994). The graduate student experience and subscription to the norms of science. *Res. High. Educ.* 35 (3), 273–299. doi:10.1007/BF02496825.
- Aust, F., Diedenhofen, B., Ullrich, S., and Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behav. Res. Methods* 45 (2), 527–535. doi:10.3758/s13428-012-0265-2
- Barnes, R. M., Johnston, H. M., MacKenzie, N., Tobin, S. J., and Taglang, C. M. (2018). The effect of ad hominem attacks on the evaluation of claims promoted by scientists. *PLoS One* 13 (1), e0192025. doi:10.1371/journal.pone.0192025
- Biesta, G. J. J. (2010). Why 'what works' still won't work: from evidence-based education to value-based education. *Stud. Philos. Educ.* 29 (5), 491–503. doi:10.1007/s11217-010-9191-x
- Borah, P. (2012). Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Commun. Res.* 41 (6), 809–827. doi:10.1177/0093650212449353
- Borko, H., Livingston, C., and Shavelson, R. J. (1990). Teachers' thinking about instruction. *Remedial Special Educ.* 11 (6), 40–49. doi:10.1177/074193259001100609
- Bromme, R., and Goldman, S. R. (2014). The public's bounded understanding of science. *Educ. Psychol.* 49 (2), 59–69. doi:10.1080/00461520.2014.921572
- Bromme, R., Prenzel, M., and Jaeger, M. (2016). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik: Zum Zusammenhang von Wissenschaftskommunikation und Evidenzbasierung in der Bildungsforschung. *Z. für Erziehungswiss. (ZfE)* 19 (1), 129–146. doi:10.1007/s11618-016-0703-5
- Brummernhenrich, B., and Jucks, R. (2019). "Get the shot, now!" Disentangling content-related and social cues in physician–patient communication. *Health Psychol. Open* 6, 2055102919833057. doi:10.1177/2055102919833057
- Carlson, E. A. (2017). Scientific feuds, polemics, and ad hominem arguments in basic and special-interest genetics. *Mutat. Res.* 771, 128–133. doi:10.1016/j.mrrev.2017.01.003
- Chan, K. H., and Yuen, K.-Y. (2020). COVID-19 epidemic: disentangling the re-emerging controversy about medical facemasks from an epidemiological perspective. *Int. J. Epidemiol.* 49, 1063–1066. doi:10.1093/ije/dyaa044
- Chen, Y.-C., Benus, M. J., and Hernandez, J. (2019). Managing uncertainty in scientific argumentation. *Sci. Educ.* 103 (5), 1235–1276. doi:10.1002/sce.21527
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New Jersey, NJ: L. Erlbaum Associates.
- Cummings, L. (2014). The "trust" heuristic: arguments from authority in public health. *Health Commun.* 29 (1), 1043–1056. doi:10.1080/10410236.2013.831685

## AUTHOR CONTRIBUTIONS

Study conception and design, JT, DK, RJ, and RB. Acquisition of data, JT. Analysis of data, JT. Interpretation of data, JT, DK, and RB. Drafting of manuscript: JT, DK, and RB. Critical revision: JT, DK, RJ, and RB. All authors contributed to the article and approved the submitted version.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2020.572503/full#supplementary-material>.

- Diedenhofen, B., and Musch, J. (2015). cocor: a comprehensive solution for the statistical comparison of correlations. *PLoS One* 10 (4), e0121945. doi:10.1371/journal.pone.0121945
- Douglas, H. (2015). Politics and science: untangling values, ideologies, and reasons. *Ann. Am. Soc. Political Soc. Sci.* 658 (1), 296–306. doi:10.1177/0002716214557237
- Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G\*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39 (2), 175–191. doi:10.3758/bf03193146
- Fisher, M., Knobe, J., Strickland, B., and Keil, F. C. (2017). The influence of social interaction on intuitions of objectivity and subjectivity. *Cognit. Sci.* 41 (4), 1119–1134. doi:10.1111/cogs.12380
- Fisher, M., Knobe, J., Strickland, B., and Keil, F. C. (2018). The tribalism of truth. *Sci. Am.* 318 (2), 50–53. doi:10.1038/scientificamerican0218-50
- Gierth, L., and Bromme, R. (2020). Attacking science on social media: how user comments affect perceived trustworthiness and credibility. *Publ. Understand. Sci.* 29 (2), 230–247. doi:10.1177/0963662519889275
- Hall, A. (1980). *Philosophers at war: the quarrel between Newton and Leibniz*. Cambridge, United Kingdom: Cambridge University Press. doi:10.1017/CBO9780511524066.
- Hendriks, F., Kienhues, D., and Bromme, R. (2015). Measuring laypeople's trust in experts in a digital age: the Muenster epistemic trustworthiness inventory (METI). *PLoS One* 10 (10), e0139309. doi:10.1371/journal.pone.0139309
- Hendriks, F., and Jucks, R. (2020). Does uncertainty in news articles affect readers' trust and decision-making? *Media Commun.* 8, 2. doi:10.17645/mac.v8i2.2824
- Hendriks, F., and Kienhues, D. (2019). "Science understanding between scientific literacy and trust: contributions of psychological and educational research," in *Handbooks of communication science*. Editors A. Lefmöllmann, M. Dascal, and T. Glöning (Berlin, Germany: de Gruyter), 17, 29–50.
- Hendriks, F., Kienhues, D., and Bromme, R. (2016). Disclose your flaws! Admission positively affects the perceived trustworthiness of an expert science blogger. *Stud. Commun. Sci.* 16 (2), 124–131. doi:10.1016/j.scoms.2016.10.003
- Hofer, B. K. (2000). Dimensionality and disciplinary difference in personal epistemology. *Contemp. Educ. Psychol.* 25 (4), 378–408. doi:10.1006/ceps.1999.1026
- Jenkins, E. W. (1999). School science, citizenship and the public understanding of science. *Int. J. Sci. Educ.* 21 (7), 703–710. doi:10.1080/095006999290363
- Jennings, F. J., and Russell, F. M. (2019). Civility, credibility, and health information: the impact of uncivil comments and source credibility on attitudes about vaccines. *Publ. Understand. Sci.* 28 (4), 417–432. doi:10.1177/0963662519837901
- Jensen, J. D. (2008). Scientific uncertainty in news coverage of cancer research: effects of hedging on scientists' and journalists' credibility. *Hum. Commun. Res.* 34 (3), 347–369. doi:10.1111/j.1468-2958.2008.00324.x
- Kardash, C. M., and Edwards, O. V. (2012). Thinking and behaving like scientists: perceptions of undergraduate science interns and their faculty mentors. *Instr. Sci.* 40 (6), 875–899. doi:10.1007/s11251-011-9195-0

- Kienhues, D., Jucks, R., and Bromme, R. (2020). Sealing the gateways for post-truthism: reestablishing the epistemic authority of science. *Educ. Psychol.* 55 (3), 144–154. doi:10.1080/00461520.2020.1784012
- Kitcher, P. (1990). The division of cognitive labor. *J. Philos.* LXXXVII (1), 5–22. doi:10.2307/2026796
- Kitcher, P. (2001). *Science, truth, and democracy*. New York, NY: Oxford University Press.
- Kolstø, S. D. (2001). Scientific literacy for citizenship: tools for dealing with the science dimension of controversial socioscientific issues. *Sci. Educ.* 85, 291–310. doi:10.1002/sce.1011
- König, L., and Jucks, R. (2019). When do information seekers trust scientific information? Insights from recipients' evaluations of online video lectures. *Int. J. Educ. Technol. Higher Edu.* 16 (1), 1. doi:10.1186/s41239-019-0132-7
- Lakatos, I., and Musgrave, A. (1970). *Criticism and the growth of knowledge*. Cambridge, United Kingdom: Cambridge University Press.
- Leitão, S. (2000). The potential of argument in knowledge building. *Hum. Dev.* 43 (6), 332–360. doi:10.1159/000022695
- Lilienfeld, S. O. (2012). Public skepticism of psychology: why many people perceive the study of human behavior as unscientific. *Am. Psychol.* 67 (2), 111–129. doi:10.1037/a0023963
- Lobato, E. J. C., and Zimmerman, C. (2019). Examining how people reason about controversial scientific topics. *Think. Reas.* 25 (2), 231–255. doi:10.1080/13546783.2018.1521870
- Lonka, K., Ketonen, E., and Vermunt, J. D. (2020). University students' epistemic profiles, conceptions of learning, and academic performance. *Higher education*. doi:10.1007/s10734-020-00575-6
- McKee, M., and Diethelm, P. (2010). How the growth of denialism undermines public health. *BMJ* 341, c6950. doi:10.1136/bmj.c6950
- Medlin, M. M., Sacco, D. F., and Brown, M. (2019). Political orientation and belief in science in a U.S. college sample. *Psychol. Rep.* 123 (5), 1688–1702. doi:10.1177/0033294119889583
- Merk, S., Rosman, T., Rueß, J., Syring, M., and Schneider, J. (2017). Pre-service teachers' perceived value of general pedagogical knowledge for practice: relations with epistemic beliefs and source beliefs. *PLoS One* 12 (9), e0184971. doi:10.1371/journal.pone.0184971
- Merton, R. K. (1942). A note on science and democracy. *J. Legal Polit. Soc.* 1, 115.
- Mitroff, I. I. (1974). Norms and counter-norms in a select group of the apollo moon scientists: a case study of the ambivalence of scientists. *Am. Socio. Rev.* 39 (4), 579–595. doi:10.2307/2094423
- Munro, G. D. (2010). The scientific impotence excuse: discounting belief-threatening scientific abstracts. *J. Appl. Soc. Psychol.* 40 (3), 579–600. doi:10.1111/j.1559-1816.2010.00588.x
- Murphy, P. K. (2015). Marking the way: school-based interventions that “work”. *Contemp. Educ. Psychol.* 40, 1–4. doi:10.1016/j.cedpsych.2014.10.003
- Mutz, D. C., and Reeves, B. (2005). The new videomalaise: effects of televised incivility on political trust. *Am. Polit. Sci. Rev.* 99 (1), 1–15. doi:10.1017/S0003055405051452
- Nau, C., and Stewart, C. O. (2013). Effects of verbal aggression and party identification bias on perceptions of political speakers. *J. Lang. Soc. Psychol.* 33 (5), 526–536. doi:10.1177/0261927X13512486
- Oreskes, N., and Conway, E. M. (2010). *Merchants of doubt: how a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. London, United Kingdom: Bloomsbury Press.
- Osborne, J. (2010). Arguing to learn in science: the role of collaborative, critical discourse. *Science* 328 (5977), 463–466. doi:10.1126/science.1183944
- Paletz, S. B. F., Chan, J., and Schunn, C. D. (2016). Uncovering uncertainty through disagreement. *Appl. Cognit. Psychol.* 30 (3), 387–400. doi:10.1002/acp.3213
- Peters, H.-P. (2013). Gap between science and media revisited: scientists as public communicators. *Proc. Natl. Acad. Sci. U.S.A.* 110 (3), 14102–14109. doi:10.1073/pnas.1212745110
- Popan, J. R., Coursey, L., Acosta, J., and Kenworthy, J. (2019). Testing the effects of incivility during internet political discussion on perceptions of rational argument and evaluations of a political outgroup. *Comput. Hum. Behav.* 96, 123–132. doi:10.1016/j.chb.2019.02.017
- Questback GmbH (2018). EFS survey (fall 2018). Cologne, Germany: Questback GmbH. Available at: <https://www.unipark.com/>.
- R Core Team (2018). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.
- National Research Council (2001). *Scientific research in education*. Washington D.C, United States: The National Academies Press.
- Rowe, I. (2015). Civility 2.0: a comparative analysis of incivility in online political discussion. *Inf. Commun. Soc.* 18 (2), 121–138. doi:10.1080/1369118X.2014.940365
- Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educ. Res.* 31 (7), 15–21. doi:10.3102/0013189X031007015
- Sodhi, M., and Etminan, M. (2020). Safety of ibuprofen in patients with COVID-19. *Chest* 158 (1), 55–56. doi:10.1016/j.chest.2020.03.040
- Tversky, A., and Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychol. Sci.* 3 (5), 305–310. doi:10.1111/j.1467-9280.1992.tb00678.x
- Zeuch, N., and Souvignier, E. (2015). Measurement of scientific thinking of pre-service teachers—development of a new instrument and identification of latent profiles. *Beltz Juventa* 43 (3), 245–262.
- Zlatkin-Troitschanskaia, O. (2016). Evidence-based actions within the multilevel system of schools – requirements, processes, and effects (EviS). *J. Educati. Res. Online* 8 (3), 5–13. Available at: <http://www.j-e-r-o.com/index.php/jero/article/view/701>.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Tkotz, Kienhues, Jucks and Bromme. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Key Information Processes for Thinking Critically in Data-Rich Environments

Jacqueline P. Leighton\*, Ying Cui and Maria Cutumisu

Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada

## OPEN ACCESS

### Edited by:

Patricia A. Alexander,  
University of Maryland,  
United States

### Reviewed by:

Sheng-Yi Wu,  
National Pingtung University, Taiwan  
Hongyang Zhao,  
University of Maryland, College Park,  
United States

### \*Correspondence:

Jacqueline P. Leighton  
jacqueline.leighton@ualberta.ca

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Education

**Received:** 13 May 2020

**Accepted:** 21 January 2021

**Published:** 24 February 2021

### Citation:

Leighton JP, Cui Y and Cutumisu M  
(2021) Key Information Processes for  
Thinking Critically in Data-  
Rich Environments.  
Front. Educ. 6:561847.  
doi: 10.3389/feduc.2021.561847

The objective of the present paper is to propose a *refined conception* of critical thinking in data-rich environments. The rationale for refining critical thinking stems from the need to identify specific information processes that direct the suspension of prior beliefs and activate broader interpretations of data. Established definitions of critical thinking, many of them originating in philosophy, do not include such processes. A refinement of critical thinking in the digital age is developed by integrating two of the most relevant areas of research for this purpose: First, the *tripartite model* of critical thinking is used to outline proactive and reactive information processes in data-rich environments. Second, a new assessment framework is used to illustrate how educational interventions and assessments can be used to incorporate processes outlined in the tripartite model, thus providing a defensible conceptual foundation for inferences about higher-level thinking in data-rich environments. Third, recommendations are provided for how a performance-based teaching and assessment module of critical thinking can be designed.

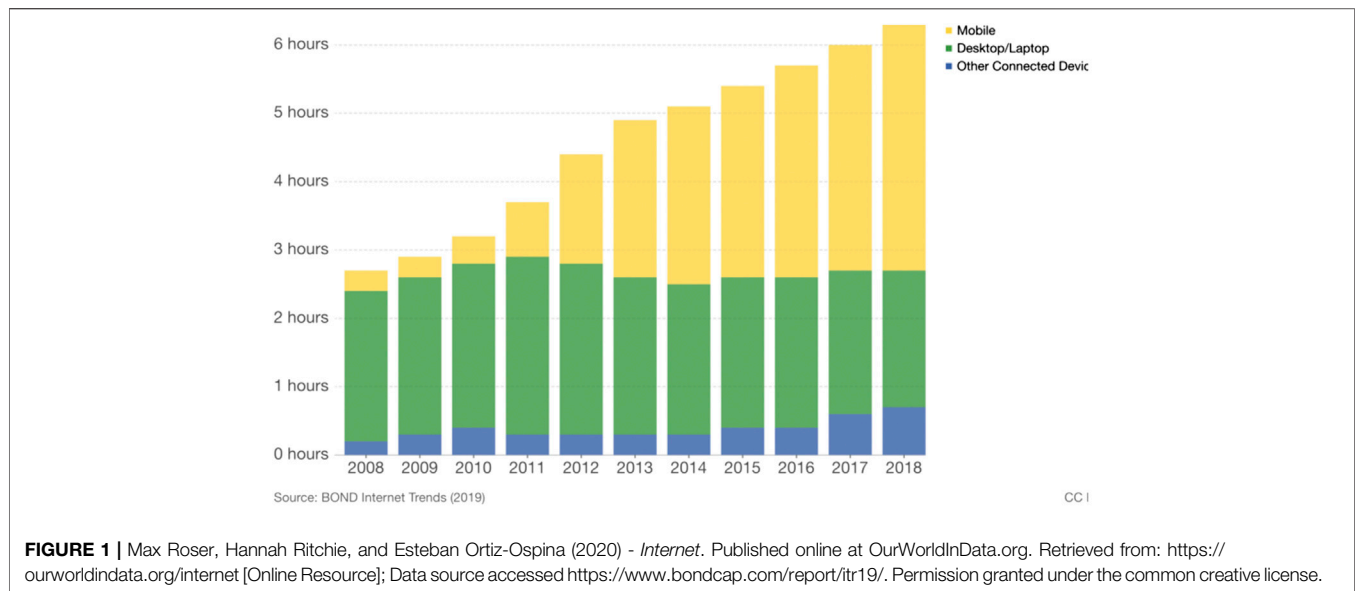
**Keywords:** post-secondary education, critical thinking, data-rich environments, cognitive biases, performance assessments

## INTRODUCTION

In response to the question, how much data are on the internet, Gareth Mitchell from Science Focus Magazine answers the question by considering the overall data held by just four companies - Amazon, Facebook, Google, and Microsoft (<https://www.sciencefocus.com/future-technology/how-much-data-is-on-the-internet/>). These four companies are estimated to hold a sum total of at least 1,200 petabytes (PB) of online data, which equals 1.2 million terabytes (TB) or 1.2 trillion gigabytes (GB). Neuroscientists propose that the average human brain holds 2.5 PB or 2.5 million GB of information in memory (Reber, 2010), or just over 7 billion 60,000-word books. However, information stored in memory is often subject to error not only from the way it is encoded but also retrieved (Mullet and Marsh, 2016).

Critical thinking requires people to minimize bias and error in information processing. Students entering post-secondary education today may be “digital natives” (Prensky, 2001) but they are still surprisingly naïve about how to critically think about the wealth of digital information available. According to Ridsdale et al. (2015), youth may be quite adept at using digital hardware such as smart phones and apps but they often lack the *mindware* to think and act critically with the information they access with their devices (Stanovich, 2012). Although this lack of mindware can be observed in the mundane activities of how some first-year undergraduates might tackle their research assignments, it is dramatically illustrated in the political narratives of radicalized young adults (Alava et al., 2017). Young adults are particularly vulnerable to misinformation because they are in





the process of developing their cognitive abilities and identities (Boyd, 2014). The objective or rationale for this paper is to propose a *refined conception* (Ennis, 2016) of critical thinking in data-rich environments. It is the authors' view that a refined conception is required because data-rich environments have ushered in many cognitive traps and the potential for personal biases to derail critical thinking as traditional understood. The research questions addressed in this conceptual paper are as follows: What can traditional definitions of critical thinking gain by considering explicit inclusion of cognitive biases? How can refined definitions of critical thinking be incorporated into theoretical frameworks for the design of performance assessments?

One of the most recommended strategies for helping young adults analyze and navigate online information is to directly and explicitly teach and assess critical thinking (Alava et al., 2017; Shavelson et al., 2019). However, teaching and assessing critical thinking is fraught with difficulties, including a multitude of definitions, improper evaluation, and studies that incorporate small samples and controls (Behar-Horenstein and Niu, 2011; El Soufi and Huat See, 2019). Aside from these predictable difficulties, new challenges have emerged. For example, the informational landscape has changed over the course of the last 30 years. The rapid increase in quantity coupled with the decrease in quality of much online information challenges the limits of human information processing.

Critical thinking today is primarily conducted in data-rich online environments, meaning that postsecondary students are searching, navigating, and thinking about a virtually limitless number of sources. Oxford University's Change Data Lab (Roser et al., 2020) writes: "adults aged 18–29 in the US are more likely to get news indirectly via social media than directly from print newspapers on news sites; and they also report being online 'almost constantly.'" As shown in **Figure 1**, not only is the total time spent online increasing but the increase is mostly the time spent on mobile phones. As mobile phones are smaller devices,

compared to desktops, laptops, and tablets, they can be expected to force even faster navigation and processing of information, which would be expected to increase the odds of error-prone thinking.

Cognitive traps are ubiquitous in online data-rich environments. For example, information can be presented as serious and credible when it is not. However, traditional critical thinking definitions have not tended to focus on avoiding cognitive traps; namely, how processing errors can be avoided. This creates a problem not only for teaching but also assessing critical thinking among postsecondary students in today's classrooms. Thus, there are at least two research opportunities in addressing this problem: 1) provide a refinement of what critical thinking entails specifically for the teaching and assessment of critical thinking in data-rich environments and 2) illustrate a framework for the design of teaching and assessment modules that can lead to stronger inferences about students' critical thinking skills in today's information world.

The present paper contributes to the literature on critical thinking in data-rich environments by providing a refinement of what critical thinking entails for teaching and assessment in data-rich environments. The refinement is rooted in cognitive scientific advancements, both theoretical and empirical, of higher-level thinking, and essentially attempts to offer test designers an update on the construct of critical thinking. In other words, this conceptual analysis does the work of translating key psychological aspects of the critical thinking construct for pragmatic purposes—student assessment. Building on the refinement of this construct, the paper also includes recommendations for the type of framework that should guide the design of teaching and assessment modules so that key aspects of students' critical thinking skills are not missed. Toward this end, this refinement can enhance the *construct representation* of assessments of critical thinking in data-rich environments. Educational assessments are only as good as their representation of the construct intended for measurement.

Without the ongoing refinement of test constructs such as critical thinking, assessments will not provide the most accurate information in the generation of inferences of student thought; refinements of test constructs are especially vital in complex informational landscapes (Leighton and Gierl, 2007). Thus, a refinement of critical thinking among young adults in data-rich environments is developed by integrating two of the most topical and relevant areas of research for this purpose: First, Stanovich and Stanovich's (2010) *tripartite model* of critical thinking is used to outline the limitations of human information processing systems in data-rich environments. Second, Shavelson et al.'s (2019) assessment framework is used to illustrate how specific educational assessment designs can be built on the tripartite model and can provide a more defensible evidentiary base for teaching and drawing inferences about critical thinking in data-rich environments. The paper concludes with an illustration of how mindware can be better integrated into teaching and performance-based assessments of critical thinking. The present paper contributes directly to the special issue on *Assessing Information Processing and Online Reasoning as a Prerequisite for Learning in Higher Education* by refining the conceptualization of critical thinking in data-rich environments among postsecondary students. This refinement provides an opportunity to guide instructive and performance-based assessment programs in the digital age.

## THEORETICAL FRAMEWORKS UNDERLYING MINDWARE FOR CRITICAL THINKING

In the 1999 science fiction movie *Matrix* Wachowski et al. (1999), human beings download computer “programs” to allow them to think and function in a world that has been overtaken by intelligent machines. Not only do these programs allow human beings to live in a dream world, which normalizes a dystopian reality, but also to effortlessly disregard their colonization. Cognitive scientists propose something analogous to these “programs” for human information processing. For example, Perkins (1995) coined the term *mindware* to refer to information processes, knowledge structures, and attitudes that can be acquired through instruction to foster good thinking. Rizeq et al. (2020, p. 2) indicate contaminated mindware as “beliefs that may be unhelpful and that may inhibit reasoning processes ... (Stanovich, 2009; Stanovich et al., 2008; Stanovich, 2016).”

Treating human information processing as analogous to computer programs, which can be contaminated, is useful and powerful because it highlights the presence of errors or bugs in thinking that can invariably distort the way in which data are perceived and understood, and instantaneously “infect” the thinking of both self and others. However, the predictability of such programs also permits anticipating when these thinking errors are likely to occur. Educational interventions and assessments can be designed to capitalize on the predictability of thinking errors to provide a more comprehensive level of thinking instruction and evaluation. Specifying what critical thinking entails in *data-rich environments* requires explicit

attention not only to the information processes, knowledge structures, and attitudes that instantiate good critical thinking but also to the thinking bugs that derail it. Hyytinen et al. (2019, p. 76) indicate that a critical thinker needs to have knowledge of what is reasonable, the thinking skills to evaluate and use that knowledge, as well as dispositions to do so (Facione, 1990; Halpern, 2014; Hyytinen et al., 2015).” We agree but we would go further in so far as critical thinkers also need to know what their own biases are and how to avoid cognitive traps (Toplak and Flora, 2020).

## Traditional Definitions of Critical Thinking

Established or traditional definitions of critical thinking have typically focused on the *proactive* processes that comprise critical thinking (Leighton, 2011). Proactive processes are positive in action. Proactive processes, such as *analyzing* and *evaluating*, are often the focus of educational objectives (e.g., Bloom's taxonomy; Bloom, 1956). Proactive processes help to identify the actions and goals of good thinking in ideal or optimal conditions. However, they are not particularly useful for creating interventions or assessments intended to diagnose faulty thinking (Leighton and Gierl, 2007). The problem is that these processes reflect only aspects of good thinking and do not reflect other processes that should be avoided for good thinking to occur. For example, *reactive* thinking processes such as *neglecting* and *confirming* must be resisted in order for proactive processes do their good work. Reactive processes are not bad in many circumstances, especially those where thinking has to be quick to avoid imminent danger (Kahneman, 2011). However, in circumstances where imminent danger is not present and actions can be enhanced by careful processing of information, it can be useful to learn about reactive processes; this is especially relevant for designing teaching interventions and assessments of critical thinking (Leighton, 2011).

The omission of reactive processes in traditional definitions of critical thinking is perhaps not surprising since many of these definitions grew out of philosophy and not out of empirical disciplines such as experimental psychology (Ennis, 2015, Ennis, 2016). Nonetheless, this section addresses established definitions in order to provide a conceptual foundation on which to build more, targeted definitions of critical thinking for specific purposes.

## Proactive Processes in Critical Thinking

Ennis (2016) provides a justification for distinguishing the *basic concept* of critical thinking from a particular *conception* of it; that is, a particular definitional instance of it in specific situations. In an analysis of the many theoretically inspired definitions of critical thinking, Ennis (2016, p. 8) explains that many established definitions share a conceptual core. To illustrate this core, consider three definitions of critical thinking outlined in Ennis (2016, p.8-9):

1. “Active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it and the further conclusions to which it tends” (Dewey, 1933, p. 9 [first edition 1910]).

2. “Purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based” (Facione 1990; Table 1).
3. “Critical thinking is skilled, active interpretation and evaluation of observations, communications, information, and argumentation as a guide to thought and action” (Fisher and Scriven 1997, p. 20).

These three examples illustrate what Ennis (2016, p. 11) considers to be the defining processes of critical thinking, namely, “the abilities to analyze, criticize, and advocate ideas” and “reach well-supported . . . conclusions.” These proactive processes represent the conceptual core.

Aside from the conceptual core, Ennis (2016) suggests that variations or distinct conceptions of critical thinking can be proposed without endangering the core concept. These variations arise from particular teaching and assessment situations to which the core concept is applied and operationalized. For example, in reviewing four different examples of particular teaching and assessment cases [i.e., Ennis’s (1996) Alpha Conception, Clemson’s (2016) Brief Conception, California State University (2011), and Edward Glaser’s (1941) Brief Conception of Critical Thinking], Ennis (2016) explains that in each case the concept of critical thinking is operationalized to have a particular meaning in a given context. Ennis (2016) concludes:

In sum, differences in the mainstream concept [of critical thinking] do not really exist, and differences in conceptions that are based on the mainstream concept of critical thinking are usually to a great extent attributable to and appropriate for the differences in the situations of the people promoting the conception. (p. 13)

Building on Ennis’ (2016) proposal, then, a *conception* of critical thinking is offered herein to serve a specific purpose: to teach and assess critical thinking skills in data-rich environments. To do this, the core concept of critical thinking must include those information processes that guard against manipulability in data-rich environments.

### Reactive Processes in Critical Thinking

Educational interventions and assessments must address reactive processes if they are to bolster critical thinking in non-idealized conditions. This is especially important in data-rich environments where information is likely to be novel, abundant (almost limitless), and quickly accessible. The tendency for people to *simplify* their information processing is amplified in data-rich environments compared to data-poor environments where information is routine and can be comfortably processed serially (e.g., writing a term paper on a familiar topic with ample time allowance). The simplification of data is necessary as the human brain only processes about 5–7 pieces of information in working memory

at any one time (Miller, 1956; see also; Cowan, 2001). This limitation exists atop the more basic limitation of what can be consciously perceived in the visual field (Kroll et al., 2010). Thus, human beings instinctively simplify the signals they receive in order to create a manageable information processing experience (Kroll et al., 2010).

Most of the information simplified and perceived will be forgotten unless it is actively processed via rehearsal and transfer into long-term memory. However, rehearsed information is not stored without error. Storage contains errors because another limitation of information processing is that memory is a *constructive process* (Schacter, 2012). What is encoded is imbued with the schemata already in memory, and what is then retrieved depends on how the information was encoded. Thus, aside from the error-prone simplification process that permits the human information process to perceive successful navigation of the environment, there is the error-prone storage-and-retrieval process that characterizes memory. Data-rich environments accentuate these significant limitations of human information processing. Consequently, identifying both *proactive* and *reactive information processes* is necessary to generate realistic educational interventions and assessments that can help 1) ameliorate thinking bugs in today’s data-rich environments while at the same time 2) cultivating better mindware for critical thinking.

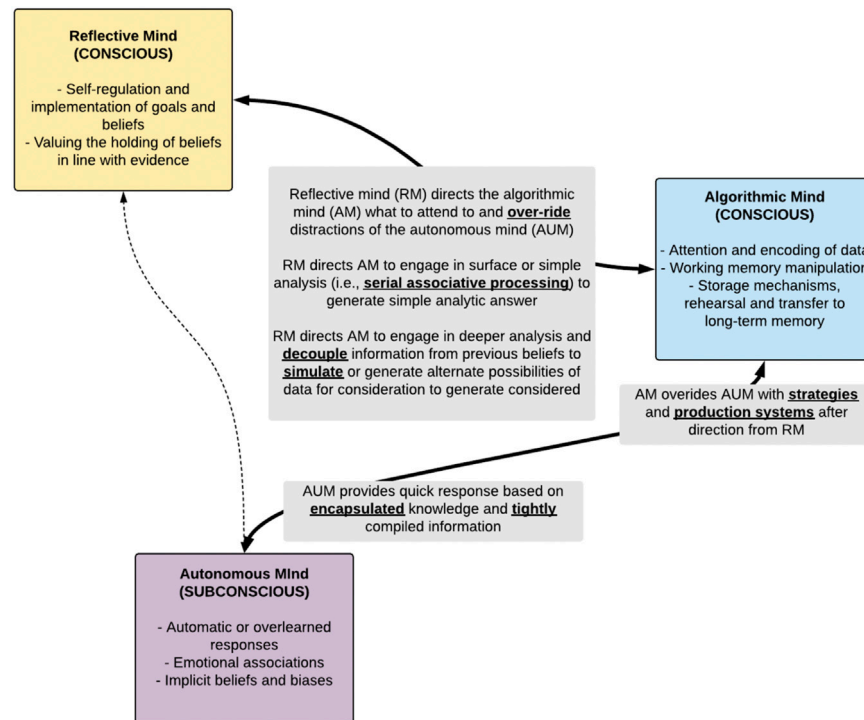
### The Tripartite Model of Critical Thinking

One of the largest problems with modern initiatives to teach and assess critical thinking in data-rich environments is the neglect of empirically based theoretical frameworks to guide efforts (Leighton, 2011). Without such frameworks, the information processes taught and measured are primarily informed by philosophical instead of psychological considerations. The former emphasizes proactive over reactive processes but both are needed. The emphases on proactive processes does not actually help educators identify and rectify the existing bugs in students’ mindware.

The conception of critical thinking that is advanced here is based on Stanovich and Stanovich’s (2010; see also Stanovich, 2021) *Tripartite Model*. The model focuses on both proactive and reactive processes. Unlike philosophical treatments of critical thinking, the tripartite model devotes significant attention to biased and error-prone information processing. According to Stanovich and Stanovich (2010, p. 219; italics added): “the tendency to process information *incompletely* has been a major theme throughout the past 30 years of research in psychology and cognitive science (Dawes, 1976; Taylor, 1981; Tversky and Kahneman, 1974).” The tripartite model does *not* provide a simple definition of what critical thinking entails given the complexity of the processes involved. Instead, it provides an outline of three levels of mindware that have been found to be constantly interacting in the process of critical thinking.

### Three Levels of the Mind

The tripartite model integrates decades of cognitive and neuroscientific research, ranging from Tversky and



**FIGURE 2 |** Adapted tripartite model (Stanovich and Stanovich, 2010) to illustrate the connections among three different aspects or minds integral to human cognition.

Kahneman's (1974) early work on biases and heuristics to the later work on dual process models of thinking (Evans, 2003). The model shown in **Figure 2** illustrates the relations between three distinct levels of information processing—the reflective mind (RM), the algorithmic mind (AM), and the autonomous mind (AUM). In **Figure 2**, the level of information processing that functions to manipulate data in working memory, store, retrieve, and generate responses is the AM. This is the level that is directly on display and observed when human beings process and respond to questions, for example, on educational assessments and tests of intelligence. The AM can be defined by its processing speed, pattern recognition and retrieval from long-term memory, and manipulation of data in working memory.

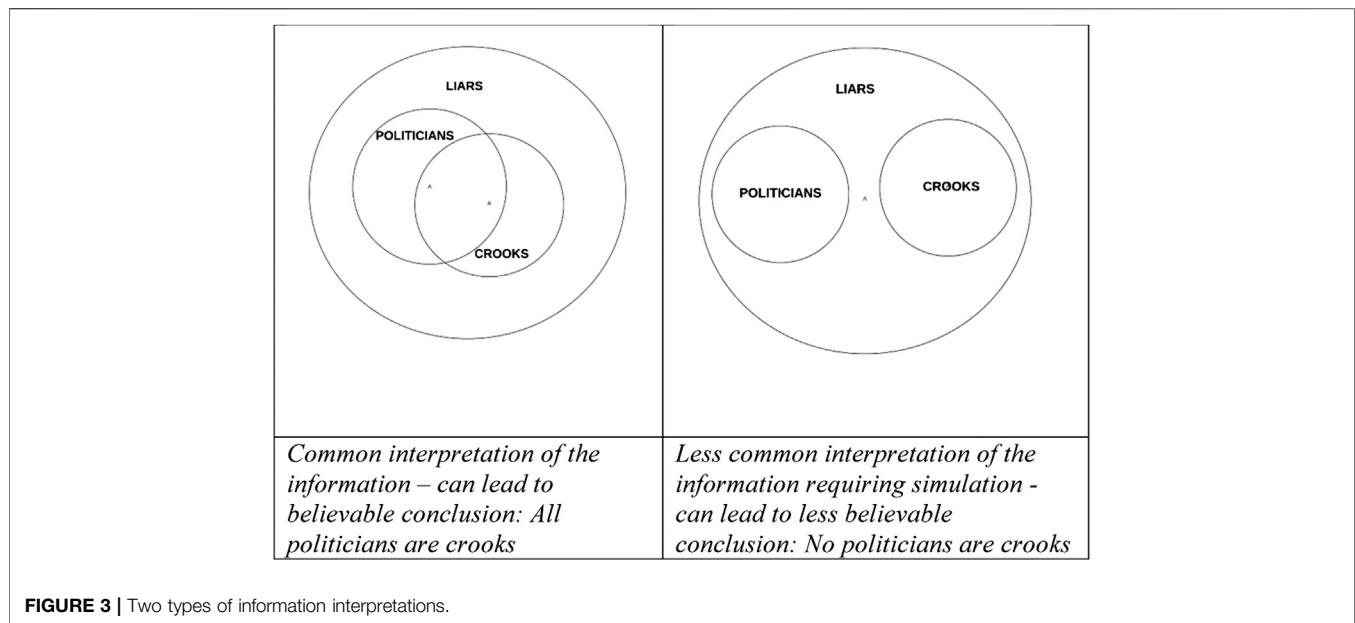
The AM takes direction from two sources—the reflective mind or RM and the autonomous mind or AUM. The AUM is the subconscious part of human information processing that retains data acquired by means of imprinting, tacit and procedural learning, and emotionally laden events, resulting in many forms of automatic responses and implicit biases. The AUM is the level at which encapsulated or modularized knowledge can be retrieved to generate a quick and simplified response, which exerts minimal load on working memory. Depending on the influence of the AUM, the AM is capable of biased or unbiased responses. For example, in view of what *appears* to be a large insect, the AUM signals the AM to focus on getting out of the way. This is a biased response but it is an expedient response that is often observed in logical tasks (see Leighton and Dawson, 2001).

Unlike the AUM, the RM is a conscious and deliberative aspect of human information processing. The RM is the part of information processing that involves goals, beliefs, and values. It is the part of the mind that provides *intentionality* to human behavior (Dennett, 1987). It directs the AM to suspend simple processing and expend the cognitive effort to deeply process information. The RM also functions to direct the AM to resist or override signals from the AUM to respond too quickly. Thus, it is the information processing directed by the RM—to engage and suspend certain processes - that needs to be the focus of most educational interventions and assessments of critical thinking.

### Decoupling and Simulation Processes

According to Stanovich and Stanovich (2010), the RM directs the AM to engage in two forms of proactive information processes. Both require cognitive effort. First, *decoupling* involves the process of suspending prior beliefs and attending to information in the context in which it is provided. For example, decoupling processes have been examined in belief bias studies (Leighton and Sternberg, 2004). In these studies, participants are typically asked to evaluate arguments that have been created to differ along two dimensions—logical soundness and believability of conclusion. For example, a *logically flawed* argument is paired with a *believable* conclusion, for example, All politicians are liars; All crooks are liars; Therefore, all politicians are crooks. In response to these types of arguments, participants have been found to accept conclusions that are believable rather





than logically sound. However, performance can be improved by instructing participants to explicitly consider the structure of the argument. In other words, the instructions are clearly designed to engage the RM. When explicit instructions are included, participants will show improved performance in correctly rejecting conclusions from flawed arguments.

Second, for decoupling to work, *simulation* is often activated in tandem. Simulation involves the process of actively considering distinct ways of interpreting information. For example, shown in **Figure 3** are two panels showing distinct interpretations of the premises of the argument provided earlier about politicians and crooks. The one on the left shows the easiest interpretation or mental model of the argument about politicians (conclusion - *All politicians are crooks*). The interpretation shown on the left is one which often may correspond to prior beliefs. On the right, an additional interpretation can be created to indicate that *no politicians are crooks*. The interpretation shown on the right may be less common but equally plausible given the premises of the argument. The effort to create additional interpretations or simulate information that contradicts prior beliefs has been found to correlate positively with working memory capacity (Johnson-Laird and Bara, 1984). In fact, both decoupling and simulation have been found to require significant working memory resources and, thus, cognitive effort for participants to willingly adopt (Johnson-Laird and Bara, 1984; Leighton and Sternberg, 2004; Stanovich, 2011; Leighton and Sternberg, 2012).

Most classroom assessments and achievement tests, even those that are purportedly designed to be cognitively complex, are not developed to evaluate whether students can decouple or simulate thinking (Leighton, 2011). Instead most tests are developed to measure whether students can reproduce what they have learned in the classroom, namely, a form of optimal performance given instruction (Leighton and Gierl, 2007; Stanovich and Stanovich, 2010). Often, then, there is little incentive for students to begin to suspend beliefs and imagine situations where what they have been

told *does not* hold. Not surprisingly, most students try to avoid “overthinking” their responses on multiple-choice or even short-answer tests precisely because such simulated thinking could lead to choosing an unexpected or non-keyed response.

### Suspending Serial Associative Processing

Unlike the thinking evoked by most classroom and achievement tests, information processing in data-rich environments calls for a different standard of evaluation. Data-rich environments typically offer students the possibility to navigate freely through multiple sites, unrestricted by time limits and/or instructions about how their performance will be evaluated. In such open, data-rich environments, individuals set their own standard of performance. According to the tripartite model, *serial associative processing* is likely to be the standard most often set by individuals. Serial associative processing is directed by the RM but it is simple processing nonetheless. It means that information is accepted as it is presented or rejected if it fails to conform with what is already known (prior beliefs). There is no decoupling or simulation. Johnson-Laird and Bara (1984; Johnson-Laird, 2004) called this simple type of processing *single-model reasoning* because information is attended and processed but goes unchallenged. Serial associative processing is different from the automatic responses originating in the AUM. Serial associative processing does involve analysis and evaluation but it does not consider multiple perspectives and so it is biased in its implementation.

### Critical Thinking as Coordinated Suspension and Engagement of Information Processes

Consider again the defining processes Ennis (2016, p. 11) proposes for critical thinking: “the abilities to analyze, criticize, and advocate ideas” and “reach well-supported . . . conclusions.” In light of Stanovich and Stanovich’s (2010) model, the processes mentioned by Ennis only reflect the AM and do not reflect the

*coordinated effort of the RM and AM* to suspend serial associative processing and engage in decoupling and simulation. In other words, what is missing in most traditional conceptions of critical thinking are reactive processes, namely, processes that lead thinking astray such as *serial associative processing*, which must be suspended for better thinking to emerge.

In data-rich environments, actively resisting serial associative processing is a necessary component of critical thinking. This form of information processing must be actively resisted because the incentive is for individuals to do the opposite in the wake of massive amounts of information. Although applying this resistance will be cognitively effortful, it can be learned by teaching students to become more meta-cognitively aware of their information processing. However, even meta-cognitive awareness training is unlikely to help students resist serial associative processing, if critical thinking is under-valued by the RM. Thus, the design of teaching interventions and assessments must consider the construct of critical thinking not as a universally accepted and desired form of thinking but as a skill that students choose to apply or ignore (Leighton et al., 2013). Consequently, interventions must persuade students of the benefits associated with critical thinking and assessments need to measure the processes that are most relevant for critical thought (e.g., decoupling and simulation). In the next section, Shavelson et al.'s (2019) assessment framework is used to illustrate how specific educational assessment designs can build on the tripartite model of critical thinking, and provide a more defensible conceptual foundation for inferences about critical thinking in data-rich environments.

## Measuring Decoupling and Simulation: Shavelson et al.'s (2019) Assessment Framework

Shavelson et al.'s (2019) assessment framework is premised on three objectives. First, performance assessments are appropriate for measuring higher-level thinking constructs; second, assessments of higher-level thinking constructs should be developed in ways that clearly link scores to claims about postsecondary students' capabilities; and third, higher-level thinking constructs, such as critical thinking, should require postsecondary students to make sense of complex information outside typical classroom environments. Each of these objectives is elaborated and connected to measuring key information processes for critical thinking.

### Performance Assessments

Performance assessments typically contain tasks (i.e., selected and constructed) that require test-takers to attend to multiple types of materials (e.g., articles, testimonials, videos) and generate responses that involve an evaluation of those materials for the purpose of providing a reasoned answer on a topic. The topic is often novel and the tasks are complex such as evaluating a claim about whether a privately funded health-care clinic should be adopted by a community. The goal of a performance assessment is to approximate the informational demands of a real-world situation, calling on individuals to have

to weigh different perspectives in the process of analyzing and evaluating materials.

The motivation to approximate real-world situations is a requirement in performance assessments. The constructs measured need to be assessed in the types of situations that justify making claims about what the test-taker can do in a context that approximates real-life. For example, performance assessments would *not* be the tool to use if the objective was to measure criterion or *optimal performance* (Stanovich and Stanovich, 2010), that is, whether someone has learned the normative timeline for the Second World War or to factor polynomials. Both of these objectives do not reflect the types of skills required in complex environments, where *typical performance* is sought in determining whether the test-taker can invoke and manage specific information processes in providing a response.

### Measuring the Mindware

In Shavelson et al.'s (2019, p. 4) framework, the environments or contexts in which to measure critical thinking are broadly conceived:

- a. contexts in which thought processes are needed for solving problems and making decisions in everyday life, and
- b. contexts in which mental processes can be applied that must be developed by formal instruction, including processes such as comparing, evaluating, and justifying.

In considering both these measurement contexts, data-rich environments satisfy both. For example, the real-life contexts in which people must solve problems and make decisions nowadays typically involve seeking, analyzing, and evaluating a lot of information. Most of this information may be online where there is almost no oversight on quantity or quality control.

However, people do not solve problems and make decisions in a cognitive vacuum. This is where Stanovich and Stanovich's (2010) tripartite model provides the necessary conceptual foundation to Shavelson et al.'s (2019) assessment framework for measuring critical thinking. The beliefs and values of the RM direct the type of information that is sought and how that information should be analyzed. Heretofore, the idea of values has not been elaborated. The valuing of critical thinking or stated differently, holding the value that beliefs should line up with evidence provides an impetus for engaging in effortful thinking. Churchland (2011) indicates that such values—what we consider good, bad, worthwhile or not—are rooted in the brain and have evolved as mechanisms to help human beings adapt and survive. Thus, the question for the reflective mind is one of why is critical thinking beneficial for me? Consequently, the design of performance assessments must include opportunities for measuring two fundamental catalytic processes for critical thinking: (a) whether the RM *values* critical thinking and for what reasons and (b) how the RM then directs the AM to engage or suspend serial associative processing for analyzing and evaluating the resources provided so that critical thinking can be achieved. The reason for measuring whether the RM values

critical thinking is to establishing that a student is indeed motivated to engage in the effort it requires. A student may value critical thinking but not know how to do it, but it is also necessary to determine whether a student does not value it and therefore, irrespective of having the skills to do it, chooses not to do it. The educational intervention for each of these scenarios will be different depending on the cognitive and affective state of the student (Leighton et al., 2013).

The question of how this engagement or suspension is measured is not trivial as it would involve finding a way to measure test-takers' epistemic values, prior beliefs, and biases about the topic. Moreover, it would involve providing confirming or disconfirming sources of data in the assessment at different levels of quality. As test-takers select data sources to analyze and evaluate, evidence of the active suspension of prior belief (i.e., decoupling) and rejection of information at face value (i.e., simulation) needs to be collected to warrant the claim that the information processes inherent to critical thinking were applied.

### Creating the Performance Assessment

According to Evidence Centered Design (ECD; Mislevy et al., 2003), an assessment is most defensibly designed by paying careful attention to the claim that is expected to be made from the assessment performance. In the case of Shavelson et al.'s (2019) assessment framework, the following high-level claim is desired:

[T]he assessment task presented here taps critical thinking on everyday complex issues, events, problems, and the like. The evidence comes from evaluating test-takers' responses to the assessment tasks and potential accompanying analyses of response processes such as think-aloud interviews or log file analyses. (p. 9)

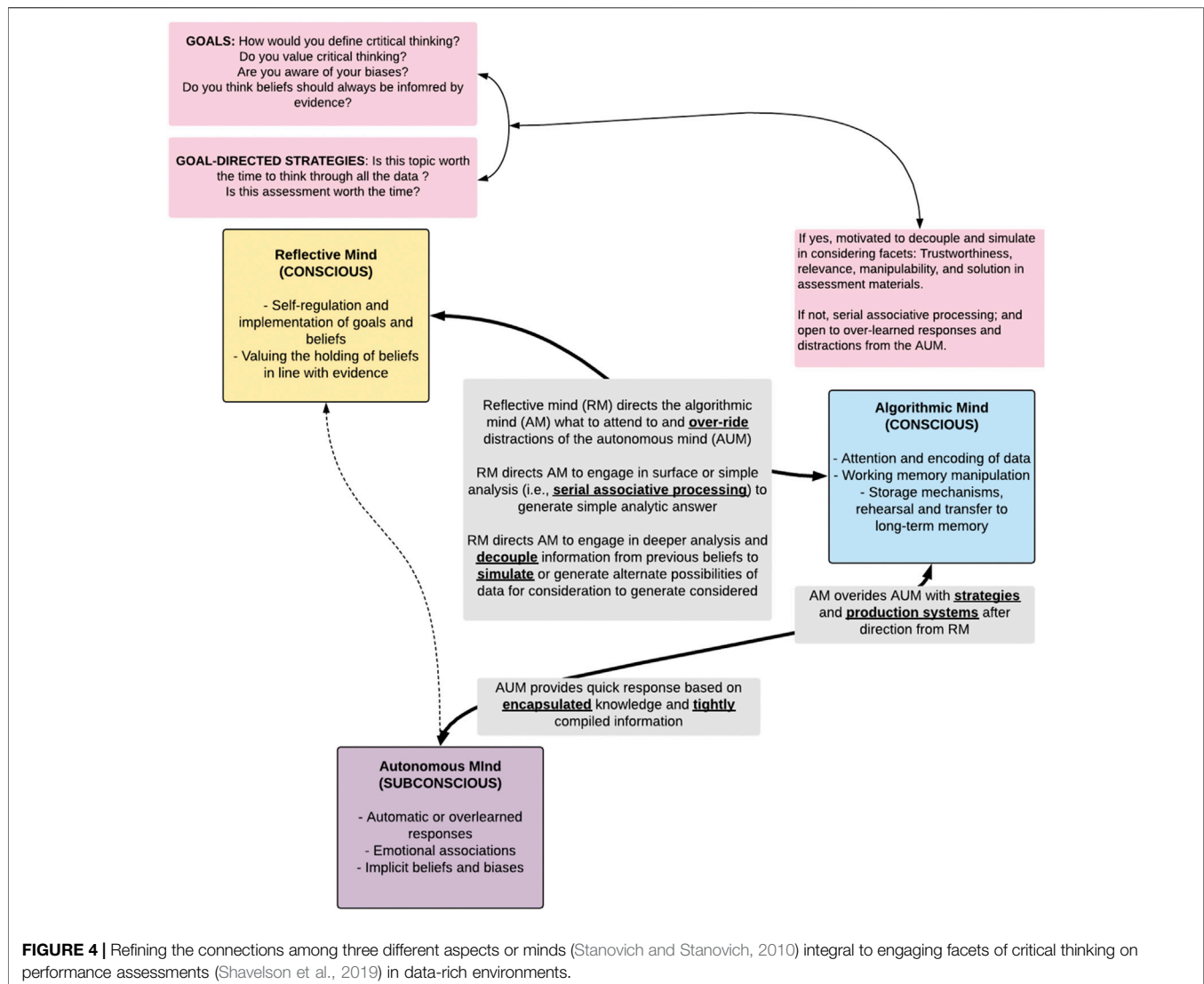
Because the claim includes '*critical thinking on everyday complex issues, events, problems, and the like*' it becomes necessary to situate this claim within the specific data-rich environment that is of most interest to the developer but also the environment that is of most interest to the test-taker. In data-rich environments, thinking will not be general but specifically guided by the relevance of topics. In particular, what is essential to consider in such environments is that individuals are unconstrained by how they search and attend to information given the vast quantity and quality of sources. Thus, test takers' *value proposition of thinking critically for a given topic* needs to be considered in their performance. If respondents do not value it, they are unlikely to engage in the effort required to suspend serial processing. And claims about what they can or cannot do will be less defensible.

At the outset of a performance assessment, a test-taker who does not value critical thinking for a given topic is unlikely to engage the critical information processes expected on the assessment. The following four facets of the data that Shavelson et al. (2019) indicate must be attended are unlikely to be invoked in depth:

1. Trustworthiness of the information or data—is it reliable, unreliable, or uncertain?
2. Relevance of the information or data—is it relevant or irrelevant to the problem under consideration?
3. Manipulability of the information to judgmental/decision/bias—is the information subject to judgmental errors and well-known biases?
4. Solution to the story problem—is the problem one where a judgment can be reached, a decision recommended, or a course of action suggested?

Each of these facets forms the basis of a question that is designed to direct the algorithmic mind (AM) to process the data in a particular way. However, the AM is an information processor that does not direct itself; it is directed by the RM. Consequently, for each of these facets, it is important to consider that both the RM and the AM are being induced and measured. For example, if critical thinking is to be demonstrated, all facets—trustworthiness, relevance, manipulability, and solution generation—require the RM to direct the AM to (a) override the autonomous mind (AU) in its reactionary response, (b) suspend serial associative processing, (c) decouple from pre-existing beliefs, and (d) simulate alternative worlds where the information is considered in the context in which it is presented. Although it is beyond the scope of the paper to illustrate the interplay of the RM and AM for each of these four facets, an example may suffice. Consider a critical thinking task that begins with a story about the delivery of a new vaccine for inoculating people against the COVID19 virus. After presentation of the story, the first item needs to probe the RM - whether the test-taker indicates importance in comprehending a story about vaccine safety. If the test-taker responds "yes," the self-report can be validated against eye tracking reaction time data to check its validity (assuming greater importance would lead to more time spent reading). The second set of items can then probe the test-taker's analysis of the trustworthiness of the information, for example, is the story reliable and how do you know? What information was irrelevant (e.g., the color of the vials) and was it decoupled from relevant information (e.g., the temperature at which the vaccine must be stored)? What variables in the story were re-imagined or simulated (e.g., transportation of a vaccine across multiple freezers might erode its integrity), leading to a different conclusion than the one stated in the story. The response to these second set of items must be evaluated, in aggregate, against the response for the first item in order to determine the rigor of AM thinking devoted to analyzing the veracity of the story and its elements. If the second response is weak, in light of a motivated RM, then one might generate the inference that the test-takers lacks the essential skills to think critically.

The induction of the RM to engage the AM *in a specific manner* in a performance assessment becomes an integral part of the critical thinking construct that is being measured in data-rich environments. In fact, one of the most important questions to be presented to test-takers before they engage with a performance measure of critical thinking might be a question that directly probes the RM to reveal the goals that drive its performance—does the RM value holding beliefs that are in line with evidence? In the



**FIGURE 4 |** Refining the connections among three different aspects or minds (Stanovich and Stanovich, 2010) integral to engaging facets of critical thinking on performance assessments (Shavelson et al., 2019) in data-rich environments.

absence of inducing the RM to accept the objective of the performance assessment, the RM's direction of the AM will simply reflect the least effortful course of thinking.

Shown in **Figure 4** are examples of preliminary questions to ask the respondent at the *initiation* of the performance assessment. These would be required to measure the meta-cognitive approach adopted by the test-taker in the specific data-rich environment in which the performance assessment is embedded. By incorporating preliminary questions into the design of the assessment such as *how do you define critical thinking* and *do you value it*, the assessment yields two sources of evidentiary data about the test-taker: First, what do they believe critically thinking entails? And second, are they motivated to demonstrate this type of thinking, namely, the construct of interest? Both these sources of data about the test-taker would help in the interpretation of their assessments results. If test-takers can define critical thinking but do not value it or are not willing to suspend associative serial processing, low scores may only reveal their lack of interest or motivation. The latter of which

becomes a key challenge for educational interventions unless the reasons for its benefits can be shown.

## MOVING BEYOND JUST TEACHING CRITICAL THINKING SKILLS

How well educators are poised to teach and assess critical thinking in data-rich environments might depend less on a specific instructional formula and more on how it is incentivized for students. In other words, there needs to be a clear message to students about what it is that they gain by suspending personal biases and engaging analytical strategies; for example, "Did you know that by becoming aware that you are reacting positively to the flashiest site of health information you may not be getting the best information? Or "Did you know that in searching for information about a political issue you will typically be drawn to information that confirms your prior beliefs? If you want to be fully prepared for debates, try



searching for information that challenges what you believe so you can be prepared for both sides of the argument.”

A shortcoming with almost all assessments of critical thinking as of the writing of this paper is that they are designed from traditional definitions of critical thinking; meaning that these assessments do not test for cognitive biases explicitly. For example, the Halpern Critical Thinking Assessments (Butler, 2012) measure five dimension of critical thinking premised on traditional conceptions of critical thinking (i.e., verbal reasoning, argument analysis, thinking as hypothesis testing, likelihood/uncertainty, and decision making and problem solving) but not cognitive biases. Another popular critical thinking test is the Cornell Critical Thinking Level Test Z (Ennis and Millman, 2005) which measures induction, deduction, credibility, identification of assumptions, semantics, definitions, and prediction in planning experiments. However, all these attributes are proactive and not reactive. Only measuring proactive attributes can almost be viewed, ironically, as yet another instance of our tendency to confirm biases. What is needed is actively falsifying what we believe—testing the limits of what we want to think is true. There are at least two notable exceptions to the typical critical thinking tests. One is the Cognitive Reflection Test (Frederick, 2005), which measures a person’s skill at reflecting on a question and resisting answering with the first response that comes to mind. In essence, this test measures reactive processes. The other is the Comprehensive Assessment of Rational Thinking (CART) by Stanovich (2016). The CART is focused on measuring the preponderance and avoidance of thinking errors or contaminated mindware. For example, the CART contains 20 subtests that assess tendencies toward overconfidence, showing inconsistent preferences and being swayed by irrelevant information. Critical thinking tests designed to measure avoidance of reactive processes are relatively new and perhaps not surprisingly there are no large-scale studies of whether it can be effectively taught. It is for this reason that the work we present here is necessary and we believe presents a contribution to the literature.

Proactive critical thinking can be taught so there is no reason to think that awareness of reactive critical thinking cannot also be taught. To be sure, most of the research on teaching critical thinking skills has been in the area of proactive skills. A meta-analysis of strategic approaches to teaching critical thinking uncovered that various forms of critical thinking can be taught with measurable positive effects (Abrami et al., 2015). However, the average effect size of educational interventions was 0.30 (Cohen’s *d*); thus, weak to moderate at best (Cohen, 1977; Abrami et al., 2015). Part of the challenge is that critical thinking, like any other disposition and/or skill, takes time to cultivate and uptake is determined by how well the audience (students) buys into what is being taught.

One would expect different approaches for teaching critical thinking depend not only on the specific goal of instruction but also how well students believe in the benefits articulated. For example, Lorencová et al. (2019) conducted a systematic review of 39 studies of critical thinking instruction in teacher education programs. The most often cited targeted skills for instruction were analysis and evaluation. A majority of the educational

interventions had the following characteristics: (a) took place during a course in one semester with an average number of 66 students, (b) were face-to-face, (c) used infusion (i.e., critical thinking added as a separate module to existing curriculum), or immersion (i.e., critical thinking integrated into the full curriculum) as the primary context for instruction with (d) discussion and self-learning as tools for pedagogy. The most frequently used standardized assessment tool for measuring learning gains was the CCTDI or California Critical Thinking Disposition Inventory, which is a measure of thinking dispositions instead of actual critical thinking performance.

In addition to the CCTDI, most instructors also developed their own assessments, including assessment of typical case studies, essays, and portfolios. Most of the 39 studies reviewed showed fully positive or some positive results; only 3 studies reported null results. Not surprisingly, however, larger effects between pre- and post-intervention were observed for studies employing *instructor-created, non-standardized* tools compared to standardized assessment tools.

One of the biggest challenges identified by Lorencová et al. (2019) is not with the interventions of critical thinking but with assessments to measure gains. Instructor-developed assessments suffer from a variety of problems such as demand characteristics, low reliability, and potentially biased grading. Thus, little can be concluded about what reliably works among the many strategies for critical thinking without good measures. A related problem is that many of these interventions do not indicate how long the effects last; good measures are also required to gauge the temporal effects of interventions. Additional problems that often plague intervention studies involve relatively small sample sizes. These challenges may be overcome in a variety of ways. For example, moving away from idiosyncratic instructor-developed critical thinking assessments and moving toward the establishment of a consortia of researchers that can pool their items for review, field-testing, refinement and ultimately leverage large enough samples to establish reliable norms for inferences. Toward this end, Shavelson et al. (2019) exemplify this work in their International Performance Assessment of Learning (iPAL) consortium.

In another recent review of critical thinking interventions in professional programs in the social sciences and STEM fields, Puig et al. (2019) noted the prevalence of unstandardized forms of assessments for measuring critical thinking, most of which were qualitative. For example, Puig et al. (2019, p. 867) indicate that most of the studies they reviewed based their results largely on “the opinions of students and/or teachers, as well as on other factors such as students’ motivation, or their level of engagement to the task... students’ perceptions, learning reflections and their participation in the task, and others even did not assess CT.” These measures may begin to probe the values and beliefs of the RM but they ignore the information processes of the AM in instantiating critical thinking.

Schmaltz et al. (2017) indicate that part of the reason educators at all levels of instruction, including postsecondary institutions, find it so challenging to teach critical thinking is that it is not well defined and there are not enough empirical studies to show what works. Although the deficits raised by Schmaltz et al.

(2017) are justified, the problem of showing what works requires measuring human behavior with minimal bias. Thus, the deficits identified by Schmaltz et al. (2017) may actually reside more with the assessments used to evaluate interventions than with the interventions themselves. Just as there many ways to teach algebra or essay composition successfully depending on the students involved, so must teaching critical thinking take on different methods as shown in the literature (e.g., Abrami et al., 2015; Lorencová et al., 2019). However, focusing so intently on the specific characteristics of educational interventions may hurt more than it helps if it distracts from the assessments that need to be designed to measure changes in thinking. In whatever form critical thinking is taught, what is certainly needed are assessments that reliably measure the construct of critical thinking, however it has been conceptualized and operationalized (Ennis, 2016).

## Teaching and Assessing Critical Thinking in Data-Rich Environments

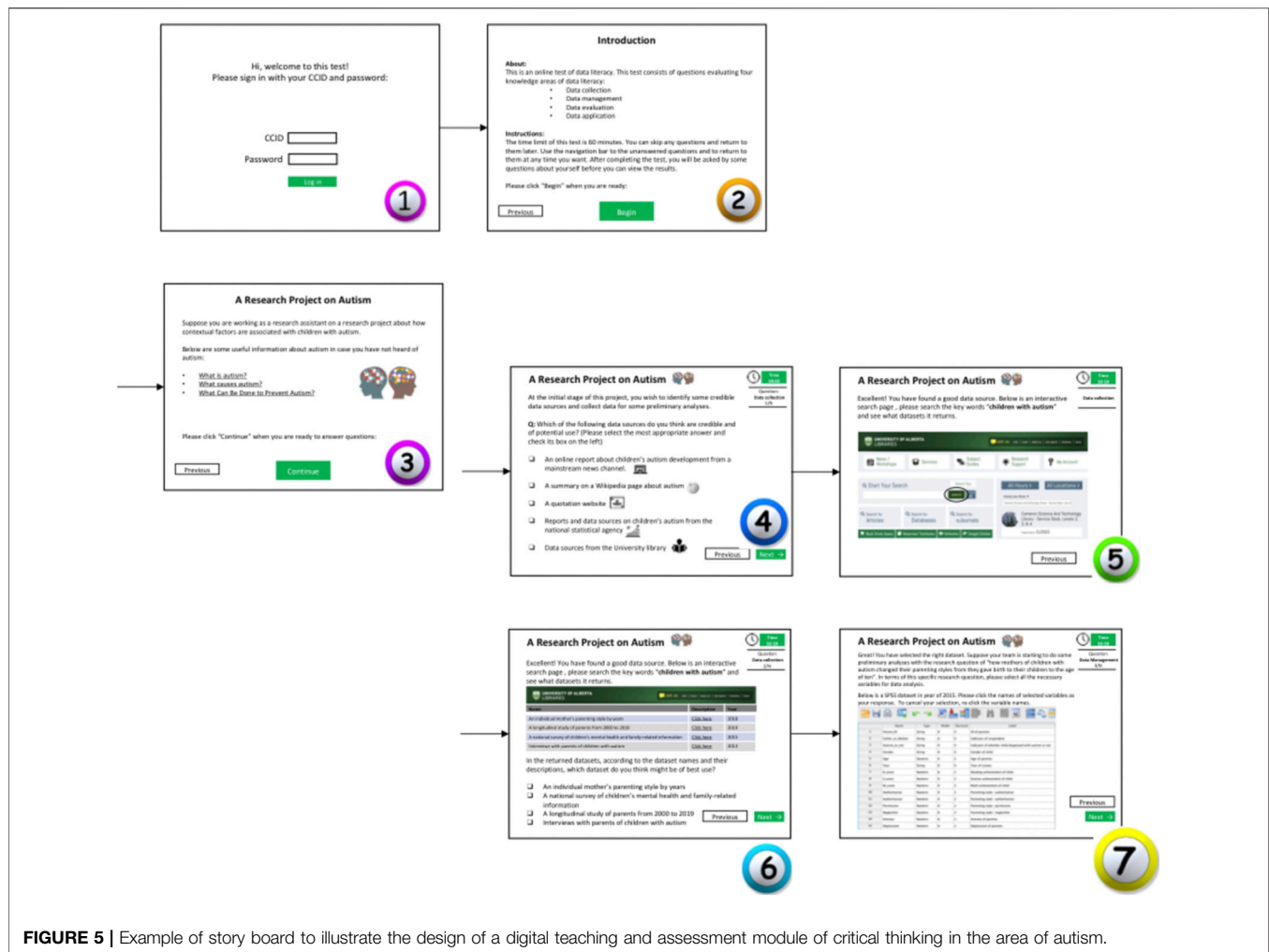
Teaching and assessing critical thinking in data-rich environments requires not only a conception of what critical thinking entails in such environments but also an adequate assessment of the information processes associated with this type of thinking. Building first on Stanovich and Stanovich's (2010) tripartite model, the instructional goals must include (a) students becoming *self-aware* of what types of thinking they value and in what circumstances and (b) students learning to apply strategies they believe are valuable in thinking critically in identified circumstances. The premise is this: If critical thinking is valued for a given topic, strategies such as decoupling and simulation can be explicitly taught, taken up by students, practiced and assessed using online information sources and tasks. This is also where Shavelson et al.'s (2019) framework provides an excellent assessment foundation to structure teaching and assessment modules. Prompts and performance tasks can be *embedded throughout* digital modules to assess students' goals for information processing, strategies for searching and analyzing data tables, reports, and graphs for the stated goals, time spent on different informational resources, and evaluation of conclusions.

Teaching and assessment modules for critical thinking must motivate students to expend the cognitive resources to suspend certain information processes (e.g., serial associative processing). As previous reviews have found (e.g., Lorencová et al., 2019), motivation is a pre-requisite to decouple and simulate as these are cognitively taxing forms of processing. In the pre-development stage of any teaching or assessment form, one of the most important tasks is to survey the population of students about interests warranting critical thinking. Then, digital teaching modules and assessments can be designed around topics that would motivate students to expend the resources needed to engage with tasks; for example, the effects of social media on mental health, the cost and value of postsecondary education or even a learning disability can be used as topics to spark the interest of students. Starting from a

position of awareness about the topics that warrant attention, students can be invited to learn about resisting serial associative processing in the collection of data (e.g., finding high-quality data that are relevant but opposed to what is believed about a topic), decoupling in the analysis of conclusions (e.g., looking at statistics that do not misrepresent the data), and simulation in evaluations of conclusions (e.g., weighing the evidence in line with its quality).

However, incentivizing students to pay attention to what they are processing does not mean it will be processed critically. Especially when topics are of interest, individuals are likely to hold strong opinions and seek to actively confirm what they already believe. Thus, teaching modules must begin with a process of having students become aware of a bias to confirm, and invoking reminders to students that this bias can surface unless it is constantly under check in their self-awareness. For example, a prompt for students to become aware of their biases can be integrated into the introductory sections of a teaching and assessment module. Prompts can also be designed to remind them of the critical thinking they have indicated they value. Previously presented information processes (e.g., suspending prior beliefs or decoupling) can be flashed as reminders in searching, assessing task information, and evaluating conclusions. Another option might be to have students choose to assume the *perspective of a professional* such as a journalist, a lawyer, or a counselor and to challenge them to process information as that professional would be expected to do.

Consider the following screenshot in **Figure 5** from a storyboard associated with the design of a teaching and assessment module on *autism*. Following an introductory screen, participating students are advised in the second screen that they are going to be learning about ways to collect, manage, evaluate, and apply data on autism. The third screen introduces them to the research project and poses initial questions they are unlikely to be able to answer critically. The fourth screen introduces them to potential data sources, such as an online report from a mainstream news channel and a report from a national statistics agency. At this point, students can be prompted to rate the trustworthiness of the sources, which reflects the first facet of Shavelson et al.'s framework. In the fifth screen, students navigate to the data source(s) selected. Irrespective of the data source selected, students are probed on the manipulability and relevance of the data source, and how it advances the investigation. At each point during the module, students are scaffolded in evidence-based learning about autism and asked to provide responses designed to reveal their chosen information processing. For example, in the fourth screen where students are asked to list the data sources for autism, the sources students indicate can be categorized according to at least two dimensions. First, is each source trustworthy? Relevant? Second, how much time and effort did students spend analyzing the sources (using reaction time data). If students appear to carefully choose what they think are trustworthy and relevant sources but do not ascribe the trustworthiness or relevance to



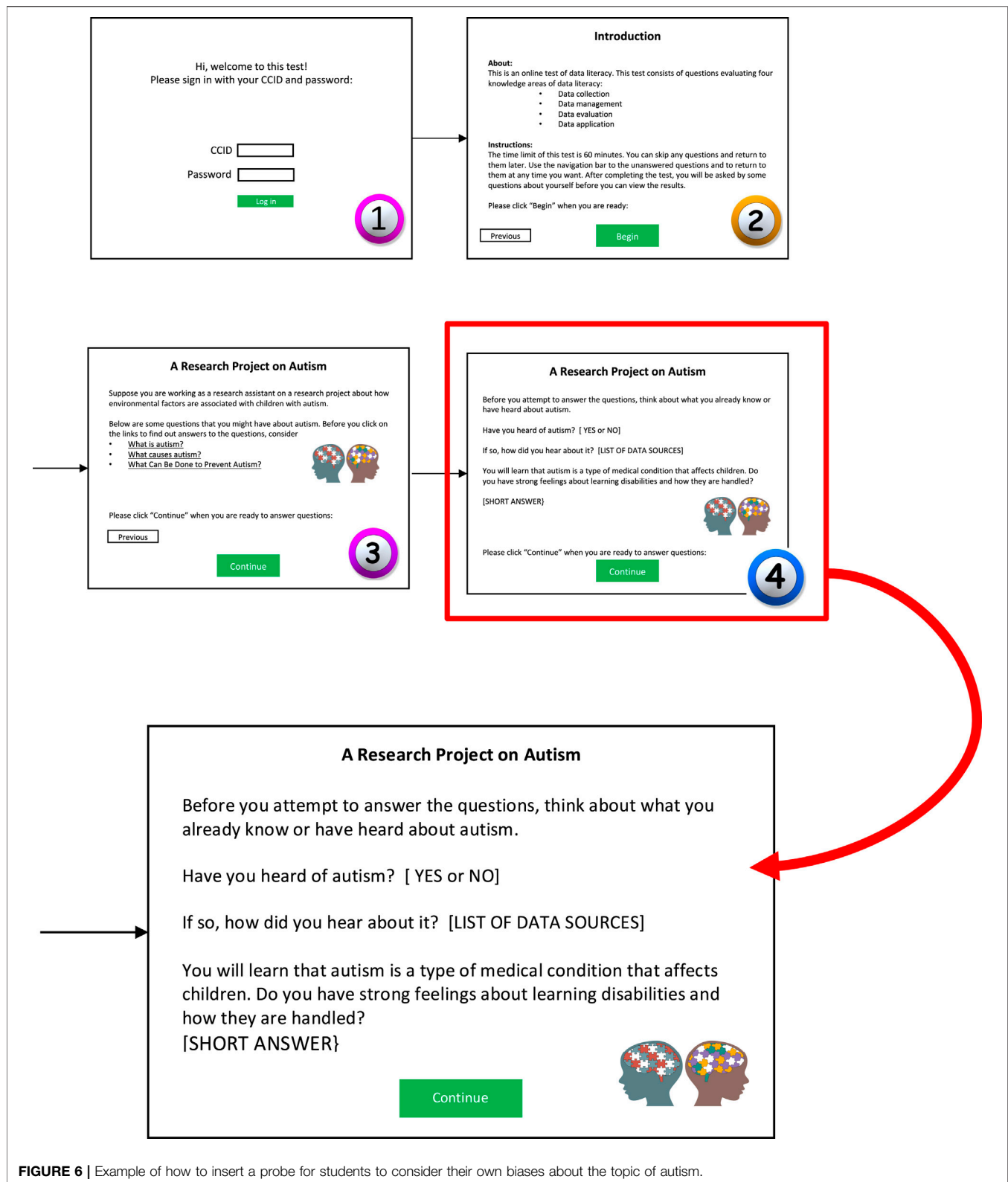
**FIGURE 5 |** Example of storyboard to illustrate the design of a digital teaching and assessment module of critical thinking in the area of autism.

substantive criteria, then students may value critical thinking (RM) but do not have the knowledge or skills to properly direct this value (RM) in their information processing. In this case, the scaffolding comes in the form of an instructional part of the module that explains the criteria that should be used for judging reliability, and relevancy in the case of neurodevelopmental disorders such as Autism.

The storyboard shown in **Figure 5** does not show how students' potential bias may be assessed at the beginning of the module. However, opportunities to bring bias into students' awareness can be inserted as is shown in the fourth screen in **Figure 6**. Following this fourth screen, another screen (not shown) could be inserted to teach students what it means to decouple and simulate in the process of information processing. For example, the instructional module can show why an uncontrolled variable (e.g., a diet supplement) should be decoupled from another variable that was controlled (e.g., age of the mother). In this way, students are reminded of their biases (e.g., diet supplements are bad for you), instructed on what it means to think critically in an information-rich environment and also prompted to decide whether such strategies should be applied in considering data during the assessment.

## DISCUSSION AND CONCLUSION

New ways of teaching and assessing critical thinking in data-rich environments are needed, given the explosion of online information. This means employing definitions of critical thinking that explicitly outline the contaminated mindware that should be avoided in data-rich environments. The democratization of information in the digital age means that anyone, regardless of qualifications or motivation, can share stories, ideas, and facts with anyone who is willing to read, watch, and be convinced. Although misinformation has always existed, never before has it been as ubiquitous as it is today and cloaked in the pretense of trustworthiness as found on the worldwide web. Errors in reasoning and bias in information processing are, therefore, central to the study of critical thinking (Leighton and Sternberg, 2004). Consequently, three lines of thinking were presented for why a refined conception of critical thinking in data-rich environments is warranted. First, traditional definitions of critical thinking typically lack connections to the information processes that are required to overcome bias. Second, data-rich environments pose cognitive traps in critical thinking that require more attention to bias. Third, personal dispositions such as



**FIGURE 6 |** Example of how to insert a probe for students to consider their own biases about the topic of autism.

motivation are more important than previously thought in the teaching and measurement of critical information processes. Because the present paper is not empirical but rather

conceptual, we end not with main findings but with essential take home ideas. The first essential idea is that explicitly articulating a refined conception of critical thinking, one that



includes reactive processes and/or mindware, must become part of how good thinking is described and taught. The second essential idea is that teaching and assessing proactive and reactive processes of critical thinking must be empirically examined.

The contemporary teaching and assessment of critical thinking must be situated within environments that are rich in data and evoke more than proactive but mechanistic information processes of analysis and evaluation. Teaching and assessment of critical thinking in data-rich environments must become more sophisticated to consider students' 1) interest in the topics that merit critical thinking, 2) self-awareness of human bias, and 3) how both interest and self-awareness are used by students' reflective minds (RM) to guide strategic application of critical-thinking processes in the AM. A conceptual refinement of critical thinking in data-rich environments, then, must be based on a strong theoretical foundation that presents a coordination of the reflective, algorithmic, and autonomous minds (Stanovich and Stanovich, 2010). This is provided by Stanovich and Stanovich's (2010) tripartite model and supported by decades of empirical research into human thinking processes (Leighton and Sternberg, 2004; Kahneman, 2011; Stanovich, 2012; Shavelson et al., 2019).

A theoretical foundation for operationalizing a new conception of critical thinking, however, is useless for practice unless there is a framework that permits the principled design of teaching and assessment modules. Shavelson et al. (2019) provides such a framework. Shavelson et al. (2019) assessment framework provide the structure for generating performance-based tasks that evoke the reflective and algorithmic information processes required of critical thinking in data-rich environments.

The mindware that students download in performance assessments of critical thinking must reflect the

sophistication of this form of information processing. Most students do not acquire these skills in secondary school or even post-secondary education (Stanovich, 2012; Ridsdale et al., 2015; Shavelson et al., 2019). Stanovich (2012, p. 356) states: "Explicit teaching of this mindware is not uniform in the school curriculum at any level. That such principles are taught very inconsistently means that some intelligent people may fail to learn these important aspects of critical thinking." He indicates that although cognitive biases are often learned implicitly, without conscious awareness, critical-thinking skills must be taught explicitly to help individuals come to know *when* and *how* to apply higher-level skills. Instruction in critical thinking thus requires domain-specific knowledge and transferable skills that allow individuals to 1) coordinate the RM and AM, 2) recognize bias, and 3) regulate the application of higher-level thinking strategies. A more sophisticated conception of critical thinking provides an opportunity to guide instructive and performance-based assessment programs in the digital age.

## AUTHOR CONTRIBUTIONS

The authors confirm being the sole contributors of this work and have approved it for publication.

## ACKNOWLEDGEMENTS

Preparation of this paper was supported by a grant to the first author from the Social Sciences and Humanities Research Council of Canada (SSHRC Grant No. 435-2016-0114).

## REFERENCES

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., and Persson, T. (2015). Strategies for teaching students to think critically: a meta-analysis. *Rev. Educ. Res.* 85 (2), 275–314. doi:10.3102/0034654308326084
- Alava, S., Frau-Meigs, D., and Hassan, G. (2017). *Youth and violent extremism on social media. Mapping the research*. Paris, France: UNESCO.
- Behar-Horenstein, L. S., and Niu, L. (2011). Teaching critical thinking skills in higher education: a review of the literature. *J. Coll. Teach. Learn.* 8 (2), 25–42. doi:10.19030/tlc.v8i2.3554
- Bloom, B. S. (1956). *Taxonomy of educational objectives. Vol. 1: Cognitive domain*. New York, NY: McKay, 20–24.
- Boyd, D. (2014). *It's complicated: the social lives of networked teens*. London: Yale University Press.
- Butler, H. A. (2012). Halpern Critical Thinking Assessment predicts real-world outcomes of critical thinking. *Appl. Cognit. Psychol.* 26 (5), 721–729. doi:10.1002/acp.2851
- California State University (2011). Executive order 1065. Available at: <https://calstate.policystat.com/policy/6741976/latest/> (Accessed April 25, 2020).
- Churchland, P. S. (2011). *Braintrust: what neuroscience tells us about morality*. Princeton, NJ: Princeton University Press.
- Clemson University (2016). Clemson Think2. Available at: <https://www.clemson.edu/academics/programs/thinks2/documents/QEP-report.pdf> (Accessed April 25, 2020).
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Revised edn. New York: Lawrence Erlbaum Associates, 490.
- Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24 (1), 87–85. doi:10.1017/s0140525x01003922
- Dewey, J. (1933). *How We Think* (1st ed.). Boston: D. C. Heath.
- Dawes, R. M. (1976). Shallow psychology. In *Cognition and social behavior*. Editors J. S. Carroll and J. W. Payne (Hillsdale, NJ: Erlbaum), pp. 3–11.
- Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
- El Soufi, N., and Huat See, B. (2019). Does explicit teaching of critical thinking improve critical thinking skills of English language learners in higher education? A critical review of causal evidence. *Stud. Educ. Eval.* 60, 140–162. doi:10.1016/j.stueduc.2018.12.006
- Ennis, R. H. (1996). *Critical thinking*. Upper Saddle River: Prentice-Hall.
- Ennis, R. H. (2015). "Critical thinking: a streamlined conception," in *The Palgrave handbook of critical thinking in higher education*. Editors M. Davies and R. Barnett (New York: Palgrave), 31–47.
- Ennis, R. H. (2016). "Definition: a three-dimensional analysis with bearing on key concepts," in Ontario society for the study of argumentation (OSSA) conference archive, University of Windsor, May 18–21, 2016. 105, 2016. Available at: <https://scholar.uwindsor.ca/ossaarchive/OSSA11/papersandcommentaries/105> (Accessed April 26, 2016).
- Ennis, R. H., and Millman, J. (2005). *Cornell critical thinking test, level X*. 5th Edn. Seaside, CA: The Critical Thinking Company.
- Evans, J. St. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends Cognit. Sci.* 7 (10), 454–459. doi:10.1016/j.tics.2003.08.012
- Facione, P. (1990). Consensus statement regarding critical thinking and the ideal critical thinker. Millbrae, CA: California Academic Press.
- Fisher, A., and Scriven, M. (1997). *Critical thinking: its definition and assessment*. Point Reyes, CA: Edgepress.

- Frederick, S. (2005). Cognitive reflection and decision making. *J. Econ. Perspect.* 19 (4), 25–42. doi:10.1257/089533005775196732
- Glaser, E. (1941). *An experiment in the development of critical thinking*. New York: Teacher's College, Columbia University.
- Halpern, D. F. (2014). *Thought and knowledge* (5th ed.). New York, NY: Psychology Press.
- Hyytinen, H., Nissinen, K., Ursin, J., Toom, A., and Lindblom-Ylänne, S. (2015). Problematising the equivalence of the test results of performance-based critical thinking tests for undergraduate students. *Stud. Educ. Evaluation* 44, 1–8. doi:10.1016/j.stueduc.2014.11.001
- Hyytinen, H., Toom, A., and Shavelson, R. J. (2019). “Enhancing scientific thinking through the development of critical thinking in higher education,” in *Redefining scientific thinking for higher education: higher-order thinking, evidence-based reasoning and research skills*. Editors M. Murtonen and K. Ballou (London: Palgrave Macmillan), 59–78.
- Johnson-Laird, P. N. (2004). “Mental models and reasoning,” in *The nature of reasoning*. Editors J. P. Leighton and R. J. Sternberg (New York, NY: Cambridge University Press), 169–204.
- Johnson-Laird, P. N., and Bara, B. G. (1984). Syllogistic inference. *Cognition* 16 (1), 1–61. doi:10.1016/0010-0277(84)90035-0
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Strauss, Giroux.
- Kroll, E. B., Rieger, J., and Vogt, B. (2010). “How does repetition of signals increase precision of numerical judgment?” in *Brain informatics. BI 2010. Lecture notes in computer science*. Editors Y. Yao, R. Sun, T. Poggio, J. Liu, N. Zhong, and J. Huang (Berlin, Heidelberg: Springer), Vol. 6334. doi:10.1007/978-3-642-15314-3\_19
- Leighton, J. P. (2011). “A cognitive model of higher order thinking skills: implications for assessment,” in *Current perspectives on cognition, learning, and instruction: assessment of higher order thinking skills*. Editors G. Schraw and D. H. Robinson (Charlotte, NC: Information Age Publishing), 151–181.
- Leighton, J. P., Chu, M.-W., and Seitz, P. (2013). “Cognitive Diagnostic assessment and the learning errors and formative feedback (LEAFF) model,” in *Informing the practice of teaching using formative and interim assessment: a systems approach*. Editor R. Lissitz (Charlotte, NC: Information Age Publishing), 183–207.
- Leighton, J. P., and Dawson, M. R. W. (2001). A parallel processing model of Wason's card selection task. *Cognit. Syst. Res.* 2 (3), 207–231. doi:10.1016/S1389-0417(01)00035-3
- Leighton, J. P., and Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educ. Meas. Issues Pract.* 26 (2), 3–16. doi:10.1111/j.1745-3992.2007.00090.x
- Leighton, J. P., and Sternberg, R. J. (2004). *The nature of reasoning* (Cambridge, MA: Cambridge University Press).
- Leighton, J. P., and Sternberg, R. J. (2012). “Reasoning and problem solving,” in *Handbook of psychology. Experimental psychology*. Editors A. Healy and R. Proctor 2nd Edn (New York: Wiley), Vol. 4, 631–659.
- Lorencová, H., Jarošová, E., Avgitidou, S., and Dimitriadou, C. (2019). Critical thinking practices in teacher education programmes: a systematic review. *Stud. Higher Educ.* 44 (5), 844–859. doi:10.1080/03075079.2019.1586331
- Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63 (2), 81–97. doi:10.1037/h0043158
- Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Res. Rep. Ser.* 2003 (1), i–29. doi:10.1002/j.2333-8504.2003.tb01908.x
- Mullet, H. G., and Marsh, E. J. (2016). Correcting false memories: errors must be noticed and replaced. *Mem. Cognit.* 44 (3), 403–412. doi:10.3758/s13421-015-0571-x
- Perkins, D. N. (1995). *Outsmarting IQ: the emerging science of learnable intelligence*. New York, NY: Free Press.
- Prensky, M. (2001). Digital natives, digital immigrants. *Horizon* 9 (5), 1–6. doi:10.1108/10748120110424816
- Puig, B., Blanco-Anaya, P., Bargiela, I. M., and Crujeiras-Pérez, B. (2019). A systematic review on critical thinking intervention studies in higher education across professional fields. *Stud. High Educ.* 44 (5), 860–869. doi:10.1080/03075079.2019.1586333
- Reber, P. (2010). What is the memory capacity of the human brain? *Sci. Am. Mind*, 21(2), 70. doi:10.1038/scientificamericanmind0510-70
- Ridsdale, C., Rothwell, J., Smit, M., Hossam, A.-H., Bliemel, M., Irvine, D., et al. (2015). *Strategies and best practices for data literacy education: Knowledge synthesis report*. Halifax, NS: Dalhousie University. Available at: <https://dalspace.library.dal.ca/handle/10222/64578> (Accessed July 1, 2018).
- Rizeq, J., Flora, D. B., and Toplak, M. E. (2020). An examination of the underlying dimensional structure of three domains of contaminated mindware: paranormal beliefs, conspiracy beliefs, and anti-science attitudes. *Thinking Reasoning* 1–25. doi:10.1080/13546783.2020.1759688
- Roser, M., Ritchie, H., and Ortiz-Ospina, E. (2020). Internet. Published online at OurWorldInData.org. Available at: <https://ourworldindata.org/internet> (Accessed May 1, 2020).
- Schacter, D. L. (2012). Adaptive constructive processes and the future of memory. *Am. Psychol.* 67 (8), 603–613. doi:10.1037/a0029869
- Schmaltz, R. M., Jansen, E., and Wenckowski, N. (2017). Redefining critical thinking: teaching students to think like scientists. *Front. Psychol.* 8, 459–464. doi:10.3389/fpsyg.2017.00459
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. P. (2019). Assessment of university students' critical thinking: next generation performance assessment. *Int. J. Test.* 19 (4), 337–362. doi:10.1080/15305058.2018.1543309
- Stanovich, K. E. (2009). *What intelligence tests miss: the psychology of rational thought*. New Haven, CT: Yale University Press.
- Stanovich, K. E. (2012). “On the distinction between rationality and intelligence: implications for understanding individual differences in reasoning,” in *The Oxford handbook of thinking and reasoning*. Editors K. Holyoak and R. Morrison (New York: Oxford University Press), 343–365.
- Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.
- Stanovich, K. E. (2016). The comprehensive assessment of rational thinking. *Educ. Psychol.* 51 (1), 23–34. doi:10.1080/00461520.2015.1125787
- Stanovich, K. E. (2021). “Why humans are cognitive misers and what it means for the great rationality debate,” in *Routledge handbook of bounded rationality*. 1st Edn, Editor R. Viale (London: Routledge), 11.
- Stanovich, K. E., and Stanovich, P. J. (2010). “A framework for critical thinking, rational thinking, and intelligence,” in *Innovations in educational psychology: perspectives on learning, teaching, and human development*. Editors D. D. Preiss and R. J. Sternberg (New York: Springer Publishing Company), 195–237.
- Stanovich, K. E., Toplak, M. E., and West, R. F. (2008). The development of rational thought: a taxonomy of heuristics and biases. *Adv. Child. Dev. Behav.* 36, 251–285. doi:10.1016/S0065-2407(08)00006-2
- Taylor, S. E. (1981). The interface of cognitive and social psychology. In *Cognition, social behavior, and the environment*. Editor J. H. Harvey (Hillsdale, NJ: Erlbaum), pp. 189–211.
- Toplak, M. E., and Flora, D. B. (2020). Resistance to cognitive biases: longitudinal trajectories and associations with cognitive abilities and academic achievement across development. *J. Behav. Decis. Making*, 1–15. doi:10.1002/bdm.2214
- Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185 (4157), 1124–1131. doi:10.1126/science.185.4157.1124
- Wachowski, L., Wachowski, L., Silver, J., Reeves, K., Fishburne III, L., Moss, C. A., et al. (1999). *The matrix*. Burbank, CA: Warner Home Video.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Leighton, Cui and Cutumisu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# The Semiotics of Test Design: Conceptual Framework on Optimal Item Features in Educational Assessment Across Cultural Groups, Countries, and Languages

Guillermo Solano-Flores\*

Graduate School of Education, Stanford University, Stanford, CA, United States

## OPEN ACCESS

### Edited by:

Olga Zlatkin-Troitschanskaia,  
Johannes Gutenberg University  
Mainz, Germany

### Reviewed by:

Sara Magdalena Lenninger,  
Kristianstad University, Sweden  
Alin Olteanu,  
RWTH Aachen University, Germany

### \*Correspondence:

Guillermo Solano-Flores  
gsolanof@stanford.edu

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Education

**Received:** 04 December 2020

**Accepted:** 10 March 2021

**Published:** 14 April 2021

### Citation:

Solano-Flores G (2021) The  
Semiotics of Test Design: Conceptual  
Framework on Optimal Item Features  
in Educational Assessment Across  
Cultural Groups, Countries,  
and Languages.  
Front. Educ. 6:637993.  
doi: 10.3389/feduc.2021.637993

This paper offers a conceptual framework on test design from the perspective of social semiotics. Items are defined as arrangements of features intended to represent information, convey meaning, and capture information on the examinees' knowledge or skills on a given content. The conceptual framework offers a typology of semiotic resources used to create items and discusses item representational complexity—the multiple ways in which the semiotic resources of an item are related to each other—and item semiotic alignment—the extent to which examinees share cultural experience encoded by items. Since the ability to make sense of items is shaped by the examinees' level of familiarity with the social conventions underlying the ways in which information is represented, unnecessary representational complexity and limited semiotic alignment may increase extraneous item cognitive load and adversely impact the performance of examinees from certain populations. Semiotic test design allows specification of optimal pools of semiotic resources to be used in creating items with the intent to minimize representational complexity and maximize semiotic alignment for the maximum number of individuals in diverse populations of examinees. These pools of semiotic resources need to be specific to the content assessed, the characteristics of the populations of examinees, the languages involved, etc., and determined based on information produced by cross-cultural frequency analyses, cognitive interviews, focus groups, and expert panels.

**Keywords:** test design, semiotics, item features, semiotic resources, cultural groups

## INTRODUCTION

Current views of assessment as evidentiary reasoning emphasize the importance of systematic approaches for determining the numbers, formats, and features of items or tasks that are to be used in assessing a given domain of knowledge (Martinez, 1999; Pellegrino et al., 2001; National Research Council, 2006; Mislevy and Haertel, 2007). In large-scale assessment, these views support the process of test development (National Research Council, 2014) and the development of item

specifications documents that prescribe the general characteristics of items to be included in a given test (e.g., Council of Chief State School Officers, 2015).

Unfortunately, given their scope and level of analysis, such documents cannot pay detailed attention to the multiple textual and non-textual features of items. At present, no methodology is available that allows systematic selection, development, and use of the hundreds of features used in items, such as graphs, lines, arrows, labels, font styles, speech balloons, abbreviations, graph axes, ways of asking questions, ways of arranging options in multiple-choice items, buttons to click, cascade menus, and boxes to type or write answers to questions. While these features may or may not be directly related to the target knowledge domain, all of them contribute to representing information and may influence examinees' understanding of items. Many of these features may be used inconsistently across items within the same assessment program and, to a large extent, their use may be shaped more by idiosyncratic factors or tradition than by principled practice.

Concerns about this lack of a principled practice are even more serious for assessment programs that test culturally and linguistically diverse populations. For example, efforts oriented to minimizing cultural bias and ensuring the comparability of measures of tests across cultural and linguistic groups focus almost exclusively on the text of tests (e.g., Hambleton, 2005; Downing and Haladyna, 2006; International Test Commission, 2017). Little is known about whether and how the non-textual features of items should be adapted for students from different countries or cultural backgrounds. Yet we know that individuals from different cultural backgrounds may differ on the level of attention they pay to focal objects or contextual and background information (Nisbett, 2003; Chua et al., 2005); that the relative frequency of some features of item illustrations vary substantially across different assessment programs (Wang, 2012); and that the extent to which item illustrations influence student performance on science items in international comparisons varies across high- and low-ranking countries (Solano-Flores and Wang, 2015). Among many other, these findings speak to the need for a perspective of test design that allows systematic, detailed selection, and examination of the features of items.

This paper offers a conceptual framework on semiotic design focused on the testing of diverse populations across cultural groups, countries, and languages. It contributes to closing an important gap in the intersection of testing and semiotics: while education has captured the attention of semioticians for decades (e.g., Lemke, 1990; Stables, 2016; Pesce, 2018), the focus has been mainly on learning, text, and the classroom; little attention has been paid to tests and testing. The goal is not to offer a semiotic theory of testing, but rather a reasoning on the ways in which key concepts from the field of semiotics can be used to systematically analyze and design the features of test items in ways intended to minimize error variance and promote fair test development practices.

The first section provides some basic concepts from the field of social semiotics—the study of the ways in which information is represented and meaning is made according to implicit and

explicit social conventions (van Leeuwen, 2004). A perspective on semiotic resources as socially made tools for conveying meaning (Kress, 2010) provides the conceptual foundation for reasoning about meaning making as cultural practice and the ways in which the features of items can be selected or created systematically. The second section offers a classification of semiotic resources used in tests and discusses their use in the testing of culturally and linguistically diverse populations. The third section offers some ideas for semiotic test design based on the notion of representational complexity—the multiple ways in which the semiotic resources of an item are related to each other—and semiotic alignment—the intersection of the cultural experience encoded by semiotic resources and the examinees' cultural experience.

## SEMIOTICS, TESTS, AND DIVERSE POPULATIONS

### Features, Semiotic Resources, and Multimodality

At the core of this conceptual framework is the concept of semiotic resource. van Leeuwen (2004) defines semiotic resources as

“the actions, materials and artifacts we use for communicative purposes, whether produced physiologically—for example, with our vocal apparatus, the muscles we use to make facial expressions and gestures—or technologically—for example, with pen and ink, or computer hardware and software—together with the ways in which these resources can be organized.”

“Semiotic resources have a meaning potential, based on their past uses, and a set of affordances based on their possible uses, and these will be actualized in concrete social contexts where their use is subject to some form of semiotic regime” (van Leeuwen, 2004, p. 285)."

This definition allows appreciation of the vastness of actions, materials, and artifacts that have the potential to communicate meaning. For example, in certain cultural contexts, the letter *A* can be a letter used in combination with other letters to create words, an option in a multiple-choice item, a grammatical article, a marker of the beginning of a sentence, a referent of hierarchy or priority, a letter denoting a variable, etc.

The definition also allows appreciation of the critical role that history plays in encoding meaning. Semiotic resources have been characterized as means for meaning making. But because they encode cultural experience, their affordances are not constant across social and cultural contexts (Kress, 2010). The ability of individuals to make meaning of semiotic resources depends on the extent to which they share that encoded cultural experience.

According to this reasoning, a test item can be viewed as an arrangement of multiple semiotic resources used in combination with the intent to represent information, convey meaning, and capture information on the examinee's knowledge or skills on a given knowledge domain. Proper interpretation of items



greatly depends on the individual's familiarity with the social conventions underlying the features of items and, therefore, their ability to make meaning of them. Those social conventions may be explicit or implicit, formally taught at school or acquired through informal experience, relevant or irrelevant to the content assessed, or specific or external to tests and testing.

Given their interrelatedness, no semiotic resource can be assumed to be intrinsically trivial. For example, a decimal point and a decimal comma are not intended to play a critical role in assessing computation skills respectively in the items  $3.1416 \times r^2 = \underline{\hspace{2cm}}$  and  $3,1416 \times r^2 = \underline{\hspace{2cm}}$ , which are intended to assess exactly the same kind of skill. Yet, since the use of decimal separators varies across countries (Baecker, 2010), in an international test comparison, not using the proper decimal separator in each country could constitute a source of measurement error.

The terms *item semiotic resource* and *item feature* are used as interchangeable in this paper. However, the former is used to emphasize purposeful design (e.g., *a team of test developers identifies the set of semiotic resources to be used in an international test*). In contrast, the latter is used more generically to refer to the characteristics of items, regardless of whether they are a result of a systematic process of design (e.g., *a researcher develops a system for coding the features identified in existing items from different countries*).

For the purposes of this conceptual framework, the term, *semiotic modes* is used to refer to broad categories of ways of representing information integrally (e.g., *textual and visual modes*) and the term, *multimodality* is used to refer to the use of semiotic resources belonging to different modalities (Kress and van Leeuwen, 2006). It is important to bear in mind that semiotic modalities should not be understood as clearly, fixed, and stable categories, but rather as interacting categories with fuzzy boundaries. For example, text contains visual features such as margins, font sizes, bold letters, etc., which contribute to conveying meaning. Also, a map has limited value as a visual device in the absence of labels and legends.

## Culture, Cultural Groups, and Cultural Experience

Broadly, *culture*, as a phenomenon, is understood here as the set of practices, views, values, attitudes, communication and socialization styles, ways of knowing, and ways of doing things among the members of a community, and which are the result of shared experience and history and learned through either formal and informal experiences or acquired through multiple forms of social participation and interaction with other individuals. The definition of culture as “the *non-hereditary memory of the community*, a memory expressing itself in a system of constraints and prescriptions,” (Lotman et al., 1978, p. 213, italics in the original) provides a perspective that is sensitive to the process of testing as a communication process (Solano-Flores, 2008). This definition is also consistent with the view, that, since it is the medium in which humans live and develop, culture “should be defined in terms of the artifacts that mediate human activity” (Packer and Cole, 2020, p. 11).

The term *cultural experience* or *cultural background* is used to refer to the set of experiences that an individual has from their contact with a given cultural context or with several cultural contexts. This set of experiences is assumed to be unique to each examinee, although multiple individuals can be regarded as a cultural group when they share many cultural experiences.

## Items as Samples of Encoded Cultural Experience

Current thinking in the field of educational measurement views the items of a test as samples of observations from a knowledge domain (Kane, 1982). According to this view, writing an item is equivalent to drawing a sample from that knowledge domain. Items are drawn (generated) systematically according to dimensions such as topic, type of knowledge, and disciplinary practice, etc. (Lane et al., 2016).

Unfortunately, item features do not receive the same level of attention in test development as these dimensions do. For example, while item specifications documents of assessment programs may provide detailed prescriptions regarding the alignment of the items to a set of standards, scant consideration is given to features beyond item format (e.g., multiple-choice or constructed-response) or text length. Such neglect dismisses the multimodal nature of disciplinary knowledge—the fact that disciplines develop elaborate ways of representing information in multiple textual and non-textual forms used in combination (see Lemke, 1998).

A wealth of evidence speaks to the influence of different item features on the examinees' performance on tests. For example, we know that the performance of students is instable across item formats (Ruiz-Primo et al., 1993); that construct equivalence may vary depending on the ways in which items are designed (Rodriguez, 2003); and that even small changes in wording may cause translated items to function differentially (Ercikan et al., 2014).

Semiotic resources effectively convey meaning to the extent that they encode cultural experience shared by the examinees. Items indeed can be viewed not only samples of a knowledge domain, but also as samples of encoded cultural experience. These samples may be biased if they predominantly reflect the cultural experience of specific segments of a society or the specific population of students for which tests are originally developed.

The amount of effort needed to minimize such bias should not be taken lightly, as the following example illustrates:

An assessment program intends to create a list of names of fictitious characters to be used in the contexts of its mathematics word problem items (e.g., *Joe and Clara need to cut a pizza into seven slices of the same size. What measure should they use to make sure that the slices have the same size?*). The intent is to have a restricted list of names that are recognizable by students with different cultural backgrounds. Using only the names included in that list should contribute to minimizing reading demands and creating equally meaningful contexts for students with different cultural backgrounds. While assembling a list of names is a simple project in principle, to serve its

intended purpose, the list should meet multiple criteria. For example: (1) female and male names should be equally represented; (2) all names should be easy to spell and read; (3) no name should have an unintended meaning in a different language; (4) all names should be familiar to many cultural groups; (5) no name should be associated to cultural stereotypes; (6) no name should be longer than ten characters; etc. Given this level of specificity, serious systematic work needs to be done to assemble a list of names that fit these rules. This work should include, among other things performing searches and asking individuals from the target populations of examinees about the suitability of the names.

Thus, even seemingly simple item features may need to be carefully designed if cultural bias is to be effectively minimized. Unfortunately, the impact on student performance of item features is yet to be investigated with this level of detail and, with some exceptions (e.g., Solano-Flores et al., 2014a), assessment programs have not paid attention to their systematic design.

## TYPES OF ITEM SEMIOTIC RESOURCES

This conceptual framework classifies item semiotic resources into six types, summarized in **Table 1**. The classification is not necessarily exhaustive. Also, the six categories and types of semiotic resources discussed should not be regarded as mutually exclusive. For the sake of simplicity, the examples provided can be viewed as basic semiotic resources—those that, in the context of design, act as building blocks of more complex semiotic resources.

Consistent with the notion that disciplinary knowledge is represented, communicated, and interpreted using multiple semiotic modes (Lemke, 1998), the categories discussed should be considered as being interconnected.

### Language Resources

For the purposes of this paper, language is understood as a *system* of socially established conventions for conveying meaning orally,

in signed language, or in written/printed form (Halliday, 1978) and language resources are defined as specific aspects of language used as semiotic resources in items. The category of language resources is vast, as it comprises resources as small and simple as a punctuation sign or a letter and as vast and complex as the language or the multiple language modes (oral, aural, textual) in which a test administered.

Because language is the vehicle through which testing takes place, examinees' limited proficiency in the language in which tests are administered or limited familiarity with the ways in which language is used constitutes a major threat to the validity of interpretations of test scores (American Educational Research Association [AERA] et al., 2014; Sireci and Faulkner-Bond, 2015). Even minimal aspects of language use may constitute important influences that shape examinees' interpretations of test items. For example, there is evidence that subtle variations on the ways in which items are worded can make a difference in the ways in which students interpret items (Ercikan, 2002). Also, the misalignment between the textual features of items in an international test and the textual features of items in national examinations (Anagnostopoulou et al., 2013) has been documented. Potentially, such misalignment could unfairly increase the difficulty of items in international tests.

Examples at three levels of complexity illustrate the wide range of language resources and their design implications. At a very basic level, text size illustrates how features of printed language may appear deceptively trivial. Because languages differ on word length and grammatical complexity (Coupé et al., 2019), the text size of items may vary considerably across different language versions of the same test. If text size ratios are not considered at a planning stage in the development of a test, the display of the items may look crowded for some of its language versions.

At another level of complexity, the ways in which vocabulary is addressed in testing illustrates the gap between what is known about language and how that knowledge is incorporated in testing practices. While there are sources that document the frequency of words in English (e.g., Nagy and Anderson, 1984; Davies and Gardner, 2010; Nation, 2014), that information is not used routinely to decide the wording and minimize the lexical complexity in items not intended to assess vocabulary knowledge.

At a higher level of complexity, issues in test translation illustrate the challenges of testing diverse populations in different languages, mainly because translation may alter the nature of the constructs assessed by items (Hambleton, 2005; Winter et al., 2006; Arffman, 2013). A great deal of the effort and time invested in the process of assessment development concerns refining the wording of items to ensure that examinees understand them as their developers intend (Abedi, 2006, 2016). Yet, compared to the time allocated for test development, assessment programs allocate considerably less time for test translation and adaptation (Solano-Flores, 2012). Tight timelines seriously limit the opportunities for examining students' interpretations of translated items (e.g., through verbal protocols and cognitive interviews) and conducting differential item functioning analyses with the purpose of detecting cultural bias. These practical constraints underscore the need for improved judgmental

**TABLE 1 |** Types of semiotic resources used in test items.

Type	Main property	Examples
(1) Language resources	Systemic	Vocabulary, grammar, syntactical structures, discourse, idiomatic expressions, quotation marks, formal language, sign language
(2) Images	Mimetic	Photographs, illustrations, drawings
(3) Metaphorical devices	Diegetic	Light bulb representing an idea, speech balloons, arrows, lines connecting labels and elements in an illustration
(4) Abstract representational devices	Analytic	Graphs, tables, symbols, formulas, schemata, flowcharts, color codes
(5) Contexts	Episodic	Characters, places, situations, stories
(6) User Interface Elements	Interactive	Text boxes, cascade menus, cursors, buttons

translation review procedures (Allalouf, 2003; Zhao and Solano-Flores, 2021).

An emerging realization concerning language resources is that language issues in testing cannot be effectively addressed without taking into consideration non-textual ways of representing information (Kopriva and Wright, 2017). Moreover, a broader view of translation as both a meaning making and meaning taking enterprise, reveals the need to recognize multiple forms of translation as intrinsic to the act of representing information (Marais, 2019, p. 122). This broader view appears to be consistent with the ultimate goal of ensuring construct equivalence across cultures and languages. A wealth of possibilities emerge. For example, in addition to replacing text in one language with text in another language, should translation concern semiotic modalities other than text (e.g., replacing illustrations used in tests)? Also, are there cases in which translation should be transmodal (e.g., replacing text with illustrations or illustrations with text)? Of course, substantial conceptual developments need to take place before these thoughts can be incorporated into testing practices.

## Images

Images are semiotic resources intended to convey meaning through mainly graphic, non-textual components. Photographs, illustrations, and drawings are examples of images. Images can be characterized as *mimetic* artifacts—they serve descriptive (rather than interpretive) purposes; they are intended to show entities, rather than to tell about their characteristics.

While images vary on their level of realism (the extent to which the representation of an object resembles the object represented as it would be seen in its presence), tangibility (the extent to which the characteristics of the object represented are concrete), and completeness (the extent to which the representation includes all the elements of the object), there is always a minimum of topological correspondence between the characteristics of the object and its representation. This topological correspondence is preserved, at least to some extent, even in cartoons—which deliberately distort, magnify, minimize, or omit components of the objects they represent. The assumption that meaning in images is self-evident neglects the role of the viewer in the communicative role of images, as there is evidence that individuals with different cultural backgrounds focus on different aspects of images (e.g., Boduroglu et al., 2009).

Research on the use of images in education has been uneven and unsystematic. Through history, images have attracted the attention of researchers at scattered points in time and the aspects investigated have not followed a coherent thematic line (e.g., Fleming, 1966; Miller, 1938; Levie and Lentz, 1982). Research on the use of images in educational assessment has been, in addition, scant (e.g., Washington and Godfrey, 1974). While assessment frameworks and other documents recognize the importance of images in assessment (e.g., NGSS Lead States, 2013), they do not provide clear conceptualizations for systematic image development. As a result, items may contain images whose intended functions (e.g., as supports of the text of items, as stimulus materials, or as decorative components) are unclear or vague, and whose characteristics (e.g., complexity, style) are not consistent across items.

An important notion in the field of social semiotics is that text and image are interconnected, in the sense that the user makes meaning based on using the textual and non-textual information in combination (Kress, 2010). Consistent with this notion, there is evidence that, in making sense of items accompanied by illustrations, examinees not only use the images to make sense of the text but also use the text of items to make sense of the images (Solano-Flores et al., 2014b). Also, evidence from international test comparisons suggests that, in making sense of items, examinees from high-ranking countries have a stronger tendency than examinees from low-ranking countries to cognitively integrate text and image (Solano-Flores et al., 2016). This evidence speaks to the importance of addressing the multiple ways in which disciplinary knowledge is represented throughout the entire process of test development. Since the inception of items, images (as well as other semiotic resources) should be developed along with the text of the items.

The use of images as potential visual supports for students to understand the text of items has originated a wide variety of types of images, such as those intended to illustrate the text of an item as a whole (Kopriva, 2008; Solano-Flores, 2011; Turkan et al., 2019) and those intended to illustrate the options of multiple choice items (Noble et al., 2020). Also, thanks to the ability of computers to interact with their users, it is possible to provide pop-up images that illustrate specific words or terms and which appear on the screen when the examinee clicks on them (Guzman-Orth and Wolf, 2017; Solano-Flores et al., 2019). Due to the recency of these innovations, empirical evidence on effective design and use is just beginning to appear.

## Metaphorical Devices

Metaphorical devices are representations of tangible or visible objects, events, actions, or conditions intended to represent invisible or intangible events, actions, or conditions figuratively. While the term, *metaphor* has a long use history in semiotics (see Eco and Paci, 1983), in this conceptual framework the word metaphorical is reserved to this type of semiotic resource.

Metaphorical devices originate from the need to overcome the limitations imposed by the medium in which information is represented. For example, the need to use lines to represent movement or the direction of actions originates from the limitations of representing certain actions in a given medium (e.g., Krull and Sharp, 2006; Lowe and Pramono, 2006). Arrows departing from labels and pointing at different parts of a flower are effective as a semiotic resource because they are associated to the idea of direction and precision. The cross section of a volcano showing its chimney and lava concretizes a hypothetical situation (*If we would cut a volcano by the half and see what is inside.*). A bubble representing the thoughts of a person is a proxy to intangibility and ephemerality; the text inside the balloon makes those thoughts accessible to the viewer.

Typically used in combination with images, metaphorical devices may have textual or non-textual components or both textual and non-textual components. Metaphorical devices are *diegetic*—they serve a narrative function, rather than a descriptive function. They inform the viewer about something being shown; they explain, clarify, or emphasize. An implicit assumption in the



use of metaphorical devices is that the viewer understands that they are not part of the objects represented. The arrows pointing at different parts of a flower in a science item are not intended to be interpreted by the viewer as being in the same place as the flower; the volcano is not supposed to be interpreted as actually being cut; the thought bubble is not part of the story told—the thoughts represented with words (although not the words) are.

While they are common in instructional materials, textbooks, tests, and other materials, it is possible that individuals do not learn to use and interpret most of the metaphorical devices through formal learning experiences. Indeed, it is possible that many metaphorical devices used in instructional materials and tests have been borrowed from popular culture. At least in the case of the representation of motion in static materials, the use of different semiotic resources tends to originate from the work of illustrators and graphic designers rather than from systematic work on visual literacy (de Souza and Dyson, 2007).

As with images, many metaphorical devices may be used in tests intuitively, under the assumption that they are universal and, therefore, their meaning is self-evident. However, while some semiotic resources can be readily used by individuals to represent and interpret abstract ideas such as sequence and causation (Heiser and Tversky, 2006), this may not apply to other metaphorical devices.

## Abstract Representational Devices

Abstract representational devices convey meaning through the interplay of multiple representational textual and non-textual components (e.g., words, symbols, and lines). Tables, graphs, and formulas are examples of abstract representational devices. Abstract representational devices are *analytic*; they present information on different aspects or parts of an object or phenomenon in ways intended to make relationships (e.g., proportion, causation, equivalence, sequence, hierarchy, magnitude, etc.) between entities explicit (e.g., through contrast or comparison).

Abstract representational devices have no topological correspondence with the objects or phenomena they represent—most of them are based on abstractions and generalizations about the objects represented. Instead, the precision in the way in which information is presented and the relevance of the information included play a critical role in their construction. For instance, the expressions  $7x + (4/3)y$  and  $(7x + 4)/3y$  have different meanings due to a difference in the location of the parentheses.

Following language resources, abstract representational devices are probably the second type of semiotic resource most commonly taught in formal instruction (Macdonald-Ross, 1977). However, this does not mean that they can be used without worrying about challenges for interpretation. For example, there is evidence that it takes a great deal of time and effort for individuals to develop the habit of communicating ideas with diagrams (Uesaka and Manalo, 2012).

The belief that, because they are part of disciplinary knowledge, formal information representation devices are universal and, therefore, everybody within the same discipline interprets and use them in the same way has been long discredited (see Pimm, 1987). For example, mathematical

notation varies considerably across countries (Libbrecht, 2010). As with images, the complexities of properly developing and using abstract representational devices in tests may have been underestimated by assessment programs and their characteristics are not discussed in detail in assessment frameworks and item specifications documents. For example, tables summarizing information provided by items as stimulus materials do not have a consistent style across items within the same assessment program (Solano-Flores et al., 2009).

While standards, assessment frameworks, and other normative documents address the use of graphs, charts, schemata, and other abstract representational devices, the prescriptions they provide focus on the interpretation of content-related data (National Research Council, 2012). Yet it is not uncommon for assessment programs such as PISA (e.g., OECD, 2019) to include, in items not intended to assess data interpretation or representation, tables as resources to provide contextual information and for examinees to provide their answers. Also, rarely do normative documents address the complexity of these devices as a factor to control for in the design of tests. There is evidence on the effectiveness of abstract representational devices in supporting examinees with different cultural backgrounds to understand the content of items (Martiniello, 2009). However, this evidence is difficult to generalize because available literature is not sufficiently explicit about the complexity of those representational devices. In addition, different authors classify abstract representational devices in different ways, for example, by referring to different representational devices with the same name or to the same representational device with different names (see Wang, 2012).

## Contexts

Contexts are plots, scenarios, or stories used with the intent to make tasks or problems meaningful to examinees. Contexts are very common in current large-scale assessment programs. For example, a study on the use of contexts in PISA 2006 and PISA 2009 items found that about one third of the sample of items examined contained contexts in the form of a narrative (Ruiz-Primo and Li, 2016).

Contexts are *episodic*—they involve a fictitious or non-fictitious event, a set of circumstances. This event and these circumstances give rise to a problem that needs to be solved. The events or objects involved are assumed to be familiar to all examinees. Contexts may vary on their degree of concreteness (the extent to which the problem resembles the kinds of problems the examinee would encounter in real life) and authenticity (the extent to which problem resembles the problems and situations that are characteristic of a given discipline or professional activity).

Although the use of contexts is not necessarily a guarantee that items tap into higher order thinking skills, their popularity may have been fueled by constructivist thinking in the field of instruction, which emphasizes situated learning (Schoenfeld, 2004). Since the 1990s, tasks situated in meaningful contexts have been regarded as potential instruments for both promoting and assessing higher order thinking skills (e.g., Shavelson et al., 1990). Yet little is known about what makes contexts effective and how



exactly they contribute to make items better (see Ruiz-Primo and Li, 2015; Ruiz-Primo et al., 2019).

At the college level, efforts to assess critical thinking have led to the development of constructed-response tasks situated in realistic, complex scenarios (Zlatkin-Troitchanskaia and Shavelson, 2019). Accomplishing context authenticity across countries takes careful work. For example, in the International Performance Assessment of Learning initiative, a great deal of the work on test development focuses on ensuring that the same context is presented in different versions according to the characteristics of each country. Also, a great effort is put into ensuring that stimulus materials such as e-mails, newspaper clips, letters, and reports that examinees are asked to read have the same appearance and style of real documents they would encounter in their countries (Shavelson et al., 2019).

While rarely is engagement mentioned in assessment normative documents, contexts are semiotic resources that potentially can capture examinees' interest during test taking (Fensham, 2009). At the same time, contexts may be distracting. There is evidence that some examinees may not be skilled enough to tell apart the problem posed by an item and the contextual information used to introduce the problem (Solano-Flores, 2011). Also, contexts may account for more for differences in student performance than the skills items are intended to assess. An investigation on inferential reading comprehension in which the narrative structure and the linguistic complexity of the texts used as stimulus materials were kept constant found that the topic of the story, more than any other factor, was the main source of score variation among second language learners in the U.S. tested in both their first language and the second language (González-Otero, 2021).

Altogether, this evidence shows that item contexts are tremendously complex, delicate semiotic resources that need to be developed carefully. If the characters, events, or objects depicted (and their appearance) are not equally familiar to all examinees, contexts may end up adding information that is irrelevant to the target construct and unnecessarily increase item difficulty.

## User Interface Elements

User interface elements are textual, visual, and auditory components embedded in a computer-administered environment and intended to facilitate the interaction of the examinee with the computer for the examinee to obtain and enter information with ease. User interface elements are *interactive*—they react to the examinees' actions. They include cursors, pointers, cascade menus, buttons, boxes, hyperlinks, icons, and navigation arrows, among many other features. They are operated or activated by actions that include hovering, clicking, dragging objects, etc.

Due to globalization, the ubiquity of some platforms, and the widespread presence of certain websites in many countries, certain user interface elements may be in the process of global standardization, may be familiar to multiple populations of examinees, and may be mimicked by many other platforms—including testing platforms. However, the influence of local and regional cultural factors in this process should not be

underestimated. There is evidence that the design of websites reflects the preferences, worldviews, and communication styles of the cultural contexts in which they originate. Important differences have been documented on attributes such as layout, color, links, navigation, etc. (Alexander et al., 2016). In addition, many user interface elements should not be assumed to be static. For example, icons tend to change with every version of the same software or platform (Familiant and Detweiler, 1993)—which potentially may be a challenge for interpretation.

The field of information technology has developed a wide variety of methods for website localization—the adaptation of the websites to the characteristics of a specific target country, cultural group, language, or region. These methods are intended to address subtle cultural differences (Aykin, 2005). Regrettably, while those methods are frequently used in marketing and business, they are yet to be adopted as part of the translation and adaptation practices in international test comparisons.

The assumption that a given user interface element is interpreted in the same way by everybody may not hold equally for different populations. Differences in the popularity and cost of certain devices and differences in access to computers and the internet (OECD, 2020) may create important differences in the examinees' level familiarity with different user interface elements. In online or computer-based testing, the characteristics of interface user elements may be determined by factors such as the technical properties of the software, processing speed, or hardware requirements, which may constrain or support the design possibilities of computer-administered tests in different ways (International Test Commission, 2005). Also, due to the specific characteristics of online tests (i.e., types of tasks, content area, skills targeted, school grade), certain user interface elements may need to be designed for specific tests (Bennett, 2015).

Since the early days of the internet, web designers have incorporated in their design practices the notion that different cultural groups ascribe different meanings to different colors and other object features. Those features may be used purposefully with the intent to communicate danger, joy, importance, etc. Indeed, it is well known that certain icons, colors, font styles, and other design elements can be used so frequently and consistently in the design of websites in a given country that they become cultural markers (e.g., Barber and Badre, 1998; Cyr et al., 2010). However, whether or how the interpretation of a specific user interface element varies across certain populations of students may be difficult to anticipate. An issue that adds to the challenges to fair, valid testing is the underrepresentation of certain cultural groups in the data that feeds the algorithms used by websites and search engines (Henrich et al., 2010; Noble, 2018).

Current cognitive-based approaches to test design pay special attention to the interplay between the characteristics of items and the characteristics of the knowledge and skills being assessed. Consistent with the notion that response processes do not take place separately for the target constructs and the means through which tests are administered (Ercikan and Pellegrino, 2017), a sound methodology for computer-administered and online test development should enable test developers to treat the constructs assessed and the characteristics of the interface in an integrated manner.

Reasoning from the field of cognitive psychology allows examination of the usability of user interface components—the ease with which they can be used or learned (see Preece et al., 1994; Norman, 2013). Because items, by definition, present examinees with novel situations, the creation of online items involves the design of microinteractions—contained product moments that involve a single use case (Saffer, 2014). To a large extent, the design of an online item is the design of a microinteraction whose complexity is shaped by the content assessed and the characteristics of the user interface.

## BASIC IDEAS FOR SEMIOTIC TEST DESIGN

### Defining Semiotic Test Design

The term, *semiotic test design* should not be confused with the term, *test design*, which is typically used in relation to the technical properties of tests (e.g., Wendler and Walker, 2006; van der Linden, 2016) and the ways in which content is covered through item and population sampling (Gonzalez and Rutkowski, 2010). While assessment frameworks and item specifications documents address the format, structure, complexity, and number of items to be included in tests (e.g., National Assessment Governing Board, 2017), they are not intended to provide detailed information on the multiple features of items.

In contrast, *semiotic test design* is concerned with the selection of optimal sets of item semiotic resources intended to meet the examinees' cultural backgrounds. Ideally, since the inception of a test, and based on the characteristics of the population of examinees, decisions should be made about the characteristics of semiotic resources to use consistently across items in ways intended to minimize challenges for interpretation due to cultural differences.

Nor the term, *semiotic test design* should be confused with *universal design* and *universal test design*, which refer to the set of basic principles and practices intended to ensure that the needs of diverse students are taken into account during the entire process of test development and to maximize accessibility to all examinees (see Lidwell et al., 2003; American Educational Research Association [AERA] et al., 2014; Thurlow and Kopriva, 2015; Sireci and O'Riordan, 2020).

While *semiotic test design* shares those goals and basic principles, it has a more explicit theoretical foundation from the field of social semiotics and its relation to cognitive science, sociolinguistics, and socio-cultural theory. More specifically, semiotic test design aims at minimizing unnecessary cognitive load in items by optimizing item representational complexity and item semiotic alignment.

### Cognitive Load

Cognitive load theory comes handy in reasoning about the design of items and its impact on the working memory that an individual needs to use in responding to an item. Cognitive load theory distinguishes three types of cognitive load—intrinsic, germane, and extraneous. While intrinsic and germane cognitive

load involve respectively mental processing of information that is needed to complete the task and mental processing of information into knowledge structures and their storage in long-term memory, extraneous cognitive load involves mental processing resulting from the manner in which information is presented (Sweller, 1988; Sweller et al., 1998).

A recurrent issue in testing is the increase in an item's extraneous cognitive load that takes place when, in addition to thinking about the problems posed, examinees need to figure out how they need to give their responses to items (Clariana and Wallace, 2002; Carpenter and Alloway, 2018). This concern arises, for example, in online testing endeavors that involve populations with varying levels of familiarity with computers. The generalizability of findings from research that compares the cognitive load imposed by paper-and-pencil and computer-administered tests (e.g., Priscari and Danielson, 2017) appears to be shaped by factors such as the content assessed, the socioeconomic characteristics of the population of examinees, and the examinees' familiarity with computers or the specific testing platform.

### Item Representational Complexity

*Item representational complexity* is defined here as the multiple ways in which the semiotic resources of an item are related to each other. It is the combination, not only the sum of semiotic resources, what influences the ways in which examinees make sense of items. A key tenet in testing is that unnecessary complexity (e.g., too much wording, a crowded item layout) is a source of construct-irrelevant variance because it contributes to increasing extraneous cognitive load. Also, information provided in different sensory modalities without proper organization of different components and pieces of information may hamper, rather than facilitate, information processing because individuals need to split their attention between information provided in disparate modalities and then mentally integrate that information (see Chandler and Sweller, 1992; Mayer et al., 2001).

One of the goals of semiotic test design is to minimize the cognitive load of items by minimizing semiotic item complexity. In online testing, the inclusion of too many features in the user interface (Norman, 2013) may lead to an unnecessary increase of extraneous cognitive load. For example, an item whose response requires from the examinee building a graph by dragging and dropping bar lines into a box, labeling the axes of the graph, and typing a number in a panel, may be too complex compared to the complexity of the specific knowledge the item is intended to assess.

Experience from item writing provides good examples of the intricacies of examining representational complexity. The work on linguistic simplification as a form of testing accommodation for second language learners has focused on minimizing the lexical and syntactical complexity of items with the intent to reduce their reading demands for students who are second language learners. While some lexical variables have been found to be good predictors of item difficulty (Shafteel et al., 2006; Martiniello, 2009), linguistic simplification has been, at best, moderately effective in minimizing limited language proficiency in the language of testing as a source of error variance

(Abedi et al., 2006; Sato et al., 2010; Haag et al., 2015; Noble et al., 2020). These moderate effects suggest that linguistic simplification does not necessarily reduce the reading demands imposed by items. For example, expressing the same idea in fewer and shorter sentences may require a higher level of encoding and the use of more precise words with lower frequencies. While a shorter sentence has fewer words to read, the level of mental processing needed to decode the sentence may be higher.

Research on images provides another set of good examples of the intricacies of examining representational complexity. Consistent with approaches to measuring visual complexity based on the number of components (Forsythe et al., 2003), the analysis of complexity of illustrations used in items has been based on counting the number of different types of features they contain (e.g., color, black and white, or grayscale tonalities; zooming; symbols), as shown in **Table 2**. Based on examining items from different assessment programs, Wang (2012) identified over a hundred features of illustrations used in science items and coded the presence and absence of illustration features as dichotomous (1–0) variables, classified into several categories of illustration features. Unlike other approaches to characterizing images (which are based on broad categories such as “chart,” “table,” or “graph”), this coding approach has allowed systematic examination of illustrations used in different assessment programs (Solano-Flores et al., 2013, 2016; Solano-Flores and Wang, 2015; Shade, 2017).

Quantifying representational complexity also makes it possible to ensure consistency in the complexity of images across items in a test or assessment program. For example, using a set of design criteria that specified the characteristics of illustrations to be added to the text of middle school science items, Wang et al. (2012) were able to create images that had, on average, about 16 features. This number contrasts with the average (rounded) number of 22, 21, and 21 different features observed in Grades 4–12 science items respectively from China, the U.S., and TIMSS (the Trends in Mathematics and Science international assessment program).

Using number of different features as a measure of item representational complexity also makes it possible to compare in detail the characteristics of items from different countries. For example, Wang (2012) compared items from Chinese science assessment programs and items from American assessment programs. She found that, while the average number of different types of features are similar across countries, the most frequent types of features were not necessarily the same across countries. For example, photographs in illustrations were 3.52 times more frequent in items from China than in items from the U.S., whereas analogic line drawings were 3.36 times more frequent in items from the U.S. than in items from China.

## Item Semiotic Alignment

Item semiotic alignment is defined here as shared cultural experience, the intersection of the cultural experience encoded

**TABLE 2 |** Segment of the list used to code non-textual components in different assessment programs.

### OBJECTS AND BACKGROUND

**Image concreteness:** photo; scanned document; text clip; realistic line drawing; schematic; map; silhouette; cartoon; logo; icon; emblem; metonymy; symbol; reference; entity; geometric shape

**Background:** with background; without background

**Zooming:** no zooming; zoom in; zoom out

**View:** external; internal; from above object; from below object; from side of object

**Dimension:** three dimensional; two dimensional

**Relative scale of objects:** proportionate; disproportionate

**Color:** black and white; multicolor; gray scale

**Composition:** single image; compound image; image in an object

### TEXT IN ILLUSTRATION

**Text unit:** non-math/scientific sign; math/scientific sign, and notation; abbreviation; Roman numeral; Arabic numeral; letter; word; phrase; sentence; paragraph; acronym

**Text function:** provide label; provide a code (legend); title/caption/heading; elaborate/explain/describe; comment/note; provide instructions; provide data; text in an object

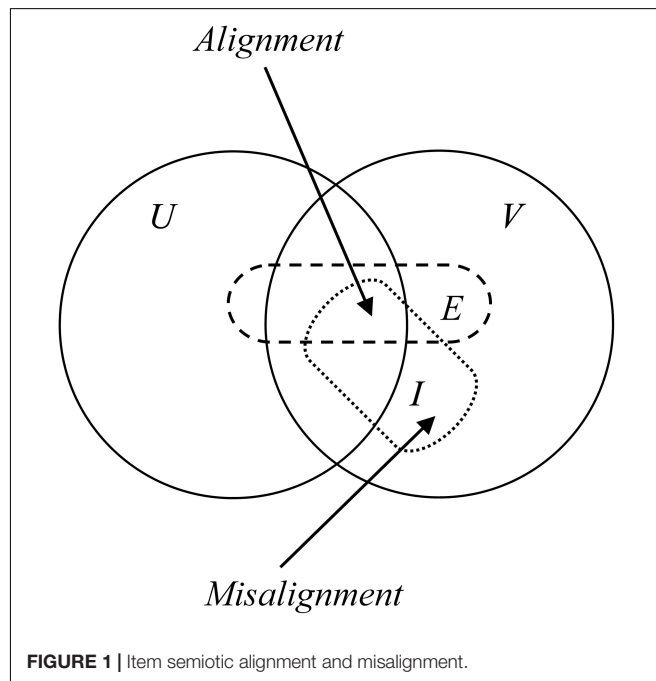
**Text emphasis:** capitalization; bolding; italicizing; underlying; circling

**Text direction:** between left and right; between top and bottom; oblique direction

### CONTEXT

**Socio-cultural focus:** an undefined person; peers/teachers; media celebrities (characters); family/home; school/class; community/neighborhood; state/province; home country; world/global

*Adapted from Wang (2012).*



**FIGURE 1 |** Item semiotic alignment and misalignment.

in the semiotic resources used in an item and the examinee's cultural experience. Conversely, semiotic item misalignment can be defined as the cultural experience encoded in the semiotic resources used in an item but not shared by the examinee.

**Figure 1** represents that intersection in a Venn diagram.  $U$  and  $V$  are different cultural contexts,  $I$  is an item originated in  $U$ , and  $E$  is an examinee's cultural experience.

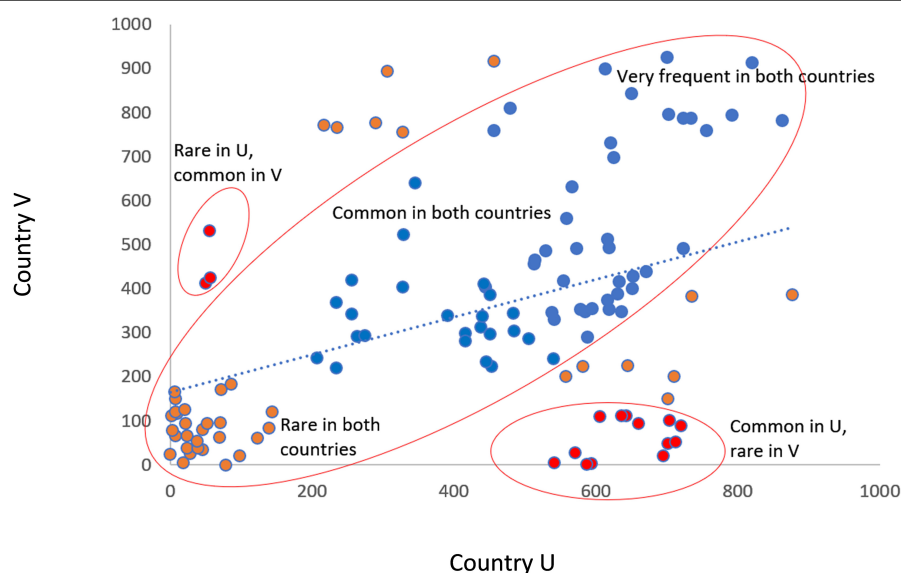
Because cultural groups are not isolated,  $U$  and  $V$  are shown as intersecting. The item is assumed to encode cultural experience predominantly from  $U$  but also from  $V$ —which is represented respectively as the intersection of  $I$  and  $U$  and the intersection of  $I$  and  $V$ . Similarly, because people do not live their lives in isolation within one single cultural context, a given individual's cultural experience is assumed to develop within both  $U$  and  $V$ —a notion that is represented in the diagram respectively as the intersection of  $E$  and  $U$  and the intersection of  $E$  and  $V$ . The figure shows misalignment as partial, as it is very unlikely for an examinee not to share any cultural experience encoded by the semiotic resources of an item.

Of course, semiotic alignment is difficult to evaluate, given the thousands of possible features of items and the uniqueness of every individual's cultural experience. Yet the notion is helpful in reasoning about the ways in which the examinees' assumed cultural experience (or the lack of knowledge on the examinees' cultural experience) needs to be taken into account when developing or examining tests. For example, experience from research examining students' interpretations of contexts indicates that semiotic misalignment increases item extraneous cognitive load. The notion that an individual's socio-cultural activity takes place at different levels of social participation (apprenticeship, guided participation, and participatory appropriation; Rogoff, 1995) is key to interpreting the findings. There is evidence that, in attempting to make sense of items, examinees make connections between the contexts of items and their own personal experiences (Solano-Flores and Li, 2009, 2013). Item contexts are more meaningful to examinees when they portray situations in which they

are actors, rather than observers or apprentices (Solano-Flores and Nelson-Barber, 2001; Le Hebel et al., 2013). An implication of this evidence is that, if the situations and lifestyles portrayed by items are predominantly those of a given cultural group, then contexts may fail to provide the same level of support to all students, even if those items are seemingly familiar to all.

The emotional impact of an excessive representation of a privileged segment of the society in tests may also adversely affect the performance on tests of students from certain cultural groups. There is evidence that the sole impression of being excluded or treated differently in a testing situation may affect the performance of examinees in a test (Steele and Aronson, 1998). Also, there is evidence that individuals with different cultural backgrounds may interact in different ways with tests (Cizek and Burg, 2006; Madaus and Russell, 2010). Given this evidence, it does not seem unreasonable to expect that examinees from certain cultural groups who do not have difficulty interpreting certain contexts may still feel alienated when the contexts used in items do not reflect their everyday lives and that feeling of alienation may adversely affect their performance on tests (Solano-Flores et al., 2014a).

Note that this reasoning on item semiotic alignment applies to all types of semiotic resources equally. While literature on testing and diversity has paid attention almost exclusively to language resources and contexts, other types of semiotic resources need to be considered in examining item semiotic alignment. For example, speech balloons and thought bubbles illustrate how and how frequently semiotic resources used in different cultural contexts may shape its effectiveness as means for item meaning making. Probably it is not an overstatement to say that these semiotic resources are used in many societies (Cohn, 2013). However, this does not



**FIGURE 2 |** Scattergram of the frequency of 120 item features in two hypothetical samples of 1,000 Grade 5 science items from two countries.



necessarily mean that their communicative value in tests is the same for any population. Due to their association with visual mass media (see Lefèvre, 2006), in some societies these metaphorical devices may be rarely used in textbooks and instructional materials; they may even be regarded as inappropriate for educational contexts. Even if they are common in textbooks and instructional materials, their use may not be customary in tests.

## Identification of Item Features and Selection of Item Semiotic Resources in International Tests

When a test involves multiple countries, two issues need to be addressed: (1) To what extent individuals from different countries are likely to interpret the features of items as intended? and (2) How similar is the frequency with which the features of items occur in different countries?

Regarding the first question, cognitive interviews, expert panels, and focus groups can produce data on response processes (e.g., Leighton, 2017; Zhao, 2018) and, more specifically, information on the ways in which the features of items influence examinees' interpretations of items. These methods have been discussed extensively (e.g., Ericsson and Simon, 1993; Megone et al., 1994) and are not discussed here. However, it is important to mention that, because these methods are costly and time consuming, their use may need to be restricted to small numbers of item semiotic resources.

Comparative frequency analyses can produce data relevant to the second question. Frequency is used as a proxy of familiarity: if a given feature occurs with similar frequencies in different countries, it is assumed that individuals from these countries are equally familiar with it and are likely to interpret it in the same way.

Lessons from investigations like Wang's (2012), discussed above, can guide actions oriented to identifying the types of semiotic resources that are or are not likely to successfully convey the intended meaning in testing culturally and linguistically diverse populations of examinees. **Figure 2** shows a hypothetical scatterplot of the frequency of 120 features in two samples of items from two countries, *U* and *V*. In this hypothetical example, each sample contained 1,000 items and the two samples were equivalent—they comprised items of the same grade and the same content area.

The trend (dotted) line shows that, in general, the features tend to appear more frequently in items from Country *U* than in items from Country *V*. Three main types of features can be identified according to the frequencies with which they appear in the two countries: Those that are more common in *U* than in *V*, those that tend to be more common in *V* than in *U*, and those that are equally common in *U* and *V*.

If a test intended to assess populations from Countries *U* and *V* were to be created, the features with substantially different frequencies (red color) would be the first candidates for exclusion from the pool of potential item semiotic resources to be used in creating the items. In contrast, features with similar frequencies (blue color) would be the first candidates for

**TABLE 3 |** Use specifications for the design parameter, *Division Notation* in a hypothetical international mathematics test.

Division notation	Use specifications
$\frac{x}{y}$	In all countries, in fill-in the blank problems. Do not use in item stems.
$x/y$	In all countries, except Country H and Country M, in item stems.
$x \div y$	Only in Country H, in item stems.
$x:y$	Only in Country M, in item stems.
$y\sqrt{x}$	Do not use.

**TABLE 4 |** Design parameters of illustrations used to create illustrations accompanying the text of items for students who were not proficient in the language in which they were tested.

Design Parameter and Categories	Value or Category Selected
Framing: Yes/No	Framing
Position relative to text: Left/Right Above/Below Text	At the right of the text of the item
Drawings: Yes/No	Drawings
Color: Full Color/Gray Tone/Black and White	Only Black and white
Realistic/Fantastic representations	Only Realistic
Cartoon: Yes/No	No cartoons
Concrete objects/Abstract ideas	Only concrete objects
View level: Horizontal/From Above/From Below	Only Horizontal view
Relative Scale of Components Preserved: Yes/No	No changes in the relative scale
Perspective: Yes/No	Perspective
Labels: Yes/No	No labels
Sequences-stages: Yes/No	No stages
Backgrounds: Yes/No	No background
Metaphorical devices: Yes/No	No metaphorical devices

*Adapted from Solano-Flores et al. (2014b).*

inclusion. After this initial selection stage, a more manageable number of features would remain yet to be examined in detail. Among these semiotic resources would be, first, those with important different frequencies—outliers in the pattern of distribution of the scatterplot—and second, those with similar but low frequencies in both countries. The viability of these two types of features as semiotic resources to be used in the test could be determined through cognitive interviews, focus groups, and expert panels.

It is important to mention that the use of this approach in international test comparisons contributes to minimizing test bias across countries, not within countries. International test comparison programs are typically silent about the tremendous cultural and socio-economic differences and countries are treated as homogeneous. Yet there is evidence of tremendous test score differences attributable to socio-economic inequalities (e.g., Carnoy and Rothstein, 2013).

## Item Design Parameters

Item design parameters are variables that specify the set of semiotic resources that are to be used in the items of a

test or assessment program and the conditions under which their values or categories are to be used (Solano-Flores et al., 2014b). The specification of design parameters is intended to ensure consistency in the characteristics of items and minimize interpretation challenges for individuals from all cultural backgrounds. Current testing practices do not reach that level of standardization because item specification documents or test translation and adaptation guidelines generated by large-scale assessment programs are not sufficiently explicit about the parameters to be used in developing items.

**Table 3** shows a design parameter and the use specifications for each of its categories for a hypothetical mathematics test involving multiple countries. **Table 4** provides an example of a set of design parameters used in an investigation that evaluated the effectiveness of vignette illustrations (illustrations added to the text of items with the intent to support students who were not proficient in the language in which the tests were administered to gain access to the content of items). The figure shows only one subset of parameters from a much larger possible set of design parameters that could be identified as relevant to creating vignette illustrations (Solano-Flores et al., 2014b).

Note that the specification of design parameters is not specific to semiotic resources clearly related to the content assessed. Also, which design parameters are relevant and which of their values or categories need to be selected need to be determined according to the characteristics of each assessment endeavor, such as the target populations of examinees, the content, and the cultural groups involved.

To date, design parameters have been used only in a few studies and programs (Kachchaf, 2018; Solano-Flores et al., 2019; Smarter Balanced Assessment Consortium, 2020) to produce pop-up illustrations glossaries (visual representations of words that appear on the screen when examinees click on words they do not understand) and other accessibility resources intended to provide support to students with special needs. These efforts show that it is possible to ensure standardization and efficiency in the selection and use of item semiotic resources.

## SUMMARY AND CONCLUDING REMARKS

Approaches to examining cultural bias in items tend to focus on the ways in which, due to cultural differences, the characteristics of items may prevent students from properly understanding the content of items. In the absence of a conceptual framework on semiotic test design, it is difficult

to link specific characteristics of items to the performance on tests of different cultural or linguistic groups or to translate the lessons learned from those experiences into improved testing practices. More specifically, in the absence of a conceptual framework on semiotic test design, it is difficult to establish the set of item features that are likely to minimize cultural bias. Item specifications documents provide coarse-grain information useful for systematically generating items according to the content and type of knowledge assessed, but they cannot provide design parameters to be used across all items within the same assessment program.

This paper has presented a conceptual framework for test design from the perspective of social semiotics. It has offered a typology for characterizing the wide variety of semiotic resources used in items and discussed challenges and possibilities in their use in the testing of culturally and linguistically diverse populations. The conceptual framework also discusses basic ideas on semiotic test design, which is intended to support the systematic selection and use of sets of semiotic resources in tests. According to the framework, differences in the frequency of semiotic resources in different societies may produce different degrees of semiotic alignment for different cultural groups. Semiotic test design allows identification of an optimal pool of semiotic resources for a test or assessment program intended to minimize extraneous cognitive load in items by minimizing item representational complexity and maximizing item semiotic alignment for the maximum number of examinees.

The conceptual framework offered makes it possible to imagine a stage in the process of test development focused on specifying design parameters that are relevant to the design of items and decide on the categories or values to apply for each design parameter. Naturally, these decisions need to be supported by information from multiple sources, such as comparative studies of tests across countries, cognitive interviews, expert panels, and focus groups with individuals from the target populations of examinees.

Semiotic test design allows development of test items based on identifying and selecting the optimal features of test items, given the cultural and linguistic characteristics of the target populations. In sum, semiotic test design offers the opportunity to address the complex representational nature of disciplinary knowledge in multicultural, multilingual contexts.

## AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

## REFERENCES

- Abedi, J. (2006). "Language issues in item development," in *Handbook of Test Development*, eds S. M. Downing and T. M. Haladyna (Mahwah, NJ: Lawrence Erlbaum Associates, Publishers), 377–398.
- Abedi, J. (2016). "Language issues in test development," in *Handbook of Test Development*, eds S. M. Downing and T. M. Haladyna (Mahwah, NJ: Lawrence Erlbaum), 355–373.
- Abedi, J., Courtney, M., Leon, S., Kao, J., and Azzam, T. (2006). *English Language Learners and Math Achievement: A Study of Opportunity to Learn and Language Accommodation*. Los Angeles, CA: University of California.
- Alexander, R., Thompson, N., and Murray, D. (2016). Towards cultural translation of websites: a large-scale study of Australian, Chinese, and Saudi Arabian design preferences. *Behav. Inf. Technol.* 36, 1–13. doi: 10.1080/14781700.2019.1664318

- Allalouf, A. (2003). Revising translated differential functioning items as a tool for improving cross-lingual assessment. *Appl. Meas. Educ.* 16, 55–73. doi: 10.1207/s15324818ame1601\_3
- American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (2014). *Standards for Educational and Psychological Testing*. Washington, DC: Joint Committee on Standards for Educational and Psychological Testing (U.S.).
- Anagnostopoulou, K., Hatzinikita, V., Christidou, V., and Dimopoulos, K. (2013). PISA test items and school-based examinations in Greece: exploring the relationship between global and local assessment discourses. *Int. J. Sci. Educ.* 35, 636–662. doi: 10.1080/09500693.2011.604801
- Arffman, I. (2013). Problems and issues in translating international educational achievement tests. *Educ. Meas.* 32, 2–14. doi: 10.1111/emip.12007
- Aykin, N. (2005). “Overview: where to start and what to consider,” in *Usability and Internationalization of Information Technology*, ed. N. Aykin (Mahwah, NJ: Lawrence Erlbaum Associates, Publishers). doi: 10.1201/b12471
- Baecker, C. R. (2010). “A cross-cultural study on the effect of decimal separator on price perception,” in *A Work Project Presented as Part of the Requirements for the Award of a Masters Degree in Management*, (Rua da Holanda: Faculdade de Economia da Universidade Nova de Lisboa).
- Barber, W., and Badre, A. (1998). “Culturability: the merging of culture and usability,” in *Proceedings of the 4th Conference on Human Factors and the Web*, June 5, 1998, Basking Ridge, NJ.
- Bennett, R. E. (2015). The changing nature of educational assessment. *Rev. Res. Educ.* 39, 370–407. doi: 10.3102/0091732X14554179
- Boduroglu, A., Shah, P., and Nisbett, R. E. (2009). Cultural differences in allocation of attention in visual information processing. *J. Cross Cult. Psychol.* 40, 349–360. doi: 10.1177/0022022108331005
- Carnoy, M., and Rothstein, R. (2013). *What do International tests Really Show about U.S. Students Erformance?*. Washington, DC: Economic Policy Institute.
- Carpenter, R., and Alloway, T. (2018). Computer versus paper-based testing: are they equivalent when it comes to working memory? *J. Psychoeduc. Assess.* 37, 382–394. doi: 10.1177/0734282918761496
- Chandler, P., and Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *Br. J. Educ. Psychol.* 62, 233–246. doi: 10.1111/j.2044-8279.1992.tb01017.x
- Chua, H. F., Boland, J. E., and Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12629–12633. doi: 10.1073/pnas.0506162102
- Cizek, G. J., and Burg, S. S. (2006). *Addressing Test Anxiety in a High-Stakes Environment*. Thousand Oaks, CA: Corwin press.
- Clariana, R., and Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *Br. J. Educ. Technol.* 33, 593–602. doi: 10.1111/1467-8535.00294
- Cohn, N. (2013). Beyond speech balloons and thought bubbles: the integration of text and image. *Semiotica* 197, 35–63. doi: 10.1515/sem-2013-0079
- Council of Chief State School Officers (2015). *Science Assessment Item Collaborative item Specifications Guidelines for the Next Generation Science Standards*. Washington, DC: Council of Chief State School Officers.
- Coupé, C., Oh, Y. M., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: comparable information rates across the human communicative niche. *Sci. Adv.* 5:eaaw2594. doi: 10.1126/sciadv.aaw2594
- Cyr, D., Head, M., and Larios, H. (2010). Colour appeal in website design with and across cultures: a multi-method evaluation. *Int. J. Hum. Comput. Stud.* 68, 1–21. doi: 10.1016/j.ijhcs.2009.08.005
- Davies, M., and Gardner, D. (2010). *Word Frequency List of American English*. Available at: <https://www.wordfrequency.info/files/entries.pdf>. (accessed February 20, 2021).
- de Souza, J. M. B., and Dyson, M. C. (2007). “An illustrated review of how motion is represented in static instructional graphics,” in *First Global Conference on Visual Literacies*, Seville: University of Seville. doi: 10.1016/b978-0-240-81010-2.50004-3
- Downing, S. M., and Haladyna, T. M. (eds) (2006). *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Eco, U., and Paci, C. (1983). The scandal of metaphor: metaphorology and semiotics. *Poetics Today* 4, 217–257. doi: 10.2307/1772287
- Ercikan, K. (2002). Disentangling sources of differential item functioning in multi-language assessments. *Int. J. Test.* 2, 199–215. doi: 10.1207/s15327574ijt023264\_2
- Ercikan, K., and Pellegrino, J. W. (2017). “Validation of score meaning using examinee response processes for the next generation of assessments,” in *Validation of Score Meaning in the Next Generation of Assessments*, eds K. Ercikan and J. Pellegrino (New York, NY: Routledge), 1–8. doi: 10.4324/9781315708591-1
- Ercikan, K., Roth, W.-M. Simon, M., Sandilands, D., and Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education* 27, 275–285. doi: 10.1080/08957347.2014.944306
- Ericsson, K. A., and Simon, H. S. (1993). *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: The MIT Press. doi: 10.7551/mitpress/5657.001.0001
- Familant, M. L., and Detweiler, M. C. (1993). Iconic reference: evolving perspectives and an organizing framework. *Int. J. Man Mach. Stud.* 39, 705–728. doi: 10.1006/imms.1993.1080
- Fensham, P. J. (2009). Real world contexts in PISA science: implications for context-based science education. *J. Res. Sci. Teach.* 46, 884–896. doi: 10.1002/tea.20334
- Fleming, M. L. (1966). *Instructional Illustrations: A Survey of Types Occurring in Print Materials for Four Subject Areas*. Washington, D.C: U.S. Department of Health, Education & Welfare. Project No. 1381 MDEA Tittle VIIA-1381, Grant OE-7-24-0210-279.
- Forsythe, A., Sheehy, N., and Sawey, M. (2003). Measuring icon complexity: an automated analysis. *Behav. Res. Methods Instrum. Comput.* 35, 334–342. doi: 10.3758/bf03202562
- Gonzalez, E., and Rutkowski, L. (2010). Principles of multiple booklet matrix designs and parameter recovery in large-scale assessments. Issues and methodologies in large-scale assessments. *IERI Monogr. Ser.* 3, 125–156.
- González-Otero, S. (2021). *Habilidades Lectoras Como Función de la Heterogeneidad Lingüística en Estados Unidos*. Tesis Doctoral, Universidad da Coruña, Spain.
- Guzman-Orth, D., and Wolf, M. (2017). “Illustration glossaries for English learners: Findings from cognitive labs,” in *Paper Presented at the Annual Conference of the American Educational Research Association*, San Antonio, TX.
- Haag, N., Heppt, B., Roppelt, A., and Stanat, P. (2015). Linguistic simplification of mathematics items: effects for language minority students in Germany. *Eur. J. Psychol. Educ.* 30, 145–167. doi: 10.1007/s10212-014-0233-6
- Halliday, M. A. K. (1978). *Language as a Social Semiotic*. London: Edward Arnold.
- Hambleton, R. K. (2005). “Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures,” in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, eds R. K. Hambleton, P. F. Merenda, and C. D. Spielberger (Mahwah, NJ: Lawrence Erlbaum Associates, Publishers). doi: 10.4324/9781410611758
- Heiser, J., and Tversky, B. (2006). Arrows in comprehending and producing mechanical diagrams. *Cogn. Sci. A Multidiscip. J.* 30, 581–592. doi: 10.1207/s15516709cog0000\_70
- Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–135. doi: 10.1017/s0140525x0999152x
- International Test Commission (2005). *ITC Guidelines on Computer-Based and Internet Delivered Testing*. Available at: [www.intestcom.org](http://www.intestcom.org). (accessed February 20, 2021).
- International Test Commission (2017). *The ITC Guidelines for Translating and Adapting Tests*, 2nd Edn. Available at: [www.IntTestCom.org](http://www.IntTestCom.org). (accessed February 20, 2021).
- Kachchaf, R. (2018). “Illustration glossaries: update on post-pilot analysis,” in *Presentation at the Smarter Balanced Technical Advisory Committee*, Minneapolis, MN.
- Kane, M. T. (1982). A sampling model for validity. *Appl. Psychol. Meas.* 6, 125–160. doi: 10.1177/014662168200600201
- Kopriva, R. J. (ed.) (2008). *Improving Testing for English Language Learners*. New York, NY: Routledge.
- Kopriva, R. J., and Wright, L. (2017). “Score processes in assessing academic content of non-native speakers: literature review and ONPAR summary,” in *Validation of Score Meaning in the Next Generation of Assessments: The use of*



- Response Processes*, eds K. Ercikan and J. Pellegrino (New York, NY: Routledge), 100–112. doi: 10.4324/9781315708591-9
- Kress, G. (2010). *Multimodality: A Social Semiotic Approach to Contemporary Communication*. New York, NY: Routledge. doi: 10.4324/9780203970034
- Kress, G., and van Leeuwen, T. (2006). *Reading Images: The Grammar of Visual Design*, 2nd Edn. New York, NY: Routledge. doi: 10.4324/9780203619728
- Krull, R., and Sharp, M. (2006). Visual verbs: using arrows to depict the direction of actions in procedural illustrations. *Inf. Des. J.* 14, 189–198. doi: 10.1075/idj.14.3.01kru
- Lane, S., Raymond, M. R., Haladyna, T. N., and Downing, S. M. (2016). “Language issues in test development,” in *Handbook of Test Development*, eds S. M. Downing and T. M. Haladyna (Mahwah, NJ: Lawrence Erlbaum), 3–18.
- Le Hebel, F., Tiberghien, A., and Montpied, P. (2013). “Sources of difficulties in PISA science items,” in *ESERA Conference 2013, Sept 2013*, eds C. P. Constantinou, N. Papadouris, and A. Hadjigeorgiou Nicosia, 76–84. Strand 11: Evaluation and assessment of students learning and development.
- Lefèvre, P. (2006). “The battle over the balloon: the conflictual institutionalization of the speech balloon in various European cultures,” in *Image Narrative: Online Magazine of the Visual Narrative*, 14. Available online at: [http://www.imageandnarrative.be/inarchive/painting/pascal\\_levivre.htm](http://www.imageandnarrative.be/inarchive/painting/pascal_levivre.htm) (accessed February 21, 2021).
- Leighton, J. P. (2017). “Collecting and analyzing verbal process data in the service of validity and interpretive arguments,” in *Validation of Score Meaning in the Next Generation of Assessments: The use of Response Processes*, eds K. Ercikan and J. Pellegrino (New York, NY: Routledge), 25–38. doi: 10.4324/9781315708591-3
- Lemke, J. L. (1990). *Talking Science: Language, Learning, and Values*. Norwood, NJ: Ablex Publishing.
- Lemke, J. L. (1998). “Multiplying meaning: visual and verbal semiotics in scientific text,” in *Reading Science: Critical and Functional Perspectives on Discourses of Science*, eds J. R. Martin and R. Veel (New York, NY: Routledge), 87–113.
- Levie, W. H., and Lentz, R. (1982). Effects of text illustrations: a review of research. *Educ. Commun. Technol. J.* 30, 195–232.
- Libbrecht, P. (2010). “Notations around the world: census and exploitation,” in *Proceedings of the 10th ASIC and 9th MKM International Conference, and 17th Calculemus Conference on Intelligent Computer Mathematics*, New York, NY: ACM, 398–410. doi: 10.1007/978-3-642-14128-7\_34
- Lidwell, W., Holden, K., and Butler, J. (2003). *Universal Principles of Design: 125 Ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions, and Teach Through Design*. Beverly, MA: Rockport Publishers, Inc.
- Lotman, Y. M., Uspensky, B. A., and Mihychuk, G. (1978). On the semiotic mechanism of culture. *New Lit. Hist.* 9, 211–232. doi: 10.2307/468571
- Lowe, R., and Pramono, H. (2006). Using graphics to support comprehension of dynamic information in texts. *Inf. Des. J.* 14, 22–34. doi: 10.1075/idj.14.1.04low
- Macdonald-Ross, M. (1977). Graphics in texts. *Rev. Res. Educ.* 5, 49–85. doi: 10.2307/1167172
- Madaus, G., and Russell, M. (2010). Paradoxes of high-stakes testing. *J. Educ.* 190, 21–30. doi: 10.1177/0022057410190001-205
- Marais, K. (2019). *A (bio)semiotic Theory of Translation: The Emergence of Social-Cultural Reality*. New York, NY: Routledge. doi: 10.4324/9781315142319
- Martinez, M. E. (1999). Cognition and the question of test item format. *Educ. Psychol.* 34, 207–218. doi: 10.1207/s15326985Sep3404\_2
- Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educ. Assess.* 14, 160–179. doi: 10.1080/10627190903422906
- Mayer, R. E., Heiser, J., and Lonn, S. (2001). Cognitive constraints on multimedia learning: when presenting more material results in less understanding. *J. Educ. Psychol.* 93, 187–198. doi: 10.1037/0022-0663.93.1.187
- Megone, M. E., Cai, J., Silver, E. A., and Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *Int. J. Educ. Res.* 21, 317–340. doi: 10.1016/s0883-0355(06)80022-4
- Miller, W. A. (1938). Reading with and without pictures. *Elem. Sch. J.* 38, 676–682. doi: 10.1086/462248
- Mislevy, R. J., and Haertel, G. D. (2007). Implications of evidence centered design for educational assessment. *Educ. Mea. Issues Pract.* 25, 6–20. doi: 10.1111/j.1745-3992.2006.00075.x
- Nagy, W. W., and Anderson, R. C. (1984). How many words are there in printed school english?. *Read. Res. Q.* 19, 304–330. doi: 10.2307/747823
- Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words. *Read. Foreign Lang.* 26, 1–16.
- National Assessment Governing Board (2017). *Mathematics Framework for the 2017 National Assessment of Educational Progress*. Washington, CG: National Assessment Governing Board.
- National Research Council (2006). “Systems for state science assessment. committee on test design for K-12 science achievement,” in *Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education*, eds M. R. Wilson and M. W. Bertenthal (Washington, DC: The National Academies Press).
- National Research Council (2012). *A Framework for Science K-12 Education: Practices, Cross-Cutting Concepts, and Core Ideas*. Washington, DC: The National Academy Press.
- National Research Council (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press, doi: 10.17226/18409
- NGSS Lead States (2013). *Next Generation Science Standards: For states, by States*. Washington, DC: The National Academies Press.
- Nisbett, R. (2003). *The Geography of Thought*. New York, NY: Free Press.
- Noble, S. U. (2018). *Algorithms of Oppression. How Search Engines Reinforce Racism*. New York, NY: New York University Press. doi: 10.2307/j.ctt1pwt9w5
- Noble, T., Sireci, S. G., Wells, C. S., Kachchaf, R. R., Rosebery, A. A., and Wang, Y. C. (2020). Targeted linguistic simplification of science test items for English learners. *Am. Educ. Res. J.* 57, 2175–2209. doi: 10.3102/0002831220905562
- Norman, D. (2013). *The Design of Everyday Things: Revised and Expanded Edition*. New York, NY: Basic Books. (accessed February 20, 2021).
- OECD (2019). *PISA Test*. Available at: <https://www.oecd.org/pisa/test/>. (accessed February 20, 2021).
- OECD (2020). *Access to Computers From Home (indicator)*. Paris: OECD, doi: 10.1787/a70b8a9f-en
- Packer, M., and Cole, M. (2020). “The institutional foundations of human evolution, ontogenesis, and learning,” in *Handbook of the Cultural Foundations of Learning*, eds N. S. Nasir, C. D. Lee, R. Pea, and M. M. de Royston (New York, NY: Routledge), 3–23. doi: 10.4324/9780203774977-2
- Pellegrino, J. W., Chudowsky, N., and Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.
- Pesce, S. (2018). “From comprehensive research to semiotic approaches to education: a subjective genealogy of educational semiotics,” in *Semiotic Theory of Learning*, eds A. Stables, W. Nöth, A. Olteanu, S. Pesce, and E. Pikkarainen (New York, NY: Routledge), 145–157. doi: 10.4324/9781315182438-11
- Pimm, C. (1987). *Speaking Mathematically: Communication in Mathematics Classrooms*. London: Routledge & Kegan Paul Ltd.
- Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., and Carey, T. (1994). *Human-Computer Interaction*. Workingham: Addison-Wesley.
- Priscari, A. A., and Danielson, J. (2017). Computer-based versus paper-based testing: investigating testing mode with cognitive load and scratch paper use. *Comput. Hum. Behav.* 77, 1–10. doi: 10.1016/j.chb.2017.07.044
- Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: a random effects synthesis of correlations. *J. Educ. Meas.* 40, 163–184. doi: 10.1111/j.1745-3984.2003.tb01102.x
- Rogoff, B. (1995). “Observing sociocultural activity on three planes: participatory appropriation, guided participation, and apprenticeship,” in *Sociocultural Studies of Mind*, eds J. V. Wertsch, P. del Río, and A. Alvarez (New York, NY: Cambridge University Press). doi: 10.1017/CBO9781139174299.008
- Ruiz-Primo, M. A., and Li, M. (2015). The relationship between item context characteristics and student performance: the case of the 2006 and 2009 PISA



- science items. *Teach. Coll. Record* 117, 1–36. Available online at: <https://www.tcrecord.org> (accessed March 21, 2021).
- Ruiz-Primo, M. A., Baxter, G. P., and Shavelson, R. J. (1993). On the stability of performance assessments. *J. Educ. Meas.* 30, 41–53. doi: 10.1111/j.1745-3984.1993.tb00421.x
- Ruiz-Primo, M. A., and Li, M. (2016). PISA science contextualized items: the link between the cognitive demands and context characteristics of the items. *RELIEVE* 22:art.M11. doi: 10.7203/relieve.22.1.8280
- Ruiz-Primo, M. A., Li, M., Minstrell, J., Kanopka, J., Hernandez, P., Dong, D., et al. (2019). Contextualized science assessments: addressing the use of information and generalization of inferences of students' performance. *Paper presented at the AERA Annual Meeting*, Toronto, ON: Canada.
- Saffer, D. (2014). *Microinteractions: Designing With Details*. Sebastopol, CA: O'Reilly.
- Sato, E., Rabinowitz, S., Gallagher, C., and Huang, C.-W. (2010). *Accommodations for English Language Learner Students: The Effect of Linguistic Modification of Math Test Item Sets*. Washington, DC: National Center for Education Evaluation and Regional Assistance. (NCEE Report 2009-4079).
- Schoenfeld, A. H. (2004). The math wars. *Educ. Policy* 18, 253–286. doi: 10.1177/0895904803260042
- Shade, C. (2017). *Mathematics Assessment in the Race to the Top era: An Exploratory Study of the Semiotic Resources in Large-Scale Assessment and Their use by Emergent and Non-Emergent Bilingual Students*. Doctoral dissertation, University of Colorado Boulder, Boulder, CO.
- Shafel, J., Belton-Kocher, E., Glasnapp, D., and Poggio, G. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educ. Assess.* 11, 105–126. doi: 10.1207/s15326977ea1102\_2
- Shavelson, R. J., Carey, N. B., and Webb, N. M. (1990). Indicators of science achievement: options for a powerful policy instrument. *Phi Delta Kappan* 71, 692–697.
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. P. (2019). Assessment of university students' critical thinking: next generation performance assessment. *In J. Test.* 19, 337–362. doi: 10.1080/15305058.2018.1543309
- Sireci, S. G., and Faulkner-Bond, M. (2015). Promoting validity in the assessment of ELs. *Rev. Res. Educ.* 39, 215–252. doi: 10.3102/0091732X14557003
- Sireci, S. G., and O'Riordan, M. (2020). "Comparability when assessing Individuals with disabilities," in *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*, eds A. I. Berman, E. H. Haertel, and J. W. Pellegrino (Washington, DC: National Academy of Education), 177–204.
- Smarter Balanced Assessment Consortium (2020). *Usability, Accessibility, and Accommodations Guidelines*. Available at: <https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf>. (accessed February 20, 2021).
- Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educ. Res.* 37, 189–199. doi: 10.3102/0013189x08319569
- Solano-Flores, G. (2011). "Assessing the cultural validity of assessment practices: an introduction," in *Cultural Validity in Assessment: Addressing Linguistic and Cultural Diversity*, eds M. D. R. Bastera, E. Trumbull, and G. Solano-Flores (New York, NY: Routledge), 3–21.
- Solano-Flores, G. (2012). *Translation Accommodations Framework for Testing English Language Learners in Mathematics*. Available at: <https://portal.smarterbalanced.org/library/en/translation-accommodations-framework-for-testing-english-language-learners-in-mathematics.pdf> (accessed September 18, 2012).
- Solano-Flores, G., Backhoff, E., and Contreras-Niño, L. A. (2009). Theory of test translation error. *Int. J. Test.* 9, 78–91. doi: 10.1080/15305050902880835
- Solano-Flores, G., Barnett-Clarke, C., and Kachchaf, R. (2013). Semiotic structure and meaning making: the performance of English language learners on mathematics tests. *Educ. Eval.* 18, 147–161. doi: 10.1080/10627197.2013.814515
- Solano-Flores, G., Chia, M. Y., and Kachchaf, R. (2019). Design and use of pop-up illustration glossaries as accessibility resources for second language learners in computer-administered tests in a large-scale assessment system. *Int. Mul. Res. J.* 13, 277–293. doi: 10.1080/19313152.2019.1611338
- Solano-Flores, G., and Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educ. Mea. Issues Pract.* 28, 9–18. doi: 10.1111/j.1745-3992.2009.00143.x
- Solano-Flores, G., and Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educ. Res. Eval.* 19, 245–263. doi: 10.1080/13803611.2013.767632
- Solano-Flores, G., and Nelson-Barber, S. (2001). On the cultural validity of science assessments. *J. Res. Sci. Teach.* 38, 553–573. doi: 10.1002/tea.1018
- Solano-Flores, G., Shade, C., and Chrzanowski, A. (2014a). *Item Accessibility and Language Variation Conceptual Framework*. Submitted to the Smarter Balanced Assessment Consortium. Available at: <https://portal.smarterbalanced.org/library/en/item-accessibility-and-language-variation-conceptual-framework.pdf>. (accessed February 20, 2021).
- Solano-Flores, G., Wang, C., Kachchaf, R., Soltero-Gonzalez, L., and Nguyen-Le, K. (2014b). Developing testing accommodations for English language learners: illustrations as visual supports for item accessibility. *Educ. Assess.* 19, 267–283. doi: 10.1080/10627197.2014.964116
- Solano-Flores, G., and Wang, C. (2015). Complexity of illustrations in PISA-2009 science items and its relationship to the performance of students from Shanghai-China, the United States, and Mexico. *Teach. Coll. Record* 117, 1–18.
- Solano-Flores, G., Wang, C., and Shade, C. (2016). International semiotics: item difficulty and the complexity of science item illustrations in the PISA-2009 international test comparison. *Int. J. Test.* 16, 205–219. doi: 10.1080/15305058.2015.1099534
- Stables, A. (2016). Edusemiotics as process semiotics: towards a new model of semiosis for teaching and learning. *Semiotica* 212, 45–58. doi: 10.1515/sem-2016-0126
- Steele, C. M., and Aronson, J. (1998). "Stereotype threat and the test performance of academically successful African Americans," in *The Black-White Test Score Gap*, eds C. Jencks and M. Phillips (Washington, DC: Brookings Institution Press), 401–427.
- Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285. doi: 10.1207/s15516709cog1202\_4
- Sweller, J., van Merriënboer, J. J., and Paas, F. G. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, 251–296. doi: 10.1023/A:1022193728205
- Thurlow, M. L., and Kopriva, R. J. (2015). Advancing accessibility and accommodations in content assessments for students with disability and English learners. *Rev. Res. Educ.* 39, 331–369. doi: 10.3102/0091732X14556076
- Turkan, S., Lopez, A., Lawless, R., and Tolentino, F. (2019). Using pictorial glossaries as an accommodation for English learners: an exploratory study. *Educ. Assess.* 24, 235–265. doi: 10.1080/10627197.2019.1615371
- Uesaka, Y., and Manalo, E. (2012). Task-related factors that influence the spontaneous use of diagrams in math word problems. *Appl. Cogn. Psychol.* 26, 251–260. doi: 10.1002/acp.1816
- van der Linden, W. J. (2016). "Optimal test assembly," in *Handbook of Test Development*, 2nd Edn, eds S. Lane, M. R. Raymond, and T. M. Haladyna (New York, NY: Routledge), 507–530.
- van Leeuwen, T. (2004). *Introducing Social Semiotics*. New York, NY: Routledge. doi: 10.4324/9780203647028
- Wang, C. (2012). *The Use of Illustrations in Large-Scale Science Assessment: A Comparative Study*. Doctoral dissertation, University of Colorado Boulder, Boulder, CO.
- Wang, C., Chia, M., Kachchaf, R., and Solano-Flores, G. (2012). "Item illustration complexity and the performance of English language learners in a science test," in *Paper Presented at the Annual Conference of the American Educational Research Association*, Vancouver.
- Washington, W. N., and Godfrey, R. R. (1974). The effectiveness of illustrated items. *J. Educ. Meas.* 11, 121–124. doi: 10.1111/j.1745-3984.1974.tb00981.x
- Wendler, C. L. W., and Walker, M. E. (2006). "Practical issues in designing and maintaining multiple test forms for large-scale programs," in *Handbook of Test Development*, eds S. M. Downing and T. M. Haladyna (New York, NY: Lawrence Erlbaum Associates, Publishers), 445–467.
- Winter, P., Kopriva, R. J., Chen, S., and Emick, J. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: results from a large-scale cognitive lab. *Learn. Individ. Differ.* 16, 267–276. doi: 10.1016/j.lindif.2007.01.001

- Zhao, X. (2018). *Test Translation Review Procedures in International Large-Scale Assessment: Sensitivity to Culture and Society*. Doctoral dissertation, University of Colorado Boulder, Boulder, CO.
- Zhao, X., and Solano-Flores, G. (2021). Testing across languages in international comparisons: cultural adaptation of consensus-based test translation review procedures. *J. Multiling. Multicult. Dev.* doi: 10.1080/01434632.2020.1852242
- Zlatkin-Troitchanskaia, O., and Shavelson, R. J. (2019). Editorial: advantages and challenges of performance assessment of student learning in higher education. *Br. J. Educ. Psychol.* 89, 413–415. doi: 10.1111/bjep.12314

**Conflict of Interest:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2021 Solano-Flores. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*



# Patterns of Domain-Specific Learning Among Medical Undergraduate Students in Relation to Confidence in Their Physiology Knowledge: Insights From a Pre-post Study

Jochen Roeper<sup>1</sup>, Jasmin Reichert-Schlax<sup>2\*</sup>, Olga Zlatkin-Troitschanskaia<sup>2</sup>, Verena Klose<sup>1</sup>, Maruschka Weber<sup>1</sup> and Marie-Theres Nagel<sup>2</sup>

<sup>1</sup> Department of Neurophysiology, University Hospital Frankfurt, Frankfurt, Germany, <sup>2</sup> Department of Business and Economics Education, Johannes Gutenberg University Mainz, Mainz, Germany

## OPEN ACCESS

### Edited by:

Tom Rosman,  
Leibniz Center for Psychological  
Information and Documentation  
(ZPID), Germany

### Reviewed by:

Jonna M. Kulikowich,  
The Pennsylvania State University  
(PSU), United States  
Pablo Ruisoto,  
Public University of Navarre, Spain

### \*Correspondence:

Jasmin Reichert-Schlax  
jaschlax@uni-mainz.de

### Specialty section:

This article was submitted to  
Educational Psychology,  
a section of the journal  
Frontiers in Psychology

**Received:** 14 May 2020

**Accepted:** 13 December 2021

**Published:** 10 February 2022

### Citation:

Roeper J, Reichert-Schlax J,  
Zlatkin-Troitschanskaia O, Klose V,  
Weber M and Nagel M-T (2022)  
Patterns of Domain-Specific Learning  
Among Medical Undergraduate  
Students in Relation to Confidence  
in Their Physiology Knowledge:  
Insights From a Pre-post Study.  
Front. Psychol. 12:562211.  
doi: 10.3389/fpsyg.2021.562211

**Research Focus:** The promotion of domain-specific knowledge is a central goal of higher education and, in the field of medicine, it is particularly essential to promote global health. Domain-specific knowledge on its own is not exhaustive; confidence regarding the factual truth of this knowledge content is also required. An increase in both knowledge and confidence is considered a necessary prerequisite for making professional decisions in the clinical context. Especially the knowledge of human physiology is fundamental and simultaneously critical to medical decision-making. However, numerous studies have shown difficulties in understanding and misconceptions in this area of knowledge. Therefore, we investigate (i) how preclinical medical students acquire knowledge in physiology over the course of their studies and simultaneously gain confidence in the correctness of this knowledge as well as (ii) the interrelations between these variables, and (iii) how they affect the *development* of domain-specific knowledge.

**Method:** In a pre-post study, 169 medical students' development of physiology knowledge and their confidence related to this knowledge were assessed *via* paper-pencil questionnaires before and after attending physiology seminars for one semester. Data from a longitudinal sample of  $n = 97$  students were analyzed using mean comparisons, regression analyses, and latent class analyses (LCAs). In addition, four types of item responses were formed based on confidence and correctness in the knowledge test.

**Results:** We found a significant and large increase in the students' physiology knowledge, with task-related confidence being the strongest predictor (apart from learning motivation). Moreover, a significantly higher level of confidence at t2 was confirmed, with the level of prior confidence being a strong predictor (apart from knowledge at t2). Furthermore, based on the students' development of knowledge and

confidence levels between measurement points, three empirically distinct groups were distinguished: knowledge gainers, confidence gainers, and overall gainers. The students whose confidence in incorrect knowledge increased constituted one particularly striking group. Therefore, the training of both knowledge and the ability to critically reflect on one's knowledge and skills as well as an assessment of their development in education is required, especially in professions such as medicine, where knowledge-based decisions made with confidence are of vital importance.

**Keywords:** domain learning, physiology knowledge, knowledge development, confidence testing, learning profiles, medical education, longitudinal study

## RELEVANCE AND RESEARCH QUESTIONS

Working in the field of medicine is all about making well-informed decisions and the focus in recent years has shifted—at least in theory—from eminence- to evidence-based frameworks (Cate et al., 2018; Custers and Cate, 2018). In the last 25 years, the principles of evidence-based medicine have been integrated into many core curricula of medical education and also drive current initiatives to improve these curricula (e.g., German National Competence-based Learning Objectives Catalog for Undergraduate Medical Education; Fritze et al., 2017a,b). While there is an agreement about the need for further solid scientific groundwork in the domain of medicine in general, the implications for the individual medical decision makers and with regard to the domain-specific knowledge, understanding, and skills to be acquired during medical training are often debated (Kelly et al., 2015). Previous research describes several personal variables, such as confidence and prior knowledge, which influence the processing and acquisition of domain-specific knowledge (Posner et al., 1982; Cordova et al., 2014). While medical students themselves recognize the importance of science-based decision-making, they often do not feel confident in their ability to do so (Pruskil et al., 2009) although there is also evidence of overconfidence effects (Borracci and Arribalzaga, 2018). This, in turn, may be due to deficits in their ability to evaluate their own knowledge and recognize deficiencies therein (Kruger and Dunning, 1999; Eva et al., 2004). With regard to the learning progressions of medical students, making professional decisions and solving problems in clinical contexts require not only the acquisition of knowledge but also an increase in confidence in their medical knowledge (Khan et al., 2001). However, in multiple-choice (MC) tests, which are commonly used to assess knowledge in medical studies as well as in the written final examinations that preclinical and clinical medical students have to pass, confidence in one's own decisions is not (routinely) assessed. Thus, participants may give correct answers despite having low confidence in their knowledge or by simply guessing (Walstad et al., 2018; Roeper et al., 2020), which may indicate that they are not ready to assume the responsibilities of a medical profession.

Physiology is one of the most basic topics in medical higher education and the basis for all applied fields of medicine. Nathial (2020, p. 2) describes anatomy and physiology as

medical fields associated with the structures (e.g., the structure of red blood cells, related to anatomy) and functions of the human body; physiology describes and investigates the functions and cooperation of the different anatomical structures (e.g., transport of oxygen *via* red blood cells, Nathial, 2020, p. 2; for another example, see Section “Test Instruments”). The particular complexity of physiological knowledge is mainly caused by many interrelated levels that interact dynamically with each other (Hmelo-Silver et al., 2007). An understanding of anatomy and physiology is, therefore, seen as fundamental for working in medicine (Nathial, 2020, p. 2). There is evidence that experts and novices differ mainly in their understanding of causal behaviors and functions and less in their knowledge of structures (Hmelo-Silver et al., 2007). Consequently, it is a matter of learning about complex systems (Burggren and Monticino, 2005), which—as Hmelo-Silver and Azevedo (2006) argue—is difficult for various reasons: the sheer volume of relevant topics, the lack of direct experience of these concepts, the violation of our intuitive assumptions and preconceptions, and thus requires cognitive but also metacognitive and affective resources.

The differences between anatomy and physiology are not only reflected in their degree of difficulty of understanding but also result in difficulties in measuring the understanding of physiological functions. In this study, it is not possible to simply ask for the names of certain structures; rather, e.g., the use of MC tests requires particularly careful development of adequate distractors that also address typical misconceptions (Roeper et al., 2020). Taking into account the specific features of the field of human physiology (functions and complex interrelationships and preconceptions), this results in a particularly challenging valid and reliable assessment of students' knowledge levels.

Especially in German-speaking countries, numerous studies have shown that difficulties in understanding and misconceptions are prevalent in this content area. Physiology is an important part of the first stage of studies and the first of two nationwide exams in Germany, with an average of 11.5% of students failing the most recent exam in autumn 2020. In this exam during autumn 2020, for example, the physiology section of the test was the part with the lowest average solution frequency (e.g., IMPP, 2020). As an example of a misconception, due to pre-academic conceptions (beliefs) that the brain is some kind of “computer,” medical students often fail to see the fundamental flaw of this metaphor and consequently struggle to understand brain functions correctly.



While software and hardware can be easily distinguished in computers, brains function as hybrids (wetware), with software (physiology) constantly altering (plasticity) hardware (anatomy). This misconception is a significant roadblock in understanding how brains function in a healthy or diseased state.

Based on the existing research on “confidence testing” and “knowledge assessment” (refer to Section “Conceptual-theoretical Background and Research Hypotheses”), the issues of (i) how preclinical medical students acquire knowledge in physiology over their course of study and simultaneously gain confidence in the correctness of their knowledge as well as (ii) the interrelations between these processes, and (iii) how they affect the *development* of domain-specific knowledge are investigated in this pre–post study. More specifically, we empirically identified certain learning profiles depending on the combination of students’ knowledge and confidence levels at both measurements, which significantly differ in terms of the considered personal characteristics (e.g., prior education). Our findings provide some insights into this complex and reciprocal relationship between knowledge development and confidence, which can lead to first suggestions on how to train preclinical medicine students to reflect on their reasoning and to initial ideas on how to improve their decision-making and problem-solving skills.

## CONCEPTUAL-THEORETICAL BACKGROUND AND RESEARCH HYPOTHESES

The acquisition of professional knowledge and skills is the central goal of higher education. Miller (1990) developed a pyramid model focusing on the basics of medical actions, with factual knowledge serving as the basis for practical knowledge and the performance of medical actions. Patel et al. (1999) distinguish between the two basic types of knowledge, which are also used in the medical field: the verbalized knowledge of facts and concepts, which are learned explicitly, and more implicit, procedural knowledge, which underlies heuristics and tends to be acquired through practical experience in hospitals or similar settings. In the following, we briefly refer to the theoretical considerations that could underlie knowledge and confidence.

Although domain-specific knowledge is of critical importance for understanding the scientific rationales of practical work in medicine (Wijnen-Meijer et al., 2013a,b), a recent meta-study (Cate et al., 2018) indicated that the issue of competence development and knowledge acquisition in medical education has not been systematically addressed in research so far. While medical education relies on a variety of admission tests (e.g., Spiel and Schober, 2018) as well as multimedia-based performance assessments, research still suggests a substantial deficit of studies that measure students’ domain-specific knowledge acquisition throughout the medical school with valid and reliable test instruments (Cox and Irby, 2007; Prediger et al., 2020).

A majority of recent studies have focused on more general competence facets such as medical problem-solving and clinical reasoning (e.g., Custers, 2018). This research is often based on the “dual process” theory (heuristic and analytic processes; Evans,

1984; refer also to Chaiken (1980) heuristic systematic model), which particularly emphasizes the role of “intuitive cognition” (Patterson and Eggleston, 2017). When solving a domain-specific task and, for example, choosing among several response options, this kind of reasoning can be associated with Kahneman (2011) “System 2,” i.e., a “subjective feeling of confidence” (Sanders et al., 2016). In this context, some studies investigated to what extent students have confidence in their acquired knowledge or, in contrast, base their domain-specific decision-making or problem-solving on “strategically selected options” (e.g., straight-out guessing rather than using their knowledge; Weier et al., 2017; Lubarsky et al., 2018). More specifically, Gardner-Medwin (1995) considers not only the factual correctness of an answer but also the participants’ confidence in their answers to be central sources of information when analyzing knowledge acquisition/development. Overall, this research indicates that the awareness of confidence levels in relation to levels of domain-specific knowledge is important in domain-specific learning and expertise development.

According to the “cognitive continuum theory” (CCT), (i.e., an extension of the dual process model, Hammond et al., 1987; Hammond, 1996), when working on a task, solving and decision-making processes can be based on prior knowledge and/or heuristic approaches and subjective “good feelings.” Based on this framework, it can be assumed that a wrong item response can also be the result of incorrect knowledge or a misconception or even an “intuition-based” decision that might also be exhibited, e.g., by guessing on individual items (refer to for modeling guessing effects in MC tests, e.g., Walstad et al., 2018). For instance, in coglabs studies (Brückner, 2017) with retrospective think-aloud interviews, students often reported making a decision in favor of a certain response option because of their “good feeling.” This can be clearly distinguished from guessing because of a lack of relevant knowledge or the selection of an incorrect answer because of a misconception. Based on these studies, it can be argued that an incorrect item response can also result from a misconception presented in an individual distractor.

Based on research on competence testing (e.g., Bruno, 1993; Davies, 2006), confidence in the correctness of one’s knowledge, and (domain-specific) task-related confidence, can be considered as an appropriate indicator of the extent to which a student’s response is based on knowledge vs. strategic guessing. Confidence testing is based on the assumption that the level of confidence in the response to an item in a knowledge test can be used to more precisely categorize individuals according to their abilities, indicating a closer reciprocal link between knowledge level and confidence (Kolbitsch et al., 2008). A recent meta-study indicates (Stankov, 2013), besides individual variables such as self-concept and self-efficacy, that confidence is found to be one of the best predictors of academic achievement (measured using domain-specific knowledge tests), with the highest predictive validity compared to the other “self-beliefs.” This research indicates a major influence of confidence on the development of knowledge.

In the context of conceptual change, this relationship was examined in a pre–post study by Cordova et al. (2014): Based on students’ prior knowledge and confidence in their knowledge (and also considering further characteristics such as self-efficacy

and interest), three empirically distinct learning profiles (“low”: low confidence and prior knowledge; “high”: high confidence and prior knowledge; and “mixed”: high confidence and low prior knowledge) were identified, which may significantly impact the conceptual learning of college students. This is further supported by the findings of Alexander and Murphy (1998), who also identified three distinct student profiles, and changes in these profiles over an academic term as well as their impact on domain learning (Alexander et al., 1997).

Based on a few existing studies with a particular focus on medical students (e.g., Khan et al., 2001; Fitzgerald et al., 2003), both a relationship between students’ performance in knowledge tests and their confidence in their response as well as the development of this relationship over the course of studies can be expected. Since the available findings are partly non-conclusive, further empirical investigations are required to analyze the interaction of confidence and learning and decision-making of medical students in preclinical courses. As some studies indicate that confidence in one’s erroneous knowledge, e.g., due to a misconception, can also increase throughout a course of study (Fitzgerald et al., 2003; Hall et al., 2007), more longitudinal research on the development of knowledge and confidence is required.

In this context, Khan et al. (2001) studied the relationship between the correctness of knowledge, the confidence in correct knowledge, and the use of correct knowledge in decision-making: “By ignoring this aspect of learning, when a student correctly responds to a question, it is not possible to determine whether the correctly identified knowledge is usable for decision making or not” (pp.160–161). The research group emphasizes that plain ignorance or misinformation can be responsible for an incorrect answer (or incorrect knowledge), whereby they see misinformation as even worse: “Misinformation is particularly dangerous because the student strongly believes that the wrong answer is correct” (pp.160–161). This finding indicates the particular relevance of the level of students’ confidence in the correctness of their knowledge for learning in a medical study domain and for uncovering possible misconceptions.

Building on the existing theoretical and methodological framework, we defined the following five hypotheses (*H*) to be tested in this study.

With regard particularly to the theory of domain learning (Alexander et al., 1995, 2018) as well as the curricular validity of the knowledge test used in this study, which comprises the contents and concepts taught in the physiology lecture and seminar series, we assume that the 1st year students have only little relevant previous knowledge from high school and vocational medical training at t1 and that the level of their knowledge increases over the course of the semester, manifesting in higher test scores at t2.

**H1:** The students’ level of domain-specific knowledge is significantly lower before (t1) than after actively participating in the physiology seminar series (t2).

Recent research on knowledge acquisition and its determinants suggested several individual student characteristics

that may significantly contribute to differences in students’ knowledge levels and their development. Cognitive variables, such as general *cognitive ability* (Brandt et al., 2019) and *prior knowledge* (Shing and Brod, 2016), as well as psychological variables, such as *motivational factors* (Rotgans and Schmidt, 2017), *personal characteristics* (as, e.g., gender, e.g., Haq et al., 2005; Firth-Cozens, 2008), and *confidence* (Rudolph et al., 2017), were repeatedly identified as significant predictors. Therefore, as *H2*, we assume that these personal characteristics significantly contribute to explaining students’ domain-specific knowledge levels at t1 and t2 and their development (difference score), as:

**H2a:** Indicators of prior knowledge [advanced school courses or vocational training] predict students’ knowledge levels.

**H2b:** Indicators of general cognitive ability (intelligence test score and grade of university entrance qualification (UEQ)) predict students’ knowledge levels.

**H2c:** (Intrinsic) learning motivation predicts students’ knowledge levels.

**H2d:** Task-related confidence predicts students’ knowledge levels.

**H2e:** Socio-biographical characteristics, e.g., gender, predict knowledge levels.

In particular, in many studies, female students show both a significantly lower degree of knowledge and confidence in MC tests (Parker, 2006; Owen, 2012), which was also reflected in a different response behavior than male students (e.g., more missing values; Walstad et al., 2018).

Task-related confidence has already been determined as a significant influencing factor for knowledge test values (Parker, 2006; Kleitman et al., 2012). Based on these studies, we further assume that task-related confidence increases with increasing knowledge levels.

**H3:** The average level of students’ confidence in their (correct as well as incorrect) responses is higher at t2 than at t1.

In addition, with reference to the reported findings, we expect that confidence is also influenced by personal characteristics (e.g., for prior knowledge, refer to Dinsmore and Parkinson, 2013), leading to *H4*,

**H4:** Personal characteristics significantly contribute to the explanation of students’ confidence levels at t1 and t2 and their development (difference score).

Therefore, the assumed relationships in *H4a–e* are formulated and tested empirically using the same strategies employed in *H2a–e* (for details, refer to Section “Statistical Methods”).

Alexander (2013a) emphasize the importance of examining not only the participants’ confidence in their responses but also and especially the issue of calibration, i.e., the relationship between their self-estimation and performance (for deeper insights, refer to Alexander, 2013b). As existing studies indicate a mutual influence between knowledge and confidence as well as corresponding learning profiles, we assume in *H5* that:

**H5a:** Students' domain-specific knowledge is positively related to their confidence when answering the test questions both at t1 and t2 and between the difference scores ( $t2 - t1$ ) in both knowledge and confidence.

**H5b:** On the basis of correlations between the change in domain-specific knowledge levels and task-related confidence, students can be empirically categorized into three groups, with specific personal attributes (such as learning motivation, general cognitive ability, and prior education) that characterize these groups.

## MATERIALS AND METHODS

### Study Design

The analyses were based on the data from a study supported by the Rhine-Main-Universities (RMU) fund, in which overall 169 students of medicine who actively participated in a 2nd semester seminar series on physiology at a university in the Rhine-Main region, Germany, were tested at the beginning (t1) and end (t2) of the seminar series (2019 summer term to 2019/2020 winter term). In total, 137 students participated at t1, 135 at t2. Out of these, 97 students participated in both measurements, whereby this longitudinal sample forms the basis of the analyses presented in this study. The surveys took place at the beginning of the mandatory seminars and under controlled test conditions in a group setting. For this purpose, test administrators were given specific training. According to the university's curriculum, the seminar series is intended for the 2nd to 3rd semester medical students. It was not possible to implement a comparable control group in the field survey as all students at this stage of medical studies are required to participate in the seminars (for limitations, refer to Section "Limitations"). Though participation in the study was voluntary, all students who attended the seminar series took part. As an incentive for test motivation, the students were given an opportunity to view their test results (on sum score level, not item level) online after both measurements and to obtain individual feedback on their knowledge development<sup>1</sup>.

### Test Instruments

Paper-pencil questionnaires were used at both measurements. In t1, in addition to their domain-specific knowledge in physiology, participants' socio-demographic data such as gender, native language, and previous education as well as further information such as general cognitive ability were assessed. In t2, the identical knowledge test was used again, and further characteristics relevant to the learning process, including motivation, were assessed. Completing the survey took on average 30 min in t1 and 25 min in t2.

<sup>1</sup> At each of the 3-h seminars, up to six physiology topics were discussed in a group comprising 20 students and a lecturer. Each topic was introduced by one of the students for about 10 min. The lecturer evaluated the quality of the presentation (satisfactory/non-satisfactory). To successfully complete this seminar series, which also includes two 30-question MC examinations, students were not allowed to have more than one non-satisfactory performance.

### Physiology Knowledge Test in t1 and t2

Knowledge of physiology was assessed using 12 newly developed single-choice items with five answer options each, a test format that students are familiar with from their medical studies (an example item: "How is water predominantly transported from the extracellular space through cell membranes? (A) Active transport *via* solvent drag, (B) carrier-mediated symport with chloride, (C) Primary active transport, (D) along an osmotic gradient *via* connexins, and (E) along osmotic gradients *via* aquaporins." The single-choice items focus on assessing primarily (declarative) knowledge about basic concepts of physiology. Due to the design of the distractors, however, a deeper understanding of the content is required to select the correct answer from the given answer options<sup>2</sup>(for further details on the assessment approach underlying the test, see Roeper et al., 2020).

To prevent the students from cheating on the test, two versions of the questionnaire were created, with the same questions but in reverse order. The same test was used at t1 and t2. Correct answers were scored with one point, while missing answers (as an indicator of non-knowledge, Baker and Kim, 2004) and incorrect answers were each scored with 0 points, i.e., a maximum of 12 points could be achieved. Dichotomous coding is a strict variant of scoring that is commonly used for MC items (e.g., Kulhavy and Anderson, 1972; Andrich and Kreiner, 2010; Lee et al., 2011; Durning et al., 2015). A total score was calculated for each test taker from the 12 responses. Due to the limited test time in this field study, a relatively small number of items was used. To cover as many basic physiological concepts as possible, the items covered a variety of topics. This led to low reliability of 0.511 (for limitations, refer to Section "Limitations"). To determine the construct validity of the test, we conducted a confirmatory factor analysis (CFA). The CFA results are consistent with the assumption of the test developers that the 12 test items cover a comprehensive construct with many facets, which are all based on one common latent factor, i.e., knowledge of physiology. The model with an assumed one-dimensional solution shows a satisfactory fit with respect to most CFA fit indices at both measurements (t1 and t2); only the standardized root mean residual (SRMR) is not optimal (Table 1). Exploratory factor analyses also indicate the one-dimensional model. With regard to the order of the questions in the test questionnaires, no significant difference in the total score could be determined at both measurements (t1:  $p = 0.573$  and t2:  $p = 0.847$ ). There are also no significant differences in the item difficulties of the 12

<sup>2</sup>Though the measured contents of physiology also involved procedural and conditional knowledge facets, the items in this test are not designed to explicitly assess them in a valid and reliable way.

**TABLE 1 |** Fit indices of confirmatory factor analysis (CFA) with a single-factor solution.

Model	$\chi^2$	df	$\chi^2/df$	RMSEA	CFI	SRMR
t1	50.777	54	0.940	< 0.001	1.000	0.138
t2	53.701	54	0.994	< 0.001	1.000	0.124



items between t1 and t2 ( $\chi^2 < 0.01 - \chi^2 = 2.29, p = 0.131 - p = 0.993$ ).

### Confidence in Responses to the Domain-Specific Test Items

To measure task-related confidence, after each task of the domain-specific test, a 4-level Likert scale was used to ask participants to what extent they were certain that their solution was correct (exact wording (translated): To what percentage are you convinced that the answer you have given is correct?). The test participants had to choose one of the four options: 0 - 25 - 75 - 100%. A mean value was calculated for all items at both measurements and used as a sum score. With a Cronbach's  $\alpha$  of 0.775, the internal reliability of the construct "confidence" can be considered acceptable.

### Indicators of General Cognitive Ability

Two variables were used as indicators of general cognitive ability. As a common and easily measured indicator, the average grade of students' UEQ was recorded. In addition, to indicate general cognitive ability, the scale "Choosing figures" of the Intelligence Structure Test (IST-2000 R, Liepmann et al., 2007) was used as an objective measure in the questionnaire. Due to the assumed stability of this construct (and the restricted test time), intelligence was only assessed at t1. The IST comprised 20 single-choice items for figural reasoning, whereby students had to work out which of the five given figures could be created by piecing together ten fragments. The test time was limited to 7 min. Correct responses were scored with one point, while missing answers (as an indicator of non-knowledge) and incorrect responses were each scored with zero points so that a maximum of 20 points could be achieved. A sum score was calculated for statistical analysis. The fit indices of the CFA indicated that the model with an assumed single factor solution fits the measured data satisfactorily regarding nearly all fit indices (root mean square error of approximation (RMSEA) = 0.041, weighted root mean square residual (WRMR) = 0.958, comparative fit index (CFI) = 0.784).

### Previous Learning Opportunities

With regard to pre-university learning opportunities, the participants were asked whether they had completed advanced courses for several relevant subjects in high school (in biology, chemistry, physics, and mathematics; (t1); in this study, multiple answers were possible. The participants were also asked about the completion of any medical vocational training (t2).

### Learning Motivation

Intrinsic learning motivation was measured with a short scale (adapted from the scale Vallerand et al., 1992; Schiefele et al., 1993), which consisted of four items (e.g., "I study for my (degree) course because I enjoy working with the content") and could be answered on a 6-point scale from 1 "applies fully" to 6 "does not apply at all" resulting in a score with an inverted scale, where lower scores signify higher motivation. The reliability analyses showed a Cronbach's  $\alpha$  of 0.833. The fit indices of the CFA show that a one-dimensional model is quite suitable for the measured data (e.g., RMSEA = 0.087, SRMR = 0.023, CFI = 0.990).

## Sample Description

The longitudinal sample and the subsamples t1 and t2 are described in Table 2.

## Statistical Methods

The matching of the longitudinal data was performed using R, packages *dplyr* and *haven* (R Core Team, 2018), based on a pseudonymized, unchangeable six-character code (e.g., including the second letter of the mother's first name) generated by the test takers. In addition to descriptive and factor-analytical analyses to investigate the internal test structure and its dimensionality (Cronbach's  $\alpha$ ; exploratory and CFA), regression analyses were also conducted to investigate the research hypotheses ( $H$ ) 2 and 4. In addition to the indicators described in Section "Test Instruments," the regression analyses always include language, age, and gender as control variables. In addition, differences in mean values were tested for significance using *t*-tests and ANOVAs ( $H1$  and  $H3$ ). Due to a relatively small sample and the correspondingly small subsamples, the effect sizes were also reported in addition to the significance levels (Cohen's  $d$  for two groups, Omega<sup>2</sup> [ $\omega^2$ ] for more than two groups, and Cramer's  $V$  for frequency distributions; refer to Cohen, 1988; Ellis, 2010; Grissom and Kim, 2012; Field, 2013).

The measurement invariance analyses for the construct "knowledge" conducted at t1 and t2 indicate a scalar invariance (WLSMV estimator for categorical variables, e.g., RMSEA = 0.036, WRMR = 1.042), whereas a metric invariance between t1 and t2 can be determined for the construct "confidence" (MLMV estimator for continuous variables, e.g., RMSEA = 0.041, SRMR = 0.093, CFI = 0.802).

**TABLE 2 |** Descriptive statistics for the samples at t1, t2, and for the matched sample.

Variables	t1 N = 134	t2 N = 132	Match N = 97
Gender, male, n (%)	37 (27.61%)	40 (30.30%)	29 (29.90%)
Preferred communication language, German, n (%)	116 (86.57%)	n/a	87 (89.69%)
Age, mean $\pm$ SD	22.01 $\pm$ 3.642	22.01 $\pm$ 3.642	22.01 $\pm$ 3.897
UEQ grade <sup>1</sup> , mean $\pm$ SD	n/a	1.45 $\pm$ 0.541	1.44 $\pm$ 0.513
IST sum score, mean $\pm$ SD	12.46 $\pm$ 3.46	n/a	12.91 $\pm$ 3.324
learning motivation <sup>2</sup> , mean $\pm$ SD	n/a	2.08 $\pm$ 0.683	2.11 $\pm$ 0.684
Educational background:			
Advanced sciences course at school, n (%)			
None	27 (40.70%)	30 (22.73%)	17 (27.87%)
Biology	35 (40.70%)	70 (53.03%)	26 (42.62%)
Chemistry	2 (2.33%)	32 (24.24%)	2 (3.28%)
Physics	1 (1.16%)	15 (11.36%)	1 (1.64%)
Mathematics	17 (19.77%)	62 (46.97%)	11 (18.03%)
More than one course	4 (4.65%)	n/a	4 (6.56%)
Medicine-related vocational training, n (%)	n/a	16 (12.12%)	12 (12.37%)

<sup>1</sup> UEQ = university entrance qualification. In Germany, lower numbers indicate better grades (1–6). <sup>2</sup> Inverted scale, lower numbers indicate higher motivation scores.



Regarding *H5a*, based on the students' confidence levels and the correctness of the item responses, in the first step, four different combinations of these variables were distinguished. Therefore, the confidence items were split into low and high, with "low" being 0% or 25% and "high" being 75% or 100%.

To test *H5b*, latent class analyses (LCAs) were calculated in the following step, indicating a model with a three-factor solution according to the Akaike information criterion (AIC) and Bayesian information criterion (BIC) fit indices (refer to Section "Results"). The fit values for the three- and the two-factor model are fairly similar. The three-factor model was chosen, which is also in line with prior research (e.g., Cordova et al., 2014).

The analyses were performed using Stata 15 (Stata Corp, 2017), and MPlus Version 7 was used for the latent analyses (Muthén and Muthén, 1998–2011). The application prerequisites for the applied methods were also checked and confirmed.

## RESULTS

### Level of Knowledge in Physiology Before (t1) and After (t2) Attending the Seminar Series (H1)

Regarding *H1*, students' levels of domain-specific knowledge before (t1) and after (t2) attending the physiology seminar series were examined. As expected, the knowledge score differs significantly between the two measurements ( $t(93) = -16.211$ ,  $p < 0.0001$ ,  $\Delta 3.45$ ,  $corr = 0.380$ , Cohen's  $d = 2.07$ ): While at t1 the students had an average score of  $4.39 (\pm 1.497)$ , the average score at t2 was  $7.84 (\pm 2.096)$  indicating a significant increase in knowledge after participating in the seminar series.

### Influencing Factors on Students' Knowledge Level and Its Development (H2)

When analyzing the students' domain-specific knowledge level at t2 as a dependent variable (*H2*), regression analyses showed that confidence in the given (correct or incorrect) response at t2 ( $\beta = 0.554$ ,  $p < 0.001$ ) was the most significant predictor in this model (*H2d*). Learning motivation was also a significant predictor of domain-specific knowledge level at t2 (*H2c*,  $\beta = -0.200$ ,  $p = 0.026$ ), whereas the two indicators of general cognitive ability (*H2b*) and gender (*H2e*) were not. In addition, the knowledge score at t1 ( $\beta = 0.199$ ,  $p = 0.024$ ) was less predictive than the level of confidence at t2. Among the assessed indicators for prior knowledge, attendance of advanced courses in biology, math, and physics in high school as well as students' socio-biographical characteristics such as gender did not contribute to the prediction (*H2a*). Overall, the regression model achieved an  $R^2$  of 50.96% (adj.  $R^2 = 40.60\%$ ).

When analyzing the difference score between the sum scores in physiology knowledge at t1 and t2 as a dependent variable ( $R^2 = 30.58\%$ , adj.  $R^2 = 17.09\%$ ), the significance of the predictor confidence (in correct as well as incorrect responses) at t2 ( $\beta = 0.414$ ,  $p = 0.001$ ) and the predictor learning motivation ( $\beta = -0.235$ ,  $p = 0.031$ ; inverted scale) was confirmed. Even when confidence was included in the model as a difference score from

t1 to t2, it showed a statistically significant effect ( $\beta = 0.235$ ,  $p = 0.038$ ), while the indicators of prior education and gender were not significant.

### Level of Confidence in Their Test Response Before (t1) and After (t2) Attending the Seminar Series (H3)

There was a positive development from an average confidence level of  $1.92 (\pm 0.404)$  at t1 to an average level of  $2.74 (\pm 0.452)$  at t2. This difference also became significant with a strong effect size ( $t(86) = -16.867$ ,  $p < 0.001$ ,  $\Delta 0.82$ ,  $r = 0.454$ , Cohen's  $d = 1.94$ ).

### Influencing Factors on Students' Confidence in Their Test Response (H4)

When analyzing the determinants of confidence at t2 (*H4*), the regression model, taking into account the knowledge scores, socio-demographic characteristics, indicators of general cognitive ability, and indicators of prior knowledge, shows that confidence at t1 was a significant predictor ( $\beta = 0.351$ ,  $p < 0.001$ ). More significant than confidence at t1 was the knowledge score at t2 with a  $\beta$  of  $0.537$  ( $p < 0.001$ ; *H4d*). Besides gender ( $\beta = 0.193$ ,  $p < 0.033$ ; *H4e*), other covariates were not significant (*H4a–c*). However, this regression model already achieved an  $R^2$  of 56% (adj.  $R^2 = 46.75\%$ ).

If the development of confidence was focused as a difference score, the knowledge score at t2 remained the only significant predictor ( $\beta = 0.434$ ,  $p < 0.001$ ,  $R^2 = 32.10\%$ , adj.  $R^2 = 18.90\%$ ).

### Relation Between Students' Domain-Specific Knowledge (Changes) and Their Confidence at t1 and t2 (H5a)

When analyzing and comparing the correlations between the mean confidence level and the sum score in the knowledge test at t1 and t2, it became evident that at t1 the correlation of  $r = -0.039$  was not significant, while it was significantly higher at t2 with a correlation of  $r = 0.529$ . At  $r = 0.208$ , the change in the confidence level was also related to the change in domain-specific knowledge. This is in line with the results of the regression analyses for *H2*, which showed that the confidence level at t2 was of particular importance, while the confidence level at t1 was negligible ( $\beta = 0.191$ ,  $p = 0.070$ ). Similar results were also found for the difference score of the knowledge test as a dependent variable (confidence level t1:  $\beta = 0.038$ ,  $p = 0.754$ ).

Furthermore, based on the students' confidence levels and the correctness of the item responses in each measurement point, four combinations were distinguished: (1) *confident and correct*, (2) *not confident and correct*, (3) *not confident and incorrect*, and (4) *confident and incorrect*. The average proportions of the four combinations at t1 and t2 are shown in **Table 3**.

Overall, the pre-post test data show that the number of *confident and correct* cases increased on average from t1 to t2 ( $x^{t1} = 1.58 \pm SD^{t1} = 1.009$ ,  $x^{t2} = 5.56 \pm SD^{t2} = 2.641$ ,  $t(88) = -15.680$ ,  $p < 0.001^*$ ,  $d = 3.68$ ) and that the number of *not confident and incorrect* cases decreased on average from t1 to t2 ( $x^{t1} = 6.33 \pm SD^{t1} = 1.700$ ,  $x^{t2} = 2.68 \pm SD^{t2} = 1.921$ ,  $t(87) = 15.355$ ,  $p < 0.001^*$ ,  $d = 1.75$ ). With regard to the

**TABLE 3 |** Average proportions (in %) of the four combinations at t1 and t2.

Case	t1	t2	$\Delta$	$p$
confident and correct	13.08	45.83	+ 32.75	< 0.001
confident and incorrect	9.92	11.83	+ 1.91	0.395
not confident and correct	24.08	19.00	−5.08	0.014
not confident and incorrect	52.92	22.50	−30.42	< 0.001

*t*-tests were used for significance testing.

not confident and correct cases, a slight decrease became evident between t1 and t2 ( $x^{t1} = 2.85 \pm SD^{t1} = 1.614$ ,  $x^{t2} = 2.315 \pm SD^{t2} = 1.669$ ,  $t(88) = 2.508$ ,  $p = 0.014^*$ ,  $d = 0.268$ ). However, no significant difference between t1 and t2 can be determined with regard to the *confident and incorrect* cases ( $x^{t1} = 1.22 \pm SD^{t1} = 1.178$ ,  $x^{t2} = 1.36 \pm SD^{t2} = 1.323$ ,  $t(87) = -0.854$ ,  $p = 0.395$ ,  $d = 0.09$ ).

### Profiles Based on Confidence and Knowledge (Change) and Specific Characteristics (H5b)

As the inferential statistical analyses showed significant differences in students' response behavior regarding confidence (in their correct as well as incorrect responses) and performance in the knowledge test at t1 and t2, an LCA was conducted to identify further possible correlations and differences within the groups depending on the confidence level.

First, an LCA was conducted on the basis of confidence at t1. As mentioned above, the three-class solution was chosen for confidence at t1 based on the LCA model parameters *AIC* and *BIC* (t1: three-class solution with *AIC* = 133.398, *BIC* = 150.368; Van Den Bergh and Vermunt, 2019). The LCA indicated three profiles: a low-confidence group (24.26%), a medium confidence group (42.01%), and a high-confidence group (33.73%). **Supplementary Table 1** shows which variables differ among the three groups at t1. While there was a difference between the groups in terms of confidence levels at t2, there was no significant difference in the knowledge scores at t1 and t2. At a marginal level of significance, there is a difference in terms of intelligence test scores in favor of the low-confidence group, which was not consistent with the User Experience Questionnaire (UEQ) grade. When comparing the groups, it became evident that a higher proportion of the medium confidence group completed medical vocational training, while members of the group with the highest confidence at t1 were significantly more likely to have taken an advanced physics course in high school.

An LCA also determined a three-cluster solution for the confidence level at t2 with the following groups (t2: three-class solution with *AIC* = 119.193, *BIC* = 134.124): low confidence (5.92%), medium confidence (21.89%), and high confidence (72.19%). The clusters for confidence at t1 and t2 show only a weak correlation of  $r = 0.179$ . **Supplementary Table 2** shows the differences among the three groups at t2. The students from the high-confidence group at t2 had a high level of confidence at t1 as well, while their knowledge level at t1 was not notably higher than that of other groups; and showed the highest increase in knowledge between t1 and t2.

Next, a *combined* latent cluster that best describes the change in knowledge and confidence from t1 to t2 was also determined. Similar to t1 and t2, the most appropriate cluster solution according to the model fit indices (three-class solution with *AIC* = 532.703, *BIC* = 558.450) resulted again in three groups: *knowledge gainers* (7.69%), *confidence gainers* (11.24%), and *overall gainers* (81.07%). **Table 4** shows the differences in the included covariates. When analyzing the difference scores for confidence and knowledge, the expected differences between the two groups became apparent, indicating a higher growth in both confidence and knowledge among the group of overall gainers. This group also has the highest confidence level at t2 and the highest knowledge level at t2, while the confidence gainers have the highest knowledge level at t1. The latter group also has the highest proportion of students who have completed medical vocational training (only marginally significant).

## DISCUSSION

### Summary and Conclusion

In a pre-post study, the medical students' knowledge development in physiology was assessed. We found a significant increase in the students' knowledge in physiology from t1 to t2 with large effect size, supporting *H1*. Task-related confidence at t2 (also when controlling for other personal covariates such as intelligence, learning motivation, and gender) was revealed to be the strongest predictor of the knowledge score at t2 and the increase in knowledge, supporting *H2d*. Learning motivation was also a significant predictor, supporting *H2c*. Overall, 51% of the variance of the knowledge score from t2 can be explained, also with regard to the difference score, even though the explained proportion of variance was 31% for the latter. The lack of predictive power of prior learning opportunities (*H2a*) may possibly be due to the relatively narrow conceptualization and empirical operationalization of "prior knowledge" in this study. For instance, current research provides further insights into the understanding of "prior knowledge" that should be considered in future work (McCarthy and McNamara, 2021). In addition, a current meta-analysis by Simonsmeier et al. (2021) shows the particular importance of analyzing the conditions under which prior knowledge has an effect on learning processes, which require further investigation.

As it is assumed that the students' knowledge increases over their course of study in physiology, a significantly higher level of confidence at t2 was expected and confirmed by the data, with a large effect size (supporting *H3*). Remarkably, a high level of confidence at t1 was a significant predictor of a high level of confidence at t2, which indicated high stability and correlation of this variable. With respect to the difference between the two confidence scores at t1 and t2, only knowledge at t2 was a significant predictor in the model explaining confidence, supporting *H4a*.

The results for *H5* indicate that at t1, confidence is less significant for predicting knowledge, while at t2, confidence is a much more significant predictor of knowledge and knowledge change that may occur through actively participating in the

**TABLE 4 |** Differences between the three profiles regarding changes in confidence and knowledge.

	knowledge gainers	confidence gainers	overall gainers	$F/\chi^2, p$	Tukey post-estimation			$\omega^2$ /Cramer's V
	$M \pm SD$ (n)/%	$M \pm SD$ (n)/%	$M \pm SD$ (n)/%		$p$ low vs. medium	$p$ medium vs. high	$p$ low vs. high	
Confidence level at t1	2.14 $\pm$ 0.351 (13)	1.92 $\pm$ 0.383 (18)	1.96 $\pm$ 0.411 (94)	1.37, 0.258	0.278	0.911	0.281	0.006
Confidence level at t2	2.19 $\pm$ 0.394 (13)	2.64 $\pm$ 0.369 (18)	2.83 $\pm$ 0.439 (93)	13.68, <0.001	0.010	0.200	<0.001	0.170
Knowledge test score at t1	3.38 $\pm$ 0.768 (13)	4.84 $\pm$ 1.675 (19)	4.26 $\pm$ 1.622 (102)	3.32, 0.039	0.030	0.308	0.142	0.033
Knowledge test score at t2	6.38 $\pm$ 1.502 (13)	5.53 $\pm$ 1.50 (19)	8.30 $\pm$ 1.850 (100)	23.50, <0.001	0.374	<0.001	0.001	0.254
Confidence difference score	0.45 $\pm$ 0.235 (13)	0.67 $\pm$ 0.264 (17)	1.03 $\pm$ 0.306 (60)	13.68, <0.001	<0.001	<0.001	<0.001	0.584
Knowledge test difference score	3.00 $\pm$ 1.732 (13)	0.68 $\pm$ 0.885 (19)	4.35 $\pm$ 1.556 (65)	46.11, <0.001	<0.001	<0.001	0.009	0.482
Learning motivation	2.13 $\pm$ 0.658 (13)	2.25 $\pm$ 0.635 (19)	2.035 $\pm$ 0.695 (100)	0.84, 0.433	0.886	0.423	0.874	-0.002
Intelligence test score	12.54 $\pm$ 3.688 (13)	11.68 $\pm$ 3.250 (19)	12.60 $\pm$ 3.48 (102)	0.56, 0.574	0.774	0.545	0.998	-0.007
UEQ grade	1.68 $\pm$ 0.650 (13)	1.35 $\pm$ 0.475 (19)	1.44 $\pm$ 0.534 (99)	1.59, 0.207	0.194	0.770	0.277	0.009
Age	22.92 $\pm$ 4.051 (13)	21.58 $\pm$ 2.694 (19)	21.97 $\pm$ 3.757 (98)	0.54, 0.582	0.565	0.905	0.651	-0.007
Sex, male	23.08%	31.58%	27.45%	0.285, 0.867	—	—	—	0.046
Language, German	15.38%	10.53%	13.73%	0.188, 0.910	—	—	—	0.038
Medical vocational training, yes	69.23%	94.74%	89.00%	5.201, 0.074	—	—	—	0.199
Advanced course in school in biology, yes	69.23%	47.37%	55.00%	0.647, 0.724	—	—	—	0.070
Advanced course in school in chemistry, yes	15.38%	26.32%	25.00%	0.631, 0.729	—	—	—	0.069
Advanced course in school in physics, yes	0.00%	15.79%	12.00%	2.076, 0.354	—	—	—	0.125
Advanced course in school in math, yes	53.85%	42.11%	47.00%	0.427, 0.808	—	—	—	0.057
Advanced course in school in sciences, no	15.38%	26.32%	23.00%	0.543, 0.762	—	—	—	0.064

<sup>1</sup>UEQ = university entrance qualification. In Germany, lower numbers indicate better grades (1–6). <sup>2</sup>Inverted scale, lower numbers indicate a higher motivation score  $\omega^2 < 0.01$  = very small, 0.01–0.06 = small, 0.06–0.14 = medium, and  $> 0.14$  = large. Cramer's V  $< 0.1$  = negligible, 0.1–0.29 = small, 0.3–0.49 = medium, and  $\geq 0.5$  = large effect.

physiology seminar series. This supports the abovementioned findings (Section “Conceptual-theoretical Background and Research Hypotheses”) that higher knowledge (at t2) may result in higher confidence (at t2) and vice versa. However, the students who are *confident but incorrect* constitute one particularly striking learning profile (i.e., *confidence gainers*) as the pre–post test results indicate no significant difference between the *incorrect responses at t1 and t2*. This finding indicates that among this group of students, incorrect knowledge or misconceptions might become established for certain domain-specific concepts captured in these test items. Identifying learning profiles of this kind enables better characterization of the possible misconceptions in medical education, and in turn, targeted interventions to correct them. Given that this group represents about 10% of the medical student population, teaching resources can now be efficiently focused on addressing potential misconceptions (refer to Section “Implications for Research and Practice”).

The emerging dissociation between confidence and knowledge might also be linked to the way medical curricula are commonly

structured in Germany, where a 2-year preclinical phase focuses on the acquisition of the knowledge and skills needed to understand and utilize the basic science underpinning medicine (Fritze et al., 2017a,b). However, in preclinical subjects like physiology, this is mostly done using canonical textbooks, and students tend to focus on memorizing and paraphrasing a long and ever-growing list of facts from textbooks instead of learning to trust their own developing sense of scientific argumentation (i.e., flexible problem-solving with confidence).

At t2, the significance of the classification in terms of the knowledge score becomes evident, which is comparable to previous results. After attending a physiology seminar over one term, students appear to have built a subject-related knowledge base. Confidence is, therefore, also a meaningful indicator when explaining and predicting knowledge scores. However, the direction of the relationships between the two constructs knowledge and confidence, and their development between t1 and t2 remains unclear. Overall, there are higher proportions of students in the high-confidence group in the cluster at t2. The correlation between the confidence clusters t1 and t2 is not very

high, which suggests that these clusters may change over the course of a seminar in physiology.

Mixed groups (LCA) of knowledge and confidence show fewer differences in terms of the covariates included. Confidence levels at t2 are the lowest in the group of *knowledge gainers*. Furthermore, the knowledge score at t2 differs significantly between the groups. The difference scores in knowledge and confidence, which were also the basis for generating and labeling this combined cluster, differ with high effects. Even if only at a marginally significant level, the group of *knowledge gainers* has the lowest proportion of students with completed vocational training. To further explain these learning profiles, more information in terms of student characteristics is needed.

When exploring the factors influencing both domain-specific knowledge and task-related confidence, as well as when considering the differences among the three clusters, intelligence—usually one of the strongest influencing factors according to other studies (e.g., Schwager et al., 2015; Wai et al., 2018)—explains only a small, non-significant amount of variance. One possible explanation may lie in the high pre-selectivity of the sample of medical undergraduates in terms of the cognitive study requirements, which leads to the high homogeneity of intellectual preconditions among this group. Indeed, the medical students scored substantially better on the IST test ( $M = 12.91 \pm SD = 3.324$ ) than a comparison group of students from business and economics ( $M = 6.57 \pm SD = 1.766$ ,  $n = 246$ ; from another project, Zlatkin-Troitschanskaia et al., 2019a).

The gender effect on confidence was only weak and marginally significant. These results were in contrast with previous research (e.g., Walstad and Robson, 1997; Hambleton, 2005; Brückner et al., 2015) and might reflect specific features of the medical student population, which has become predominantly female. In the context of knowledge assessment and competence testing, this finding can be interpreted as an indicator of discriminate validity and test fairness and supports the implementation of assessments as presented in this study in medical education practice (Zlatkin-Troitschanskaia et al., 2019b).

## Limitations

The presented results should be interpreted in consideration of the limitations of this study. First, the sample size is relatively small (even though an entire student cohort was examined in the specific context of a physiology seminar series). Consequently, the subsamples (e.g., students who took an advanced course in chemistry at school) were also small, which might have caused sampling effects. Although this sample can be considered representative for medical students according to the German official statistics (in terms of the students' main descriptive characteristics), it would be premature to draw general conclusions.

Second, the final sample of 97 students is limited as a basis for latent analysis and should also be considered cautiously with regard to the number of statistical analyses (for power determination, refer to Muthén and Muthén, 2002). To gain a comprehensive first insight into the relationship between confidence and knowledge development as well as into influencing factors, the multiple linear regression and LCA were

carried out. Therefore, the results should be interpreted with caution and require testing in larger samples. Larger samples would also allow for more intensive parallelism checks (using item rasch theory (IRT) procedures), which were not carried out in this study.

Third, due to the nature of the field survey, and as all students at this stage of their medical studies had to attend the seminar series, it was not possible to establish a control group. This limits the interpretability of the causality of the identified effects. However, this is a general concern in higher education field research, since in field studies it is almost impossible to conduct a study with a control group. Therefore, the robustness of the results should be examined in similar follow-up studies with larger cohorts and more measurement points, and the study should also be conducted at other universities and in other countries.

Fourth, due to time restraints in this field study, only a short version of the knowledge test was used in the study, which reflected only a small part of both physiology curricula and the cognitive requirements in the preclinical phase. However, our aim with this short test was not to evaluate the teaching effects in medical education, but rather to examine medical students' fundamental developmental tendencies and above all the relationships between confidence and knowledge, as demonstrated in the analyses presented in this study (refer to Sections "Relation between students' domain-specific knowledge (changes) and their confidence at t1 and t2 (H5a)" and

"Groups based on confidence and knowledge (change) and specific characteristics (H5b)"). The low reliability of the test might be improved in the future by implementing polytomous scoring (e.g., Embretson and Reise, 2000); partial crediting for the short test is currently in progress and may be used for future improvement. Due to the low number of test items, it was not possible to design two parallel test versions, and carry-over effects cannot be fully ruled out by this study design. However, since there is an interval of about 6 months between t1 and t2, significant test-induced learning effects seem unlikely (e.g., Scharfen et al., 2018).

Fifth, the predictive validity of the assessed personal characteristics may be limited as short scales of general cognitive ability and learning motivation were used in this study. As some studies already suggest (e.g., Kruger and Dunning, 1999; Klymkowsky et al., 2006), the assessed "task-related" students' confidence (students' confidence in their responses) is not necessarily indicative of students' self-confidence in their (metacognitive) abilities such as critically reflecting on their knowledge, problem-solving, and decision-making.

Despite these limitations, the examination of the relationship between knowledge and confidence as well as of the development of this relationship over time in a pre-post design contributes to the internal validity of the study results. A significant contribution to the still limited existing research in medical education as well as providing important insights into the seemingly reciprocal relationship between knowledge and confidence is made herewith (Section "Conceptual-theoretical Background and Research Hypotheses"). Based on prior research and the findings presented in this study; however, we cannot reach a satisfactory conclusion as to what particular



(meta)cognitive and/or affective processes and trait- and/or state-like abilities underlie task-related confidence. Further, more differentiated research is needed.

## Implications for Research and Practice

Through a particular focus on the relationship between domain-specific knowledge and confidence dynamics, and the examination of the development of these variables in a longitudinal analysis, this study makes contributions to bridging current research gaps. Overall, with regard to the regression modeling of knowledge and task-related confidence at t2, we already explained more than 50% of the variance in students' knowledge test scores and identified its most significant influencing factors. Compared to existing studies, which usually explain a relatively small amount of variance in students' knowledge (refer to Section "Conceptual-theoretical Background and Research Hypotheses"), this large share of explained variance is particularly remarkable. This also indicates the high practical importance of the included influencing factors and, in particular, supports the claim that the valid assessment of confidence levels in relation to the levels of domain-specific knowledge and their development is important in domain-specific learning. At the same time, when looking at the still unexplained variance, additional research is required to explain the complex relationship between the development of (prior) knowledge and (task-related) confidence as well as the dynamics of this relationship in more detail.

To date, only little is known about the development of students' confidence. In particular, the relationship between knowledge development and confidence requires more in-depth research. Moreover, the aspect of confidence should also be taken into account in formal knowledge assessments. Examining this relationship at the item level and, in particular, with a view to the item contents has the potential to contribute to a better understanding of how knowledge and confidence develop in relation to certain domain-specific concepts and/or types of tasks and problems. Therefore, future research should overcome the abovementioned limitations of the present study. In particular, in-depth (qualitative or mixed methods) analyses at the level of individual items and case studies of individual learning or development profiles have the potential to contribute to a more precise understanding of knowledge acquisition processes and their relation to confidence development.

This study offers initial insights; nevertheless, the chicken-or-egg dilemma remains: are students more confident because they have a higher level of knowledge, or do students choose the correct answers because they are more confident? More research on students' mental processes, using think-aloud protocols and eye-tracking studies, is required to further investigate this reciprocal relationship.

Despite the limitations of this study, the results indicate that confidence is of particular importance and that a stronger focus should be placed on this aspect in education and training, especially in professions such as medicine, where fast (spontaneous) decision-making with confidence is essential. In this study, we argue that integrating the practice of asking students to critically reflect on their level of confidence in

their task responses and their decision-making as early as during preclinical physiology courses is a useful exercise in several ways, for both students and instructors. For instance, in (self-)assessments, students will be trained to reflect on their reasoning and to improve their metacognitive ability to assess whether their confidence is justified. If combined with knowledge assessments, targeted interventions, and feedback (Butler et al., 2008), this will likely become an effective tool to increase student understanding and conceptual change. In addition to being a preparatory activity that will become more central (and complex) later on, when it comes to clinical decision-making, where multiple dimensions—including the needs and preferences of individual patients—need to be integrated, it also provides information about individual learning progressions in physiology. Confidence testing can provide teachers with valuable feedback about students' learning difficulties, and identify certain content that students are uncertain about or areas in which they are misinformed (for deeper insights into current developments and perspectives in medical higher education, refer to e.g., Kopp et al., 2008; Blohm et al., 2015; Heitzmann et al., 2019).

*In summation*, a higher level of confidence in one's own decisions can develop together with a higher level of understanding of physiological processes—thus providing a richer, denser, and more interconnected mental landscape of qualitative and quantitative checkpoints. However, our study also identified students who developed a high level of confidence in incorrect solutions to physiological problems over the course of the seminar series, which hints at preconceptions or even misconceptions. This provides a new starting point for targeted interventions as well as for a critical assessment focused on which resources and strategies these students utilized.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

JR provided the idea for this study, developed the instrument for knowledge assessment, and was involved in the data collection and in preparing and reviewing the manuscript. JR-S conducted the analyses and co-wrote the manuscript. OZ-T developed, in collaboration with JR, the study design, the instrument for confidence testing, was involved in the analyses, and co-wrote the manuscript. VK and MW were involved in the development of the instrument for knowledge assessment and the data collection and supported the analyses. M-TN was involved

in the organization of the data collection and supported the analyses. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was part of an Rhine-Main-Universities (RMU) project, which was funded by the Rhine-Main-Universities (RMU) fund.

## ACKNOWLEDGMENTS

We would like to thank all physicians and medical students from the Medical Faculty of Goethe University

Frankfurt who participated in this study. Furthermore, we would also like to thank the reviewers for their differentiated feedback and their contribution to the improvement of this manuscript. Detailed comments, efforts, and expertise of the reviewers were outstanding and certainly contributed significantly to the quality of this manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.562211/full#supplementary-material>

## REFERENCES

- Alexander, P. A. (2013a). Calibration: what is it and why it matters? An introduction to the special issue on calibrating calibration. *Learn. Instr.* 24, 1–3. doi: 10.1016/j.learninstruc.2012.10.003
- Alexander, P. A. (ed.) (2013b). Calibrating calibration: creating conceptual clarity to guide measurement and calculation. *Learn. Instr.* 24, 1–66. doi: 10.1016/j.learninstruc.2012.10.003
- Alexander, P. A., Jetton, T. L., and Kulikowich, J. M. (1995). Interrelationship of knowledge, interest, and recall: assessing a model of domain learning. *J. Educ. Psychol.* 87, 559–575. doi: 10.1037/0022-0663.87.4.559
- Alexander, P. A., Murphy, K., and Sun, Y. (2018). “Knowledge and belief change in academic development,” in *The Model of Domain Learning. Understanding the Development of Expertise*, eds H. Fives and D. Dinsmore (New York, NY: Routledge Taylor & Francis), 157–174. doi: 10.4324/9781315458014-10
- Alexander, P. A., and Murphy, P. K. (1998). Profiling the differences in students’ knowledge, interest, and strategic processing. *J. Educ. Psychol.* 90, 435–447. doi: 10.1037/0022-0663.90.3.435
- Alexander, P. A., Murphy, P. K., Woods, B. S., Duhon, K. E., and Parker, D. (1997). College instruction and concomitant changes in students’ knowledge, interest, and strategy use: a study of domain learning. *Contemp. Educ. Psychol.* 22, 125–146. doi: 10.1006/ceps.1997.0927
- Andrich, D., and Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Appl. Psychol. Measur.* 34, 181–192. doi: 10.1177/0146621609360202
- Baker, F. B., and Kim, S. H. (2004). *Item Response Theory: Parameter Estimation Techniques*. New York, NY: Dekker. doi: 10.1201/9781482276725
- Blohm, M., Lauter, J., Branchereau, S., Krautter, M., Köhl-Hackert, N., Jünger, J., et al. (2015). “Peer-assisted learning” (PAL) in the skills-lab – an inventory at the medical faculties of the Federal Republic of Germany. *GMS Zeitschrift für medizinische Ausbildung*. 32:Doc10.
- Borracci, R. A., and Arribalzaga, E. B. (2018). The incidence of overconfidence and underconfidence effects in medical student examinations. *J. Surg. Educ.* 75:5. doi: 10.1016/j.jsurg.2018.01.015
- Brandt, N. D., Lechner, C. M., Tetzner, J., and Rammstedt, B. (2019). Personality, cognitive ability, and academic performance: differential associations across school subjects and school tracks. *J. Pers.* 88:2. doi: 10.1111/jopy.12482
- Brückner, S. (2017). *Prozessbezogene Validierung anhand von mentalen Operationen bei der Bearbeitung wirtschaftswissenschaftlicher Testaufgaben [Process-Related Validation on the Basis of Mental Operations in the Processing of Economic Test items]*. (Empirische Berufsbildungs- und Hochschulforschung, Bd. 6). Landau: Verlag Empirische Pädagogik.
- Brückner, S., Förster, M., Zlatkin-Troitschanskaia, O., and Walstad, W. B. (2015). Effects of prior economic education, native language, and gender on economic knowledge of first-year students in higher education. A comparative study between Germany and the USA. *Stud. High. Educ.* 40, 437–453. doi: 10.1080/03075079.2015.1004235
- Bruno, J. E. (1993). “Using testing to provide feedback to support instruction: a reexamination of the role of assessment in educational organizations,” in *Item Banking: Interactive Testing and Self-Assessment*, eds D. A. Leclercq and J. E. Bruno (New York, NY: Springer), 190–209. doi: 10.1007/978-3-642-58033-8\_16
- Burggren, W. W., and Monticino, M. G. (2005). Assessing physiological complexity. *J. Exp. Biol.* 208, 3221–3232. doi: 10.1242/jeb.01762
- Butler, A. C., Karpicke, J. D., and Roediger, H. L. (2008). Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 918–928. doi: 10.1037/0278-7393.34.4.918
- Cate, O. T., Custers, E. J. F. M., and Durning, S. J. (2018). *Principles and Practice of Case-based Clinical Reasoning Education. Innovation and Change in Professional Education*. Berlin: Springer Nature. doi: 10.1007/978-3-319-64828-6
- Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *J. Pers. Soc. Psychol.* 39, 752–766. doi: 10.1037/0022-3514.39.5.752
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cordova, J. R., Sinatra, G. M., Jones, S. H., Taasobshirazi, G., and Lombardi, D. (2014). Confidence in prior knowledge, self-efficacy, interest and prior knowledge: influences on conceptual change. *Contemp. Educ. Psychol.* 39, 164–174. doi: 10.1016/j.cedpsych.2014.03.006
- Cox, M. D. M., and Irby, D. M. (2007). Assessment in medical education. *N. Engl. J. Med.* 356, 387–396. doi: 10.1056/NEJMra054784
- Custers, E. J. F. M. (2018). The script concordance test: an adequate tool to assess clinical reasoning? *Perspect. Med. Educ.* 7, 145–146. doi: 10.1007/s40037-018-0437-6
- Custers, E. J. F. M., and Cate, O. T. (2018). The history of medical education in Europe and the United States, with respect to time and proficiency. *Acad. Med.* 93, 49–54. doi: 10.1097/ACM.0000000000002079
- Davies, P. (2006). “There’s no confidence in multiple-choice testing,” in *Proceedings of the 6th CAA Conference*, Loughborough.
- Dinsmore, D. L., and Parkinson, M. M. (2013). What are confidence judgments made of? Students’ explanations for their confidence ratings and what that means for calibration. *Learn. Instr.* 24, 4–14. doi: 10.1016/j.learninstruc.2012.06.001
- Durning, S. J., Dong, T., Artino, A. R., van der Vleuten, C., Holmboe, E., and Schuwirth, L. (2015). Dual processing theory and experts’ reasoning: exploring thinking on national multiple-choice questions. *Perspect. Med. Educ.* 4, 168–175. doi: 10.1007/s40037-015-0196-6
- Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511761676
- Embretson, S. E., and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Eva, K. W., Cunningham, J. P. W., Reiter, H. I., Keane, D. R., and Norman, G. R. (2004). How can I know what I don’t know? Poor self assessment in a

- well-defined domain. *Adv. Health Sci. Educ.* 9, 211–224. doi: 10.1023/B:AHSE.0000038209.65714.d4
- Evans, J. (1984). Heuristic and analytic processes in reasoning. *Br. J. Psychol.* 75:4. doi: 10.1111/j.2044-8295.1984.tb01915.x
- Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. London: Sage Publications.
- Firth-Cozens, J. (2008). Effects of gender on performance in medicine. *BMJ* 336:7647. doi: 10.1136/bmj.39526.359630.BE
- Fitzgerald, J. T., White, C. B., and Gruppen, L. D. (2003). A longitudinal study of self-assessment accuracy. *Med. Educ.* 37, 645–649. doi: 10.1046/j.1365-2923.2003.01567.x
- Fritze, O., Griewatz, J., Narciß, E., Shiozawa, T., Wosnik, A., Zipfel, S., et al. (2017b). How much GK is in the NKLM? A comparison between the catalogues of exam-relevant topics (GK) and the German National Competence-based Learning Objectives Catalogue for Undergraduate Medical Education (NKLM). *GMS. J. Med. Educ.* 34:1.
- Fritze, O., Lammerding-Koepfel, M., Giesler, M., Narciss, E., Steffens, S., Wosnik, A., et al. (2017a). Benchmarking for research-related competencies—a curricular mapping approach at medical faculties in Germany. *Med. Teach.* 40:2. doi: 10.1080/0142159X.2017.1395403
- Gardner-Medwin, A. R. (1995). Confidence assessment in the teaching of basic science. *Res. Learn. Technol.* 3:1. doi: 10.1080/0968776950030113
- Grissom, R. J., and Kim, J. J. (2012). *Effect Sizes for Research: Univariate and Multivariate Applications*. New York, NY: Routledge. doi: 10.4324/9780203803233
- Hall, C. C., Ariss, L., and Todorov, A. (2007). The illusion of knowledge: when more information reduces accuracy and increases confidence. *Organ. Behav. Hum. Decis. Process.* 103, 277–290. doi: 10.1016/j.obhdp.2007.01.003
- Hambleton, R. K. (2005). “Issues, designs, and technical guidelines for adapting tests into multiple languages,” in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, eds R. K. Hambleton, P. F. Merenda, and C. D. Spielberger (Mahwah, NJ: L. Erlbaum Associates), 3–38. doi: 10.4324/9781410611758
- Hammond, K. R. (1996). *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. New York, NY: Oxford University Press.
- Hammond, K. R., Hamm, R. M., Grassia, J., and Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Trans. Syst. Man Cybern.* 17, 753–770. doi: 10.1109/TSMC.1987.6499282
- Hag, I., Higham, J., Morris, R., and Dacre, J. (2005). Effect of ethnicity and gender on performance in undergraduate medical examinations. *Med. Educ.* 39:11. doi: 10.1111/j.1365-2929.2005.02319.x
- Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M. R., et al. (2019). Facilitating diagnostic competences in simulations in higher education: a framework and a research agenda. *Frontline Learn. Res.* 7:4. doi: 10.14786/flr.v7i4.384
- Hmelo-Silver, C. E., and Azevedo, R. (2006). Understanding complex systems: some core challenges. *J. Learn. Sci.* 15, 53–61. doi: 10.1207/s15327809jls1501\_7
- Hmelo-Silver, C. E., Marathe, S., and Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: expert-novice understanding of complex systems. *J. Learn. Sci.* 16, 307–331. doi: 10.1080/1058400701413401
- IMPP (2020). *Prüfungsergebnisse Des Ersten Abschnitts der Ärztlichen Prüfung: Herbst 2020 [Results of the First Part of the Medical Examination: Autumn 2020]*. Available online at: <https://www.impp.de/pruefungen/medizin/l%C3%B6sungen-und-ergebnisse.html> (accessed November 2, 2020).
- Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.
- Kelly, M. P., Heath, I., Howick, J., and Greenhalgh, T. (2015). The importance of values in evidence-based medicine. *BMC Med. Ethics* 16:1–8. doi: 10.1186/s12910-015-0063-3
- Khan, K. S., Davies, D. A., and Gupta, J. K. (2001). Formative self-assessment using multiple true-false questions on the internet: feedback according to confidence about correct knowledge. *Med. Teach.* 23, 158–163. doi: 10.1080/01421590031075
- Kleitman, S., Stankov, L., Allwood, C. M., Young, S., and Mak, K. K. L. (2012). “Metacognitive self-confidence in school-aged children,” in *Self-Directed Learning Oriented Assessments in the Asia-Pacific*, ed. M. Mo Ching Mok (Dordrecht: Springer), 139–153. doi: 10.1007/978-94-007-4507-0\_8
- Klymkowsky, M. W., Taylor, L. B., Spindler, S. R., and Garvin-Doxas, R. K. (2006). Two-dimensional, implicit confidence tests as a tool for recognizing student misconceptions. *J. Coll. Sci. Teach.* 36, 44–48.
- Kolbitsch, J., Ebner, M., Nagler, W., and Scerbakov, N. (2008). “Can confidence assessment enhance traditional multiple-choice testing?” in *Proceedings of the ICL Interactive Computer Aided Learning 24 Sep 2008 to 26 Sep 2008*, Villach, 1–5.
- Kopp, V., Stark, R., and Fischer, M. R. (2008). Fostering diagnostic knowledge through computer-supported, case-based worked examples: effects of erroneous examples and feedback. *Med. Educ.* 42:8. doi: 10.1111/j.1365-2923.2008.03122.x
- Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* 77, 1121–1134. doi: 10.1037/0022-3514.77.6.1121
- Kulhavy, R. W., and Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *J. Educ. Psychol.* 63:5. doi: 10.1037/h0033243
- Lee, H. S., Liu, O. L., and Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Appl. Measur. Educ.* 24, 115–136. doi: 10.1080/08957347.2011.554604
- Liepmann, D., Beauducel, A., Brocke, B., and Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R [Intelligence structure test 2000 R] (I-S-T 2000 R)*. Göttingen: Hogrefe.
- Lubarsky, S., Dory, V., Meterissian, S., Lambert, C., and Gagnon, R. (2018). Examining the effects of gaming and guessing on script concordance test scores. *Perspect. Med. Educ.* 7, 174–181. doi: 10.1007/s40037-018-0435-8
- McCarthy, K. S., and McNamara, D. S. (2021). The multidimensional knowledge in text comprehension framework. *Educ. Psychol.* 56, 196–214. doi: 10.1080/00461520.2021.1872379
- Miller, G. (1990). The assessment of clinical skills/competence/performance. *Acad. Med.* 65, 63–67. doi: 10.1097/00001888-199009000-00045
- Muthén, L. K., and Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equ. Model.* 4, 599–620. doi: 10.1207/S15328007SEM0904\_8
- Muthén, L. K., and Muthén, B. O. (1998–2011). *Mplus User's Guide*, 6th Edn. Los Angeles, CA: Muthén & Muthén.
- Nathali, M. S. (2020). *Anatomy & Physiology of Physical Education*. New Delhi: Friends Publications. doi: 10.1111/j.1813-6982.2006.00054.x
- Owen, A. L. (2012). “Student characteristics, behavior, and performance in economic classes,” in *International Handbook on Teaching and Learning Economics*, eds G. M. Hoyt and K. McGoldrick (Cheltenham: Edward Elgar), 341–350.
- Parker, K. (2006). The effect of student characteristics on achievement in introductory microeconomics in South Africa. *S. Afr. J. Econ.* 74, 137–149. doi: 10.1177/1555343416686476
- Patel, V. L., Arocha, J. F., and Kaufman, D. R. (1999). “Expertise and tacit knowledge in medicine,” in *Tacit Knowledge in Professional Practice: Researcher and Practitioner Perspectives*, eds R. J. Sternberg and J. A. Horvath (Hillsdale, NJ: Erlbaum), 75–99.
- Patterson, R. E., and Eggleston, R. G. (2017). Intuitive cognition. *J. Cogn. Eng. Decis. Mak.* 11:1. 1555343416686476 doi: 10.1002/sce.3730660207
- Posner, G. J., Strike, K. A., Hewson, P. W., and Gertzog, W. A. (1982). Accommodation of a scientific conception: toward a theory of conceptual change. *Sci. Educ.* 66, 211–227. doi: 10.1080/01421590802637925
- Preidger, S., Berberat, P. O., Kadmon, M., and Harendza, S. (2020). “Measuring medical competence and entrusting professional activities in an assessment simulating the first day of residency,” in *Student Learning in German Higher Education: Innovative Measurement Approaches and Research Results*, eds O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, and C. Lautenbach (Cham: Springer), 317–332.
- Pruskil, S., Burgwinkel, P., Georg, W., Keil, T., and Kiessling, C. (2009). Medical students' attitudes towards science and involvement in research activities: a comparative study with students from a reformed and a traditional curriculum. *Med. Teach.* 31, 254–259. doi: 10.34297/AJBSR.2020.07.001166
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Available online at: <https://www.R-project.org/> (accessed April 15, 2020).

- Roeper, J., Zlatkin-Troitschanskaia, O., Klose, V., Nagel, M.-T., and Schlax, J. (2020). A new approach to analyzing the development of domain-specific knowledge among undergraduate medical students using learning scores. *Am. J. Biomed. Sci. Res.* 7, 319–323. doi: 10.34297/AJSR.2020.07.001166
- Rotgans, J. I., and Schmidt, H. G. (2017). The relation between individual interest and knowledge acquisition. *Br. Educ. Res. J.* 43:2. doi: 10.1016/j.neuron.2016.03.025
- Rudolph, J., Niepel, C., Greiff, S., Goldhammer, F., and Kröner, S. (2017). Metacognitive confidence judgments and their link to complex problem solving. *Intelligence* 63, 1–8. doi: 10.1016/j.intell.2018.01.003
- Sanders, J. I., Hangya, B., and Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron* 90, 499–506. doi: 10.1037/t64341-000
- Scharfen, J., Peters, J. M., and Holling, H. (2018). Retest effects in cognitive ability tests: a meta-analysis. *Intelligence* 67, 44–66. doi: 10.1111/ijsa.12096
- Schiefele, U., Krapp, A., Wild, K., and Winteler, A. (1993). Der “fragebogen zum studieninteresse” (FSI). [The study interest questionnaire (SIQ)]. *Diagnostica* 39, 335–351. doi: 10.1111/mbe.12110
- Schwager, I. T. L., Hülsheger, U. R., Bridgeman, B., and Lang, J. W. B. (2015). Graduate student selection: graduate record examination, socioeconomic status, and undergraduate grade point average as predictors of study success in a western European University. *Int. J. Select. Assess.* 23:1. doi: 10.1080/00461520.2021.1939700
- Shing, Y. L., and Brod, G. (2016). Effects of prior knowledge on memory: implications for education. *Mind Brain Educ.* 10:3. doi: 10.1007/978-3-319-74338-7\_4
- Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., and Schneider, M. (2021). Domain-specific prior knowledge and learning: a meta-analysis. *Educ. Psychol.* 1–24. doi: 10.1016/j.paid.2013.07.006
- Spiel, C., and Schober, B. (2018). “Challenges for evaluation in higher education: entrance examinations and beyond: the sample case of medical education,” in *Assessment of Learning Outcomes in Higher Education*, eds O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, and C. Kuhn (Cham: Springer), 59–71.
- Stankov, L. (2013). Noncognitive predictors of intelligence and academic achievement: an important role of confidence. *Pers. Individ. Differ.* 55, 727–732. doi: 10.1177/0013164492052004025
- Stata Corp (2017). *Stata Statistical Software: Release 15*. College Station: StataCorp LLC. doi: 10.1080/10705511.2018.1550364
- Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., and Vallieres, E. F. (1992). The academic motivation scale: a measure of intrinsic, extrinsic, and amotivation in education. *Educ. Psychol. Measur.* 52, 1003–1017. doi: 10.3390/jintelligence6030037
- Van Den Bergh, M., and Vermunt, J. K. (2019). Latent class trees with the three-step approach. *Struct. Equat. Model.* 26, 481–492. doi: 10.1080/00220489709595917
- Wai, J., Brown, M. I., and Chabris, C. F. (2018). Using standardized test scores to include general cognitive ability in education research and policy. *J. Intell.* 6, 1–16.
- Walstad, W. B., and Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in economics. *J. Econ. Educ.* 28, 155–171. doi: 10.1371/journal.pone.0182460
- Walstad, W. B., Schmidt, S., Zlatkin-Troitschanskaia, O., and Happ, R. (2018). “Pretest-posttest measurement of the economic knowledge of undergraduates – estimating guessing effects,” in *Proceedings of the Annual AEA Conference on Teaching and Research in Economic Education*, Philadelphia, PA. doi: 10.4300/JGME-D-11-00324.1
- Weier, N., Thursky, K., and Zaidi, S. T. R. (2017). Antimicrobial knowledge and confidence amongst final year medical students in Australia. *PLoS One* 12:e0182460. doi: 10.1007/s40037-013-0090-z
- Wijnen-Meijer, M., van der Schaaf, M., Nillesen, K., Harendza, S., and Ten Cate, O. (2013a). Essential facets of competence that enable trust in graduates: a delphi study among physician educators in the Netherlands. *J. Grad. Med. Educ.* 5, 46–53.
- Wijnen-Meijer, M., van der Schaaf, M., Nillesen, K., Harendza, S., and Ten Cate, O. (2013b). Essential facets of competence that enable trust in medical graduates: a ranking study among physician educators in two countries. *Perspect. Med. Educ.* 2, 290–297.
- Zlatkin-Troitschanskaia, O., Schlax, J., Jitomirski, J., Happ, R., Kühling-Thees, C., Brückner, S., et al. (2019b). Ethics and fairness in assessing learning outcomes in higher education. *High. Educ. Policy* 32, 537–556. doi: 10.1057/s41307-019-00149-x
- Zlatkin-Troitschanskaia, O., Jitomirski, J., Happ, R., Molerov, D., Schlax, J., Kühling-Thees, C., et al. (2019a). Validating a test for measuring knowledge and understanding of economics among university students. *Z. Pädagog. Psychol.* 33, 119–133. doi: 10.1024/1010-0652/a000239

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Roeper, Reichert-Schlax, Zlatkin-Troitschanskaia, Klose, Weber and Nagel. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Advantages of publishing in Frontiers



## OPEN ACCESS

Articles are free to read  
for greatest visibility  
and readership



## FAST PUBLICATION

Around 90 days  
from submission  
to decision



## HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,  
and constructive  
peer-review



## TRANSPARENT PEER-REVIEW

Editors and reviewers  
acknowledged by name  
on published articles

## Frontiers

Avenue du Tribunal-Fédéral 34  
1005 Lausanne | Switzerland

Visit us: [www.frontiersin.org](http://www.frontiersin.org)

Contact us: [frontiersin.org/about/contact](http://frontiersin.org/about/contact)



## REPRODUCIBILITY OF RESEARCH

Support open data  
and methods to enhance  
research reproducibility



## DIGITAL PUBLISHING

Articles designed  
for optimal readership  
across devices



## FOLLOW US

@frontiersin



## IMPACT METRICS

Advanced article metrics  
track visibility across  
digital media



## EXTENSIVE PROMOTION

Marketing  
and promotion  
of impactful research



## LOOP RESEARCH NETWORK

Our network  
increases your  
article's readership