

On the automatic link between affect and tendencies to approach and avoid: Chen and Bargh (1999) revisited

Mark Rotteveel^{1,2*}, Alexander Gierholz¹, Gijs Koch¹, Chelle van Aalst¹, Yair Pinto^{2,3}, Dora Matzke⁴, Helen Steingroever⁴, Josine Verhagen⁴, Titia F. Beek¹, Ravi Selker⁴, Adam Sasiadek⁴ and Eric-Jan Wagenmakers^{2,4}

¹ Social Psychology Program, Department of Psychology, Faculty of Behavioral and Social Sciences, University of Amsterdam, Amsterdam, Netherlands, ² Amsterdam Brain and Cognition, University of Amsterdam, Amsterdam, Netherlands, ³ Cognitive Neuroscience Group, Amsterdam, Netherlands, ⁴ Psychological Methods, Amsterdam, Netherlands

OPEN ACCESS

Edited by:

Rene Zeelenberg,
Erasmus University Rotterdam,
Netherlands

Reviewed by:

Guido P. H. Band,
Leiden University, Netherlands
Christoph Stahl,
University of Cologne, Germany

*Correspondence:

Mark Rotteveel,
Social Psychology Program,
Department of Psychology, Faculty of
Behavioural and Social Sciences,
University of Amsterdam,
Weesperplein 4, 1018 XA,
Amsterdam, Netherlands
m.rotteveel@uva.nl

Specialty section:

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

Received: 31 May 2013

Accepted: 09 March 2015

Published: 02 April 2015

Citation:

Rotteveel M, Gierholz A, Koch G, van
Aalst C, Pinto Y, Matzke D,
Steingroever H, Verhagen J, Beek TF,
Selker R, Sasiadek A and
Wagenmakers E-J (2015) On the
automatic link between affect and
tendencies to approach and avoid:
Chen and Bargh (1999) revisited.
Front. Psychol. 6:335.
doi: 10.3389/fpsyg.2015.00335

Within the literature on emotion and behavioral action, studies on approach-avoidance take up a prominent place. Several experimental paradigms feature successful conceptual replications but many original studies have not yet been replicated directly. We present such a direct replication attempt of two seminal experiments originally conducted by Chen and Bargh (1999). In their first experiment, participants affectively evaluated attitude objects by pulling or pushing a lever. Participants who had to pull the lever with positively valenced attitude objects and push the lever with negatively valenced attitude objects (i.e., congruent instruction) did so faster than participants who had to follow the reverse (i.e., incongruent) instruction. In Chen and Bargh's second experiment, the explicit evaluative instructions were absent and participants merely responded to the attitude objects by either always pushing or always pulling the lever. Similar results were obtained as in Experiment 1. Based on these findings, Chen and Bargh concluded that (1) attitude objects are evaluated automatically; and (2) attitude objects automatically trigger a behavioral tendency to approach or avoid. We attempted to replicate both experiments and failed to find the effects reported by Chen and Bargh as indicated by our pre-registered Bayesian data analyses; nevertheless, the evidence in favor of the null hypotheses was only anecdotal, and definitive conclusions await further study.

Keywords: emotion, approach, avoidance, replication, affect

Introduction

Several prominent psychological theories (Frijda, 1986, 2007; Lang and Bradley, 2008) state that at the core of an emotion is the tendency to act. In order to survive, organisms need to approach reward and avoid danger or punishment. These tendencies to act, it is argued, manifest themselves in part through emotions. Other psychological theories also assume a link between action and evaluation (e.g., Lang et al., 1990; Cacioppo and Gardner, 1999; Neumann et al., 2003; Strack and Deutsch, 2004). Empirically, there is ample evidence for an association between affective evaluation and some kind of approach and avoidance behavior (e.g., Solarz, 1960; Chen and Bargh, 1999;

Rotteveel and Phaf, 2004; Krieglmeyer and Deutsch, 2010). What remains unclear, however, is the precise nature of this association; some researchers have argued that it is basic, direct, and default (e.g., Chen and Bargh, 1999; Duckworth et al., 2002), whereas others have argued that the association is more flexible and goal orientated (e.g., Bamford and Ward, 2008). This debate was initiated with the publication of the article by Chen and Bargh (1999) entitled “Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus.”

In the field of emotion and approach-avoidance behavior, the Chen and Bargh (1999) article (henceforth CB) has attracted a lot of attention (cited 866 times, January 19th 2015). In their first experiment, one that may be considered a conceptual replication of an earlier experiment (Solarz, 1960, cited 214 times, January 19th 2015), CB instructed participants to evaluate attitude objects affectively by pulling or pushing a lever. Participants who had to pull the lever for positively valenced attitude objects and push the lever for negatively valenced attitude objects did so faster than participants who had to pull for negatively valenced attitude objects and push for positively valenced attitude objects. In other words, the results showed a congruency effect between the affective valence of attitude objects and the direction of the lever movement—it is easier to pull (i.e., execute a “toward yourself” movement) for positively valenced objects and to push (i.e., execute an “away-from yourself” movement) for negatively valenced objects.

In their second experiment, CB manipulated congruency within participants and eliminated the explicit evaluative instruction. Specifically, participants were instructed to respond to the mere presentation of the attitude objects; in one block of trials, participants had to push the lever (i.e., execute an “away-from yourself” movement), and in another block they had to pull the lever (i.e., execute a “toward yourself” movement). The results again demonstrated a congruency effect: pulling was faster for positively valenced attitude objects, and pushing was faster for negatively valenced attitude objects. On basis of these results, CB concluded that (1) attitude objects are automatically evaluated; and (2) attitude objects automatically trigger a behavioral tendency to approach or avoid.

Since the publication of CB, numerous papers have been published in which approach and avoidance behavior was studied; however, the automatic link between affective evaluation and approach-avoidance tendencies was often simply taken for granted. To complicate matters, different results have been obtained using different experimental paradigms such as the manikin task (e.g., De Houwer et al., 2001), the joystick task (e.g., Eder and Rothermund, 2008), the joystick with zoom task (e.g., Rinck and Becker, 2007), and a button stand task (e.g., Rotteveel and Phaf, 2004). One issue with such conceptual replications of approach and avoidance behavior concerns construct validity. That is, different operationalizations in conceptual replications can tap different psychological processes reflecting partially different constructs that vary in their relation with the conceptualization of approach and avoidance as in the CB studies.

Replication is at the core of the scientific effort to further our understanding of the empirical world. Many effects do replicate reliably across laboratories in psychology (e.g., Simons, 2014), but some prominent effects are now under doubt (e.g., Pashler and Wagenmakers, 2012). Although opinions differ with regard to the extent of this “replication crisis” (e.g., Pashler and Harris, 2012; Stroebe and Strack, 2014), the scientific community seems to be shifting its focus more toward direct replication. For instance, several journals recently proposed special issues on replication (e.g., Nosek and Lakens, 2014, this issue of *Frontiers in Psychology*) or even launched a new type of article (i.e., Registered Replication Reports in *Cortex*, *Perspectives on Psychological Science*, *Attention, Perception, and Psychophysics*, and other journals (see Chambers, 2013; Wolfe, 2013); for an overview see <https://osf.io/8mpji/wiki/home/>).

Direct replications benefit from preregistration of design and analysis plan, ensuring a clean separation between which analyses are pre-planned (i.e., confirmatory, hypothesis-testing) and which analyses are *post-hoc* (i.e., exploratory, hypothesis-generating; see e.g., De Groot, 1956/2014; Wagenmakers et al., 2012). Such separation is also required for the proper statistical interpretation of the results. When an initial finding replicates successfully in a preregistered setting, this raises researchers’ confidence that the effect is real and can form the basis for more empirical as well as theoretical work. When an initial finding fails to replicate, however, scientific effort may be re-oriented toward other, more promising avenues of investigation—at least when null results are published and do not disappear in the file drawer (e.g., Rosenthal, 1979; Francis, 2013). Direct replications not only affect one’s confidence about the veracity of the phenomenon under study, but they also increase our knowledge about effect size (see also Simons, 2014; but see also Stroebe and Strack, 2014).

Our decision to replicate the CB studies was motivated in part by a recent meta-analysis on approach and avoidance behavior including 29 published studies and 81 effect sizes (Phaf et al., 2014), which indicated a moderate publication bias for congruency effects with explicit affective evaluation as obtained in Experiment 1 of CB. More importantly, to the best of our knowledge the CB findings were never replicated directly. This is remarkable, particularly in light of the central importance of the CB findings in the literature on emotion and approach and avoidance behavior. For these reasons we attempted to replicate the original CB findings using a similar experimental setup (i.e., a lever, see **Figure 1**), similar stimuli, and similar instructions. To remove all researcher’s degrees of freedom in the analysis stage we used a preregistered protocol on the Open Science Framework¹ (e.g., Open Science Collaboration, 2012). This protocol detailed the design, method, hypotheses, as well as the entire analysis plan.

In direct replication studies it is essential to be able to quantify evidence in favor of the null hypothesis. In addition, it is desirable to collect data until the results are compelling. Neither desideratum can be accomplished within the framework of frequentist statistics, and this is why our analysis of both experiments will rely on hypothesis testing using the Bayes factor (e.g., Edwards et al., 1963; Berger and Mortera, 1999; Wagenmakers,

¹<https://osf.io/wxigz/>.



FIGURE 1 | The experimental setup of Experiment 1 and Experiment 2.

The 100 cm lever is fixed to base with a hinge. Two identical weak springs make sure that the lever will return to mid position after responding. Responses were recorded whenever the lever reached 15.6° of movement backwards and 15.3° of movement forwards.

2007; Rouder et al., 2009, 2012; Wagenmakers et al., 2012). The method section below provides the details of our design and analysis methodology. This research follows a strictly confirmatory protocol as described in Wagenmakers et al. (2012).

Experiment 1

Method

Pre-Registered Sampling Plan

A frequentist analysis would start with an assessment of the effect size of Experiment 1 from CB which would then form the basis of a power analysis to determine the number of participants that yields a specific probability for rejecting the null hypothesis when it is false. This frequentist analysis plan is needlessly constraining and potentially wasteful: the experiment cannot continue after the planned number of participants has been tested, and it cannot stop even when the data yield a compelling result earlier than expected (e.g., Wagenmakers, 2007). Here we circumvent these frequentist limitations by calculating and monitoring the Bayes factor (e.g., Edwards et al., 1963; Berger and Mortera, 1999; Wagenmakers et al., 2012; Rouder et al., 2012). The Bayes factor quantifies the change from prior model odds to posterior model odds; in other words, the Bayes factor quantifies the extent to which the data shift our opinion away from one hypothesis and toward another. A Bayes factor of 5 in favor of the null hypothesis, for example, indicates that the data are 5 times more likely to occur under the null hypothesis than under the alternative hypothesis. For the interpretation of evidence in the Bayesian paradigm, the intention with which the data are collected is irrelevant; hence, the Bayes factor can be monitored as the data come in, and data collection can be terminated at any point (Berger and Wolpert, 1988; Rouder, 2014; see also Figures 2, 3).

Based on the above considerations, our sampling plan was as follows: We set out to collect a minimum of 20 participants in

each between-subject condition (i.e., the congruent and incongruent condition, for a minimum of 40 participants in total). Next we planned to monitor the Bayes factor and stop the experiment whenever both critical hypothesis tests (detailed below) reached a Bayes factor that could be considered “strong” evidence (Jeffreys, 1961); this meant that the Bayes factor should be either 10 in favor of the null hypothesis, or 10 in favor of the alternative hypothesis. The experiment was also to be stopped whenever the maximum number of participants was reached. This number was set to 50 participants per condition (i.e., a maximum of 100 participants in total). Additionally, the experiment was to be stopped by January 1st, 2014 if neither criteria were met. The latter date was, however, amended on OSF on January the 7th and reset to January 31st 2014 because the Bayes factor had not reached the pre-set level of strong evidence and the maximum number of participants had not been reached either. From a Bayesian perspective the specification of this sampling plan was needlessly precise; we nevertheless felt the urge to be as complete as possible. In the end, data collection was terminated because the maximum number of participants was reached.

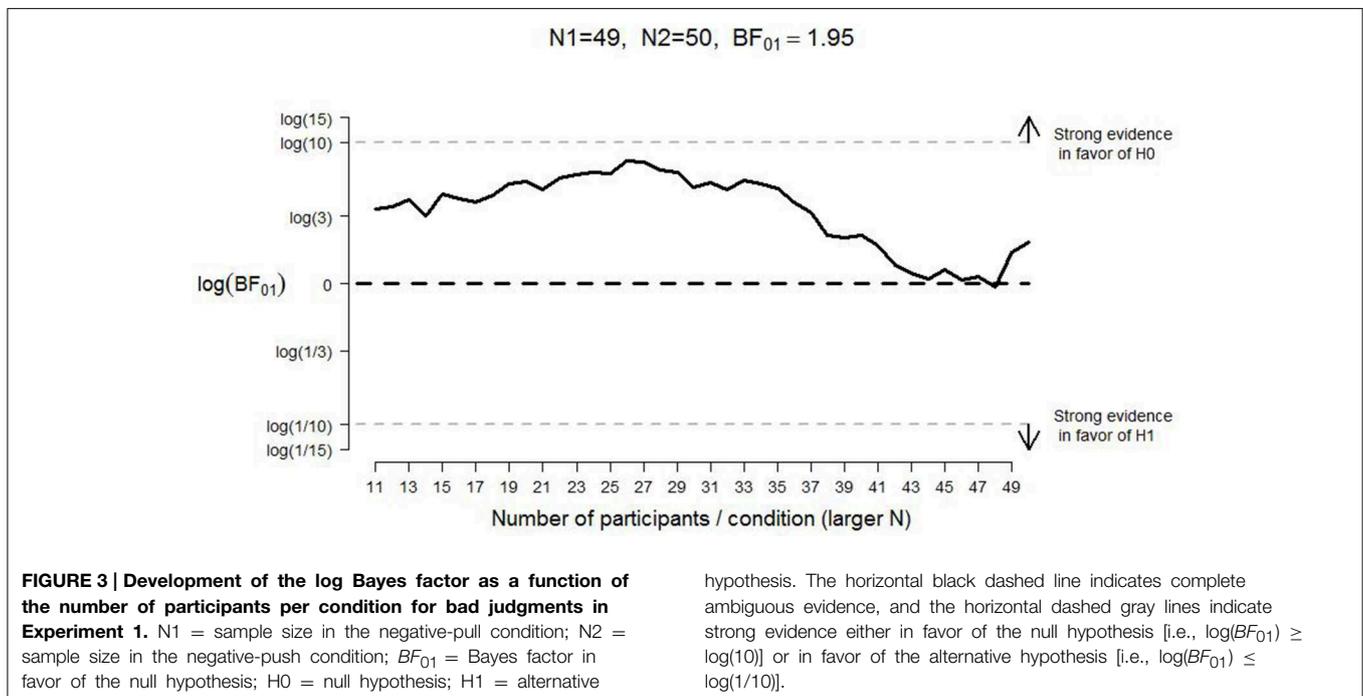
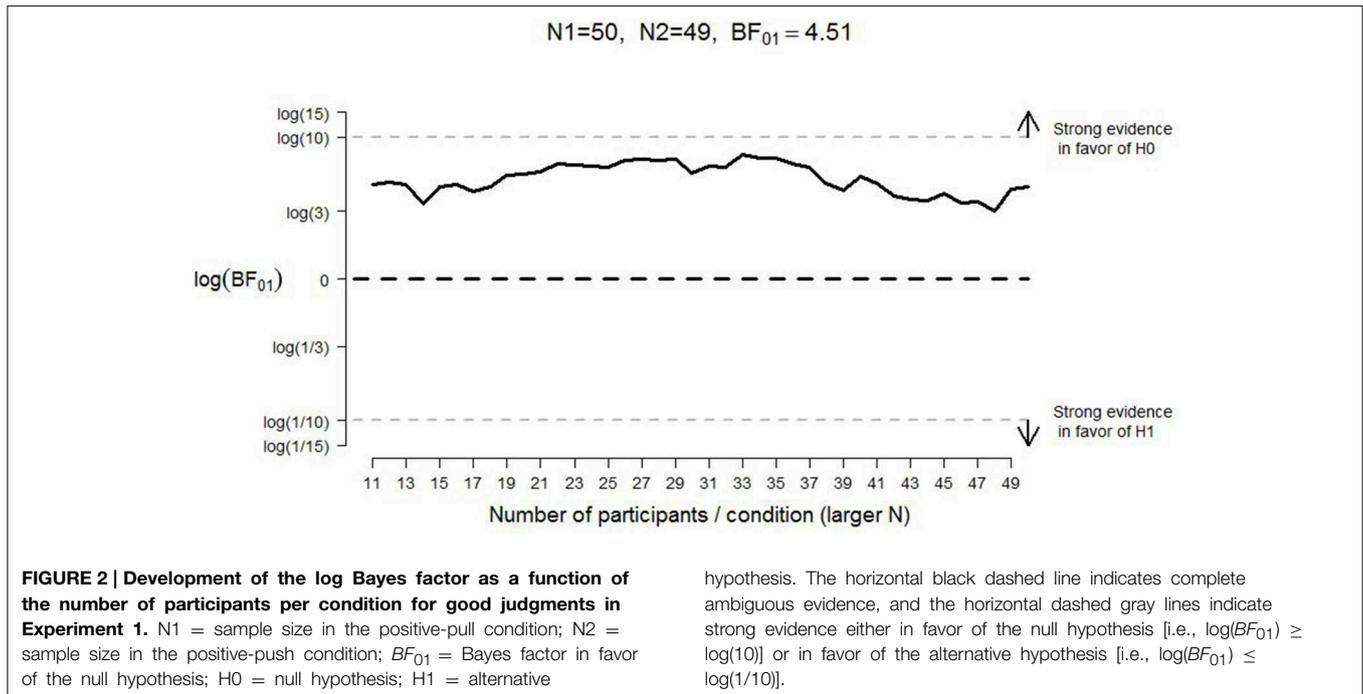
Participants

We recruited 100 students (23 male, mean age = 21.2 year, SD = 0.42; Congruent: 10 male; Incongruent: 13 male) from the University of Amsterdam. All participants were rewarded with course credits or €5. Only students with Dutch as their native language were allowed to participate. One participant did not meet this criterion and was excluded from further analysis. All participants were informed about the procedure with an information brochure and subsequently signed an informed consent form.

Materials

Participants were seated in a dimly lit room at approximately one arm length's distance in front of a computer monitor. A 100 cm long lever (see Figure 1) triggered one of two switches upon being pulled or pushed for, respectively 15.6 or 15.3° (Note that CB report a movement of 10° in both directions with a lever measuring 92 cm). Participants held the lever at a marked height of approximately 68 cm resulting in actual hand movement of 18.7 cm forwards and 19 cm backwards. In CB's experiments actual hand movements were 12 cm in both directions assuming participants were holding the lever at the same height as in our experiment². Both switches were connected to the computer so that response latencies as well as movement direction could be recorded. Responses were recorded using a mechanism based on a Logitech G400 gaming mouse. Polling rate of the mouse was set at 500 S/s, response latency was less than 3 ms. Start of lever movement was recorded by polling the mouse position every screen refresh (16.667 ms). This means start-of-response latency could be anywhere between 3 and 16.67 ms. To ensure that the lever returned to the central position after each response, two identical weak springs were connected to the lever at the front and at the back. These springs were added to the experimental setup to make sure that the lever would return to mid position after responding. On the few occasions that the lever was not

²We contacted John A. Bargh for help in designing our experiments.



returning to mid position by itself, participants were instructed to move the lever back to mid position themselves. We used almost the same attitude objects (henceforth “targets”) as were used in the original CB study and reported in Bargh et al. (1992). The targets were first translated (see Tables 1, 2 in Supplementary Material) from English to Dutch, and then back-translated to check for any inconsistencies. Inconsistent targets were replaced with comparable alternates in order to approach the original stimuli

as closely as possible in Dutch (i.e., “hangover” was replaced with “misselijkheid” which means “nausea”; “Reagan” was replaced with the name of the current Dutch prime-minister “Rutte”). Students from the University of Amsterdam ($n = 130$) then evaluated these targets in a test session on an 11 point scale (from very bad -5 to very good $+5$) and the resulting mean evaluations were compared to the mean evaluations of the original stimuli (Bargh et al., 1992). Targets that differed substantially in their

affective evaluation (e.g., “beer”; “priest”; “clown”) were excluded, as well as targets that differed more than 1.5 points on average from the original ratings of Bargh et al. (1992; see Table 2 in Supplementary Material). As can be expected, the valence of the remaining 78 targets (i.e., 39 positive targets and 39 negative targets) correlated highly with the affective valence of the original stimulus list ($r = 0.98$). Due to a programming mistake, the negative target “wormen” (“worms”) was replaced by the positive target “stereo” (“stereo”), a word from the practice session. Data obtained with “stereo” was excluded from further analysis so in the end responses to 39 positive words and 38 negative words were included in the results. These targets were presented in a random order and every target was presented only once in a lower case “Times” font on a white background.

Procedure

After reading the information brochure and signing an informed consent form, participants were seated in front of the computer screen with the lever next to their dominant hand, after which they read the procedure of the experiment off the screen. Participants were asked to classify the targets presented on the screen as either “good” or “bad” (for the exact wording of the specific lever movement instructions see Appendix A in Supplementary Material). Participants had to do this by either pushing the lever away from themselves, or pulling the lever toward themselves. The participants were alternately assigned to either the congruent or the incongruent condition. In the congruent condition, participants were instructed to pull the lever toward themselves if the target had a “good” meaning and push it away from themselves to indicate a “bad” meaning. In the incongruent condition, these instructions were reversed and participants had to pull the lever toward themselves if the target had a “bad” meaning and push it away from themselves to indicate a “good” meaning. Participants were instructed to release the lever after responding in order for it to return to its starting position. In case the lever did not return to its starting mid position itself participants were asked to return the lever themselves to its starting position.

Before the start of the actual experiment we confirmed that participants understood the instructions correctly by making them perform 10 practice trials with 10 separate targets that were not part of the 78 target words (except for “stereo,” see Table 2 in Supplementary Material). After the practice trials the experimenter left the room, so that the actual trials were completed without the experimenter present.

During the experiment each target appeared on screen until the participant pulled or pushed the lever above the 15.6 and 15.3° angle, respectively, necessary to trigger the response-switches. The computer recorded the time delay between the appearance of the target, the onset of the lever movement, and the triggering of the response-switch as specified before. The computer also recorded whether the lever had been pulled or pushed. After every response the target disappeared and it took 4 s until the next trial commenced and a new target appeared again at the center of the screen. The targets were presented in a random order with every target appearing once only. After responding on the last trial the experimenter returned to the room to thank and debrief the participant.

Preregistered Data Analysis and Presentation of Results

Based on the reasoning of CB and our own pilot tests, all trials with latencies greater than 3000 ms or smaller than 300 ms were excluded from further analysis (pulling with “good” judgments: 1.7%; pulling with “bad” judgments: 2.2%; pushing with “good” judgments: 1%; pushing with “bad” judgments: 1.2%). These criteria for outlier removal had been specified in the preregistration document. Whereas CB removed only latencies greater than 4000 ms, we had to reduce this value because pilot testing showed that in our setup one can easily push/pull the lever under 4000 ms. As in CB, and as specified in the preregistration document, the dependent measure for all analyses was the mean \log_{10} -transformed response latency for every participant. Results are reported as untransformed response latencies. The crucial hypothesis concerns the interaction that describes the congruency effect. Specifically, the congruency effect can be decomposed in two directional hypotheses: the first hypothesis states that participants respond faster to a positive target by pulling instead of pushing a lever; the second hypothesis states that participants respond faster to a negative target by pushing instead of pulling a lever. As specified in the preregistration document, the two crucial hypotheses will be assessed separately by means of two default Bayes factors for unpaired, one-sided t -tests as outlined in Rouder et al. (2009) and Wetzels et al. (2009). Specifically, the effect sizes under the alternative hypothesis are assumed to follow a folded Cauchy(0,1) distribution. Exploratory analyses will vary the shape of this prior to probe the robustness of our conclusions.

As described above, Bayes factors quantify the support that the data provide for the null hypothesis vis-a-vis the alternative hypothesis. Support in favor of the alternative hypotheses constitutes support in favor of the effects reported by CB in their Experiment 1.

Results

Bayes Factor Hypothesis Tests

For “good” evaluations, pulling the lever was a little faster ($M = 1147$ ms, $SE = 29$) than pushing ($M = 1165$ ms, $SE = 30$, see Table 1) whereas for “bad” evaluation, pushing the lever was ($M = 1204$ ms, $SE = 35$) faster than pulling ($M = 1267$, $SE = 39$). The direction of these effects is consistent with the results reported by CB. However, the Bayes factor (assuming equal variances, as was done for all analyses reported in this manuscript) indicated that the observed data were more likely under the null hypothesis than under the alternative hypothesis; specifically, $BF_{01} = 4.51$ for “good” evaluations (i.e., the data for “good” evaluations are almost five times more likely under the null hypothesis than under the alternative hypothesis) and $BF_{01} = 1.95$ for “bad” evaluations (i.e., the data for “bad” evaluations are almost twice as likely under the null hypothesis than under the alternative hypothesis). Figures 2, 3 display the development of the log Bayes factor as a function of the number of participants per condition for the “good” and the “bad” evaluations, respectively. Log Bayes factors larger than zero provide evidence for the null hypothesis; log Bayes factors smaller than zero provide evidence for the alternative hypothesis. For “good”

TABLE 1 | Response latencies in ms (SE) for lever movement in Experiment 1 and for Experiment 2 (* in Experiment 1 response latencies reflect good vs. bad judgments whereas response latencies in Experiment 2 reflect responses to good vs. bad words).

	Experiment 1		Experiment 2	
	Pull	Push	Pull	Push
Good*	1147 (29)	1165 (30)	562 (13)	571 (14)
Bad*	1267 (39)	1204 (35)	574 (12)	562 (13)

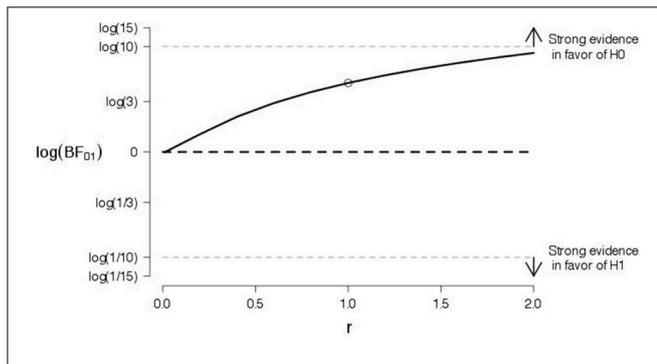


FIGURE 4 | A robustness analysis for the data of good judgments from Experiment 1. The log Bayes factor $\log(BF_{01})$ is plotted as a function of the scale parameter r of the Cauchy prior for the effect size under the alternative hypothesis. The dot indicates the result from the default prior, the horizontal black dashed line indicates complete ambiguous evidence, and the horizontal dashed gray lines indicate strong evidence either in favor of the null hypothesis [i.e., $\log(BF_{01}) \geq \log(10)$] or in favor of the alternative hypothesis [i.e., $\log(BF_{01}) \leq \log(1/10)$].

evaluations, after testing 11 participants per condition, the Bayes factor fluctuated around three, indicating anecdotal to moderate evidence in favor of the null hypothesis (see Jeffreys, 1961, for a categorization of the evidential strength provided by the Bayes factor). For “bad” evaluations, the Bayes factor initially indicated moderate evidence in favor of the null hypothesis. However, after testing 27 participants per condition, the Bayes factor gradually decreased, and indicated only anecdotal evidence in favor of the null hypothesis at the end of the data collection.

Exploratory Analysis

To probe the robustness of our conclusions, we varied the shape of the prior for the effect size under the alternative hypothesis. **Figures 4, 5** show the log Bayes factor as a function of the scale parameter r of the Cauchy prior for the “good” and “bad” evaluations, respectively. The dot indicates the result from the default prior used in the preregistered data analysis. It is evident that, as the scale parameter r increases (i.e., the prior becomes progressively wider), the Bayes factor increasingly favors the null hypothesis. In addition, it is evident that, even under the prior setting that favors the alternative hypothesis most (i.e., scale parameter r very close to zero), the log Bayes factor is close to zero indicating ambiguous evidence.

For completeness, we also analyzed the data using a frequentist repeated measures 2 (Evaluation: “Good” vs. “Bad”) × 2

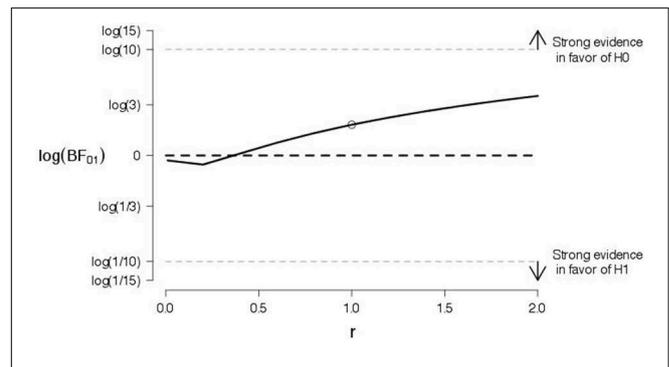


FIGURE 5 | A robustness analysis for the data of bad judgments from Experiment 1. The log Bayes factor $\log(BF_{01})$ is plotted as a function of the scale parameter r of the Cauchy prior for the effect size under the alternative hypothesis. The dot indicates the result from the default prior, the horizontal black dashed line indicates complete ambiguous evidence, and the horizontal dashed gray lines indicate strong evidence either in favor of the null hypothesis [i.e., $\log(BF_{01}) \geq \log(10)$] or in favor of the alternative hypothesis [i.e., $\log(BF_{01}) \leq \log(1/10)$].

(Instruction: Congruent vs. Incongruent) analysis of variance (ANOVA). Although congruent lever movements were faster ($M = 1176$ ms, $SE = 31$) than incongruent lever movements ($M = 1216$ ms, $SE = 33$), this difference did not reach significance [$F_{(1, 97)} < 1$, *n.s.*]. Additionally, “good” evaluations were given faster than “bad” evaluations [$F_{(1, 97)} = 36.52$, $p < 0.005$, $\eta_p^2 = 0.274$; $M_{good} = 1156$ ms, $SE_{good} = 21$; $M_{bad} = 1236$ ms, $SE_{bad} = 26$]. This main effect of judgment was the opposite of that obtained by CB (i.e., “bad” evaluations were faster than “good” evaluations). As shown in **Table 1**, this main effect of evaluative judgment was qualified by a two-way interaction between evaluative judgment and lever movement that almost reached the 0.05 level of significance [$F_{(1, 97)} = 3.02$, $p = 0.085$; $\eta_p^2 = 0.030$]: Pulling the lever with “good” evaluations was somewhat faster than pushing [$F_{(1, 97)} = 0.18$, *n.s.*], whereas pulling the lever with negative words was somewhat slower than pushing [$F_{(1, 97)} = 1.38$, $p = 0.243$], perhaps providing a weak indication of congruency (i.e., the alternative hypothesis) as obtained in CB. No other effects reached the 0.10 level of marginal significance.

The importance of the Two-Way interaction was also assessed with the help of a Bayesian ANOVA (Rouder et al., 2012) with participants as a random factor, which equals a repeated measures ANOVA. The Bayes factor of interest contrasts the full model that includes both the main effects and the interaction to a simpler model that includes only the main effects. The Bayes factor indicates that the data support the two models to an equal extent (i.e., $BF_{01} = 1.20$).

In sum, our preregistered Bayesian hypothesis tests yielded evidence in favor of the null hypothesis, although the strength of this evidence was not compelling. The exploratory Bayesian ANOVA suggested that the data do not favor the alternative hypothesis over the null hypothesis. From a Bayesian perspective, the data certainly did not support the hypothesis as proposed by CB although our experiment included almost twice as many participants ($n = 52$ in Experiment 1 of CB). Our *post-hoc* frequentist data analysis, however, did indicate some weak evidence

in favor of the alternative hypothesis, consistent with the original CB findings. Although we did not find a general congruency effect as reported by CB we did obtain an interaction indicative of a similar congruency between affective evaluation and lever movement.

The discrepancy between the outcome of the frequentist and Bayesian hypothesis tests arguably reflects the shortcomings of p -value based null hypothesis significance testing. Despite its widespread use, most psychologists fail to recognize that p -values overestimate the amount of statistical evidence against the null hypothesis (e.g., Berger and Delampady, 1987; Wagenmakers, 2007; Wetzels et al., 2011; Johnson, 2013). When researchers compute p -values, they only consider the plausibility of the data given the null hypothesis and ignore the possibility that the data may be similarly implausible given the alternative hypothesis (Berkson, 1938; Wagenmakers et al., 2015). Note also that, contrary to Bayes factors, p -values cannot be used to quantify evidence in favor of the null hypothesis; within the frequentist framework one can only fail to reject the null. The fact that Bayes factors can be used to obtain evidence for the absence of a hypothesized effect makes Bayes inference particularly useful for assessing the success of replication studies.

In comparison with the (corrected for publication bias) small to medium sized effect size reported in Phaf et al. (2014) the effect size in this experiment seems very low even though we tested almost twice as many participants than CB did. **Figure 6** shows the posterior distribution of the effect size for the two pre-registered comparisons (i.e., for “good” evaluations in the left panel, and for “bad” evaluations in the right panel). The posterior distribution quantifies the uncertainty about the effect size given the observed data. It is evident, that in the case of the “good” evaluations, most posterior mass is around 0; in the case of the “bad” evaluations, the posterior distribution is slightly shifted to positive values.

Experiment 2

Method

Pre-Registered Sampling Plan

A frequentist analysis would start with an assessment of the effect size of Experiment 2 from CB which would then form the basis of a power analysis. As for Experiment 1, however, our analysis is based on monitoring the Bayes factors of the critical hypothesis tests (detailed below).

Specifically, our sampling plan was as follows: We first set out to collect a minimum of 30 participants in a within-subject design. Next we planned to monitor the Bayes factors and stop the experiment whenever both critical hypothesis tests (detailed below) reached a Bayes factor that could be considered “strong” evidence (Jeffreys, 1961); this means that the Bayes factor should be either 10 in favor of the null hypothesis, or 10 in favor of the alternative hypothesis. The experiment would also stop whenever we would first reach the maximum number of participants, which we set to 50 participants. Finally, the experiment would also stop on January 1st, 2014, in case neither of the two other criteria had been met. As was the case for Experiment 1, data collection for Experiment 2 was terminated because the maximum number of participants was reached.

Participants

We recruited 56 students from the University of Amsterdam. Six participants were excluded for the following reasons: three participants did not operate the lever as instructed; two participants did not receive the correct instructions due to technical failure; and one left-handed participant was excluded because the experimental setup was not positioned correctly (i.e., at the left side). The remaining 50 participants (10 male, mean age = 21.3 year, $SD = 3.5$) were all native Dutch speakers and had not participated in Experiment 1. Participants were rewarded with course credits or €5. All participants were informed about the procedure with

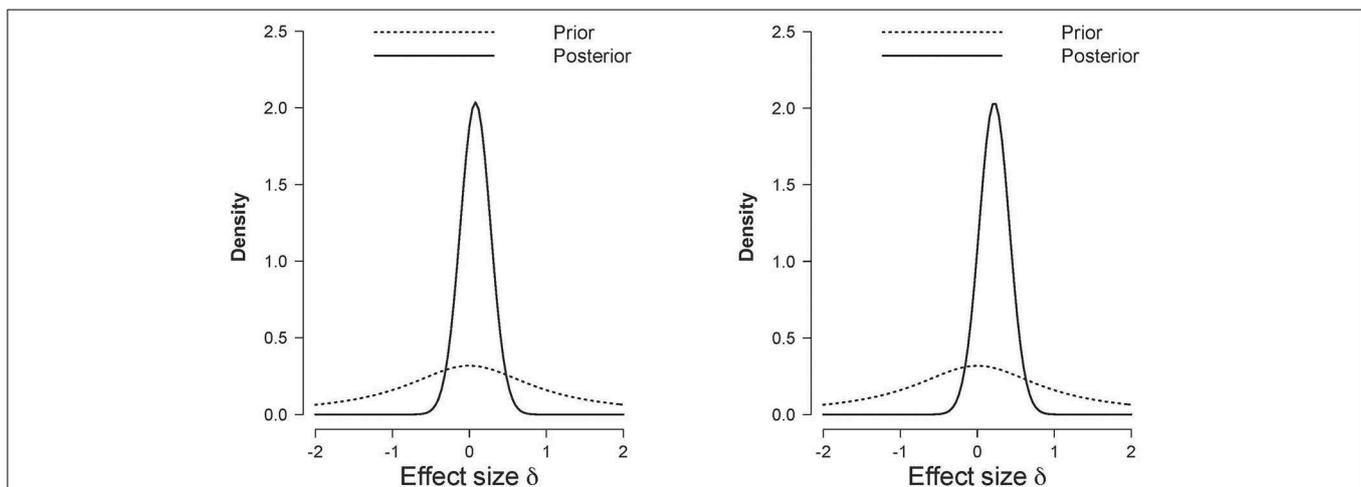


FIGURE 6 | Prior (dashed line) and posterior (solid line) distribution of the effect size δ in Experiment 1 for positive judgments (left panel) and negative judgments (right panel). The horizontal black dashed line

indicates complete ambiguous evidence, and the horizontal dashed gray lines indicate strong evidence either in favor of the null hypothesis [i.e., $\log(BF_{01}) \geq \log(10)$] or in favor of the alternative hypothesis [i.e., $\log(BF_{01}) \leq \log(1/10)$].

an information brochure and subsequently signed an informed consent form.

Materials and Procedure

The same materials used in Experiment 1 were also used in Experiment 2, except that “worms” was now included in the stimulus set, resulting in a total of 78 targets (i.e., 39 positive targets and 39 negative targets). The procedure differed only with respect to instructions given (See Appendix B in Supplementary Material, instructions a and b). In contrast to Experiment 1, participants were not instructed to evaluate the targets affectively. Instead, and in accordance with CB, participants were told that the experiment was about responding as quickly as possible to the mere presentation of the words. We alternately assigned participants to a condition in which they either always pushed the lever away from themselves (Instruction a) or always pulled the lever toward themselves (Instruction b). To discourage anticipation based on timing, targets were presented after a random delay from 2 to 7 s. The original study by CB did not include such random delays. We added this feature to the design in order to prevent timed responding based on fixed delays.

After half of the trials had been completed, a text appeared on screen to inform the participants that instructions would now change and that they had to switch lever movement direction from pushing to pulling (or vice versa). Additionally, the experimenter returned to the room to explain the new instructions and to ensure that the participants had understood them. Across all participants in CB’s Experiment 2, the targets were presented in a fixed order. Although not explicated in CB, this may have been done to ensure the presence of an equal number of positive and negative objects as well as an equal number of weak attitude objects in both conditions (for details see CB). We followed these constraints but presented our targets in a semi-random fashion; targets were randomly drawn without replacement from two different lists containing 19 positive targets and 20 negative targets or 20 positive targets and 19 negative targets, respectively. For every participant the order of both lists was the same.

Preregistered Data Analysis and Presentation of Results

Our data analysis closely followed that of Experiment 1, the main exception being that the design was fully within-subjects instead of between-subjects with regard to the association between affective valence of the targets and specific lever movement. As outlined in the preregistration document, we followed the reasoning of CB and treated response times above 1500 ms and below 300 ms as outliers, and excluded them from the analysis (pulling with positive words: 2.4%; pulling with negative words: 1.2%; pushing with positive words: 1.6%; pushing with negative words: 1.7%). As in Experiment 1, and as outlined in the preregistration document, the dependent measure was the mean \log_{10} -transformed response latency for every participant. The crucial hypothesis (i.e., alternative hypothesis) concerned the interaction that describes the congruency effect. Specifically, the congruency effect can be decomposed in two directional hypotheses: the first hypothesis states that participants respond faster to a positive

target by pulling instead of pushing a lever; the second hypothesis states that participants respond faster to a negative target by pushing instead of pulling a lever. Both hypotheses were assessed separately by means of two default Bayes factors for paired, one-sided *t*-tests as outlined in Rouder et al. (2009) and Wetzels et al. (2009). Specifically, for the distribution for effect size under the alternative hypothesis we used a folded Cauchy(0,1) distribution. Exploratory analyses will vary the shape of this prior to probe the robustness of our conclusions.

Bayes factors quantify the support that the data provide for the null hypothesis vis-a-vis the alternative hypothesis. Support in favor of the alternative hypotheses constitutes support in favor of the effects reported by CB in their Experiment 2.

Results

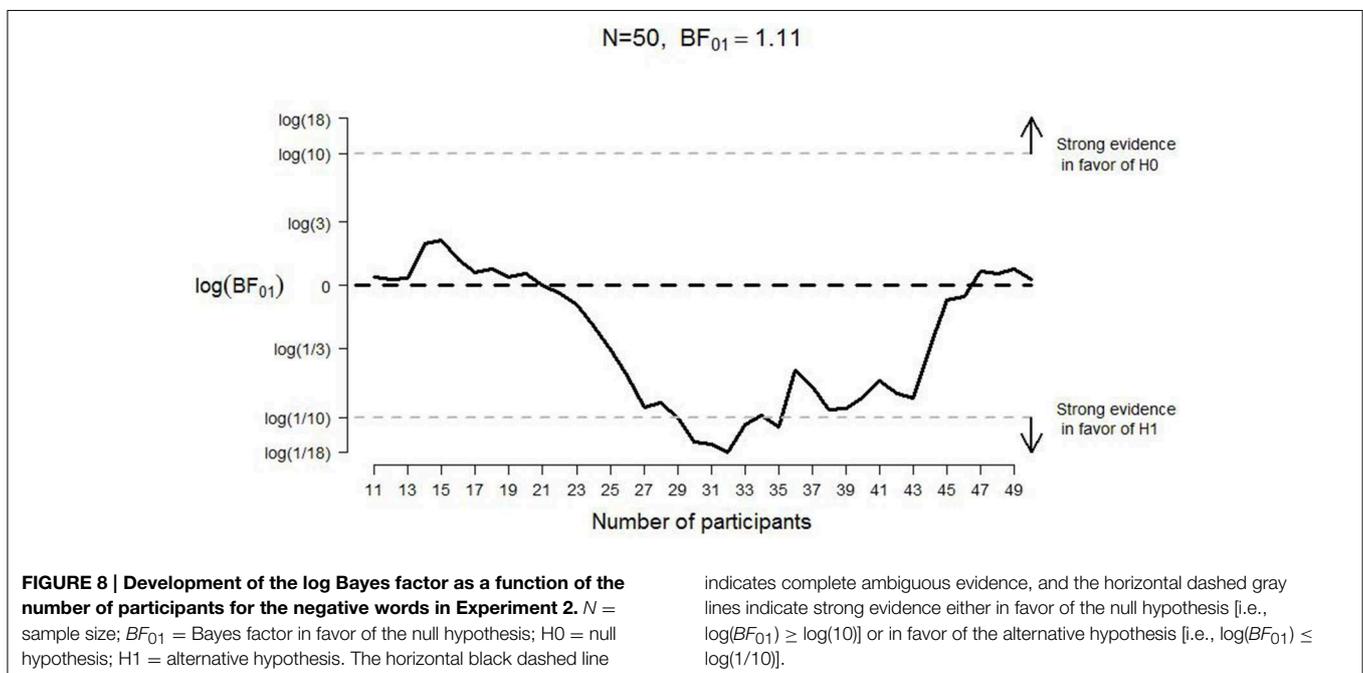
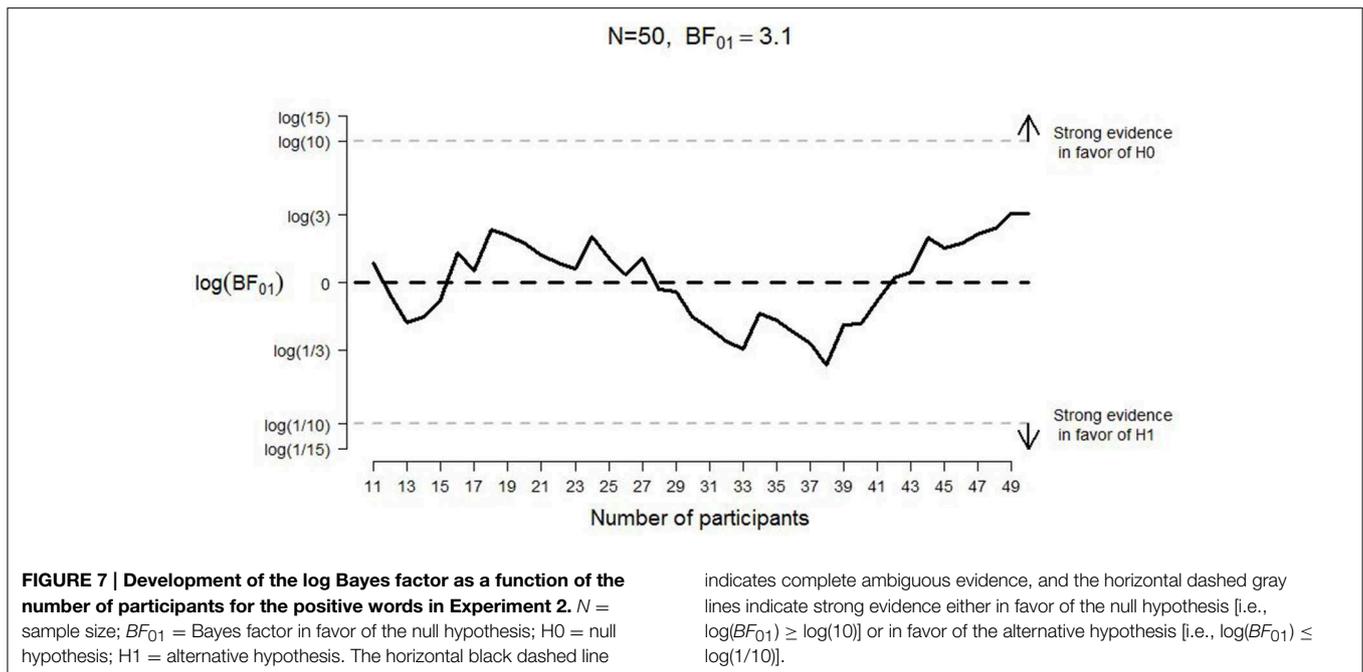
Bayes Factor Hypothesis Tests

As **Table 1** shows, for positive words participants were somewhat faster to pull ($M = 562$, $SE = 13$) the lever than to push it ($M = 571$, $SE = 14$). For negative words participants were somewhat faster to push the lever ($M = 562$, $SE = 13$) than to pull it ($M = 574$, $SE = 12$). For positive words, the comparison of the two lever movements yielded $BF_{01} = 3.10$; for negative words, it yielded $BF_{01} = 1.11$. In other words, for both positive and negative words we obtained “anecdotal” evidence (Jeffreys, 1961) in favor of the null hypothesis: the data are only about twice as likely under the null hypothesis as under the alternative hypothesis. **Figures 7, 8** display the development of the log Bayes factor as a function of the number of participants for the positive words and the negative words, respectively. For positive words, the Bayes factor fluctuated heavily throughout the experiment, sometimes indicating anecdotal evidence in favor of the null hypothesis, and sometimes indicating anecdotal to moderate evidence for the alternative hypothesis. At the end of the data collection, the Bayes factor indicated moderate evidence for the absence of the congruency effect. For negative words, the Bayes factor initially indicated moderate to strong evidence in favor of the alternative hypothesis. However, after testing 33 participants, the Bayes factor support in favor of the alternative hypothesis started to lessen; at the end of data collection, the Bayes factor indicated that the evidence is almost perfectly ambiguous.

Exploratory Analysis

To probe the robustness of our conclusions, we varied the shape of the prior for the effect size under the alternative hypothesis. **Figures 9, 10** show the log Bayes factor as a function of the scale parameter r of the Cauchy prior for the positive and negative words, respectively. The dot indicates the result from the default prior used in the preregistered data analysis. It is evident that, as the scale parameter r increases (i.e., the prior becomes progressively wider), the Bayes factor increasingly favors the null hypothesis. In addition, it is evident that, even under the prior setting that favors the alternative hypothesis most (i.e., scale parameter r very close to zero), the log Bayes factor is close to zero indicating ambiguous evidence.

For completeness, we also analyzed the data using a frequentist ANOVA (**Table 1**). Congruent lever movements were somewhat faster ($M = 562$ ms, $SE = 13$) than incongruent lever



movements [$M = 573$ ms, $SE = 13$; $t_{(49)} = 1.713$, $p = 0.093$]. In line with this observation, a repeated measures 2 (Affective valence: Positive vs. Negative) \times 2 (Lever movement: Pull vs. Push) ANOVA indicated a marginally significant Two-Way interaction [$F_{(1, 49)} = 2.93$, $p = 0.09$, $\eta_p^2 = 0.057$]: Pulling the lever with positive words was a little faster [$t_{(49)} = 1.054$, $p = 0.297$.] than pushing, whereas pulling the lever with negative words [$t_{(49)} = 1.742$, $p = 0.088$] was slower than pushing. This result perhaps provides a weak indication of congruency

(i.e., the alternative hypothesis) as obtained in CB's Experiment 2. No other effects reached the 0.10 level of marginal significance.

The importance of the Two-Way interaction was also assessed with the help of a Bayesian ANOVA (Rouder et al., 2012) with participants as a random factor, which equals a repeated measures ANOVA. The Bayes factor of interest contrasts the full model that includes both the main effects and the interaction to a simpler model that includes only the main effects. The Bayes factor slightly favored the model without the interaction term (i.e.,

$BF_{01} = 4.28$), that is, the observed data are 4.28 times more likely under the model without the interaction compared to the model with the interaction.

In sum, as was the case for Experiment 1, our preregistered Bayesian hypothesis tests yielded evidence in favor of the null hypothesis, although the strength of this evidence was rather modest. **Figure 11** shows the posterior distribution of the effect size for the two preregistered comparisons (i.e., for positive words in the left panel, and for negative words in the right panel). It is evident, that in the case of both hypotheses, the posterior distribution is only slightly shifted to positive values. From a Bayesian perspective, the data certainly did not support the CB hypotheses. Our exploratory frequentist data analysis indicated some weak evidence in line with the original CB findings. This finding,

however, is in contrast with the results obtained in the meta-analyses reported in Phaf et al. (2014, see also Laham et al., 2014) that seems to suggest the absence of any automatic congruency effect.

Discussion and Concluding Comments

Our attempts to replicate the CB experiments did not succeed: for both replication attempts, the preregistered Bayesian hypothesis tests showed that the data provided more evidence for the null hypotheses than for the alternative hypotheses. The strength of this evidence is certainly not compelling, but the results do suggest that additional direct preregistered replications of the CB experiments are called for.

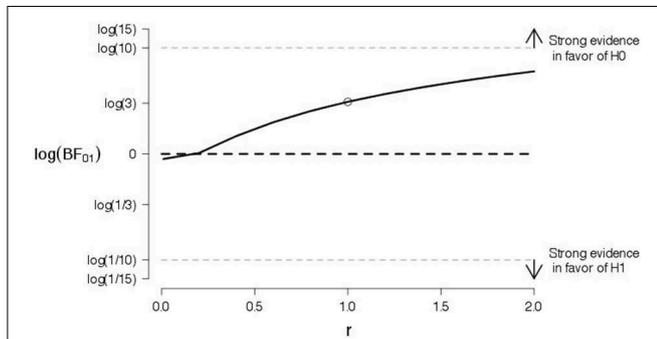


FIGURE 9 | A robustness analysis for the data of positive words from Experiment 2. The log Bayes factor $\log(BF_{01})$ is plotted as a function of the scale parameter r of the Cauchy prior for the effect size under the alternative hypothesis. The dot indicates the result from the default prior, the horizontal black dashed line indicates complete ambiguous evidence, and the horizontal dashed gray lines indicate strong evidence either in favor of the null hypothesis [i.e., $\log(BF_{01}) \geq \log(10)$] or in favor of the alternative hypothesis [i.e., $\log(BF_{01}) \leq \log(1/10)$].

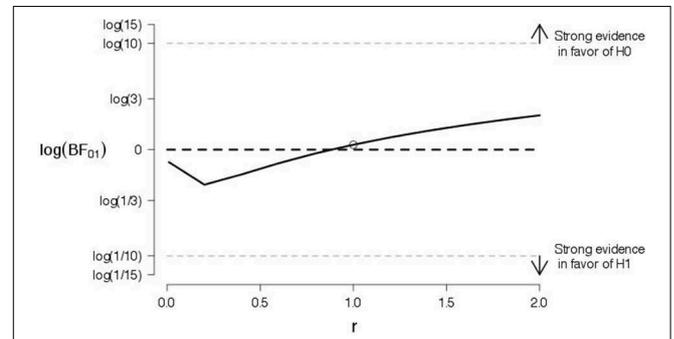


FIGURE 10 | A robustness analysis for the data of negative words from Experiment 2. The log Bayes factor $\log(BF_{01})$ is plotted as a function of the scale parameter r of the Cauchy prior for the effect size under the alternative hypothesis. The dot indicates the result from the default prior, the horizontal black dashed line indicates complete ambiguous evidence, and the horizontal dashed gray lines indicate strong evidence either in favor of the null hypothesis [i.e., $\log(BF_{01}) \geq \log(10)$] or in favor of the alternative hypothesis [i.e., $\log(BF_{01}) \leq \log(1/10)$].

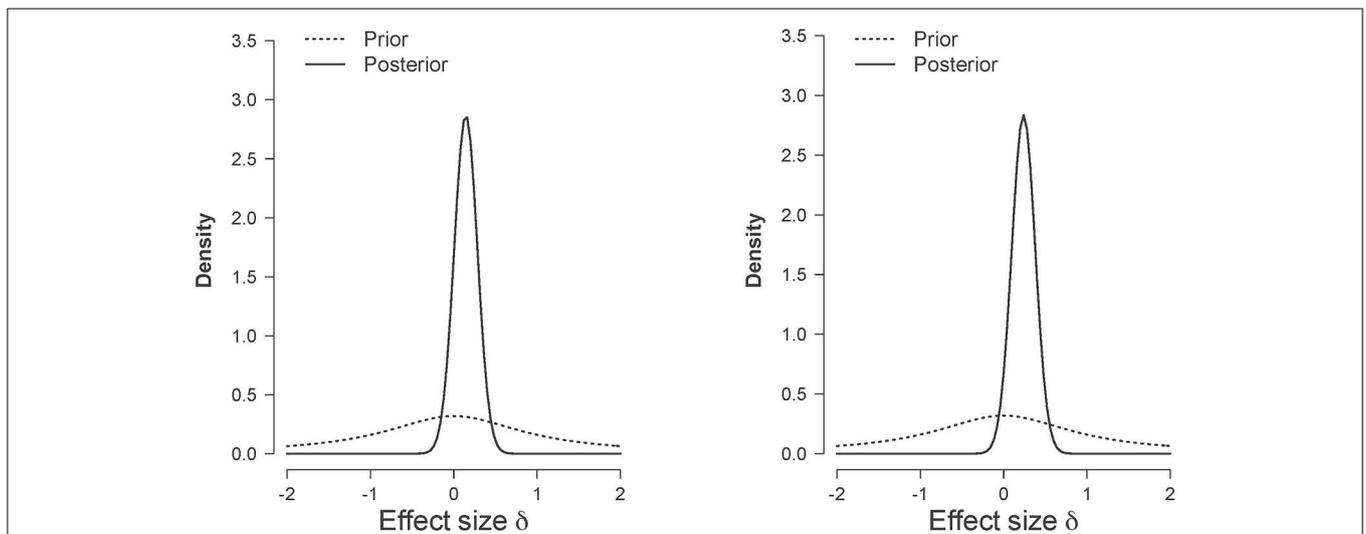


FIGURE 11 | Prior (dashed line) and posterior (solid line) distribution of the effect size δ in Experiment 2 for positive words (left panel) and negative words (right panel).

Using a frequentist ANOVA, exploratory analyses of Experiment 1 and Experiment 2 revealed a weak indication for congruency, expressed through interaction effects that were both marginally significant. These results were not, however, corroborated by a Bayesian ANOVA, which again provided weak evidence in favor of the absence of an interaction. This inconsistency arises as a result of the statistical peculiarities of p -values. As explained earlier, p -value-based inference overstates the amount of evidence against the null hypothesis because it fails to consider the extremeness of the data under the alternative hypothesis. Nevertheless, the evidence for the presence of the effects is weak to non-existent, even if we focus on the most favorable analysis (i.e., the exploratory frequentist ANOVA producing marginally significant results).

Although we attempted to duplicate the original experimental setup as accurately as possible there were of course small differences in the experimental setup and procedure that can maybe account for the differences in results. First, in our experimental setup participants had to move their hand a bit more than in the original setup. This could have resulted in less easy movements, for instance, maybe interfering with the congruency effect. When trying out both trajectories ourselves though we could not feel any more interference in the longer one we used than in the shorter one used by CB; hence, we do not believe this difference can explain the discrepant results. Moreover, latencies obtained in our experiment were faster than the original latencies suggesting that when any of such interference took place it certainly did not slow down our participants. Second, we used fewer words in our experiments (resp., 77 and 78 out of 92) than CB (i.e., 82 out of 92) assuming that targets used for practice trials in the original experiment were not used in the actual experiments and reported results. Of course this difference was due to our efforts to get our stimulus set to resemble the original stimulus set used by CB as close as possible so we do not think this difference can explain the discrepant results either. Third, in our experiments targets were randomly presented (Experiment 1) as well as semi-randomly (Experiment 2) whereas the original authors used a single randomly ordered list of words in both experiments. If this difference could explain the discrepant results we should probably conclude that the original findings are due to experimental noise alone but this would contrast again with our findings and the findings in Phaf et al. (2014, see also Laham et al., 2014). In sum, we do not think that these differences in experimental setup and procedure can account for the differences in results obtained

by us and CB. But we are aware that we of course do not know all differences since we could rely only on specifics provided in the original report.

It seems clear that although we failed to replicate CB using our pre-registered Bayesian analyses we cannot conclude that there is no link between affective evaluation and approach-avoidance behavior. First, the evidence in favor of the null hypothesis is not compelling and stems from only two experiments. Second, using exploratory frequentist statistics we found weak evidence for this link. Third, a recent meta-analysis (Phaf et al., 2014; see also Laham et al., 2014) shows ample evidence for the presence of this link. Phaf et al. included studies on approach and avoidance behavior using different experimental paradigms (e.g., joystick, manikin) and, after correcting for publication bias, reported evidence for the presence of a moderate congruency effect between explicit affective evaluation and approach and avoidance behavior. Additionally, no evidence was obtained in this meta-analysis for such a direct link in case of implicit affective evaluation as in our Experiment 2.

In addition, we were also unable to replicate the main effect of evaluative judgment (Experiment 1) and affective meaning (Experiment 2) that was obtained by CB. In Experiment 1 of CB, negative evaluations were faster than positive evaluations; in Experiment 2 of CB, participants responded faster to negative than to positive words. In contrast, in our Experiment 1 we found that positive evaluations were faster than negative evaluations; in our Experiment 2, there was no evidence that affective meaning influenced lever movement in the absence of explicit affective evaluation. This finding and the aforementioned findings suggests that the pattern of results obtained by CB may be more fragile than previously thought.

Acknowledgments

This work was funded in part by the ERC grant “Bayes or Bust.” We thank Bert Molenkamp and Coos Hakvoort for their help with the hardware and software for the experiments.

Supplementary Material

The Supplementary Material for this article can be found online at: <http://www.frontiersin.org/journal/10.3389/fpsyg.2015.00335/abstract>

References

- Bamford, S., and Ward, R. (2008). Predispositions to approach and avoid are contextually sensitive and goal dependent. *Emotion* 8, 174–183. doi: 10.1037/1528-3542.8.2.174
- Bargh, J. A., Chaiken, S., Gendler, R., and Pratto, F. (1992). The generality of the automatic attitude activation effect. *J. Pers. Soc. Psychol.* 62, 893–912. doi: 10.1037/0022-3514.62.6.893
- Berger, J. O., and Delampady, M. (1987). Testing precise hypotheses. *Stat. Sci.* 2, 317–352. doi: 10.1214/ss/1177013238
- Berger, J. O., and Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *J. Am. Stat. Assoc.* 94, 542–554. doi: 10.1080/01621459.1999.10474149
- Berger, J. O., and Wolpert, R. L. (1988). *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics.
- Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *J. Am. Stat. Assoc.* 33, 526–536. doi: 10.1080/01621459.1938.10502329
- Cacioppo, J. T., and Gardner, W. L. (1999). Emotion. *Annu. Rev. Psychol.* 50, 191–214. doi: 10.1146/annurev.psych.50.1.191
- Chambers, C. D. (2013). Registered reports: a new publishing initiative at cortex. *Cortex* 49, 609–610. doi: 10.1016/j.cortex.2012.12.016
- Chen, M., and Bargh, J. A. (1999). Consequences of automatic evaluation: immediate behavioral predispositions to approach or avoid the stimulus. *Pers. Soc. Psychol. Bull.* 25, 215–224. doi: 10.1177/0146167299025002007

- De Groot, A. D. (1956/2014). The meaning of “significance” for different types of research. [Translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angélique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas.] *Acta Psychol.* 148, 188–194. doi: 10.1016/j.actpsy.2014.02.001
- De Houwer, J., Crombez, G., Baeyens, F., and Hermans, D. (2001). On the generality of the affective Simon effect. *Cogn. Emot.* 15, 189–206. doi: 10.1080/02699930125883
- Duckworth, K. L., Bargh, J. A., Garcia, M., and Chaiken, S. (2002). The automatic evaluation of novel stimuli. *Psychol. Sci.* 13, 513–519. doi: 10.1111/1467-9280.00490
- Eder, A. B., and Rothermund, K. (2008). When do motor behaviors (mis)match affective stimuli? An evaluative coding view of approach and avoidance reactions. *J. Exp. Psychol.* 137, 262–281. doi: 10.1037/0096-3445.137.2.262
- Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* 70, 193–242. doi: 10.1037/h0044139
- Francis, G. (2013). Replication, statistical consistency, and publication bias. *J. Math. Psychol.* 57, 153–169. doi: 10.1016/j.jmp.2013.02.003
- Frijda, N. H. (1986). *The Emotions*. Cambridge: Cambridge University Press.
- Frijda, N. H. (2007). *The Laws of Emotion*. Mahwah, NJ: Erlbaum.
- Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press.
- Johnson, V. E. (2013). Revised standards for statistical evidence. *Proc. Natl. Acad. Sci. U.S.A.* 110, 19313–19317. doi: 10.1073/pnas.1313476110
- Krieglmeyer, R., and Deutsch, R. (2010). Comparing measures of approach avoidance behaviour: the manikin task vs. two versions of the joystick task. *Cogn. Emot.* 24, 810–828. doi: 10.1080/02699930903047298
- Laham, S. M., Kashima, Y., Dix, J., and Wheeler, M. (2014). A meta-analysis of the facilitation of arm flexion and extension movements as a function of stimulus valence. *Cogn. Emot.* doi: 10.1080/02699931.2014.968096. [Epub ahead of print].
- Lang, P. J., and Bradley, M. M. (2008). “Appetitive and defensive motivation is the substrate of emotion,” in *Handbook of Approach and Avoidance Motivation*, ed A. J. Elliot (New York, NY: Psychology Press), 51–65.
- Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychol. Rev.* 97, 377–395. doi: 10.1037/0033-295X.97.3.377
- Neumann, R., Förster, J., and Strack, F. (2003). “Motor compatibility: the bidirectional link between behavior and evaluation,” in *The Psychology of Evaluation*, eds J. Musch and K. C. Klauer (Mahwah, NJ: Lawrence Erlbaum Associates, Inc.), 371–391.
- Nosek, B., and Lakens, D. (2014). Registered reports: a method to increase the credibility of published results. *Soc. Psychol.* 45, 137–141. doi: 10.1027/1864-9335/a000192
- Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. Psychol. Sci.* 7, 657–660. doi: 10.1177/1745691612462588
- Pashler, H., and Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspect. Psychol. Sci.* 7, 528–530. doi: 10.1177/1745691612465253
- Pashler, H., and Wagenmakers, E.-J. (2012). Editors’ introduction to the special section on replicability in psychological science: a crises of confidence? *Perspect. Psychol. Sci.* 7, 528–530. doi: 10.1177/1745691612465253
- Phaf, R. H., Mohr, S. E., Rotteveel, M., and Wicherts, J. (2014). Approach, avoidance, and affect: a meta-analysis of approach-avoidance tendencies in manual reaction time tasks. *Front. Psychol.* 5:378. doi: 10.3389/fpsyg.2014.00378
- Rinck, M., and Becker, E. S. (2007). Approach and avoidance in fear of spiders. *J. Behav. Ther. Exp. Psychiatry* 38, 105–120. doi: 10.1016/j.jbtep.2006.10.001
- Rosenthal, R. (1979). An introduction to the file drawer problem. *Psychol. Bull.* 86, 683–641. doi: 10.1037/0033-2909.86.3.638
- Rotteveel, M., and Phaf, R. H. (2004). Automatic affective evaluation does not automatically predispose for arm flexion and extension. *Emotion* 4, 156–172. doi: 10.1037/1528-3542.4.2.156
- Rouder, J. N. (2014). Optional stopping: no problem for Bayesians. *Psychon. Bull. Rev.* 21, 301–308. doi: 10.3758/s13423-014-0595-4
- Rouder, J. N., Morey, R. D., Speckman, P. L., and Province, J. M. (2012). Default bayes factors for ANOVA designs. *J. Math. Psychol.* 56, 356–374. doi: 10.1016/j.jmp.2012.08.001
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16, 225–237. doi: 10.3758/PBR.16.2.225
- Simons, D. J. (2014). The value of direct replication. *Perspect. Psychol. Sci.* 9, 76–80. doi: 10.1177/1745691613514755
- Solarz, A. K. (1960). Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *J. Exp. Psychol.* 59, 239–245. doi: 10.1037/h0047274
- Strack, F., and Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Pers. Soc. Psychol. Rev.* 8, 220–247. doi: 10.1207/s15327957pspr0803_1
- Stroebel, W., and Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* 9, 59–71. doi: 10.1177/1745691613514450
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev.* 14, 779–804. doi: 10.3758/BF03194105
- Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Bakker, M., Lee, M., Matzke, D., et al. (2015). A power fallacy. *Behav. Res. Methods*. doi: 10.3758/s13428-014-0517-4. [Epub ahead of print].
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., and Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* 7, 632–638. doi: 10.1177/1745691612463078
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., and Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: an empirical comparison using 855 t tests. *Perspect. Psychol. Sci.* 6, 291–298. doi: 10.1177/1745691611406923
- Wetzels, R., Raaijmakers, J. G. W., Jakab, E., and Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: a flexible WinBUGS implementation of a default Bayesian t test. *Psychon. Bull. Rev.* 16, 752–760. doi: 10.3758/PBR.16.4.752
- Wolfe, J. M. (2013). Registered reports and replications in Attention, Perception, and Psychophysics. *Atten. Percept. Psychophys.* 75, 781–783. doi: 10.3758/s13414-013-0502-5

Conflict of Interest Statement: 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 © 2015 Rotteveel, Gierholz, Koch, van Aalst, Pinto, Matzke, Steingroever, Verhagen, Beek, Selker, Sasiadek and Wagenmakers. 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) or licensor 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.