Journal of Articles in Support of the Null Hypothesis
Vol. 14, No. 2
 1539-8714
www.jasnh.com

Incidental Haptic Sensations May Not Influence Social Judgments:

A Purely Confirmatory Replication Attempt of Study 1 by Ackerman, Nocera, and Bargh (2010)

Titia F. Beek
Dora Matzke
Yair Pinto
Mark Rotteveel
Alexander Gierholz
Josine Verhagen
Ravi Selker
Adam Sasiadek
Helen Steingroever
Nils B. Jostmann
Eric-Jan Wagenmakers
University of Amsterdam

This preregistered replication attempt focuses on the finding from Ackerman, Nocera, and Bargh (2010; ANB) that holding a heavy object triggers concepts related to importance. ANB reported that participants who were holding a heavy clipboard rated a job candidate as better overall and more seriously interested in the job than participants holding a light clipboard. We failed to replicate ANB's results; instead, Bayes factor hypothesis tests indicated evidence for the absence of a difference between the heavy and the light condition in the overall evaluation and perceived seriousness of the candidate, and in participants' perceived task importance. Our findings highlight the importance of conducting preregistered replication research and illustrate the theoretical and practical advantages of Bayesian inference in psychological research.

Introduction

Do incidental haptic sensations influence social judgments and decisions? When we interact with a heavy object, do ideas of importance become activated? Does holding a heavy object influence our evaluations, even if the heavy object is not what we are evaluating? These questions were addressed by Ackerman, Nocera, and Bargh (2010; henceforth ANB) who argued that holding objects triggers the application of associated concepts.

ANB's research questions were inspired by embodiment theories (e.g., Barsalou, 2008) which hold that cognitive representations are grounded in the brain's sensorimotor systems (e.g., Jostmann, Lakens, & Schubert, 2009). An empirical prediction from this is that cognitive states (e.g., feeling confident) can trigger corresponding bodily states (e.g., walking upright) and vice versa (Barsalou, 2008). In support of embodiment theories, previous work has shown that stimulating facial muscles to facilitate smiling can induce positive affect (e.g., Strack, Martin, & Stepper, 1988; but see Wagenmakers et al., 2016), whereas having people make a pushing-away movement with their arms can induce negative affect (e.g., Cacioppo, Priester, & Berntson, 1993; but see Rotteveel et al., 2015). Furthermore, embodiment theories suggest that haptic information can also become linked to more abstract concepts, such as importance and seriousness. This is thought to originate from early physical interactions with one's environment, when children use their hands to acquire information and manipulate their world (Ackerman et al., 2010; Barsalou, 2008). These sensorimotor experiences are thought to provide a scaffold for the development of conceptual knowledge (Ackerman et al., 2010). Heavy objects have more impact on our bodies than light objects, leading to an association between heaviness and importance (Jostmann et al., 2009). Informal support for the association between heaviness and importance comes from metaphors such as "thinking about weighty matters," "gravity of the situation," and "weighing the pros and cons of a decision" (Ackerman et al., 2010; Chandler, Reinhard, & Schwarz, 2012; Jostmann et al., 2009; Maglio & Trope, 2012). Later in life, a certain bodily state can trigger or prime people with a related concept; for example, touching something heavy can trigger associations of importance (Jostmann et al., 2009). The reader is referred to Reimann et al. (2012) for an overview of research on embodiment in judgment and choice.

This link between heaviness and importance was the focus of ANB's study. In particular, ANB set out to demonstrate that heaviness can trigger the concept of importance and that weight can unconsciously influence evaluations of a person. In Experiment 1, participants evaluated a job candidate for a postdoctoral position by reviewing resumes presented on either a light (340.2 grams) or a heavy (2041.2 grams) clipboard. ANB hypothesized that the weight of the clipboard unconsciously activates the associated concept of importance (e.g., "thinking about weighty matters"). In line with the authors' prediction, the results showed that participants holding a heavy clipboard rated the job candidate as better overall and displaying more serious interest in the position than participants holding a light clipboard. In contrast, participants with heavy clipboards did not rate the job candidate as more likely to "get along" with co-workers, presumably because the social dimension is not associated with "weighty" matters. In addition, participants with heavy clipboards rated their own accuracy on the task as more important, but did not report devoting more effort to the task than participants with light clipboards, suggesting that participants' impressions were not due to a self-perception effect (Ackerman et al., 2010).

The Current Study

The current study describes an attempt to replicate ANB's result (Study 1) that participants holding heavy clipboards rate a job candidate as better on aspects related to seriousness than participants holding a light clipboard. We believe that it is important to attempt and replicate ANB's results because the study has been cited many times (367 citations according to Google Scholar, 9 March 2017) and was published in the prestigious journal *Science*.

We tested the following hypotheses from ANB: (1) Participants holding a heavy clipboard evaluate the job candidate as better overall than participants holding a light clipboard; (2) participants holding a heavy clipboard judge the candidate as displaying more serious interest in the position than participants holding a light clipboard; (3) participants holding a heavy clipboard do not rate the job candidate as more likely to "get along" with co-workers than participants holding a light clipboard; (4) participants holding a heavy clipboard rate their own accuracy on the task as more important than participants with the light clipboard; and (5) participants holding a heavy clipboard do not report devoting more effort to the task than participants with the light clipboard. Hypotheses 1, 2, and 4 are the central hypotheses; these test the idea that heaviness triggers concepts related to importance and seriousness. Hypothesis 3 was included by ANB to demonstrate the specificity of their theory; heaviness is metaphorically linked *only* to concepts related to seriousness and importance, but not to other concepts, such as social likeability. Hypothesis 5 serves to rule out the possibility that – if participants in the heavy condition indeed rate their own accuracy as more important (Hypothesis 4) – these impressions are influenced by a self-perception effect; that is participants would see their own increased effort as indicative of participation in an especially important study (Ackerman et al., 2010).

We also attempted to replicate the well-established associative-priming effect using a visual lexical decision task (e.g., de Groot, 1984, 1987; Neely, 1967, 1977). This task was included to confirm the reliability of our experimental procedure, such as the statistical methodology and the seriousness of the participants. The task and the stimuli were provided by Matzke et al. (2015). As explained by Matzke and colleagues, the associative-priming tasks required participants to categorize letter strings as words or nonwords. Each stimulus (target word) was preceded by a prime word that is either a semantic associate of the target (e.g., dog – cat) or is unrelated to the target (e.g., uncle – cat). The dependent variable of interest was the mean response time (RT) for correct responses to target words. Effect sizes for associative priming are reported to vary considerably across studies, with an average Cohen's d of 0.59 (Lucas, 2000). Typically, mean correct RTs in the associative-priming task are faster for target words preceded by related primes than for target words preceded by unrelated primes, which constitutes Hypothesis 6 for the present study.

Prior to data collection, the hypotheses, the methods, and the analysis plan were pre-registered on the Open Science Framework (https://osf.io/fwhbu/; the preregistration was frozen on 2013-04-01). Pre-registration forces researchers to specify a priori which analyses are confirmatory and which ones are exploratory and may help to eliminate questionable research practices (QRPs; Bakker, van Dijk, & Wicherts, 2012; Kerr, 1998, MacCallum, Roznowski, & Necowitz, 1992, Nosek, Spies, & Motyl, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

Bayesian Statistics

In 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 a point has been proven or disproven. Neither desideratum can be accomplished within the framework of frequentist statistics; using *p* value null-hypothesis testing, one can only fail to reject a null hypothesis and there is no possibility to monitor the data and stop data collection when the evidence is sufficiently compelling (Wagenmakers, 2007). This is why our analysis relied on Bayesian hypothesis testing. In Bayesian statistics, competing hypotheses, in this case the null hypothesis H_0 and the alternative hypothesis H_+, are evaluated against each other; the hypothesis that predicts the observed data best is preferred. In our case, we want to know how well H_0 predicts the data in relation to H_+. Note that we use subscripts to indicate the directional nature of the alternative hypothesis: H_1 denotes an undirected hypothesis, H_- denotes the directed hypothesis that the effect is negative, and H_+ denotes the directed hypothesis that the effect is positive. The hypotheses under scrutiny clearly specifies a direction and therefore we use H_+ throughout.

The weight of the evidence provided by the data is given by a quantity known as the Bayes factor (BF; e.g., Berger & Mortera, 1999; Edwards, Lindman, & Savage, 1963; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007; Wagenmakers et al., 2012; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009). For instance, a BF_{0+} of 11 means that the observed data are 11 times more likely to have occurred under H_0 than under H_+; a BF_{0+} of .2 indicates that the observed data are $1/.2 = 5$ times more likely to have occurred under H_+ than under H_0. A Bayes factor hypothesis test thus prefers the hypothesis under which the observed data are most likely. In contrast to null hypothesis significance testing, with Bayes statistics researchers can quantify evidence for the null hypothesis.

Method

Participants

Participants were recruited via the University of Amsterdam research website (www.test.uva.nl) and were rewarded with course credits or 5 euros.

Materials

All (Dutch) materials can be found on the OSF project page at https://osf.io/nm9vb/.

Information brochure. The information brochure described the procedure and goal of the experiment. It read that the experiment consisted of four separate, unrelated tasks (only the first two are relevant for the present study) and that participants would receive information about the other tasks later. Furthermore, it was stated that the goal of the first (job candidate evaluation) task was twofold; to (1) look at the possible influence of body position on information processing; and (2) compare expert versus non-expert judgment. The first set of instructions was adapted from Jostmann et al. (2009) and ensured that participants stood upright. This was done because the original study by ANB took place outdoors and recruited passersby who were naturally in a standing position. The second set of instructions mimicked ANB. The instructions continued to state that participants

would read and evaluate the resume of a job applicant in one of four cubicles, to which they would be assigned by picking a closed envelope containing the number of one of the cubicles (1, 2, 3, or 4). Furthermore, participants read from the instructions that in this cubicle, they would complete the task while standing on a cross that was taped on the floor (see Figure 1).

Candidate Evaluation Task

We attempted to stay as close as possible to the original experiment; all materials we used (except the clipboards) were provided to us by ANB and were then translated to Dutch.

- *Information sheet.* This was the document on top of the stack attached to the clipboard. These additional instructions about the task were translations of ANB's instructions.
- *Clipboard.* We used a bottom-opening clipboard with a storage container. In the light condition, the storage container was empty and weighed 550 g. In the heavy condition, the storage container was filled with paper and weighed 2175 g. Note that our empty clipboard was about 210 g and our heavy clipboard was about 134 g heavier than ANB's clipboards.
- *Resume.* We made a few minor adaptations to ANB's original resume in order to make the evaluation task suitable and relevant for our Dutch sample. Firstly, we translated the resume to Dutch. Secondly, we changed the names of the Universities to Dutch Universities (e.g., University of Oregon vs. Universiteit Utrecht). Third, we changed the names of some of the colleagues to more Dutch-sounding names (e.g., Sara Hodgkiss vs. Sarah Hogenaar). Fourth, we changed some of the conference locations to Dutch cities (e.g., Memphis, TN vs. Utrecht). Fifth, we changed some of the professional memberships (e.g., American Psychological Association vs. Nederlands Instituut van Psychologen [the Dutch Institute of Psychologists]). Lastly, since the original study was conducted in 2010 and the current study was conducted in 2013, 3 years were added to every date mentioned in the original resume (e.g., 2005 vs. 2008). We believe that these small changes were necessary to create a realistic resume.
- *Evaluation scale.* We administered the same eight evaluation items used by ANB. Participants responded on a 7-point Likert scale, ranging from 1 (very negative) to 7 (very positive). An example item is: "What is your impression of the candidate's application materials?" ("Wat is jouw indruk van de sollicitatiematerialen van de sollicitant?"). We averaged six of these items to calculate the composite job candidate evaluation, as was done by ANB. The other two items were analyzed separately, again as was done by ANB: one served as measure for social compatibility with future colleagues, the other served as measure of serious interest of the candidate.
- *Perceived task importance.* Following ANB, perceived task importance was measured by asking participants: *"How important is it for you to make a correct evaluation?"* ("Hoe belangrijk is het voor jou om een correcte evaluatie te maken?"). Participants indicated their response on a 7-point Likert scale ranging from 1 (*not at all*) to 7 (*very*).
- *Perceived effort on task.* Following ANB and to rule out a possible self-perception effect, we asked participants the following question: "How much effort did you devote to the evaluation task?" ("Hoeveel moeite heb je gedaan om de evaluatietaak uit te voeren?"). Participants indicated this on a 7-point Likert scale ranging from 1 (*no*) to 7 (*very much*).

Associative-Priming Task

The associative-priming task was taken from Matzke et al. (2015). The detailed description of the task is available on the OSF project page.

Procedure

Participants arrived in a central hall that was connected to four cubicles. In the center of each cubicle, we taped a cross on the floor. As the cross was in the center of the cubicle, participants could not reach the desk, nor could they let the clipboard rest on it. The heavy (cubicles 1 and 2) and the light (cubicles 3 and 4) clipboards were already present in the room. All materials needed for the evaluation of the job applicant (instructions, resume, evaluation form, and pen) were attached to the clipboard (see Figure 1). When participants arrived in the central hall, the experimenter welcomed them and gave them the information brochure that participants read. Participants then had the opportunity to ask questions to the experimenter, signed the informed consent and were invited to choose an envelope from a randomly shuffled set. Because the assignment to the cubicles was carried out using closed, randomly shuffled envelopes, the experimenter was unaware of participants' condition at the time participants received the instructions. This way, we reduced the possibility of experimenter bias (Rosenthal, 1963). The envelope also contained the instructions to go to the cubicle and reminded the participants to pick up the evaluation materials from the desk and to complete the task while standing on the cross, facing the computer. During the experiment, the cubicle door remained open. While participants completed the task[1], the experimenter unobtrusively checked whether participants performed the experiment in a standing position. When participants were finished with the evaluation task, the experimenter gave the participants a chair and asked them to fill out the exit-interview asking them to guess the goal of the study. Participants then completed the associative-priming task. This was followed by two unrelated tasks, after which participants received their reward. Debriefing took place via e-mail, after the experimental period was over.

Preregistration

A preregistration document was submitted to OSF with information about the hypotheses, the materials and methods of the experiment (see above), and our sampling plan, stopping rule, exclusion criteria, and statistical analyses.

Sampling Plan

We set out to test a minimum of 20 participants in each between-subject condition (i.e., the light and the heavy condition), for a minimum of 40 participants in total. We would then monitor the Bayes factor and stop the experiment whenever all critical hypothesis tests (detailed above) reach a Bayes factor that can be considered "strong" evidence (Jeffreys, 1961); this means that the Bayes factor is either 10 in favor of H_0, or 10 in favor of H_+. The experiment would also stop whenever we reach the maximum number of participants, which we set to 50 participants per condition. Finally, the experiment would also stop on

1 The instructions upon entering the room were slightly changed after two days of testing to ensure that participants completed the task while standing *and* holding the clipboard, see the addendum on OSF for details (https://osf.io/7y9br/).

Figure 1. *The experimental set-up, with a picture one of the four cubicles (top left), a more detailed view inside the cubicle with the cross on the floor (top right), full (heavy) clipboard (bottom left), and an empty (light) clipboard (bottom right). Figure available at* https://flic.kr/p/299gWnr*, under CC license* https://creativecommons.org/licenses/by/2.0/

October 1st, 2013. Even though we preregistered a sampling plan, note that the sampling plan is irrelevant for Bayesian inference (e.g., Berger & Mortera, 1999; Berger & Wolpert, 1988; Rouder, 2014); as summarized by Edwards et al. (1963, p. 193): "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience."

Exclusion Criteria

We set out to exclude participants from the analysis if they met one or more of the following criteria: (1) They did not follow the instructions; (2) They did not stand up while holding the clipboard; (3) They closed the cubicle door so that the experimenter could not verify whether they completed the task standing up; (4) They let the clipboard rest on something (e.g., the desk) so that they did not carry its full weight; (5) They expressed suspicion about the hypotheses for the evaluation task in the exit interview (i.e., weight of the clipboard influences evaluation); (6) They had a mean RT longer than 1200 ms and error rate higher than 20% in the associative-priming task. In the associative-priming task, RTs for incorrect responses and RTs faster than 250 ms and slower than 1500 ms would be also excluded from the analyses.

Statistical Analyses

The evaluation form consisted of 10 questions; eight were related to the evaluation of the resume, and two related to the role of the participant him-or herself. We set out to average six of these items (items 1, 2, 5, 6, 7, and 8) to calculate the composite job candidate evaluation, as was done by ANB. The other two resume-related items (item 3 measuring social compatibility with future colleagues and item 4 measuring the seriousness of the candidate) would be analyzed separately. Similarly, the perceived-importance and perceived-effort items would be analyzed separately.

For each of the corresponding five hypotheses, we set out to use a Bayesian hypothesis test to monitor the evidence for H_0 versus H_+. Specifically, we planned to use default Bayes factors for unpaired, one-sided *t*-tests as outlined in Rouder et al. (2009) and Wetzels et al. (2009), that is, a folded Cauchy distribution with a mode at 0 and a scale of 1 (see also Jeffreys, 1961; and Ly, Verhagen, & Wagenmakers, 2016). We would monitor the Bayes factor as the data come in, and report the results as a function of the number of participants, using sequential analysis plot (see Figure 2 in Wagenmakers et al., 2012; see also Berger & Mortera, 1999, Table 1). This study was planned in 2013, and since then Morey and Rouder (2015) have advocated a default Cauchy scale of $r = 0.707$. Below we first describe our pre-registered analyses and then, in a separate section, report exploratory analyses that include the results from the new Cauchy scale.

In addition to the five critical hypotheses, we also tested the associative-priming effect in lexical decision. The corresponding statistical analysis was identical to the one for the five hypotheses of interest (i.e., a default Bayes factor for a one-sided *t*-test) except that the *t*-test is paired (i.e., within-subjects) instead of unpaired.

Results

Pre-registration plan, materials, data, and analysis scripts can be found on the OSF project page (https://osf.io/nm9vb/). The analyses were executed in JASP (jasp-stats.org; JASP team, 2017; Wagenmakers et al., in press; 2017). On the OSF, JASP output can be viewed without having JASP installed.

We deviated from the preregistration document on two counts. First, in the preregistration document we specified that we would recruit psychology students from the University of Amsterdam (UvA). In practice, any UvA student, also students outside of psychology, could make an appointment via the UvA-participant website and participate in the experiment. An additional 13 participants were not UvA students and participated purely for monetary rewards. Hence, our sample is more diverse than our initial sampling plan had specified. As a more diverse sample is more desirable (for problems related to having a narrow sample, see for example Sears, 1986), we decided not to exclude the data from the non-(psychology) student population.

The second deviation concerned our stopping rule; as specified in our preregistration document, our plan was to monitor the Bayes factor and stop the experiment whenever all critical hypothesis tests reached a Bayes factor that can be considered "strong" evidence (Jeffreys, 1961); this means that the Bayes factor is either of 10 in favor of H_0, or 10 in favor of H_+. Alternatively, the experiment would also stop whenever we reached the maximum number of participants, which we set to 50 participants per condition (i.e., a maximum of

100 participants in total). However, after testing 50 participants per condition, the Bayes factor had still not reached the critical threshold for all of the six hypotheses. As it was unclear how many participants would have to be excluded based on our preregistered exclusion criteria, we deviated from our initial sampling plan and tested a total of 132 participants.

Exclusion of Participants

We excluded seven participants who did not complete the evaluation-task as intended; they took the materials off the clipboard before or during reading the resume and evaluating the candidate. Based on the results of the associative priming task, we excluded another eight participants; six participants had error rate higher than 20%, one participant could not finish the task because the building was evacuated, and one participant accidently completed an older version of the task. In total we excluded 15 participants. The final sample consisted of 117 participants (81 women), 63 (44 women) in the heavy condition and 54 (37 women) in the light condition, with an average age of 22.6 years (range 17–51). Most participants were students (N = 106).

Confirmatory Analyses

The results reported in this section were obtained by executing the preregistered data analysis plan that is available at the OSF (https://osf.io/nm9vb/).

Evaluation Task (Weight)

Table 1 shows the means and standard deviations in the two conditions, and the Bayes factors for the five preregistered hypotheses of interest. Below we discuss the results

Table 1. *The Number of Participants (N), Means (M), Standard Deviations (SD), and Bayes Factors for the Five Critical Hypotheses.*

	Heavy Condition		Light Condition		Bayes Factor (Preregistered Prior)	Bayes Factor (Morey & Rouder Prior)
	N	*M* (*SD*)	*N*	*M* (*SD*)	BF	BF
Overall Evaluation	63	5.92 (0.70)	54	6.10 (0.60)	16.41	11.75
Serious Interest	63	6.19 (0.91)	54	6.15 (1.07)	5.78	4.22
Social Compatibility	62	5.19 (1.01)	53	5.53 (0.93)	18.74	13.42
Perceived Task Importance	63	5.54 (0.96)	54	5.85 (0.81)	19.15	13.59
Effort	63	5.57 (0.69)	53	5.68 (0.73)	11.84	8.50

Note. The preregistered prior specifies the alternative hypothesis H+ by assigning effect size a folded Cauchy distribution with mode 0 and scale 1 (Jeffreys, 1961; Rouder et al., 2009); the Morey and Rouder (2015) prior reduces the scale to 0.707. The results for the latter, exploratory analysis are discussed in a separate section. The Social Compatibility question (heavy and light condition) and the Effort question (light condition) feature one missing value.

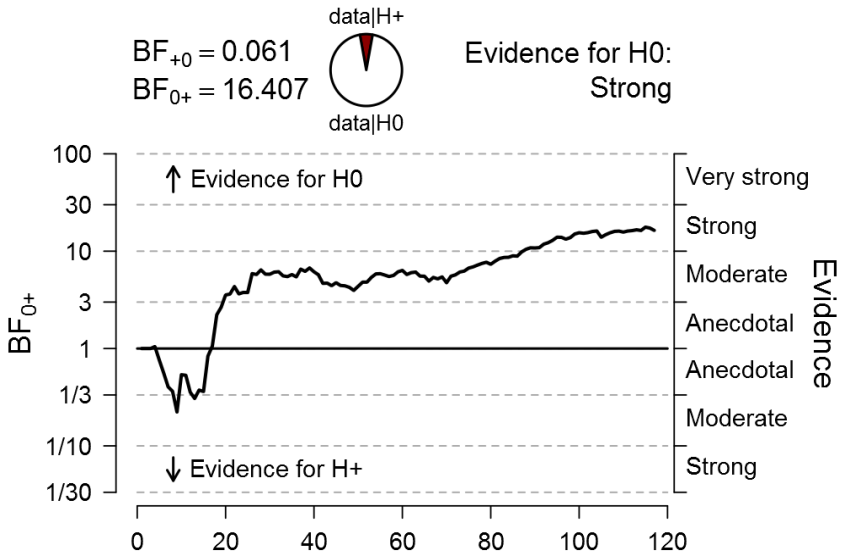

Figure 2. *Bayes factors for the comparison of overall candidate evaluation between the heavy and light condition. Figure from JASP (*jasp-stats.org*).*

for each analysis separately and show the Bayes factor as a function of the number of participants per condition using sequential analysis plots.

First, we tested the hypothesis that participants who are holding a heavy clipboard evaluate the job candidate as better overall than participants who are holding a light clipboard. As shown in Figure 2, we found strong evidence for the absence of a difference in composite candidate evaluation between the heavy and the light condition (i.e., BF_{0+} = 16.41). The corresponding mean ratings for the heavy and the light condition are 5.92 (*SD* = 0.70) and 6.10 (*SD* = 0.60), respectively, such that the observed effect size is slightly in the direction opposite to that observed by ANB.

Secondly, we tested whether participants who are holding a heavy clipboard perceive the job candidate as expressing more serious interest in the position than participants who are holding a light clipboard. As shown in Figure 3, we found moderate evidence for the absence of difference in perceived seriousness between the heavy and the light condition (i.e., BF_{0+} = 5.78). The means in perceived seriousness for the heavy and the light condition are 6.19 (*SD* = 0.91) and 6.15 (*SD* = 1.07), respectively, such that the observed effect sizes in the two conditions are highly similar.

Thirdly, we tested whether participants who are holding a heavy clipboard rate the candidate higher on social compatibility with future colleagues than participants who are holding a light clipboard. Since social compatibility is unrelated to concepts of importance or seriousness, ANB did not expect (nor did they find) a difference between the two conditions. As shown in Figure 4, we found strong evidence for the absence of difference in social compatibility rating between the heavy and the light condition (i.e., BF_{0+} = 18.74). This result is in agreement with ANB.

Fourth, we tested whether participants who are holding heavy clipboards find it more important to accurately evaluate the candidate than participants who are holding light clipboards. As shown in Figure 5, we found strong evidence for the absence of difference in perceived importance between the heavy and the light condition (i.e., BF_{0+} = 19.15).

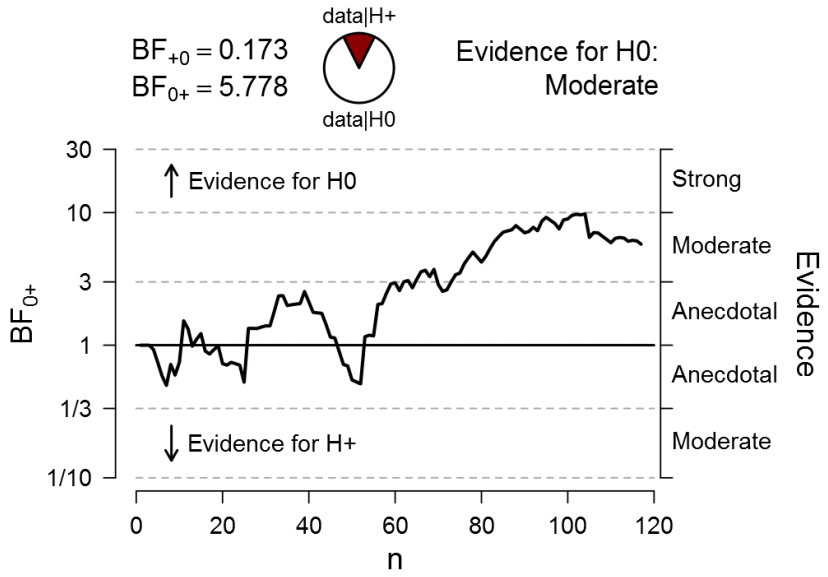

Figure 3. *Bayes factors for the comparison of perceived seriousness between the heavy and light condition. Figure from JASP (jasp-stats.org).*

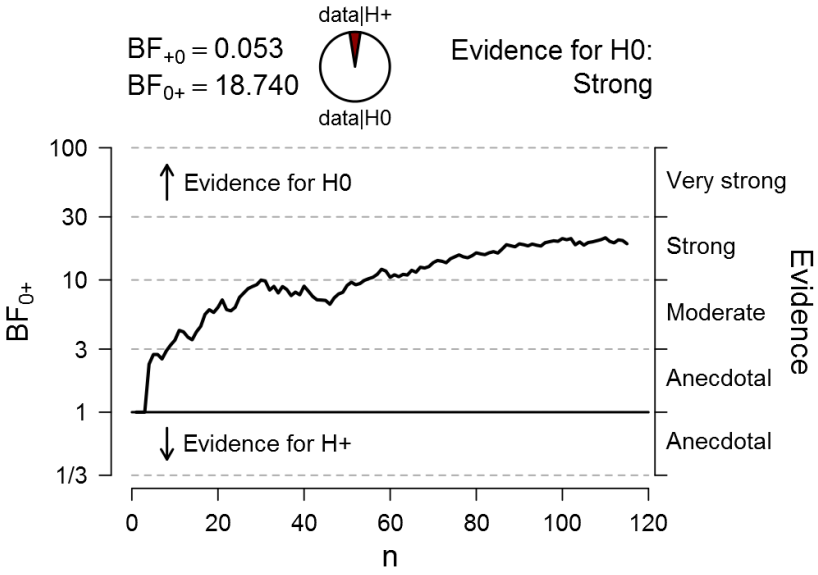

Figure 4. *Bayes factors for the comparison of social compatibility rating between the heavy and light condition. Figure from JASP (jasp-stats.org).*

The means in perceived task importance for the heavy and the light condition are 5.54 (*SD* = 0.96) and 5.85 (*SD* = 0.81), respectively, such that the observed effect size is slightly in the direction opposite to that observed by ANB.

Lastly, the fifth critical test quantified evidence for the hypothesis that the two conditions do not differ in self-reported ratings of the effort participants devoted to the task. As shown in Figure 6, we found strong evidence for the absence of difference in self-

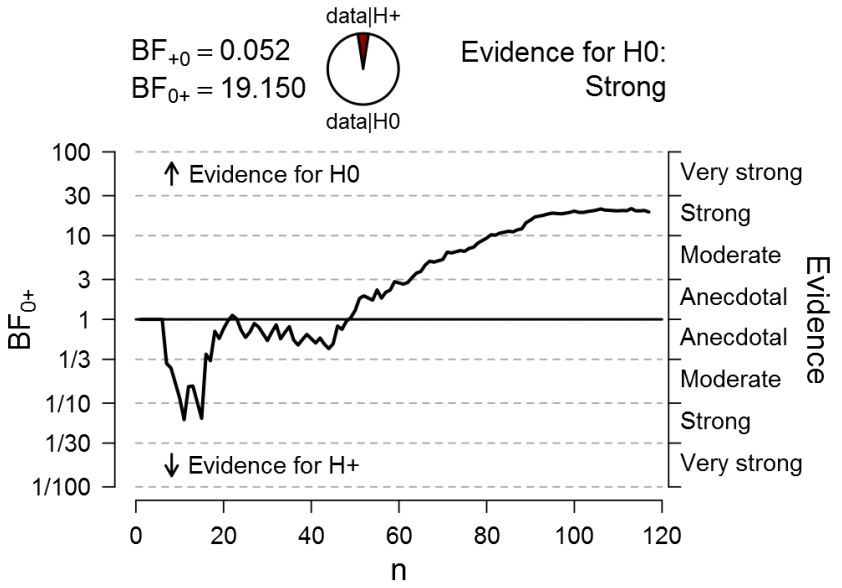

Figure 5. *Bayes factors for the comparison of rating of importance of making an accurate evaluation between the heavy and light condition. Figure from JASP (jasp-stats.org).*

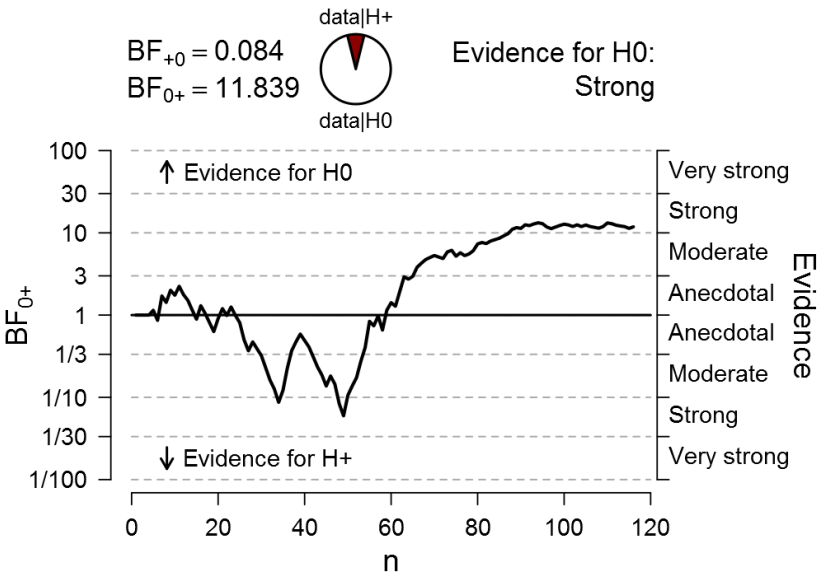

Figure 6. *Bayes factors for the comparison of self-reported effort devoted to the task between the heavy and light condition. Figure from JASP (jasp-stats.org).*

reported effort between the heavy and the light condition (i.e., BF_{0+}= 11.84). This result is in agreement with ANB.

Associative-Priming Task

Figure 7 shows the sequential Bayes factors from the preregistered one–sided paired Bayesian *t*-test (e.g., Rouder et al., 2009). The Bayes factor indicates extreme evidence for the presence of the associative-priming effect (i.e., BF_{0+}= 2.79e+17). The fact that we were

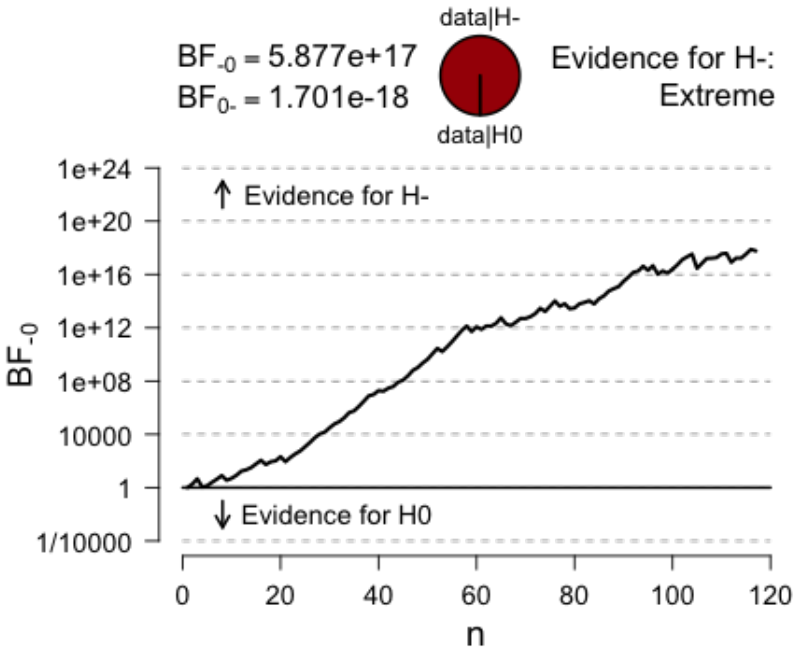

Figure 7. *Bayes factor for the comparison of mean RT for related vs. unrelated prime-target pairs. Figure from JASP (jasp-stats.org).*

able to replicate the robust associative-priming effect demonstrates that our participants were sufficiently motivated and illustrates that if an effect exists, our Bayesian methodology enables one to find evidence in its favor.

Exploratory Analyses

The results reported in this section were not preregistered but may nonetheless be of interest.

Posterior distributions.

The appendix shows the posterior distributions of effect size under H_+, for each of the six hypothesis tests reported in the previous section. For all effects except the associative-priming effect, the posterior distributions are relatively peaked near 0. This pattern remains true if the directionality of the hypotheses is ignored, as the reader can confirm by conducting the analyses in JASP using the materials provided on the OSF project page.

Morey and Rouder (2015) prior.

When this study was planned, in 2013, the default prior distributions for effect size were those proposed by Jeffreys (1961) and Rouder et al. (2009): A Cauchy distribution with mode 0 and scale 1. In recent years, Morey & Rouder (2015) have advocated a scale value that is less wide (i.e., 0.707), meaning that the predictions of the alternative hypothesis are less extreme and more similar to the predictions from the null hypothesis. The results for this new prior are shown in the last column of Table 1. The OSF project page provides the JASP files and JASP output. The reader is invited to conduct his or her own analyses to explore the robustness of the results.

As Table 1 shows, the Morey & Rouder prior yields results that are qualitatively consistent with the earlier Jeffreys's default prior. The alternative H_+ is now specified so

as to make predictions closer to those of H_0, and consequently the data are somewhat less diagnostic than before. Nevertheless, the overall pattern remains the same and indicates moderate to strong evidence in favor of the absence of an effect.

More Stringent Exclusion Criteria

In the exit-interviews, none of the participants guessed that the goal of our experiment was to test whether holding a heavy clipboard would result in higher ratings of the job candidate on aspects related to seriousness and importance than holding a light clipboard. However, seven participants correctly assumed that the weight of the clipboard differed between the two conditions. Although we did not preregister this as an exclusion criterion, participants' awareness of the experimental manipulation may have lead them to respond in a different manner than participants unsuspicious about the weight difference. To rule out this possibility, we conducted a series of exploratory Bayes hypothesis tests where we excluded these seven participants, and found that the conclusions did not change. The reader may confirm this using the materials provided on the OSF project page.

Frequentist Statistics

We conducted frequentist one-sided *t*-tests to explore if this would alter our conclusions regarding the five hypotheses about the influence of weight on evaluation. Unsurprisingly, all five *t*-tests yielded nonsignificant results (all *p*-values > .40). As suggested by the descriptive statistics reported in Table 1, our failure to replicate the ANB's effect does not appear to hinge on our statistical approach.

Discussion

Our results fail to provide support for ANB's central thesis about the link between heaviness and importance. In particular, our results do not support the hypotheses that compared to a light clipboard, holding a heavy clipboard leads to higher ratings for the overall evaluation of the candidate (Hypothesis 1), higher ratings for the perceived serious interest of the candidate in the job (Hypothesis 2), and higher ratings for participants' self-reported importance of accurately evaluating the candidate (Hypothesis 4). In fact, we found strong (Hypothesis 1 and 4) and moderate (Hypothesis 2) evidence for the absence of a difference between the conditions.

In line with their hypothesis, ANB did not find a difference between the two conditions in ratings of perceived compatibility with future colleagues (Hypothesis 3); compatibility is a social trait and is unrelated to seriousness or importance. We replicated this finding and found strong evidence for the absence of a difference between the light and heavy conditions.

We strove to follow the original design as closely as possible, but we deviated from ANB's approach on a few counts. First, ANB conducted the study in the field (on a university campus), whereas we opted for a controlled laboratory environment, allowing us to eliminate experimenter bias and administer the computerized associative-priming task. As a result, we had to rely on an additional cover story (i.e., goal of the study was to look at the influence of bodily positions on information processing; Jostmann et al., 2009) to justify why participants had to carry out the evaluation task in a standing position. Note that in order to adhere to the original design, our instructions also included ANB's original cover story (i.e., the goal of the study is to compare expert and non-expert evaluations). Our

experimental setup was more artificial than ANB's and included an additional cover story. Possibly this resulted in our participants becoming more aware of their surroundings and/or bodily position, which might have prevented the effect from occurring. Although other researchers have found the effect in a lab setting (e.g., Jostmann et al., 2009), perhaps the effect reported by ANB is more likely to be found when people are in a natural setting (in the streets), as they are perhaps slightly absent-minded (as there is much distraction when one completes the task outdoors on a college campus). If so, the artificial setting and/or the additional cover story may (partially) explain our failure to replicate. However, if this is the case, there are boundaries to when the effect occurs. The specific circumstances under which the effect can potentially occur should then be clarified. Additionally, one can question how much weight we should place on an effect that only occurs under (currently unknown) specific circumstances.

Second, the relative weight difference between the heavy and light condition was smaller in our study (where the heavy clipboard was almost 4 times heavier than the light clipboard; the weight difference was 1625 grams) compared to that of the original study (where the heavy clipboard was almost 6 times heavier than the light clipboard; the weight difference was 1701 grams). This was due to a difference in the weight of the light (unfilled) clipboards: in our study the unfilled clipboards were about 210 grams heavier and our heavy (i.e., maximally filled with paper) clipboards were about 130 grams heavier than the ones used by ANB. One may argue that these weight difference are responsible for the disappearance of the effect. However, if such modest differences in weight suffice to eliminate the effect, then the effect is extremely fragile and much less robust than suggested by ANB.

To the best of our knowledge, no other direct replications of ANB's results have been published so far. Several studies, however, have tried to conduct a similar experiment, one that examined the influence of weight on aspects related to seriousness (Jostmann et al., 2009). In their original study, Jostmann and colleagues found that compared to participants who were holding a light clipboard, participants who were holding a heavy clipboard (1) believed foreign currencies were more valuable (Jostmann et al., Experiment 1); (2) indicated that it was more important for students to have a say in decisions that affected them (Jostmann et al., Experiment 2); and (3) showed more polarization between agreement with strong arguments and disagreement with weak arguments, presumably indicating more elaborate thinking (Jostmann et al., Experiment 4). Psych File Drawer (www.psychfiledrawer.org) shows that out of the six replication attempts targeting these experiments, five were unsuccessful. One failed replication was conducted by the original first author. Recently, a Many Labs project (https://osf.io/csygd/) with 2,285 participants also attempted to replicate Experiment 2 from Jostmann et al., again without success (Ebersole et al., 2016). In response to these failures to replicate, Jostmann, Schubert, and Lakens (2016) commented: "We have had to conclude that there is actually no reliable evidence for the effect" (p. 93).

The current study presents another failed replication of the hypothesized link between heaviness and importance. Our results suggest that the effect of weight on aspects related to importance and seriousness is not as strong as originally assumed or – in the most extreme case – the effect may not exist at all.

More generally, the current study underlines the importance of conducting and publishing replication work. Replications (failed or successful) ought to become a structural part of the academic literature, as they enable us to obtain a clearer indication of the strength of a reported effect. It is important to know about failed replications and non-significant

findings, because researchers build on work conducted by others (Simmons, Nelson, & Simonsohn, 2011). Because reported effects are often not as strong as presented (or false-positives), researchers may be wasting valuable time, money, and energy chasing findings that are either absent or much smaller than reported. It is important to add, however, that replications would become even more informative if the validity of the manipulation and the theoretical understanding of the original finding were known. Ideally, such information is provided by the original authors. If that is not the case, replications should involve careful pretesting (see Jostmann et al., 2016).

In our experience, the combination of replication research, preregistered analysis plans, careful pre-testing, and Bayesian statistics constitutes a productive way of learning about the presence and the strength of published effects that, although appealing, may nevertheless not stand up to close experimental scrutiny.

Acknowledgments

The preregistration document, materials, data, JASP analysis scripts, figures, and supplementary material can be found on the Open Science Framework: https://osf.io/nm9vb/. We thank Ackerman, Nocera, and Bargh for providing us with the materials from their original study. This work was supported by the ERC grant "Bayes or Bust" from the European Research Council. DM is supported by a Veni grant (451-15-010) from the Netherlands Organization of Scientific Research (NWO).

References

Ackerman, J. M., Nocera, C. C., & Bargh, J. A. (2010). Incidental haptic sensations influence social judgments and decisions. *Science, 328*(5986), 1712-1715.

Bakker, M., van Dijk, A., & Wicherts, J. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review Psychology, 59*, 617-645.

Berger, J.O., & Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association, 94*(466), 542-554.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle*. Hayward, CA: Institute of Mathematical Statistics.

Cacioppo, J. T., Priester, J. R., & Berntson, G. G. (1993). Rudimentary determinants of attitudes: II. Arm flexion and extension have differential effects on attitudes. *Journal of Personality and Social Psychology, 65*(1), 5-17.

Chandler, J. J., Reinhard, D., & Schwarz, N. (2012). To judge a book by its weight you need to know its content: Knowledge moderates the use of embodied cues. *Journal of Experimental Social Psychology, 48*(4), 948-952.

De Groot, A. M. B. (1984). Primed lexical decisions: Combined effects of the proportion of related prime-

target pairs and the stimulus-onset asynchrony of prime and target. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 36*(2), 253-280.

De Groot, A. M. B. (1987). The priming of word associations: A levels-of-processing approach. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 39*(4), 721-756.

Ebersole, D., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., … (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68-82.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193-242.

JASP Team. JASP (Version 0.8.1)[Computer software].

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press, Oxford, UK.

Jostmann, N. B., Lakens, D., & Schubert, T. W. (2009). Weight as an embodiment of importance. *Psychological Science, 20*(9), 1169-1174.

Jostmann, N. B., Lakens, D., & Schubert, T. W. (2016). A short history of the weight-importance effect and a recommendation for pre-testing: Commentary on Ebersole et al.(2016). *Journal of Experimental Social Psychology, 67*, 93-94.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Lee, M. D., & Wagenmakers, E. J. (2013). Bayesian cognitive modeling: A practical course. retrieved from www.ejwagenmakers.com

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review, 7*(4), 618-630.

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72*, 19-32.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504.

Maglio, S. J., & Trope, Y. (2012). Disembodiment: Abstract construal attenuates the influence of contextual bodily state in judgment. *Journal of Experimental Psychology: General, 141*(2), 211-216.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H.A, van der Molen, M.W, & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General, 144*(1), e1-e15.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor 0.9.11-1. *Comprehensive R Archive Network*.

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence forfacilitatory and inhibitory processes. *Memory and Cognition, 4*(5), 648-654.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading of activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106*(3), 226-254.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615-631.

Reimann, M., Feye, W., Malter, A. J., Ackerman, J. M., Castano, R., Garg, N., … (2012). Embodiment in judgment and choice. *Journal of Neuroscience, Psychology, and Economics, 5*(2), 104.

Rosenthal, R. (1963). On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results, *American Scientist, 51*(2), 268-283.

Rotteveel, M., Gierholz, A., Koch, G., van Aalst, C., Pinto, Y., Matzke, D., … (2015). On the automatic link between affect and tendencies to approach and avoid: Chen and Bargh (1999) revisited. *Frontiers in Psychology, 6*, e1-e12.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301-308.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237.

RStudio (2012). RStudio: Integrated development environment for R (Version 0.97.551) [Computer software]. Boston, MA.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*(3), 515-530.

Simmons, J. P, Nelson, L.D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*(5), 768-777.

Wagenmakers, E. -J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., … (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*(6), 917-928.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, ... (2017). Bayesian statistical inference for psychological science. Part II: Example applications with JASP. OSF: https://osf.io/m6bi8/.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., Selker, R., … (in press). Bayesian statistical inference for psychological science. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*. OSF: https://osf.io/m6bi8/.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E. J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t* test. *Psychonomic Bulletin & Review, 16*(4), 752-760.

Appendix: Posterior Distributions Under H^+

Here we present the posterior distributions under H_+ for the six preregistered hypothesis tests described in the main text. For all effects except the associative-priming effect, the posterior distributions are relatively peaked near 0.

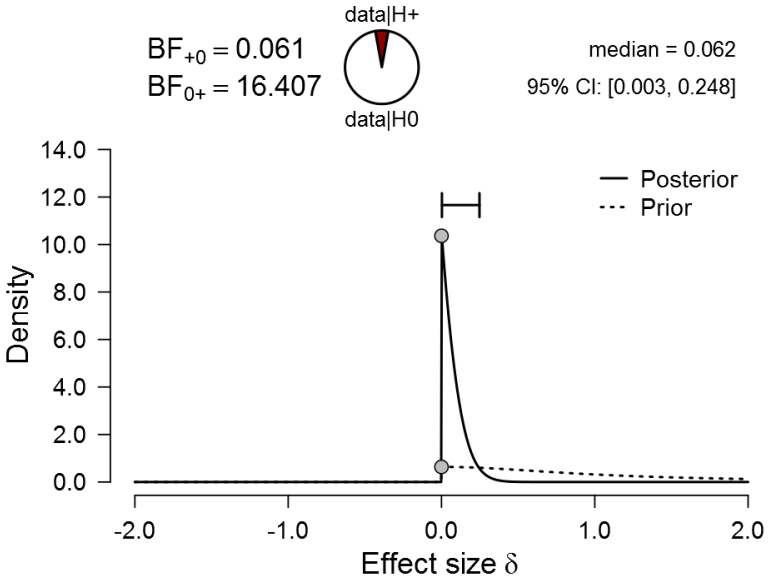

Figure A1. *Posterior distribution of effect size for the comparison of overall candidate evaluation between the heavy and light condition. Figure from JASP (jasp-stats.org).*

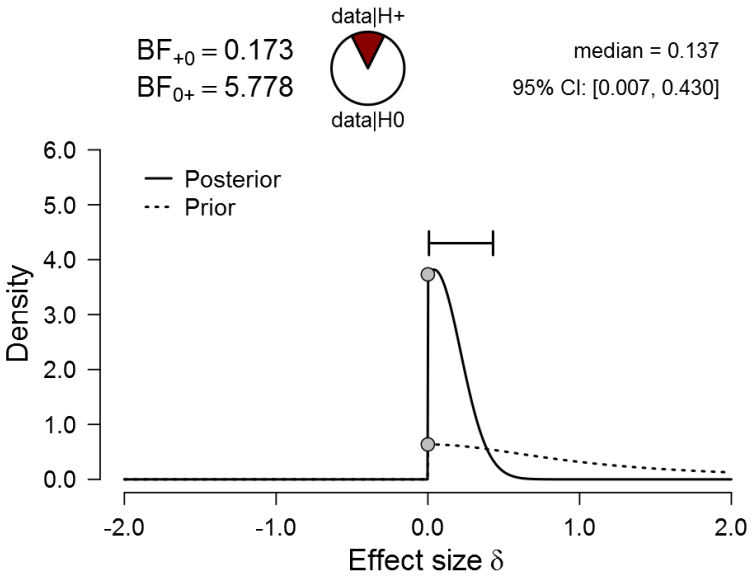

Figure A2. *Posterior distribution of effect size for the comparison of perceived seriousness between the heavy and light condition. Figure from JASP (jasp-stats.org).*

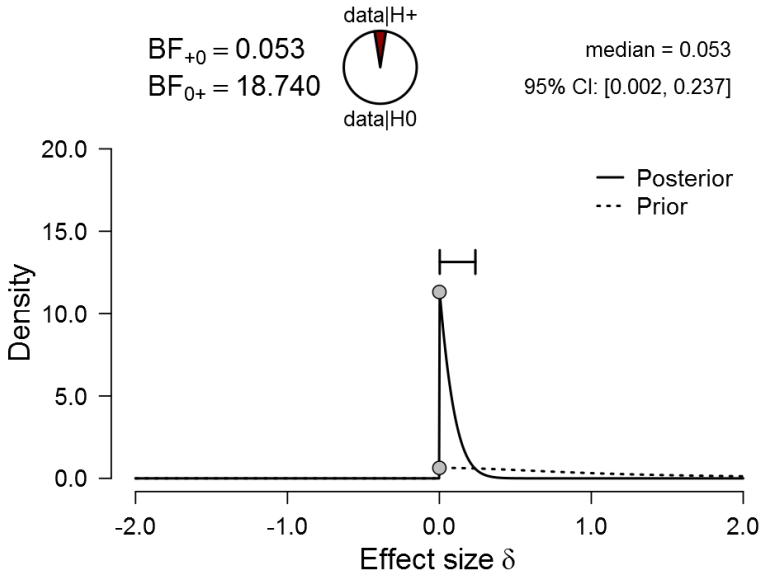

Figure A3. *Posterior distribution of effect size for the comparison of overall candidate evaluation between the heavy and light condition. Figure from JASP (jasp-stats.org).*

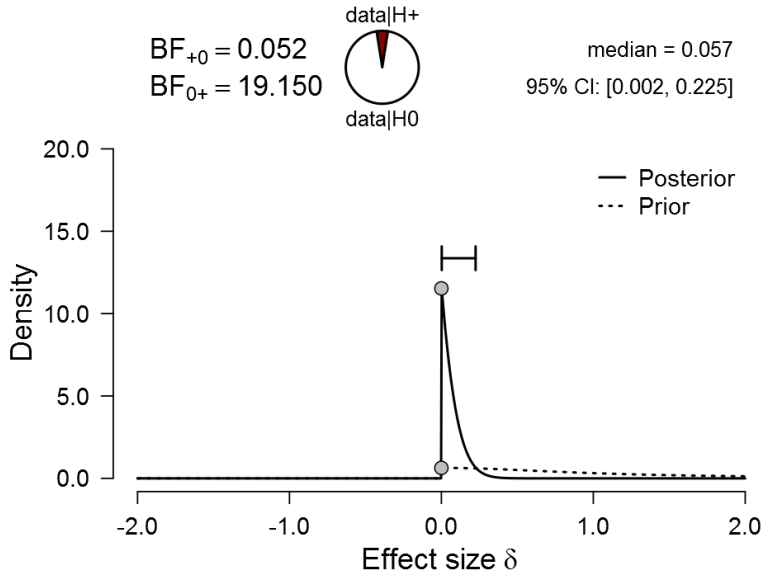

Figure A4. *Posterior distribution of effect size for the comparison of rating of importance of making an accurate evaluation between the heavy and light condition. Figure from JASP (jasp-stats.org).*

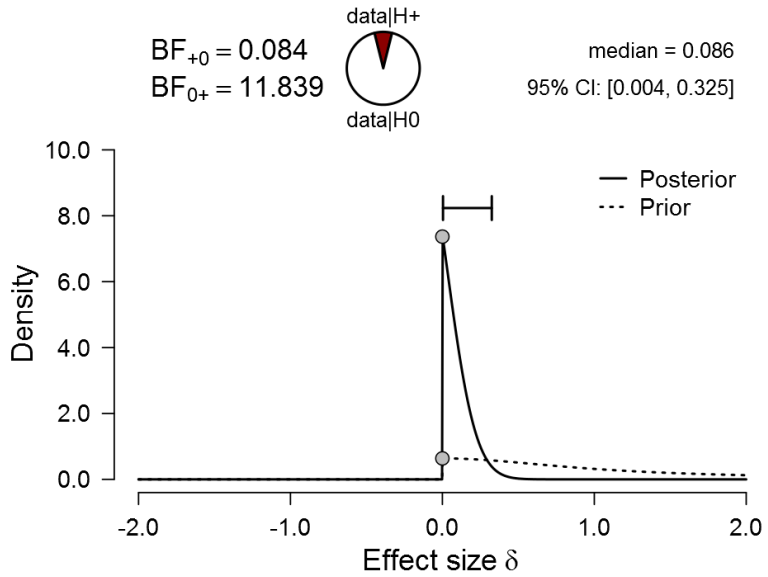

Figure A5. *Posterior distribution of effect size for the comparison of self-reported effort devoted to the task between the heavy and light condition. Figure from JASP (jasp-stats.org).*

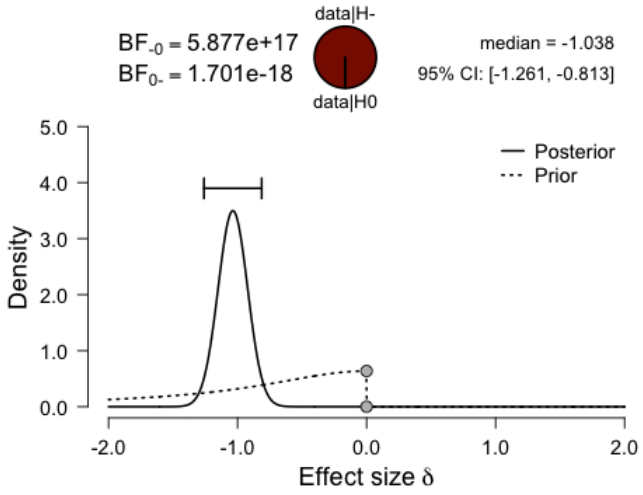

Figure A6. *Posterior distribution of effect size for the comparison of mean RT for related vs. unrelated prime–target pairs. Figure from JASP (*jasp-stats.org*).*

Received: 4.3.2017
Revised: 8.15.2017
Accepted: 8.16.2017