Errata to the Examples Section of: "Bayesian Tests to Quantify the Result of a Replication Attempt"

Josine Verhagen & Eric-Jan Wagenmakers
University of Amsterdam

Abstract

Three problems in the Examples section of the published paper were detected and corrected:

• **Problem 1**: The effect sizes printed in the text were computed assuming a one-sample instead of a two-sample t-test.

Change: The effect sizes in the text were changed to match the correct effect sizes in Table 5.

• **Problem 2**: The studies used to compute the professor priming results (i.e. Studies 3 and 5 from Shanks et al., 2013) did not match the studies described in the text (i.e. Studies 5 and 6 from Shanks et al., 2013).

Change: Table 5 and Figure 3 were adjusted to display the results from Studies 5 and 6 from Shanks et al., 2013. The Bayes factor results in the text and the corresponding conclusions were adjusted.

• **Problem 3**: The meta-analysis Bayes factor R code contained a bug which falsely assumed equal sample sizes in all replication studies.

Changed: The meta-analysis Bayes factors were recomputed and changed in Table 5 and in the text. The results only differ substantially for the first example (Red, Rank, and Romance). In the text below, the corrections are indicated in yellow.

Keywords: Effect Size, Prior Distribution, Bayes Factor.

Examples

We now apply the above Bayesian t tests to three examples of replication attempts from the literature. These examples cover one-sample and two-sample t tests in which the outcome is replication failure, replication success, and replication ambivalence. The

This research was supported by an ERC grant from the European Research Council. Correspondence concerning this article may be addressed to Josine Verhagen, University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, the Netherlands. Email address: A.J.Verhagen@uva.nl.

examples also allow us to visualize the prior and posterior distributions for effect size, as well as the test outcome. The examples were chosen for illustrative purposes and do not reflect our opinion, positive or negative, about the experimental work or the researchers who carried it out. Moreover, our analysis is purely statistical and purposefully ignores the many qualitative aspects that may come into play when assessing the strength of a replication study (see Brandt et al., 2013). For simplicity, we also ignore the possibility that experiments with nonsignificant results may have been suppressed (i.e., publication bias, e.g., Francis, 2013b, in press). In each of the examples, it is evident that the graded measure of evidence provided by the Bayesian replication tests is more informative and balanced than the "significant-nonsignificant" dichotomy inherent to the p value assessment of replication success that is currently dominant in psychological research.

Example 1: Red, Rank, and Romance in Women Viewing Men

In an attempt to unravel the mystery of female sexuality, Elliot et al. (2010) set out to discover what influences women's attraction to men. Inspired by findings in crustaceans, sticklebacks, and rhesus macaques, Elliot et al. (2010) decided to test the hypothesis that "viewing red leads women to perceive men as more attractive and more sexually desirable" (p. 400). In a series of experiments, female undergraduate students were shown a picture of a moderately attractive man; subsequently, the students had to indicate their perception of the man's attractiveness. The variable of interest was either the man's shirt color or the picture background color (for a critique see Francis, 2013a).

The first experiment of Elliot et al. (2010) produced a significant effect of color on perceived attractiveness ($t(20) = 2.18$, $p < .05$, $\delta = 0.95$): the female students rated the target man as more attractive when the picture was presented on a red background ($M = 6.79$, $SD = 1.00$) than when it was presented on a white background ($M = 5.67$, $SD = 1.34$). The second experiment was designed to replicate the first and to assess whether the color effect generalised to male students. The results showed the predicted effect of color on perceived attractiveness for female students ($t(53) = 3.06$, $p < .01$, $\delta = 1.33$) but not for male students ($t(53) = 0.25$, $p > .80$, $\delta = .11$). In the third experiment, featuring female participants only, the neutral background color was changed from white to grey. The results again confirmed the presence of the predicted color effect ($t(32) = 2.44$, $p < .05$, $\delta = 1.07$): the students rated the target man as more attractive when the picture was presented on a red background ($M = 6.69$, $SD = 1.22$) than when it was presented on a grey background ($M = 5.27$, $SD = 2.04$).

We now re-analyze these results with our Bayesian replication tests, assuming that Experiment 1 from Elliot et al. (2010) is the original study and the others are the replication attempts. The results are summarized in Table 1.[1]

Before discussing the extant Bayes factor tests we first discuss and visualize the results from our new Bayes factor test for replication. The left panel of Figure 1 shows the results for the data from the female students in Elliot et al.'s Experiment 2. The dotted line indicates the prior distribution for the replication test, that is, the proponent's posterior distribution for effect size after observing the data from the original experiment, $p(\delta \mid Y_{orig})$. The solid line indicates the posterior distribution for effect size after observing the additional data

[1] R code to reproduce the analyses for all of the examples is available from the first author's webpage.

		Rep B_{r0}	JZS B_{10} (one-sided)		Equality B_{01}	Meta B_{10}
Red and Romance						64.75 (976*)
Original	$\delta =$ 0.95		1.79			
Female	$\delta =$ 1.33	39.73	10.51	(20.96)	3.29	
Male	$\delta =$ 0.11	0.13	0.21	(0.25)	1.07	
Gray	$\delta =$ 1.07	9.76	2.75	(5.42)	3.10	
Professor priming						0.16
Original	$\delta =$ 0.91		8.92			
Replication 1	$\delta =$ -0.07	0.05	0.22	(0.18)	0.39	
Replication 2	$\delta =$ -0.38	0.03	0.50	(0.13)	0.11	
Negative priming						0.10
Original	$\delta =$ 0.74		10.45			
Replication 1	$\delta =$ 0.46	2.36	0.97	(1.89)	2.95	
Replication 2	$\delta =$ -0.54	0.01	1.67	(0.04)	0.00	

* Result for the meta-analysis without the Male study.

Table 1: Results for four different Bayes factor tests applied to three example studies. Bayes factors higher than 1 favor the hypothesis that an effect is present. Note: "Rep B_{10}" is the new Bayes factor test for replication; "JZS B_{10}" is the two-sided independent default Bayes factor test, with the result of the one-sided test between brackets; "Equality B_{01}" is the equality-of-effect-size Bayes factor test; and "Meta B_{10}" is the fixed-effect meta-analysis Bayes factor test.

from the replication experiment, $p(\delta \mid Y_{rep}, Y_{orig})$. This posterior distribution assigns more mass to values of δ higher than zero than did the prior distribution, a finding that suggests replication success. The extent of this success is quantified by the computation of B_{r0}, which yields 39.73: the data from the female students in Experiment 2 are about 40 times more likely under the proponent's replication hypothesis $\mathcal{H}_r$ than under the skeptic's null hypothesis $\mathcal{H}_0$.

The outcome of the new Bayesian replication t test is visualized by the ordinates of the prior and posterior distributions at $\delta = 0$, indicated in the left panel of Figure 1 by the two filled circles. Intuitively, the height of the prior distribution at $\delta = 0$ reflects the believability of $\mathcal{H}_0$ before seeing the data from the replication attempt, and the height of the posterior distribution at $\delta = 0$ reflects the believability of $\mathcal{H}_0$ after seeing those data. In this case, observing the replication data decreases the believability of $\mathcal{H}_0$. It so happens that the ratio of the ordinates exactly equals the Bayes factor B_{r0} (i.e., the Savage-Dickey density ratio test, see e.g., Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010), providing a visual confirmation of the test outcome.

The middle panel of Figure 1 shows the results for the data from the male students in Elliot et al.'s Experiment 2. The posterior distribution assigns more mass to $\delta = 0$ than does the prior distribution, indicating evidence in favor of the null hypothesis $\mathcal{H}_0$ over the replication hypothesis $\mathcal{H}_r$. The Bayesian replication test yields $B_{r0} = 0.13$, indicating that the replication data are $1/.13 = 7.69$ times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_r$.

Finally, the right panel of Figure 1 shows the results for the data from Elliot et al.'s Experiment 3. After observing the data from this replication attempt the posterior distribution assigns more mass to values of δ higher than zero than did the prior distribution; hence the Bayesian replication test indicates support for the replication hypothesis $\mathcal{H}_r$ over the null hypothesis $\mathcal{H}_0$. The Bayes factor equals 9.76, indicating that the data are almost 10 times more likely to occur under $\mathcal{H}_r$ than under $\mathcal{H}_0$. Because the sample size in this replication attempt is smaller than in the first replication attempt, the posterior is less peaked and the test outcome less extreme.

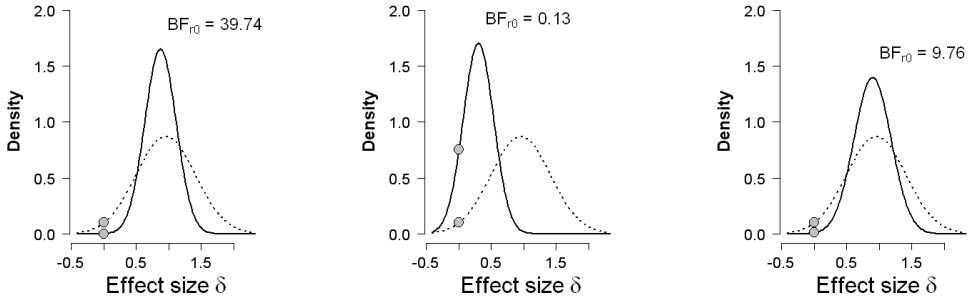

Figure 1. Results from the Bayes factor replication test applied to Experiment 2 and 3 from Elliot et al. (2010). The original experiment had shown that female students judge men in red to be more attractive. The left and middle panel show the results from Elliot et al.'s Experiment 2 (for female and male students, respectively), and the right panel shows the results from Elliot et al.'s Experiment 3 that featured a different control condition. In each panel, the dotted lines represent the posterior from the original experiment, which is used as prior for effect size in the replication tests. The solid lines represent the posterior distributions after the data from the replication attempt are taken into account. The grey dots indicate the ordinates of this prior and posterior at the skeptic's null hypothesis that the effect size is zero. The ratio of these two ordinates gives the result of the replication test.

In contrast to the analysis above, the three extant Bayes factor tests produce results that are more unequivocal. The default JZS Bayes factor equals 1.79 for the original experiment, indicating that the data are uninformative, as they are almost as likely to have occurred under the null hypothesis as under the alternative hypothesis. The study with female students yields strong support in favor of the presence of an effect ($B_{10} = 10.51$), whereas the study with male students yields moderate evidence in favor of the absence of an effect ($B_{01} = 1/B_{10} = 1/0.21 = 4.76$). The study with the gray background, however, yields only anecdotal support for the presence of an effect ($B_{10} = 2.75$; a one-sided test, however, almost doubles the support, $B_{10} = 5.42$).

The equality-of-effect-size Bayes factor test does not yield strong support for or against the hypothesis that the effect sizes are equal, with Bayes factors that never exceed 3.29. The fixed-effect meta-analysis Bayes factor test that pools the data from all four studies yields $B_{10} = 64.75$ in favor of the presence of an overall effect, including the hypothesis that the effect generalises to men. When the experiment with male students is omitted the pooled Bayes factor is $B_{10} = 976$, indicating extreme evidence in favor of an effect.

Example 2: Dissimilarity Contrast in the Professor Priming Effect

Seminal work by Dijksterhuis and others (e.g., Dijksterhuis & van Knippenberg, 1998; Dijksterhuis et al., 1998) has suggested that priming people with intelligence-related concepts (e.g., "professor") can make them behave more intelligently (e.g., answer more trivial questions correctly). The extent to which this effect manifests itself is thought to depend on whether the prime results in assimilation or contrast (e.g., Mussweiler, 2003). Specifically, the presentation of general categories such as "professor" may lead to assimilation (i.e., activation of the concept of intelligence), whereas the presentation of a specific exemplar such as "Einstein" may lead to contrast (i.e., activation of the concept of stupidity).

However, LeBoeuf and Estes (2004) argued that the balance between assimilation and contrast is determined not primarily by whether the prime is a category or an exemplar, but rather by whether the prime is perceived as relevant in the social comparison process. To test their hypothesis, LeBoeuf and Estes (2004) designed an experiment in which different groups of participants were first presented with either the category prime "professor" or the exemplar prime "Einstein". To manipulate prime relevance participants were then asked to list similarities and differences between themselves and the presented prime. Subsequently, a test phase featured a series of multiple-choice general knowledge questions.

The results showed that performance was better in the difference-listing condition than in the similarity-listing condition. LeBoeuf and Estes (2004) interpreted these findings as follows: "As hypothesized, when participants were encouraged to perceive themselves as unlike a prime, behavior assimilated to prime activated traits, presumably because the prime was rejected as a comparison standard. When participants contemplated how they were similar to a prime, that prime was seemingly adopted as a relevant standard for self-comparison. With the current primes, such a comparison led to negative selfevaluations of intelligence and to lower test performance. Counterintuitively, participants who considered similarities between themselves and an intelligent prime exhibited *worse* performance than did participants who considered differences between themselves and an intelligent prime." (pp. 616-617, italics in original).

Among the various conditions of Experiment 1 in LeBoeuf and Estes (2004), the best performance was attained in the "differences to Einstein" condition (56.2% correct) whereas the worst performance was observed in the "similarities to professors" condition (45.2%), a performance gap that was highly significant ($t(42) = 3.00$, $p = .005$, $\delta = .91$). These two cells in the design were the target of two recent replication attempts by Shanks et al. (2013). Specifically, Experiment 5 from Shanks et al. (2013) was designed to be "as close to LeBoeuf and Estes' study as possible" and yielded a nonsignificant effect that was slightly in the opposite direction ($t(47) = -.25$, $p = .60$, $\delta = -.07$, one-sided t test). Experiment 6 from Shanks et al. (2013) tried to maximize the possibility of finding the effect by informing participants beforehand about the hypothesized effects of the primes; however, the results again yielded a nonsignificant effect that was slightly in the opposite direction ($t(30) = -1.25$, $p = .89$, $\delta = -.38$, one-sided t test).

We now re-analyze these results, assuming that Experiment 1 from LeBoeuf and Estes (2004) is the original study and Experiment 5 and 6 from Shanks et al. (2013) are the replication attempts. The resulting Bayes factors are presented in Table 1.

Before discussing the extant Bayes factor tests we first discuss and visualize the results

from our new Bayes factor test for replication. In both panels of Figure 2, the dotted line indicates the prior distribution for the replication test, that is, the proponent's posterior distribution for effect size after observing the data from the original experiment, $p(\delta \mid Y_{orig})$. The left panel of Figure 2 shows the results for the data from Shanks et al.'s Experiment 5. The solid line indicates the posterior distribution for effect size after observing the additional data from the replication experiment, $p(\delta \mid Y_{rep}, Y_{orig})$. The posterior distribution assigns more mass to $\delta = 0$ than does the prior distribution, indicating evidence in favor of the null hypothesis $\mathcal{H}_0$ over the replication hypothesis $\mathcal{H}_r$. The Bayesian replication test yields $B_{r0} = 0.05$, indicating that the replication data are $1/.05 = 20$ times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_r$. This constitutes strong evidence against H_r.

The right panel of Figure 2 shows the results for the data from Shanks et al.'s Experiment 6. The solid line again indicates the posterior distribution after observing the data from the replication experiment. As in the left panel, the posterior distribution assigns more mass to $\delta = 0$ than does the prior distribution, and the Bayesian replication test yields $B_{r0} = 0.03$, indicating that the replication data are $1/.03 = 33$ times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_r$. This constitutes strong evidence against H_r.

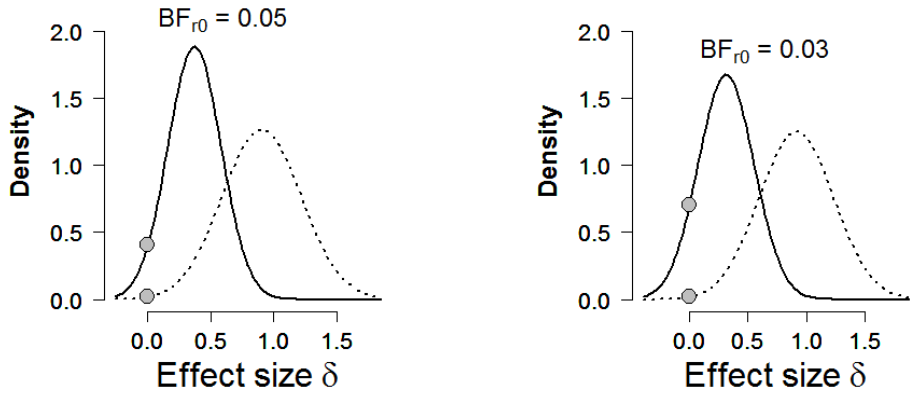

Figure 2. Results from the Bayes factor replication test applied to Experiment 5 (left panel) and Experiment 6 (right panel) from Shanks et al. (2013). In each panel, the dotted lines represent the posterior from the original experiment by LeBoeuf and Estes (2004), which is used as prior for effect size in the replication tests. The solid lines represent the posterior distributions after the data from the replication attempt are taken into account. The grey dots indicate the ordinates of this prior and posterior at the skeptic's null hypothesis that the effect size is zero. The ratio of these two ordinates gives the result of the replication test.

The results from the three extant Bayes factor tests are as follows. The default JZS Bayes factor equals 8.92 for the original experiment, indicating moderate to strong support for the alternative hypothesis. The two replication attempts, however, show the opposite pattern. In both studies, the data are five to six times more likely under the null hypothesis than under the alternative hypothesis.

The equality-of-effect-size Bayes factor test does yield some support against the null

hypothesis that the effect sizes are equal: for the first replication attempt the data are only 2.75 (1/.39) times more likely, but for the second replication attempt the data are 9.09 (1/.11) times more likely under the hypothesis that the effect sizes are unequal rather than equal. The fixed-effect meta-analysis Bayes factor test that pools the data from all three studies yields $B_{10} = 0.16$ in favor of the presence of an overall effect, which indicates that the combined data are about six times (i.e., 1/.16) more likely under the null hypothesis of no effect.

Example 3: Negative Priming

Negative priming refers to the decrease in performance (e.g., longer response times or RTs, more errors) for a stimulus that was previously presented in a context in which it had to be ignored. For instance, assume that on each trial of an experiment, participants are confronted with two words: one printed in red, the other in green. Participants are told to respond only to the red target word (e.g., by indicating whether or not it represents an animate entity, *horse*: yes, *furnace*: no) and ignore the green distractor word. Negative priming is said to have occurred when performance on the target word suffers when, on the previous trial, this same word was presented as a green distractor. The theoretical relevance of negative priming is that it is evidence for inhibitory processing – commonly, negative priming is attributed to an attentional mechanism that actively suppresses or inhibits irrelevant stimuli; when these inhibited stimuli later become relevant and need to be re-activated, this process takes time and a performance decrement is observed.

However, the standard suppression account of negative priming was called into question by Milliken, Joordens, Merikle, and Seiffert (1998), who showed that negative priming can also be observed in situations where the first presentation of the repeated item does not specifically call for it to be ignored. Specifically, we focus here on Experiment 2A from Milliken et al. (1998), where participants had to name a target word printed in red, while ignoring a distractor word printed in green.

Prior to the target stimulus display, a prime word was presented for 33 ms, printed in white. No action was required related to the prime word. In the unrepeated condition, the prime was not related to the words from the target display; in the repeated condition, the prime was identical to the target word that was presented 500 ms later. Despite the fact that the prime was so briefly presented that most participants were unaware of its presence, and despite the fact that no active suppression of the prime was called for, Milliken et al. (1998) nonetheless found evidence for negative priming: RTs were 8 ms slower in the repeated than in the unrepeated condition, $t(19) = 3.29$, $p < .004$, $\delta = .74$.

The experiment by Milliken et al. (1998) was the target for two nearly exact replication attempts. In Experiment 1A from Neill and Kahan (1999), the negative priming effect was again observed: RTs in the repeated condition were 13 ms slower than those in the unrepeated condition, $t(29) = 2.06$, $p = .048$, $\delta = .46$. However, the results from Experiment 1B showed the opposite result, with RTs in the repeated condition being 7 ms faster than those in the unrepeated condition, $t(43) = -2.40$, $p = .021$, $\delta = -.54$.

We now re-analyze these results with our Bayesian replication t test, assuming that Experiment 2A from Milliken et al. (1998) is the original study and Experiments 1A and 1B from Neill and Kahan (1999) are the replication attempts. The results are presented in Table 1.

Before discussing the extant Bayes factor tests we first discuss and visualize the results from our new Bayes factor test for replication. Similar to the previous examples, in both panels of Figure 3 the dotted line indicates the prior distribution for the replication test, that is, the proponent's posterior distribution for effect size after observing the data from the original experiment, $p(\delta \mid Y_{orig})$. The left panel of Figure 3 shows the results for the data from Neill and Kahan's Experiment 1A. The solid line indicates the posterior distribution for effect size after observing the additional data from the replication experiment, $p(\delta \mid Y_{rep}, Y_{orig})$. Although barely discernable from the plot, the posterior distribution assigns less mass to $\delta = 0$ than does the prior distribution, indicating evidence in favor of the replication hypothesis $\mathcal{H}_r$ over the null hypothesis $\mathcal{H}_0$. The Bayesian replication test yields $B_{r0} = 2.36$, indicating that the replication data are 2.36 times more likely under $\mathcal{H}_r$ than under $\mathcal{H}_0$. This constitutes evidence in favor of the replication hypothesis H_r, albeit weak and, according to Jeffreys, "not worth more than a bare mention".

The right panel of Figure 3 shows the results for the data from Neill and Kahan's Experiment 1B. The solid line again indicates the posterior distribution after observing the data from the replication experiment. Now the posterior distribution assigns much more mass to $\delta = 0$ than does the prior distribution, and the Bayesian replication test yields $B_{r0} = 0.01$, indicating that the replication data are $1/.01 = 100$ times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_r$. This constitutes compelling evidence against H_r.

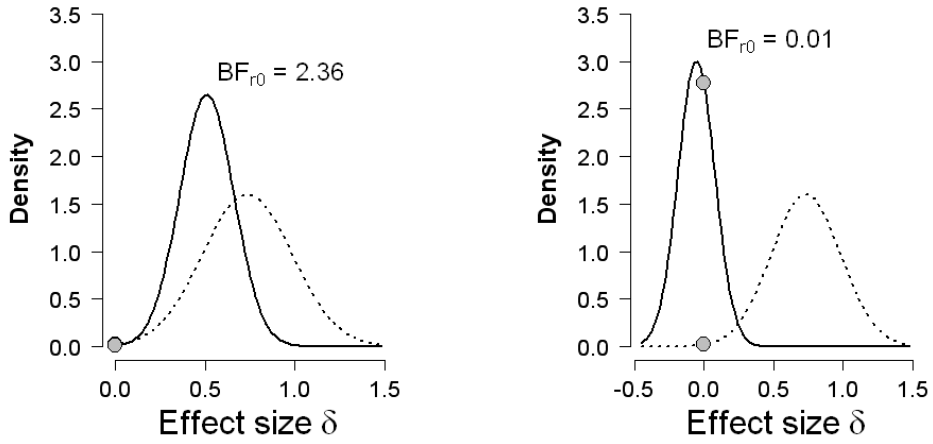

Figure 3. Results from the Bayes factor replication test applied to the replication studies of Neill and Kahan (1999) 1A (left panel) and 1B (right panel). In each panel, the dotted lines represent the posterior from the original experiment by Milliken et al. (1998), which is used as prior for effect size in the replication tests. The solid lines represent the posterior distributions after the data from the replication attempt are taken into account. The grey dots indicate the ordinates of this prior and posterior at the skeptic's null hypothesis that the effect size is zero. The ratio of these two ordinates gives the result of the replication test.

The results from the three extant Bayes factor tests are as follows. The default JZS Bayes factor equals 10.45 for the original experiment, indicating strong support for the

alternative hypothesis. The results for the two replication attempts, however, are unequivocal. In both studies, the data are about equally likely under the null hypothesis as under the alternative hypothesis. For the second replication attempt, this outcome is artificial, brought about by the fact that the analyses are two-sided instead of one-sided. A one-sided default Bayes factor test for the second replication attempt provides strong evidence in favor of the null hypothesis versus the one-sided alternative hypothesis that postulates a decrease in performance in the repeated condition, $B_{01} = 26.87$ (Morey & Wagenmakers, 2014; Wagenmakers et al., 2010).

For the first replication attempt, the equality-of-effect-size Bayes factor test provides anecdotal support for the equality of effect sizes; however, for the second replication attempt the test strongly supports the hypothesis of unequal effect sizes – consistent with the fact that the effect is in the opposite direction. Finally, the fixed-effect meta-analysis Bayes factor test that pools the data from all three studies yields $B_{10} = 0.10$ in favor of the presence of an overall effect, which indicates that the combined data are about ten times (1/.10) more likely under the null hypothesis of no effect.

References

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2013). The replication recipe: What makes for a convincing replication? *Manuscript submitted for publication*.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Dijksterhuis, A., Spears, R., Postmes, T., Stapel, D., Koomen, W., van Knippenberg, A., & Scheepers, D. (1998). Seeing one thing and doing another: Contrast effects in automatic behavior. *Journal of Personality and Social Psychology*, *75*, 862–871.

Dijksterhuis, A., & van Knippenberg, A. (1998). The relation between perception and behavior, or how to win a game of Trivial Pursuit. *Journal of Personality and Social Psychology*, *74*, 865–877.

Elliot, A. J., Kayser, D. N., Greitemeyer, T., Lichtenfeld, S., Gramzow, R. H., Maier, M. A., & Liu, H. (2010). Red, rank, and romance in women viewing men. *Journal of Experimental Psychology: General*, *139*, 399–417.

Francis, G. (2013a). Publication bias in "Red, rank, and romance in women viewing men," by Elliot et al. (2010). *Journal of Experimental Psychology: General*, *142*, 292–296.

Francis, G. (2013b). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*, 153–169.

Francis, G. (in press). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*.

LeBoeuf, R. A., & Estes, Z. (2004). "Fortunately, I'm no Einstein": Comparison relevance as a determinant of behavioral assimilation and contrast. *Social Cognition*, *22*, 607–636.

Milliken, B., Joordens, S., Merikle, P. M., & Seiffert, A. E. (1998). Selective attention: A reevaluation of the implications of negative priming. *Psychological Review*, *105*, 203–229.

Morey, R. D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Manuscript submitted for publication*.

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, *110*, 472–489.

Neill, W. T., & Kahan, T. A. (1999). Response conflict reverses priming: A replication. *Psychonomic Bulletin & Review*, *6*, 304–308.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., Kavvadia, F., & Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*, *8*, e56515.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.