Still No Evidence that Risk-Taking and Consumer Choices can be Primed by Mating Motives: Reply to Sundie, Beal, Neuberg, and Kenrick (2018)

David R. Shanks

University College London, England

Miguel A. Vadillo

Universidad Autónoma de Madrid, Spain

Address correspondence to:

David R. Shanks

Division of Psychology and Language Sciences

University College London

26 Bedford Way

London WC1H 0AP

d.shanks@ucl.ac.uk

Acknowledgments: This work was supported by Grant ES/P009522/1 from the Economic and Social Research Council to David R. Shanks and by grants 2016-T1/SOC-1395 from Comunidad de Madrid (Programa de Atracción de Talento Investigador) and PSI2017-85159-P from Ministerio de Economía y Competitividad to Miguel A. Vadillo. The R script for the analysis shown in Figure 2 is available via the Open Science Framework (OSF) at osf.io/fjkea/.

Abstract

Shanks et al. (2015) challenged the evidence that various forms of decision making can be influenced by romantic/mating primes. In their comment, Sundie, Beal, Neuberg, and Kenrick (in press) question both the meta-analysis and the 8 studies Shanks et al. reported, and describe an alternative *p*-curve analysis which they interpret as showing that romantic priming is a genuine phenomenon. In this reply we comment on several contradictions in Sundie et al.'s article. First, they suggest that Shanks et al.'s replication experiments yielded different results from the original studies because we failed to appreciate the contextual sensitivity of romantic priming effects, but this argument rests largely on evidence from the very studies we were unable to replicate, and a wealth of other evidence suggests that social priming effects are largely invariant across samples and settings. Secondly, Sundie et al. criticize the selection rule by which Shanks et al. identified relevant priming studies, but then go on to include exactly the same set of studies in their *p*-curve analysis. Thirdly, they criticize Shanks et al.'s selection of statistical results from these studies and propose a much wider selection, but then acknowledge that their selection process is poorly suited to assessing publication bias and *p*-hacking. Fourthly, we show that their *p*-curve analysis, far from demonstrating that this literature is unaffected by *p*-hacking, in fact shows the exact opposite. Sundie et al. claim that Shanks et al.'s priming manipulation was demonstrably weak, but their argument is based on a confusion between different dependent measures. We conclude that romantic priming remains unproven, and urge researchers in this field to undertake high-powered pre-registered replication studies.

KEY WORDS: risk, consumer behavior, decision making, priming, meta-analysis, *p*-curve

Can various forms of decision making, including risk-taking and consumer choices, be influenced by romantic/mating primes? Shanks et al. (2015) questioned the published evidence on this issue and made two major proposals, that past research on romantic priming should be viewed with considerable caution because it showed strong evidence of publication bias and/or *p*-hacking, and that in any case several of the key findings could not be replicated. Meta-analyses and replication studies have cast doubts on the reproducibility of other forms of social priming such as 'money' (e.g., Lodder, Ong, Grasman, & Wicherts, in press) 'flag' (e.g., Klein et al. 2014), 'intelligence' (e.g., O'Donnell et al., 2018), and 'religious' priming (e.g., Billingsley, Gomes, & McCullough, 2018). Shanks et al.'s findings suggest that the reality of romantic/mating priming is equally doubtful.

In their Comment, Sundie et al. (in press) argue that the conclusions of our article (Shanks et al., 2015) are flawed for a number of reasons. These criticisms are undermined, we argue, by a series of profound contradictions:

(1) Sundie et al. point to a failure to appreciate the contextual sensitivity of romantic priming effects as a reason why Shanks et al.'s replication experiments failed to obtain evidence of romantic priming, but beyond being a speculation, this argument rests to a substantial degree on the very effects (e.g., gender effects) we were unable to replicate. Moreover we present statistical evidence showing that effects highly similar to romantic priming are not particularly sensitive to contextual variation.

(2) Sundie et al. criticize us for failing to provide an adequate definition of the effect of interest and for "including a hodgepodge of variables that stretched the limits of what can and should be compared meta-analytically… Treating such disparate effects as belonging to a single distribution of the same effect calls into question whether any sort of meaningful interpretation is possible" (p. 8). Yet the identical set of studies formed the basis of their own *p*-curve analysis.

(3) Sundie et al. take issue with the method by which we selected effects from the set of 15 studies we identified, and conclude that our assessment of publication bias or *p*-hacking is therefore invalid. But after describing a meta-analysis based on a different selection rule – namely including all 144 available effects – they concede that "examining a funnel plot that includes all of the simple and main effects from a set of studies designed to examine interactions is poorly suited to providing any definitive information about publication bias or *p*-hacking in a literature."

(4) We critically evaluate Sundie et al.'s new *p*-curve analysis, demonstrating that it serves if anything to bolster the case for being extremely cautious about romantic priming.

In addition to highlighting these contradictions, we also point out a fallacious argument Sundie et al. make regarding the priming manipulation Shanks et al. employed. The claim that this manipulation was demonstrably weak rests on a confusion between different dependent measures. We end this Reply by making a proposal about how this field of research could profitably move forward in the future.

Contextual sensitivity of priming effects

If a given result is highly contextually sensitive, then there is a real risk that a replication will fail to recapitulate the precise context in which the effect was originally observed and fail to replicate it. Research on contextual sensitivity highlights time, location, culture and population as the major contextual variables (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). Sundie et al. begin their Comment by arguing that any conclusions drawn from the meta-analysis and experiments Shanks et al. conducted must be weak because we failed to respect the contextual sensitivity of the research we evaluated.

We do not dispute that romantic priming effects, if real, are likely to be sensitive to a range of factors, just like any other psychological measurements. But the case that Sundie et al. make for the supposed high contextual sensitivity of the original effects, beyond being a speculation, rests largely on the very evidence that is in dispute. Sundie et al.'s Table 1 lists interactions of gender (a contextual variable) with manipulations such as public versus private display, consumption versus benevolence decisions, and conspicuousness of the decision, for example, but these are all interactions we were unable to replicate. It is plainly circular to base a criticism of a failed replication on the very result that cannot be replicated.

Is there any other concrete evidence regarding the contextual sensitivity of phenomena like romantic priming? Indeed there is, although not mentioned by Sundie et al. Several multi-lab replication projects have found that priming effects very similar to romantic priming are not especially contextually sensitive, at least in the sense that they do not vary systematically across different countries, cultures, languages, samples, and so on. As shown in Table 1 (to which we return later), in the multi-lab replication projects by Ebersole et al. (2016), Klein (2014), Klein et al. (in press), and O'Donnell et al. (2018), the degree of heterogeneity across laboratories for 7 types of priming, measured by the conventional statistics Q and I^2, was small and only statistically significant in one case. For example, Klein (2014) found no heterogeneity in money priming across 36 laboratories in countries as diverse as Malaysia, Brazil, Turkey and the United States and across which the mean age of the tested participant samples varied from 18-35. The common appeal to hidden moderators as an explanation for replication failures is strongly challenged by these and related (e.g., Caruso, Shapira, & Landy, 2017) results, at least in the case of social priming.

A further contradiction is that Sundie et al.'s Table 1 lists other factors which are manipulated independent variables characterizing the stimuli or the type of decision being investigated, not contextual variables. For instance, manipulations of tactile versus visual or more versus less attractive stimuli, or of temporal versus physical length decision making measures, do not fall under the definition of contextual factors.

This is not the only inconsistency in Sundie et al.'s claims about contextual sensitivity. They state, for instance, that "in many of the original research studies critiqued by Shanks et al., participants were pre-qualified as being heterosexual in orientation, before being asked to respond to manipulations designed only for heterosexuals (e.g., men asked to review dating profiles of attractive women). Shanks et al.'s replication studies had no such prequalification, and they retained substantial percentages of participants in their replication samples who had reported a non-heterosexual orientation. Moreover, whereas most participants in the original research papers were young undergraduate students, Shanks et al. recruited participants up to age 60."

Sexual orientation and age are certainly contextual variables (relating to the population studied), but the first part of this statement is a dubious claim: Participants were pre-qualified as heterosexual in only 8 of the 15 studies Shanks et al. evaluated. Nearly half (7/15) of the studies, including two in which Sundie et al. themselves participated (Griskevicius et al., 2007; Sundie et al., 2011), made no reference to such pre-screening. In several of our experiments we did ask participants to report their sexual orientation. For example, in Experiments 7a and 7b, replicating experiments by Greitemeyer, Kastenmüller, and Fischer (2013), only 12/235 participants (5.1%) self-identified as homosexual. Excluding them makes no difference to the results, as re-analysis of the dataset (https://osf.io/ytvj7/) reveals.

Likewise, the age profile of our participant samples was nothing like as different from the original studies as Sundie et al. imply. The 120 participants in our Study 7a, for example, had a mean age of 21.7 years, very close to that in Greitemeyer's experiment (mean = 20.3). In Experiments 7a and 7b, less than 10% of participants were older than 35. Again, excluding these participants does not alter the conclusions. Moreover, contrary to Sundie et al.'s speculation, age has very little effect on behavior. In Figure 1 we plot data from Experiment 8 which was pre-registered with a total sample of 650 participants, endeavoring to replicate a study by Li, Kenrick, Griskevicius, and Neuberg (2012). From this sample we have extracted loss aversion scores for all 222 participants allocated to the romantic prime group and all 222 participants allocated to the control group (for details of the experiment and the loss aversion score, see Shanks et al., 2015). It can clearly be seen that neither age nor gender moderates loss aversion scores, nor is there any suggestion that the difference between the prime and control groups is larger for young male participants. Priming is absent across the age range[1].

In sum, the argument that our replication experiments should be discounted because we failed to recognize the contextual sensitivity of romantic priming effects is a weak one. Sundie et al. provide no evidence that romantic priming is any more contextually sensitive than any other psychological phenomena, and indeed the heterogeneity of similar priming effects is low. Their argument rests to a large extent on the very results that we were unable to replicate, and many of the factors listed in their Table 1 and described as contextual variables are actually manipulated independent variables. Moreover, nearly half the studies included in the meta-analysis reported no pre-screening for sexual orientation, and restricting our analyses on the basis of age and sexual orientation does not alter the conclusions.

Specifying the effect of interest

It goes without saying that an important aspect of meta-analysis is to specify with as much precision as possible what the effect of interest is. We described in some detail how we conducted our literature search for studies assessing the effects of mating primes, using a variety of different priming manipulations to activate mating/romantic motives/goals, on decision-making dependent variables. We highlighted the boundaries of this effect of interest, excluding for instance non-decision-making behaviors such as creativity and aggression. We defined our effect carefully, not least in terms of an explicit specification of the search terms we used to identify studies from a range of databases.

Sundie et al. (in press) argue that our meta-analysis is in effect uninterpretable because our selection of effects to include was flawed and "stretched the limits of what can and should be compared meta-analytically" (p. 8). Yet they included precisely the same set of 15 studies in their own *p*-curve analysis (see below). If the effect was ill-defined for one form of meta-analysis then it must be for other types too. Sundie et al. seem unaware of the contradiction between asserting that we should not have compared this set of studies meta-analytically while simultaneously conducting their own analysis on these same studies and drawing positive conclusions from it.

It is curious that despite criticizing our specification of the effect of interest, Sundie et al. neither identify any additional studies that we failed to include in our meta-analysis nor any that were included inappropriately. If we "used a flawed procedure to identify and exclude effects…" (p. 4) then why have Sundie et al. not provided concrete evidence, in the form of examples, to bolster this claim? The fact that the included studies employed a range of different decision-making dependent variables is consistent with standard practice, provided that due heed is paid to the magnitude of observed heterogeneity. Just to give one example, the Hagger, Wood, Stiff, and Chatzisarantis (2010) meta-analysis of the ego depletion effect included measures of the control of attention, emotion, thoughts, impulses, cognitive

processing, choice and volition, and social processing, all employing distinct dependent variables.

Regarding the effects we included, Sundie et al. say that "treating such disparate effects as belonging to a single distribution of the same effect calls into question whether any sort of meaningful interpretation is possible" (p. 8). But the measurement of heterogeneity in meta-analysis (e.g., via the I^2 statistic) is aimed precisely at determining whether a set of effects is too disparate to be analysed as a unitary set, or whether instead a moderation analysis needs to be conducted. If our selection of studies was so questionable, why was the observed heterogeneity so low (I^2 = 19.6%)? (We return to this issue below when we comment on Sundie et al.'s own meta-analysis).

Method for effect-size selection

In our meta-analysis we endeavored to include all reported tests of the hypothesis that romantic or sexual primes would affect some target decision making behaviour. We did not select effect sizes depending on the observed results (whether they were significant or not) but instead depending on the predictions of the authors (whether they were predicted to be significant).

As an example, Greitmeyer et al. (2013) gave participants in their experimental condition a priming task where they were asked to rate the attractiveness of opposite sex photos and to imagine and briefly write about a perfect date with one of the individuals. In the control condition participants rated different pictures (e.g., of streets). The effect of these primes on various risk-taking behaviors (e.g., sexual, gambling) was measured.

In our meta-analysis we included the (significant) effects Greitmeyer et al. obtained for male participants but we excluded the (null) effects for female participants. The rationale was straightforward: Greitmeyer et al. predicted the former but not the latter:

> "It was predicted that the mating prime would have differential effects on women's and men's intentions to engage in risky sex: whereas a mating prime should increase men's intentions to engage in risky sex (relative to a control prime), it should not affect women's risk-taking." (pp. 22-23).

Why did we follow the standard practice of other recent meta-analyses of social priming effects (e.g., Lodder et al., in press) and select effects on the basis of the researcher's predictions? The reason is clear: if this literature has been contaminated by publication bias/p-hacking, then *that will only be detectable by looking at the primary outcome measures* – that is, the effects that the researchers set out to test. It will not be detectable on secondary outcome measures (such as effects on females, or in males with high investment strategies)

because publication does not depend (or at least does not depend so strongly) on the results obtained with secondary measures. This in a nutshell is why our selection method paid heed to the researchers' key experimental predictions.

Sundie et al. (in press) argue (pp. 9-11) that our policy for including or excluding effects in our meta-analysis was arbitrary and applied inconsistently. In support of this charge, they offer the following example:

> Moreover, there was great inconsistency in how this "key experimental prediction" selection criteria was applied. For example, sometimes this operationalization was based on the hypotheses of the original study authors and sometimes based on the judgment of Shanks et al. (2015). Consider that Chan (2015, Experiment 1) predicted that a manipulation of romantic motivation using same-sex photos would influence risk-taking, and this effect was included in Shanks et al.'s meta-analysis, but that Hill and Durante (2011, study 1) predicted an effect of romantic motivation on risk-taking using the same manipulation as Chan—and this effect was not included by Shanks et al. in their meta-analysis. In their supplement, Shanks et al. suggest that they did not include the effect from Hill and Durante because they were only interested in opposite-sex effects. (p. 9).

Did we exclude the Hill and Durante result because we were only interested in opposite-sex effects (in which case, why include the Chan results)? Far from it. At no point did we state that our meta-analysis was restricted to opposite-sex effects (and this is not stated anywhere in our article or Supplemental materials, contrary to Sundie et al.'s assertion). So why was one of these effects included and the other excluded? Was it because we applied the selection rule inconsistently? The reason for excluding the Hill and Durante effect was, again, based on the authors' theoretical perspective. Hill and Durante included both opposite- and same-sex conditions in their experiment and conceptualized the same-sex condition as constituting a 'competition' prime. The opposite-sex condition plainly provides a better estimate of the effect of mating/sexual primes on decision making, which was our focus. By contrast, Chan took a different theoretical perspective, regarding same-sex photographs as being mating primes: "when the average heterosexual man sees attractive males, he likely perceives himself to be less physically-attractive and less desirable as a mating partner to women. Compensatory theories in psychology suggest that this perceived lack should motivate him to increase his desirability as a mating partner to women" (Chan, 2015, p. 408). Thus by the stated criterion, it was appropriate to include one of these effects but not the other (and of course note that excluding the Chan result makes a negligible difference to the asymmetry of the funnel plot).

Sundie et al. also claim that we "included a number of effects in [our] meta-analysis of "key experimental predictions" that were not even reported in the original articles… It seems doubtful that authors would fail to present tests of their "key experimental predictions" in their articles, and that editors would agree to publish such articles" (p. 10). In reality our selection procedure was anything but arbitrary, and we submit that any neutral person would choose the same effect sizes as we did if she was following our stated selection rules.

To illustrate how wide of the mark this criticism is, in the context of trying to extract statistical results for a meta-analysis, consider one of the examples Sundie et al. give. In their study of priming effects on temporal discounting, Kim and Zauberman (2013, Study 4) say:

> Specifically, for delayed monetary rewards, participants in the hot condition demonstrated decreased happiness after the manipulation (M_{before} = 102.01 mm, SD = 57.76 vs. M_{after} = 84.83 mm, SD = 48.89), $F(1, 52) = 4.74$, $p < .05$, $\omega^2 = .06$, whereas happiness ratings in the control condition were the same before and after the manipulation (M_{before} = 105.49 mm, SD = 60.95 vs. M_{after} = 109.35 mm, SD = 60.13), $F(1, 52) = 0.50$, $p = .48$, $\omega^2 = 0$. Supporting our hypothesis that <u>sexual cues induce impatience by making delayed rewards seem even less attractive</u>, we found that preference for delayed rewards decreased after the sexual-cue presentation, but immediate rewards did not become more attractive. (p. 333, underlining added).

Although the design and analysis of the study is somewhat complex, it could hardly be clearer that they key result is the reduced happiness in the hot condition (which we calculated by comparing the M_{after} ratings in the hot versus control condition), and to see its transparent link to Kim and Zauberman's hypothesis (underlined).

To conclude that an effect such as this could not have been the authors' key experimental prediction because it was not reported in the original research paper is misleading. Kim and Zauberman chose to report statistical analyses comparing scores after versus before the manipulation whereas our meta-analysis required a comparison of scores after the manipulation in the experimental versus control groups, but this is a detail. Every datapoint included in the meta-analysis required some data transformation to derive an appropriate effect size.

Sundie et al. go on to claim (p.11) that our ""key experimental prediction" inclusion criterion is… likely to produce the appearance of bias even if none exists" but this assertion rests on an incorrect characterization of our selection rule. They say:

> Suppose that an aspiring replicator was conducting a meta-analysis of a set of effects that, in truth, had no publication bias. Let us then suppose that the replicator decided on an inclusion criterion of only selecting effects that were positive and significantly

> different from zero. Because this has the same effect as publication bias (i.e., both include only significant effects), it would yield an affirmative test for publication bias. Notably, the appearance of bias in this example is due entirely to the nature of the "only significant effects" inclusion criterion employed. (p. 11).

In implying that this scenario applies to our meta-analysis, Sundie et al. are misrepresenting our selection rule. To be clear, the criterion we adopted did not require effects to be significant (indeed some were not, at the conventional $p < .05$ level), it required them to be key experimenter predictions, and there is no necessary reason why this criterion must yield a biased outcome. If a set of researchers independently conducted the same experiment in a Many Labs study, with the key experimental prediction being pre-registered, the ensuing results would be perfectly unbiased estimates. There would be no bias in selecting these effects for a meta-analysis.

It is perhaps natural that Sundie et al. think we selected studies on the basis of statistical significance. If that indeed had been our decision rule, we would have ended up selecting effect sizes highly overlapping with the effect sizes that we did extract based on the authors' predictions. This coincidence between the experimental predictions and the resulting statistically significant effects is exactly what is suspicious.

Sundie et al. criticize our calculation of bias in our funnel plot (Shanks et al., Figure 2) on the basis that "if the true effect size were zero, and non-significant effects were suppressed due to publication bias or *p*-hacking, then publications capitalizing on these biases should just as often show negative effects as they do positive effects, resulting in a biased distribution on the left and right side of zero" (p. 13). The logic of this criticism is hard to follow. If a researcher, predicting on theoretical grounds that a romantic prime would make male participants more risk-seeking, instead obtained a statistically-significant result in the wrong direction, then this result would probably be just as likely to be consigned to a file drawer as a nonsignificant result. Publication bias and/or *p*-hacking are wholly consistent with the funnel plot asymmetry we observed.

The Sundie et al. effect-size selection method

Sundie et al.'s response to this supposed selection bias is to conduct an alternative meta-analysis (Sundie et al., Figure 1) in which they apply no selection at all. They take every effect, even ones where the authors explicitly predicted a null result. Their meta-analysis includes, for instance, the effect for females (mentioned above) that Greitmeyer et al. explicitly predicted would be absent. But this approach radically changes the question the meta-analysis is seeking to address. Imagine that we're interested in knowing whether statins are effective treatments for individuals with high blood cholesterol. It would be

patently inappropriate to include in a meta-analysis studies which tested the (mainly null) effects of statins in individuals with normal cholesterol levels. Yet this is precisely what Sundie et al.'s approach amounts to, yielding a theoretically-uninterpretable meta-analysis.

This problem is illustrated by many other effects included in Sundie et al.'s meta-analysis but excluded from that of Shanks et al. Consider a study (Kim & Zauberman, 2013, Study 2) designed to test the hypothesis that "Sexual cues induce impatience through their ability to lengthen the perceived temporal distance to delayed rewards" (p. 328). These authors first showed participants either sexual or control photographs to rate, and then administered a time-perception task in which participants adjusted the length of a line on the computer display to represent a time interval (e.g., 6 months). In our meta-analysis we included the effect of priming condition on time perception. Kim and Zauberman also included a control task, described as follows: "to examine whether the effect of sexual cues is specific to judgment of future time or applies more generally, we measured participants' perception of a line length" (p. 330) – in other words to check that the priming manipulation affected time perception *per se* rather than having a general effect on perceptual judgments. In this control task, which was not affected by the prime, participants judged the length of a line shown on the screen. Obviously, we did not include this control task in our meta-analysis. *Yet Sundie et al. did*. It is hard to see what possible logic might justify this inclusion. The outcome measure was not predicted by Kim and Zauberman to show a priming effect, nor was it a decision making measure.

Hill and Durante (2011, Study 2) primed women with a mating-related or control task and then measured willingness to take risks to test

> "the hypothesis… that activating mating goals would lead to suppressed beliefs about the likelihood of incurring negative side-effects from the target health risk behaviors... Accordingly, participants filled out two types of measures for each of the target risk behaviors. First, they were asked to estimate the likelihood of experiencing negative health side-effects from the two attractiveness-enhancement risks — going tanning and using diet pills — and two control risks — using cough syrup as a sleep aid (an off-label use) and painting in a nonventilated room to avoid outside noise." (p. 389).

They observed a positive priming effect for tanning and diet pills but not for cough syrup or painting, "lending support for these effects being specific to risks associated with attractiveness enhancement" (p. 391). Again we excluded these control measures on the grounds that they are irrelevant to the focal question, as well as being predicted by the authors not to show a priming effect, while Sundie et al. included them.

It is not necessary to labor the point further. If one includes effects such as these (many of which were null effects) then it is hard to see how the outcome could be any different from what Sundie et al. obtained – a dilution of both the mean effect size and the funnel plot asymmetry, and an increase in heterogeneity. In sum, the logic for conducting an unselective meta-analysis is very hard to discern. Whatever the question is that such a meta-analysis addresses, it is not the question that Shanks et al. were concerned with: Whether the primary outcome measures reported in this literature, for which a priming effect was predicted, show evidence of publication bias/*p*-hacking. Put differently, if one is interested in testing for publication bias/*p*-hacking, then one has to focus on the dependent measures on which it is plausible to imagine these factors might have had an influence.

The inappropriateness of the meta-analysis shown in their Figure 1 is clearly not lost on Sundie et al., because they acknowledge that their selection process is poorly suited to assessing publication bias and *p*-hacking. They say: "There are, however, problems with this approach to examining bias as well… even with all of the effects included in the funnel plot, it still does a poor job of representing the effects that would most likely be subject to publication bias, or the target of *p*-hacking… Consequently, examining a funnel plot that includes all of the simple and main effects from a set of studies designed to examine interactions is poorly suited to providing any definitive information about publication bias or *p*-hacking in a literature" (pp. 14-15). We could not agree more, but are left wondering how Sundie et al. intend the meta-analysis reported in their Figure 1 to clarify rather than confuse the discussion.

Heterogeneity

Sundie et al. argue that the effects of mating motives are heterogeneous and that the rules we followed for selecting effect sizes to include in our meta-analysis artificially reduced this true heterogeneity. When a much broader selection of effects is allowed in Sundie et al.'s meta-analysis, the degree of heterogeneity increases markedly and minimal evidence of funnel plot asymmetry is obtained. Thus the argument is that in reality the effects measured across these studies are highly diverse and they should not therefore be pooled in a meta-analysis.

Against this argument, it is important to emphasize that the heterogeneity observed in their larger meta-analysis (I^2 = 58.9%) does not demonstrate that it is inappropriate to pool these effects. In fact this level of heterogeneity is in line with the average heterogeneity observed in meta-analyses generally, including in ones in which all contributing studies have highly similar designs and dependent measures. For example, recent analyses of published meta-analyses by Rubio-Aparicio, Marín-Martínez, Sánchez-Meca, and López-López (2018) and

Stanley, Carter, and Doucouliagos (2018) found median I^2 values of 60% and 74%, respectively. More directly pertinent, Table 1 lists a sample of recent meta-analyses of other priming effects. Like studies on romantic priming, these effects are investigated by measuring the effect of a prime (such as money or religious words) on a dependent measure, often in the decision making domain. The heterogeneity Sundie et al. observe in their larger meta-analysis is not at all out of line with the figures reported in these other meta-analyses. So there is no evidence that this body of effects is any more heterogeneous than is typical in behavioural research generally or priming research specifically, or that it was inappropriate of us to aggregate them for meta-analysis.

Moreover Sundie et al.'s conclusions regarding heterogeneity between their and our meta-analyses are virtually pre-ordained. Their method of challenging a meta-analysis that reveals funnel-plot asymmetry and homogeneity of effect sizes (the one reported by Shanks et al.) is to merge it with a large set of additional effect sizes yielding a meta-analysis that (a) Sundie et al. themselves concede makes little conceptual sense, and which (b) shows little asymmetry and a considerable amount of heterogeneity. But this pattern obviously lends itself to a very different interpretation: Our selection process was more appropriate than Sundie et al.'s *precisely because it yielded low heterogeneity*. By mixing together primary and secondary outcome measures, the increase in heterogeneity that Sundie et al. obtained is entirely unsurprising. The fact that the effects we studied were homogeneous exactly supports our decision to select those effects.

To make this point concrete, consider the following simulation. In this model we mix two sets of studies. In both sets the real effect size is $d = 0$ and all variation comes from sampling error (sampling random values from a central t-distribution with N-2 degrees of freedom). Therefore, in truth, there is absolutely no non-random heterogeneity across the studies nor any priming effect. Set 1, which represents the studies' primary outcome measures, undergoes publication bias. For the sake of simplicity, this is modeled by running 1000 experiments and retaining only the ones with significantly positive results (usually 2.5% of them in a two-tailed test)[2]. Set 2, which represents all the secondary outcome measures that Sundie et al. included in their meta-analysis, is completely unbiased: everything is published. This is modeled by just sampling 100 studies. Mixing both sets results in a combined dataset of about 125 studies where 100 are unbiased and 25 have been subject to publication bias. Each of these simulations yields a funnel plot like the one shown in Figure 2.

Several noteworthy results emerge in this model. First, there is significant heterogeneity across studies. If we repeat this process 100 times, the mean I^2 across simulations is 40.5% and the associated Q statistic is statistically significant in all iterations. This is interesting because, actually, there is no true heterogeneity in these datasets: They all come from a

population with a true $d = 0$ and with $SD = 0$ in the true effect. But mixing biased and unbiased studies induces a substantial amount of observed heterogeneity. Therefore, Sundie et al.'s finding that there is too much variability across effect sizes to run a meaningful meta-analysis is likely to be the result of mixing biased effect sizes from primary outcomes and unbiased effect sizes from secondary outcomes. This is also important for a second reason. The fact that heterogeneity is large in Sundie et al.'s meta-analysis might invite the reader to think that there must be some true effects. After all, if effects are different from each other it follows that not all of them can be null effects. But these simulations show that, if a subset of studies is biased, they can. Another interesting result is that Egger's regression test becomes less able to detect bias when the 100 unbiased studies are added. Across the 100 simulations, Egger's test is only significant 32% of the time.

It could be argued that the fact that our simulation produces results similar to the ones obtained by Sundie et al. does not necessarily mean that this is the process that actually produced their data set. To put this hypothesis to the test, ideally, one would need a meta-analytic model that could assume that the submitted data come from a population of effects where the true effect is close to zero but both heterogeneity and the average observed effect size have been inflated artificially by mixing (a) a set of studies affected by publication or reporting biases and (b) a set of unbiased studies. Fortunately, such models exist. For instance, the selection model devised by Vevea and Hedges (1995) includes the assumption that non-significant studies may have a lower (but nonzero) probability of being published and entered into a meta-analysis and can correct for the inflation of effect sizes and heterogeneity produced by this 'selection' process.

When this model is applied to the set of effect sizes computed by Sundie et al. (summarized in their Figure 2), it returns an average effect size that is still significant, $d = 0.12$, 95% CI [0.04, 0.19], but only half the magnitude of the effect computed by an equivalent standard random-effects meta-analysis applied to the same data set, $d = 0.24$, 95% CI [0.18, 0.31]. Heterogeneity is also halved from $\tau^2 = 0.08$ to 0.04. And, most importantly, the likelihood ratio comparing the fit of these two models is significant, $\chi^2(1) = 13.89$, $p < .001$, confirming that the selection model, which assumes publication bias, fits the data better than an unadjusted model ignoring such a possibility[3].

In sum, Sundie et al.'s meta-analysis does not remotely challenge the claim that the primary outcome measures in the target studies manifest publication bias/*p*-hacking. Because publication bias/*p*-hacking would affect the primary but not secondary measures, it is a statistical necessity that when the unbiased secondary measures are added to a meta-analysis, heterogeneity will increase and funnel-plot asymmetry will be diluted. This is both what Sundie et al.'s results and the model shown above reveal. In fact, far from indicating

that this literature is unbiased, the data set gathered by Sundie et al. shows exactly the opposite when analysed with a sufficiently powerful method.

Sundie et al.'s *p*-curve analysis

As an alternative to the Shanks et al. meta-analysis and their own one, Sundie et al. report a *p*-curve analysis (Simonsohn, Nelson, & Simmons, 2014). They state that – in contrast to our findings – their *p*-curve analysis "suggests a different, more positive conclusion with no evidence of *p*-hacking" (Abstract). In showing that the data included in this analysis has evidential value, this analysis implies that romantic priming is a true phenomenon. We make several observations about this analysis.

First, Sundie et al.'s conclusions are strikingly at variance with a completely independent analysis with a similar method reported by Schimmack (2016) which they do not refer to. Schimmack's results and conclusions are identical to those of Shanks et al. – namely that there is irregularity in the romantic priming literature. Secondly, this analysis highlights a self-contradiction. Sundie et al. criticize our protocol for selecting studies to include in our meta-analysis (see above), yet include in their *p*-curve analysis exactly the same set of 15 articles.

Thirdly, we described previously Sundie et al.'s criticism of our method of selecting effect sizes to include in our meta-analysis (based on the researchers' predictions). Ironically, this is exactly the selection rule required by *p*-curve: "Included *p* values must meet three criteria: (a) test the hypothesis of interest… the researcher's stated hypothesis determines which *p* values can and cannot be included in *p*-curve" (Simonsohn et al., 2014, p. 540). Thus Sundie et al.'s own approach contradicts their assertion that it is inappropriate to pay heed to the original research article's key experimental predictions. The mere fact that Sundie et al. used *p*-curve at all and that they followed Simonsohn et al.'s rules implicitly supports our decision to select only some contrasts.

Fourthly, another outcome of the *p*-curve analysis is the strong evidence that the body of studies is woefully under-powered, not having significantly greater power than 33% (the actual estimate of power is 37%). Sundie et al. do not comment on what this says about the published romantic priming literature, which despite this lack of power yields almost exclusively significant results.

In fact, there are good reasons to suspect that the actual average power must be even lower than 37%. By far the most delicate part of *p*-curve analysis is the selection of the crucial statistical contrast from each study. Simonsohn et al. (2014) provided detailed guidelines for the selection of *p*-values in different experimental designs. In general, the selection rules are straightforward for simple designs, such as those involving a correlation or comparing just two conditions. But as the complexity of designs and hypotheses grows, the chances of

selecting the wrong statistic increase substantially and, not surprisingly, the creators of *p*-curve analysis have expressed concerns about recurrent failures to follow their guidelines in published studies (Nelson, Simmons, & Simonsohn, 2017; Simmons, Nelson, & Simonsohn, 2017). Sundie et al. recognise explicitly that in some cases they did not follow these guidelines because the authors of the papers they assessed failed to report the appropriate contrasts. Furthermore, their disclosure table reveals that they included many omnibus tests in their analysis which, in the words of Nelson et al. (2017) "are almost never the right test to select in psychological research".

The consequences of this decision are not trivial. In Figure 3 we plot the (reciprocal) *p*-values of each of the contrasts in the main analysis of Sundie et al. As can be seen, some of the smallest *p*-values come from contrasts that are easily identifiable as omnibus tests, because they refer to *F*-tests with more than one degree of freedom in the numerator. In fact, the smallest effect size, an order of magnitude lower than the next smallest *p*-value, comes from one such contrast. (Note that Figure 3 plots reciprocal *p*-values on a logarithmic scale.) Given that *p*-curve analysis (particularly the continuous test) is highly sensitive to outliers, including or excluding these *p*-values in the analysis makes a substantial difference[4]. Once all the omnibus tests in Figure 3 are removed from the analyses, the evidential value of studies fails to reach statistical significance in a binomial test, $p = .076$, and in the half *p*-curve, $z = -0.60$, $p = .275$, although it remains significant in the continuous test, $z = -2.21$, $p = .014$. More importantly, the average estimated power after correcting for publication bias falls to 22% with 95% CI [7%, 42%], which again leads us to wonder how so many studies could yield significant support for their main hypotheses given their puzzlingly low statistical power.

Perhaps more interestingly, the selection of contrasts made by Sundie et al. for their *p*-curve analysis provides yet another opportunity to test for publication and reporting bias in this literature with a data set that we did not select and that, consequently, cannot possibly be influenced by our "agenda" (Sundie et al., p. 3). We converted each of the contrasts selected for the main *p*-curve analysis to a correlation per degree of freedom, following the same set of equations and code used in the Supplementary Material of Open Science Collaboration (2015) to compare effect sizes and test for funnel plot asymmetry. Of note, this method also excludes data from *F*-tests with more than one degree of freedom in the numerator, in this case because their standard errors cannot be computed. There is no reason why these effect sizes, coming from the analysis of disparate effects, should follow any particular distribution and, definitely, it makes no sense to wonder what the average effect size of these tests is. Yet when one plots these effect sizes against their standard errors (see Figure 4) an all too familiar shape arises, which, we think, speaks for itself.

Whatever the merits of these various arguments about study selection, the key point is that all of the proposed datasets reveal evidence of irregularity. Added to the original Shanks et al. (2015) dataset, we now have the larger set that Sundie et al. aggregated (144 effect sizes, plotted in their Figure 1), plus the dataset they deemed suitable for *p*-curve (their Figure 2), and our version of this dataset that excludes 6 effects derived from omnibus tests. As Table 2 summarises, there is evidence of publication bias and/or *p*-hacking in all of them, across a range of convergent analytic techniques.

Shanks et al.'s replication studies

The second half of Sundie et al.'s Comment argues that a series of flaws in our 8 replication studies render them largely uninformative about the original effects they sought to examine.

We acknowledge at the outset that replication experiments, like all experiments, are rarely perfect and would almost invariably, with the benefit of hindsight, be conducted differently and better. The contextual factors discussed above, such as sexual orientation and age, are good examples. Data collection is rarely unconstrained by resources. But how fair is it to claim, as Sundie et al. do, that "each Shanks et al. replication study contained multiple methodological deviations from the parallel original study" (p. 24)? To give just one counterexample, Study 7 replicated Greitmeyer et al.'s (2013) experiments on sexual risk-taking and gambling nearly exactly and indeed did so with considerably larger sample sizes, by a factor of about 3. Whereas Greitmeyer et al. obtained effect sizes of around 0.8, our replication estimates were zero. Sundie et al. make no mention of these striking replication failures. They also (p. 22) criticize our use of predominantly online samples, yet Study 7a included more participants tested in laboratory cubicles than Greitemeyer et al. themselves tested and had substantial statistical power even in this sub-sample to detect the effect Greitemeyer et al. obtained – and again, analysing data from just these participants reveals no change in the outcome.

We explicitly acknowledged differences between our studies and the originals (indeed including a Table listing the main differences), and presented considerable information (such as Bayesian analyses) relevant to judging the adequacy of our sample sizes. One of our studies was fully pre-registered, to date one of only two such studies in the entire romantic priming literature (the other is by Chiou, Wu, & Cheng, 2017), and had a sample size of 650, far larger than any other study in the field. Whatever flaws individual experiments may have had, the sheer weight of evidence across our studies makes it highly unlikely that all our results are false negatives.

Manipulation checks

Sundie et al. (pp. 23-24) make the perfectly valid point that manipulation checks are important for ensuring that a replication study is successfully manipulating the key variables in the selected participant sample. As they acknowledge, we reported data from one particularly important check which obtained a satisfactory medium-sized effect. Nonetheless, Sundie et al. criticize us for not doing more to pre-test our priming manipulations and dependent measures to confirm, for example, that items assumed to represent conspicuous consumer items were indeed judged as such by our participants.

Although it is hard to imagine that items like a car or holiday, taken from the original studies, could be viewed otherwise, a key point to emphasize is that most of the studies included in the meta-analysis reported no pre-testing of their manipulations or dependent measures whatsoever. We showed that primes increased desire for a romantic partner, but the majority of studies in the meta-analysis included no such checks.

By comparing the effect of our priming manipulation against one of their own checks (Griskevicius et al., 2006), Sundie et al. imply (p. 19) that ours may have been too weak to be an effective prime. This is an erroneous conclusion based on an elementary confusion. They cite a check (Griskevicius et al., 2006) from a study not included in the meta-analysis which obtained a very large effect size ($d > 7$) of a prime. Our own effect was much smaller ($d = 0.41$). *But the dependent measure in the Griskevicius et al. study was different from that used in our experiment*: They measured sexual arousal whereas we measured desire for a romantic partner. These are completely different dependent variables. Someone can feel sexually aroused as a result of looking at pictures of attractive individuals without any change in her/his desire for a romantic partner. Sundie et al.'s confusion is akin to the following scenario: Imagine two employees X and Y are given pay increases of unknown amounts. X's subjective life satisfaction improves by 0.01% whereas Y's bank balance increases by 50%. Therefore Y received a larger pay increase than X.

None of the studies in the meta-analysis estimated an effect size for the manipulation we tested, but if they had, what would they have found? Does the Griskevicius et al. estimate or our estimate give a better guide? There is simply no way of knowing, and crucially it cannot be inferred that our effect was weaker. Put differently, there is no reason to think that if we had used sexual arousal as our dependent measure we would have obtained a weaker effect than Griskevicius et al. did. Sundie et al.'s argument about the strength of our manipulation is fallacious and rests on comparing apples and oranges.

Conclusion

Sundie et al. claim that we misapplied the selection method we adopted (namely to base our inclusion rule on the original authors' predictions). We have argued, instead, that almost

anyone following our rule would have made the same selections that we did, and our decisions about which effects to include in our meta-analysis, based on experimenters' predictions, were appropriate. We regard it as illogical to conduct a meta-analysis which includes all effects from the studies in question. We have suggested that the new meta-analysis reported by Sundie et al. is both conceptually uninterpretable and also fails to demonstrate anything new or unexpected. It is uninterpretable because it includes effects (such as the results of a control condition measuring a dependent variable unconnected with decision making and showing, as the authors expected, a null result) that have no bearing on the original authors' predictions. Its results are unsurprising because even if it is the case (as Shanks et al. claimed) that the primary outcome measures from the target articles reveal evidence of publication bias/*p*-hacking, one would expect to see the overall effect size as well as the funnel plot asymmetry become diluted and heterogeneity increased with the inclusion of secondary effects much less likely to have been affected by publication bias/*p*-hacking. This is exactly the outcome Sundie et al. obtained and is perfectly consistent with Shanks et al.'s conclusion.

Sundie et al.'s *p*-curve analysis highlights the internal contradictions in their approach and serves, if anything, to strengthen our conclusion that publication bias and/or *p*-hacking may be a genuine problem in this field of research. Despite condemning our study selection protocol, Sundie et al. include exactly the same set of studies in their *p*-curve analysis. Despite condemning our method for selecting effects from within these studies – based on the experimenters' key predictions – they employ exactly the same method in selecting effects to include. And contrary to their claim that the analysis reveals no evidence of publication bias and/or *p*-hacking, Figure 4 shows the exact opposite: A striking and highly significant correlation between effect size and sample size in the very sub-set of data Sundie et al. thought it reasonable and appropriate to include in their analysis.

Regarding the second main part of the Comment, the points raised (i) are mostly pure speculations, (ii) were acknowledged at length by Shanks et al., (iii) and have minimal empirical support (e.g., the results do not change if sub-groups are selected based on age).

If Sundie et al. are confident that mating motives can prime decision-making measures, one might wonder whether debate about our meta-analysis and empirical results is the best way for the field to proceed. Better would be to conduct new and preferably pre-registered studies employing all the instructions, manipulation checks and so on that Sundie et al. and other evolutionary psychologists working in this field regard as critical to obtaining the key priming effects. Indeed a successful study of exactly this sort has recently been published (Chiou et al., 2017). Better still would be a multi-lab registered replication project.

Footnotes

1. Though note that the experiment was sensitive to loss aversion which across both groups and genders is significantly negative, loss aversion score = -10.3, 95% CI [-17.7, -2.9].

2. We are not implying that bias in this literature is entirely due to selective publication and that only 2.5% of the studies ever conducted are published. Questionable research practices can yield similar levels of bias more "efficiently", that is, without condemning most studies to the file drawer (e.g., Yu, Sprenger, Thomas, & Dougherty, 2014).

3. This analysis does not take into account the fact that some of these effects were statistically dependent. We are not aware of any selection model that explicitly addresses this issue.

4. It is worth briefly expanding on the pattern of results which yielded the contrast with the smallest *p* value in Sundie et al.'s dataset. This comes from an experiment (Greitemeyer et al., 2013, Experiment 3) which yielded a 2-way interaction [$F(2, 111) = 12.31$] between participant gender and condition (short-term prime vs. long-term prime vs. control). The basis of this interaction was an overall priming effect in male participants (Cohen's $d = 1.65$, combining the short- and long-term prime conditions, which did not differ). Females showed no priming effect. However Shanks et al.'s Experiment 5 failed to replicate this effect in males, despite having very high power ($1 - \beta = 1.00$) to detect an effect of the magnitude observed by Greitemeyer et al. and high power to detect an effect of half the size (power = 0.97). There is therefore a major question mark over Greitemeyer et al.'s result and by extension over any dataset which includes it.

References

Billingsley, J., Gomes, C. M., & McCullough, M. E. (2018). Implicit and explicit influences of religious cognition on Dictator Game transfers. *Royal Society Open Science, 5*. doi:10.1098/rsos.170238

Caruso, E. M., Shapira, O., & Landy, J. F. (2017). Show me the money: A systematic exploration of manipulations, moderators, and mechanisms of priming effects. *Psychological Science, 28*, 1148-1159. doi:10.1177/0956797617706161

Chan, E. Y. (2015). Physically-attractive males increase men's financial risk-taking. *Evolution and Human Behavior, 36*, 407-413.

Chiou, W.-B., Wu, W.-H., & Cheng, W. (2017). Self-control, generosity and honesty depend on exposure to pictures of the opposite sex in men but not women. *Evolution and Human Behavior, 38*, 616-625. doi:10.1016/j.evolhumbehav.2017.02.001

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68-82. doi:10.1016/j.jesp.2015.10.012

Greitemeyer, T., Kastenmüller, A., & Fischer, P. (2013). Romantic motives and risk-taking: An evolutionary approach. *Journal of Risk Research, 16*, 19-38.

Griskevicius, V., Tybur, J. M., Sundie, J. M., Cialdini, R. B., Miller, G. F., & Kenrick, D. T. (2007). Blatant benevolence and conspicuous consumption: When romantic motives elicit strategic costly signals. *Journal of Personality and Social Psychology, 93*, 85-102.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin, 136*, 495-525. doi:10.1037/a0019486

Klein, R. A. (2014). Investigating variation in replicability: A “Many Labs” replication project. *Social Psychology, 45*, 142-152.

Klein, R. A. (in press). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*.

Li, Y. J., Kenrick, D. T., Griskevicius, V., & Neuberg, S. L. (2012). Economic decision biases and fundamental motivations: How mating and self-protection alter loss aversion. *Journal of Personality and Social Psychology, 102*, 550-561.

Lodder, P., Ong, H. H., Grasman, R. P. P. P., & Wicherts, J. M. (in press). A comprehensive meta-analysis of money priming. *Journal of Experimental Psychology: General.*

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2017). Forthcoming in JPSP: A non-diagnostic audit of psychological research. Retrieved from http://datacolada.org/60

O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., . . . Zrubka, M. (2018). Registered Replication Report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science, 13*(2), 268-294. doi:10.1177/1745691618755704

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*. doi:10.1126/science.aac4716

Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., & López-López, J. A. (2018). A methodological review of meta-analyses of the effectiveness of clinical psychology treatments. *Behavior Research Methods, 50*, 2057-2073. doi:10.3758/s13428-017-0973-8

Schimmack, U. (2016). Replicability report no.2: Do mating primes have a replicable effects on behavior? Retrieved from https://replicationindex.wordpress.com/2016/05/21/replicability-report-no-2-do-mating-primes-have-a-replicable-effects-on-behavior/

Shanks, D. R., Vadillo, M. A., Riedel, B., Clymo, A., Govind, S., Hickin, N., . . . Puhlmann, L. M. C. (2015). Romance, Risk, and Replication: Can Consumer Choices and Risk-Taking Be Primed by Mating Motives? *Journal of Experimental Psychology-General, 144*(6), E142-E158. doi:10.1037/xge0000116

Shariff, A. F., Willard, A. K., Andersen, T., & Norenzayan, A. (2016). Religious priming: A meta-analysis with a focus on prosociality. *Personality and Social Psychology Review, 20*, 27-48.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2017). Outliers: Evaluating a new p-curve of power poses. Retrieved from http://datacolada.org/66

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file drawer. *Journal of Experimental Psychology: General, 143*, 534-547.

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144*, 1325-1346. doi:10.1037/bul0000169

Sundie, J. M., Beal, D. J., Neuberg, S. L., & Kenrick, D. T. (in press). Moving beyond unwise replication practices: The case of romantic motivation. *Journal of Experimental Psychology: General*.

Sundie, J. M., Kenrick, D. T., Griskevicius, V., Tybur, J. M., Vohs, K. D., & Beal, D. J. (2011). Peacocks, Porsches, and Thorstein Veblen: Conspicuous consumption as a sexual signaling system. *Journal of Personality and Social Psychology, 100*, 664-680.

Vadillo, M. A., Hardwicke, T. E., & Shanks, D. R. (2016). Selection bias, vote counting, and money-priming effects: A comment on Rohrer, Pashler, and Harris (2015) and Vohs

(2015). *Journal of Experimental Psychology: General, 145*, 655-663. doi:10.1037/xge0000157

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, 113*, 6454-6459. doi:10.1073/pnas.1521897113

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*, 419-435. doi:10.1007/bf02294384

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., & Albarracín, D. (2016). From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin, 142*, 472-497.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review, 21*, 268-282. doi:10.3758/s13423-013-0495-z

Table 1

Measures of heterogeneity (*Q* and I^2) in the Shanks et al. (2015) meta-analysis of romantic priming and published meta-analyses of related priming effects.

Study	Domain	*Q*	*df*	*p*	I^2
Meta-analyses of published literature					
Shanks et al. (2015)	**Romantic priming**	**53.7**	**42**	**.11**	**19.6%**
Billingsley et al. (2018)	Religious priming	43.4	8	< .001	88.5%
Lodder et al. (in press)	Money priming	1048.7	245	< .001	81.3%
Shariff et al. (2016)	Religious priming	195.2	91	< .001	53.4%
Weingarten et al. (2016)	Action priming	934.8	351	< .001	62.5%
Vadillo et al. (2016)	Money priming	441.7	99	< .001	81.5%
Meta-analyses of multi-lab replication studies					
Ebersole et al. (2016)	Metaphoric restructuring	21.9	19	.29	18.2%
	Warmth perceptions	16.9	20	.66	0.0%
Klein et al. (2014)	Money priming	28.4	35	.78	0.0%
	Flag priming	30.3	35	.69	0.0%
Klein et al. (in press)	Consumption priming	63.8	53	.15	12.0%
	Warmth priming	73.0	46	.01	37.0%
O'Donnell et al. (2018)	Intelligence priming	28.1	22	.17	17.4%

Table 2

Datasets of romantic priming effects and relevant evidence of publication bias/*p*-hacking in each.

Source	Dataset	No. of articles	No. of effects	Evidence of publication bias/p-hacking?	Support
Shanks et al. (2015)	Figure 2	15	43	Yes	Funnel plot asymmetry, Egger's test, $t(41) = 6.24$, $p < .0001$
Sundie et al. (in press)	Meta-analysis (Figure 1)	15	144	Yes	Vevea and Hedges (1995) selection model fits better than random-effects meta-analysis, $\chi^2(1) = 13.89$, $p < .001$
	p-curve (Figure 2)	15	32	Yes	Power = 37%
	Omnibus tests removed	15	26	Yes	(i) Power = 22% (ii) Nonsignificant right skew, $p_{Half} = .275$ (iii) Funnel plot asymmetry, Egger's test, $t(23) = 6.48$, $p < .0001$ (Figure 4)

Figure Captions

Figure 1

Scatterplot of loss aversion scores as a function of age for male (left panel) and female (right panel) participants in the prime and control groups of Shanks et al.'s (2015) Experiment 8. If romantic primes make males but not females less averse to losses and do so in an age-dependent manner then the best-fitting regression lines in the left but not right panel should be more widely separated (scores lower in the control [red dotted line] than in the prime [blue dotted line] group) in young than older participants. This is clearly not the case, with the lines virtually superimposed (no priming) across the age range for both male and female participants. See the online article for the color version of this figure.

Figure 2

Funnel plot based on the model described in the text. Each dot depicts the effect size (Cohen's *d*) of a simulated study plotted against the inverse of that study's SE. The studies come from two sets, a biased one (red dots) and an unbiased one (green dots). In both sets the real effect size is $d = 0$ but publication bias is applied to the biased set such that only statistically significant effects are included. See the online article for the color version of this figure.

Figure 3

Reciprocal *p* values (log scale) for all effects included by Sundie et al. in their *p*-curve analysis. Red bars depict omnibus tests (*F*-tests with more than one degree of freedom in the numerator), green bars depict the remaining tests.

Figure 4

Funnel plot of the data selected by Sundie et al. (in press) for their *p*-curve analysis, omitting results from omnibus tests. Each dot depicts effect size (Fisher's *z*-transformed correlation) against the inverse SE. See the online article for the color version of this figure.