# The Prior Odds of Testing a True Effect in Cognitive and Social Psychology

Brent M. Wilson and John T. Wixted

University of California, San Diego

## Author Note

Correspondence concerning this article should be addressed to Brent M. Wilson, Department of Psychology, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093. Email: b6wilson@ucsd.edu

**Abstract**

Efforts to increase replication rates in psychology generally consist of recommended improvements to methodology, such as increasing sample sizes to increase power and using less flexible statistical analyses. However, little attention has been paid to how the prior odds (*R*) that a tested effect is true affect the probability that a significant result will replicate. The lower *R* is, the less likely a published result will replicate even if power is high. It follows that if *R* is lower in one set of studies than in another, then all else being equal, published results will be less replicable in the set with lower *R*. We illustrate this approach by analyzing data from the social psychology and cognitive psychology studies that were replicated as part of the Open Science Collaboration (2015). We find that *R* is lower for the social psychology studies than for the cognitive psychology studies, which might explain the difference in replication rates. This difference may reflect the degree to which the two fields value risky, but potentially groundbreaking, research. We show that if a field prefers risky research, then in order to achieve replication rates comparable to fields that prefer less risky research, it needs to either use an especially low alpha level and/or conduct experiments that have especially high power.

# The Prior Odds of Testing a True Effect in Cognitive and Social Psychology

In years gone by, it was often assumed that if a low-powered (e.g., small-*n*) experiment yielded a significant result, then not only was the effect probably real, its effect size was probably large as well. After all, how would the effect have been detected if that were not the case? If anything, low-powered studies reporting significant effects once seemed like a good thing, not a bad thing. In recent years, it has become apparent that there is a problem with this line of thinking. As it turns out, if a field typically runs low-powered experiments, the significant effects reported in its scientific journals will often be false positives, not real effects with large effect sizes (Button et al., 2013).

The probability that a significant effect is real is known as the positive predictive value (*PPV*). Button et al. (2013) pointed out that the equation specifying the relationship between *PPV* and power is:

$$PPV = [(1 - \beta) \times R] \,/\, [(1 - \beta) \times R + \alpha] \qquad (1)$$

where $1 - \beta$ represents power, $\alpha$ represents the Type I error rate (usually .05), and $R$ represents the pre-study odds (i.e., the odds that real effects are investigated by the scientific field in question). The potential importance of $R$ – that is, the potential importance of the base rate of tested effects with non-zero effect-sizes among the totality of effects subjected to empirical investigation in a given scientific field – is the main focus of our article.

Button et al. (2013) showed that a significant effect is more likely to be real if it was obtained with a high-powered study than with a low-powered study. All else being equal, if Field A runs high-powered studies and Field B runs low-powered

studies, then significant effects from Field A will be more likely to replicate than significant effects from Field B (which is to say that *PPV* will be higher for Field A than Field B). In their calculations, Button et al. held *R* constant to illustrate the point that power affects *PPV*. However, others have noted the importance of differing prior odds (e.g., Overall, 1969; Lakens & Evers, 2014).

### *R* matters too

It is well known in the medical literature that base rates play a critical role in the likelihood of a positive test result actually indicating a true positive result as opposed to a false positive result (Gigerenzer, 2015; Hoffrage & Gigerenzer, 1998). Consider a disease so rare that for every one person who has the disease, 100,000 do not, and assume a test so diagnostic of the disease that it is correct 99.99% of the time (true positive rate = .9999, false positive rate = .0001). What are the odds that a person who tests positive actually has the disease? The answer is provided by Bayes' rule, which, in odds form, is given by

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}$$

In this example, the prior odds are 1 / 100,000, and the likelihood ratio is the true positive rate divided by the false positive rate, or .9999 / .0001. Multiplying the prior odds by the likelihood ratio yields a posterior odds of ~.1. In other words, despite the test being incredibly diagnostic, a positive test result means that there is only a one in ten chance that the disease is actually present. Expressed as a probability, *PPV* = .1 / (.1 + 1) = .09.

What if, instead of having no a priori reason to believe that the rare disease was present, the next tested individual was already showing signs of the disease

(e.g., a distinctive skin-coloring pattern)? Imagine that prior research has shown that among people with that symptom, the odds that they have the disease are 50/50. Under those conditions, the prior odds are even, in which case Bayes' rule indicates that the posterior odds for a positive test result would now equal 9999-to-1 (i.e., *PPV* = .9999).

This example illustrates the fact that when there is a good reason to believe that someone actually has the disease *before* running the test, a positive test result can strongly imply that a person has the disease (*PPV* is high). By contrast, when there is no reason to believe that someone actually has the disease before running the test, a positive test result – even from a highly diagnostic test – may strongly imply that a person does not have the disease (*PPV* is low). Although such a test will update the prior odds from being extremely low to being much higher than they were before, the posterior odds (and *PPV*) can still weigh heavily against the disease being present.

Just as the prior odds that someone has a disease affect the meaning of a positive test, the prior odds that an experiment is testing a true effect (i.e., *R*) should influence our belief in that effect following a significant result. In the above example, even an extremely diagnostic test did not strongly imply a true positive result when the test was run without any prior reason for conducting the test in the first place. As strong as the research methodology in psychology may be, no one would argue that psychology experiments are anywhere near 99.99% diagnostic of a true effect (correctly detecting true effects 99.99% of the time and generating false positives 0.01% of the time). It is therefore important to consider the factors that determine

which specific hypotheses among the entire population of hypotheses scientists choose to test when they conduct an experiment. The fewer signs in advance of the experiment that the effect might be true, the lower *R* is likely to be (just as in the medical example above).

### *R* may differ across types of studies

To investigate the potential role of *R*, we compared two sets of studies from different subfields of experimental psychology – cognitive psychology and social psychology – which the Open Science Collaboration (2015) project found to have different replication rates (and, therefore, different *PPV*s) . They conducted replications of 100 quasi-randomly sampled studies published in three psychology journals. Of the 100 replicated studies, 57 were from social psychology and 43 were from cognitive psychology[1]. They reported that 50% of findings in cognitive psychology and 25% of findings in social psychology replicated at the $p < .05$ level. This difference in replication rates was significant, $z = 2.49$, $p = .013$. A larger percentage of the originally reported effects were likely real because not every real effect will be detected when studies are replicated, and the miss rate likely exceeds the false positive rate. One way to estimate how many of the reported effects were real is to consider how many findings yielded a replicated effect in the same direction as the original study. The proportion of replicated effects that were in the same direction as the original effect was significantly higher for cognitive psychology (.905) than for social psychology (.745), $z = 2.00$, $p = .046$. Because this

[1] Three studies were excluded from all analyses (two in social and one in cognitive). In all three cases, the *p*-values in the original papers were greater than 0.10 and were interpreted as an effect not being present in the original publications.

was a relatively small-*n* study, it seems fair to say that this statistical comparison does not provide strong evidence for a true difference between the two subfields. Nevertheless, we use these point estimates to illustrate how *R* might help to account for the differing replication rates of the cognitive and social psychology studies included in the Open Science Collaboration.

Using binary logic for the time being, we assume that the observed proportion of studies yielding effects in the same direction, $\omega$, is equal to the proportion of true effects, *PPV*, plus half of the remaining 1 – *PPV* non-effects, which would be expected to yield an effect in the same direction as the original study 50% of the time due to chance. In other words, $\omega \approx PPV + .50(1 - PPV)$. Solving for *PPV* yields $PPV \approx 2\omega - 1$ (see Box 1 for more details). For cognitive psychology, $\omega = .905$, so $PPV = 2(.905) - 1 = .81$. For social psychology, $\omega = .745$, so $PPV = 2(.745) - 1 = .49$. In other words, using this measure of the proportion of "real" effects originally reported as being significant, 81% of reported cognitive psychology effects were real, whereas only 49% of reported social psychology effects were real. Why the difference?

As noted in the Open Science Collaboration (2015) article, "[r]eproducibility may vary by subdiscipline in psychology because of differing practices. For example, within-subjects designs are more common in cognitive than social psychology, and these designs often have greater power to detect effects with the same number of participants" (p. aac4716-2). Their analysis focuses on experimental design issues (and resulting differences in statistical power) to explain the difference in reproducibility rates between cognitive and social psychology studies. Such an

interpretation carries with it the implication that *R* is the same for both sets of studies and that increasing statistical power in social psychology would have yielded replication rates comparable to those in cognitive psychology. We next consider whether a difference in statistical power in the cognitive and social psychology studies included in the Open Science Collaboration is sufficient to explain that outcome or whether a difference in *R* – the odds that tested effects are real – is also needed to explain the difference. In this analysis, we treat the point estimates from the Open Science Collaboration as being the true values even though there is likely to be error in those estimates.

**Can a difference in power explain the difference in *PPV*?**

According to Equation 1, *PPV* is affected by three variables: power (1- $\beta$), base rate or prior odds (*R*), and alpha level ($\alpha$). Both social and cognitive psychology generally follow the same $p < .05$ convention for their alpha level, so the difference in that variable likely does not explain the difference in *PPV*. This leaves us with two other variables that might explain differences in the percentage of claimed discoveries that actually are true: power (1 - $\beta$) and the base-rate of true effects (*R*). If either (or both) of these variables is lower in social psychology, the percentage of claimed discoveries that are actually true (*PPV*) would also be lower. The usual assumption is that studies in social psychology are underpowered compared to cognitive psychology, thereby explaining the lower replication rate. As discussed above, Equation 1 predicts that low power alone is theoretically sufficient to have that effect. However, another possibility is that the cognitive and social psychology studies differed in *R*, not power. Assume for the sake of argument that power was

actually the same for the two sets of studies and, for both, was equal to 80%. With power specified at 80% and $\alpha$ set to .05 for both, and with *PPV* set to .49 for social psychology and .81 for cognitive psychology (based on the analysis presented earlier), we can use Equation 1 to solve for *R*, separately for the cognitive and social psychology studies. Doing so yields an estimate of 0.06 ($R \approx 1 / 16$) for social psychology and 0.27 ($R \approx 1 / 4$) for cognitive psychology. That is, if these point estimates from the Open Science Collaboration are representative of the field at large, then in cognitive psychology, for every real effect that is subjected to empirical test, there are about 4 non-effects that are subjected to empirical test. For social psychology, for every real effect that is subjected to empirical test, there are about 16 non-effects that are subjected to empirical test. Keep in mind that these estimates of *R* pertain to the number of true effects out of the totality of effects that are subjected to empirical investigation (prior odds), not to the number of true effects in the published literature out of the totality of published effects (i.e., not to posterior odds or to *PPV*).

Although we equated power at 80% in the example above, the underlying base rate of true effects would be lower for the social psychology studies than for the cognitive psychology studies for any level of (equated) power we might choose. Thus, although the observed difference in replication rates for the cognitive and social psychology studies can be accounted for by assuming that the original studies differed in power (which is the usual assumption), those differences can also be accounted for by assuming that the studies were equated in terms of power but differed in the probability of true effects tested.

How can we determine whether one explanation actually accounts for the observed difference in the replication rates (namely, differential power or differential *R*) or whether both explanations play a role? Some insight into the question of whether or not the originally published cognitive and social psychology studies differed in power can be obtained by examining the *p*-curves for those studies, separately for cognitive and social psychology. Simonsohn, Nelson, and Simmons (2014) performed simulations for various levels of power and showed what the distribution of *p*-values look like for 50% and 80% power. The distributions will be right-skewed in both cases, but as power increases, the proportion of significant *p*-values less than .01 increases. In their simulations, with 50% power, about 50% of the significant *p*-values were less than .01. With 80% power, about 70% of the significant *p*-values were less than .01. Simonsohn et al. (2014) found that these patterns held true for a wide range of effect sizes and sample sizes. Thus, for example, the *p*-curve for 50% power was similar whether Cohen's $d = 0.64$ and $n = 20$ or Cohen's $d = 0.28$ and $n = 100$. Moreover, they concluded that *p*-curve is robust to heterogeneity of effect size (see "Supplement 2. Robustness to heterogeneity of effect size" in Simonsohn et al., 2014), however, others disagree with that claim (Schimmack & Brunner, 2017). Although the issue remains unresolved, we used *p*-curves to obtain tentative estimates of power for illustrative purposes.

Figure 1 shows the *p*-curves (distributions of significant *p*-values) for the original studies published in cognitive and social psychology reported by the Open Science Collaboration (2015). These *p*-curves are consistent with the widely-held

view that social psychology studies have lower power than cognitive psychology studies. More specifically, the *p*-curve for the originally published cognitive psychology studies in the Open Science Collaboration shown here in Figure 1A resembles the hypothetical *p*-curves associated with 80% power in Figure 1 of Simonsohn et al. (2014). The *p*-curve for the originally published social psychology studies in the Open Science Collaboration shown here in Figure 1B resembles the hypothetical *p*-curves associated with 50% power in Figure 1 of Simonsohn et al. (2014). Whether or not these exact power estimates are correct, the *p*-curve data are consistent with the prevailing view that power was lower in the original social psychology studies than in the original cognitive psychology studies.

Although a difference in power may provide part of the explanation for the difference in replication rates, it does not seem to explain the entire difference. If we set $PPV = .49$, $\alpha = .05$, and $1 - \beta$ (i.e., power) = .50, for social psychology and we set the corresponding values for cognitive psychology to $PPV = .81$, $\alpha = .05$, and $1 - \beta = .80$, we can use Equation 1 to solve for $R$ (separately for cognitive and social psychology). When rearranged to solve for $R$, Equation 1 becomes:

$$R = \alpha PPV \,/\, [(1 - \beta) \times (1 - PPV)] \qquad (2)$$

Using Equation 2 and the values specified above, $R$ works out to be 0.24 (odds = 1 to ~4) for cognitive psychology and 0.10 (odds = 1 to 10) for social psychology. In other words, according to this analysis, even allowing for the fact that the cognitive and social psychology studies differed substantially in statistical power (50% for social vs. 80% for cognitive in this example), they also differed in the likelihood that a tested effect is real. Although this analysis is based on simplified binary logic

(according to which effect sizes are real or not) and assumes these point estimates for *R* and power are the true values, essentially the same conclusion applies under the more realistic assumption that effect sizes are continuously distributed (see Box 2).

An alternative way of making this same point is to assume that *R* was actually the same for the cognitive and social psychology studies and then to consider just *how* much lower power would have to be for the social psychology studies for the two sets of studies to yield *PPV* values of .81 and .49, respectively. For this analysis, we rearrange Equation 1 again, this time solving for power:

$$(1 - \beta) = \alpha PPV / [R \times (1 - PPV)] \quad (3)$$

Once again, we set *PPV* in Equation 3 to .81 for the cognitive psychology studies and .49 for the social psychology studies, with alpha set to .05 for both fields. For a given value of *R* (which is constrained to be equal for cognitive and social psychology), we can then estimate power. For example, if $R = 0.266$ for both, then Equation 3 indicates that power for the cognitive psychology studies was .80, whereas power for the social psychology studies was only .18. Did power actually differ between the cognitive and social psychology studies to that great of an extent? Perhaps, but then one would need to explain why the *p*-curve for social psychology (in Figure 1) is closer to the 50% power *p*-curve than the 25% power *p*-curve in the Simonsohn et al. (2014) simulations.

Figure 2A shows a generalization of this example. More specifically, the figure shows what statistical power would have to have been in social psychology for any level of power in cognitive psychology under the assumption that the

underlying base rates are the same in the two sets of studies. If the cognitive psychology studies had 100% power (which occurs if $R = 0.213$ according to Equation 3), then, assuming an equivalent $R$, Equation 3 indicates that power for the social psychology studies would be only 23%. In other words, 23% would be the maximum power that the social psychology studies could have had if the two sets of studies were equated with respect to $R$. As far as we can tell, no one has argued (and the *p*-curve analysis does not suggest) that the power differential between the cognitive and social psychology studies was that extreme. Instead, it would seem more reasonable to assume that although power may differ, $R$ differed as well. Similarly, Figure 2B shows that $R$ would have to be vastly lower in social psychology than cognitive psychology in order for the two sets of studies to have been equated with respect to power. Given the large effects that one would otherwise have to assume, it seems reasonable to assume that both factors (lower power and lower $R$) played a role in the lower rate of replication observed for social psychology. We cannot conclusively state that the differences in $R$ between cognitive and social psychology that were observed in the Open Science Collaboration (2015) dataset will generalize more broadly to those subfields of psychology, but it seems reasonable to suppose that they might.

**Factors Affecting *R***

What factors determine the prior odds that a tested hypothesis is true? Our assumption is that established knowledge is the key consideration, just as it is in the medical domain when considering the prior odds that a disease is present. Testing a hypothesis based on established knowledge – acquired either from scientific

research or from common experience (e.g., going without sleep makes you tired) – increases *R* compared to relying on less dependable sources of knowledge, such as idiosyncratic hunches or theories grounded in a scientist's own personal experience.

To illustrate an extreme example of low-*R* research, imagine we were walking around the grocery store one afternoon and eyed some balloons on our way to the checkout line. We decide to buy them and randomly place them outside the doors of some classrooms and not others based on a novel theory that doing so would boost morale and help to address the problem of scholastic underachievement. This theory might be based on the experimenter's own childhood experiences with the effect of ambient balloons on his or her motivation to excel in math. We then measure student learning in the classrooms with and without balloons at the entranceways. Let's suppose we even get a statistically significant difference ($p < .05$) between the learning scores in the two conditions. The danger with rushing to publish these surprising results and starting our new Balloon Brain business designed to harness the power of balloons in classrooms is that the prior odds of a hypothesis being true based on a theory generated in the way that the balloon idea was generated (i.e., an idea largely untethered to established knowledge) were low.

The results of low-*R* research will be more surprising than the results of high-*R* research because, by definition, a result is surprising to the extent that it violates one's priors. In the medical context, for example, if someone who is already showing signs of a rare disease (high prior odds) tests positive, that outcome would probably not be viewed as very surprising. By contrast, if someone who is showing

no signs of a rare disease (low prior odds) nevertheless tests positive, that outcome probably *would* be viewed as surprising. Conceivably, social psychology places higher value on surprising findings – that is, findings that reflect a greater departure from what is already known – than cognitive psychology (e.g., editors of social psychology journals might place greater weight on the surprisingness factor). All else being equal, a difference in preference along that dimension would lead to a difference in *R* between the two fields (lower for social psychology). In agreement with this line of reasoning, The Open Science Collaboration (2015) asked independent coders to rate how surprising and how exciting/important the findings reported in the original studies were using a 6-point scale[2]. When these two measures were averaged together, the original cognitive psychology studies were rated as being lower on surprising and exciting (mean = 3.04) than the original social psychology studies (mean = 3.33), $t(110) = -2.25$, $p = .026$.

A difference in preference for more surprising vs. less surprising findings would not be an automatic indictment of either field. Indeed, there is an inherent tension between the degree to which a study would advance knowledge and the likelihood that a reported effect is replicable. Different fields may have different preferences for where they would like to operate on the *R* continuum. If *R* for a particular field is 0, the published literature would likely be very exciting to read (i.e., large apparent leaps in knowledge, such as people having ESP), but none of it would be true. At the other extreme, if *R* for a particular field is infinitely high, the published literature would all be true, but the results of most experiments might be

[2] This analysis includes some additional studies that were coded but for which replications were not completed.

so pre-experimentally obvious as to be useless. Maximizing *PPV* solely by maximizing *R* would serve only to "…enshrine trivial, safe science" (Mayo & Morey, 2017, p. 26). We can imagine pages full of findings reporting that people are hungry after missing a meal or that people are sleepy after staying up all night. Neither of these scenarios ($R = 0$ vs. $R = \infty$) would be ideal for advancing our understanding of the world. The ideal point on the *R* continuum lies somewhere in between, but specifying the optimal point is difficult even though it follows that increasing *R* increases replicability (see Miller & Ulrich, 2016).

**Implications**

In response to concerns regarding replication rates in experimental psychology, many methodological recommendations have been made, but they focus on factors that affect power rather than *R*. For example, *Psychological Science* now asks submitting authors "to explain why they believe that the sample sizes in the studies they report were appropriate" (Association for Psychological Science, 2016). The Attitudes and Social Cognition section of *Journal of Personality and Social Psychology* now requires authors to include "a broad discussion on how the authors sought to maximize power" (American Psychological Association, 2017). *Psychonomic Bulletin & Review* now instructs submitting authors that, "It is important to address the issue of statistical power. . .Studies with low statistical power produce inherently ambiguous results because they often fail to replicate"

(Psychonomic Society, 2017). Equation 1 makes it clear why that is, but power is not the only factor that will affect replication rates.[3]

In addition to a growing understanding of the importance of power, the field is becoming increasingly aware of how the alpha level affects the likelihood of a false positive. For example, Lindsay (2015) suggested that studies where *p* is just barely below .05 should be regarded with more skepticism than is typically the case. Along the same lines, Benjamin et al. (2017) recently proposed that the standard for statistical significance for claims of new discoveries should be $p < .005$ rather than $p < .05$ in order to make published findings more replicable. These developments underscore the fact that, in addition to running studies with higher power, another way to increase *PPV* would be to lower the alpha level (Equation 1).

If our analysis is correct, then neither of these approaches, if applied non-differentially, would do away with the difference in replicability for the cognitive and social psychology studies we analyzed. If *R* differs for the studies included in the Open Science Collaboration (2015), then even if power were set to 80% and alpha were set to .005, the cognitive psychology studies would still be more likely to replicate than social psychology studies. Given our prior estimate of *R* for the cognitive psychology studies, Equation 1 indicates that these standards would result in 97.7% of published cognitive psychology effects being true. For the social psychology studies, the result would be 93.9% being true (a result that could be achieved in cognitive psychology using an alpha level of .014). Therefore, if a field

[3] Our analysis is based on the factors included in Equation 1 and is therefore predicated on the assumption that other factors that might affect replicability (e.g., differences in the frequency of *p*-hacking) are equated across the two fields.

prefers to engage in riskier (low-*R*) research, either higher power or a lower alpha (or some combination of the two) would be needed in order to achieve the same *PPV* as a field that engages in less risky research.

If, instead, the same methodological standards are applied to both fields (e.g., 80% power and alpha of .005), and if both fields strive to achieve the same high *PPV* (i.e., to achieve the same posterior odds in Bayes' rule), then *R* would need to be increased for the low-*R* discipline. One way to increase *R* would be to base new experiments more directly on knowledge derived from prior scientific research rather than testing hypotheses that move farther away from what is scientifically known. A change in focus like that would presumably happen only if editors of top journals in a high-risk (low-*R*) field placed slightly less emphasis than they usually do on the novelty of new findings (in which case scientists themselves might do the same). Low-*R* research is likely to be surprising and exciting (e.g., ESP is real), but unless it has differentially high power or a differentially low alpha level, it is likely to be less replicable than high-*R* research.

**Open Practices**

All analysis code has been made publicly available via the Open Science Framework and can be accessed at osf.io/qrykc.

## References

American Psychological Association. (2017). Attitudes and social cognition section. Retrieved from http://www.apa.org/pubs/journals/psp/?tab=4

Association for Psychological Science. (2016). Submission guidelines. Retrieved from http://www.psychologicalscience.org/publications/psychological_science/ps-submissions

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, *R*., ... Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behavior*. http://dx.doi.org/10.1038/s41562-017-0189-z.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. *R*. (2009). *Introduction to meta-analysis*. West Sussex, UK: John Wiley & Sons, Ltd.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. *R*. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376.

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science, 12* (1), 46-61.

Gelman, A. & Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science, 9*, 641–651.

Gigerenzer, G. (2015). *Calculated risks: How to know when numbers deceive you*. New York: Simon and Schuster.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*(5), 538-540.

Johnson, V. E., Payne, *R*. D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association*, *112*(517), 1-10.

Lakens, D., & Evers, E. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives in Psychological Science*, *9*, 278–292.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*(12), 1827–1832. doi:10.1177/0956797615616374.

Mayo, D., & Morey, *R*. D. (2017, July 26). A Poor Prognosis for the Diagnostic Screening Critique of Statistical Tests. Retrieved from osf.io/ps38b

Miller, J. & Ulrich, *R*. (2016). Optimizing research payoff. *Perspectives on Psychological Science, 11*, 664–691.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi:10.1126/science.aac4716.

Overall, J. E. (1969). Classical statistical hypotheses testing within the context of Bayesian theory. *Psychological Bulletin*, *71*(4), 285.

Park, J.H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K. B., Chanock, S. J., & Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics, 42*(7), 570–575.

Psychonomic Society. (2017). Instructions for authors statistical guidelines. Retrieved from http://www.springer.com/psychology/cognitive+psychology/journal/13423

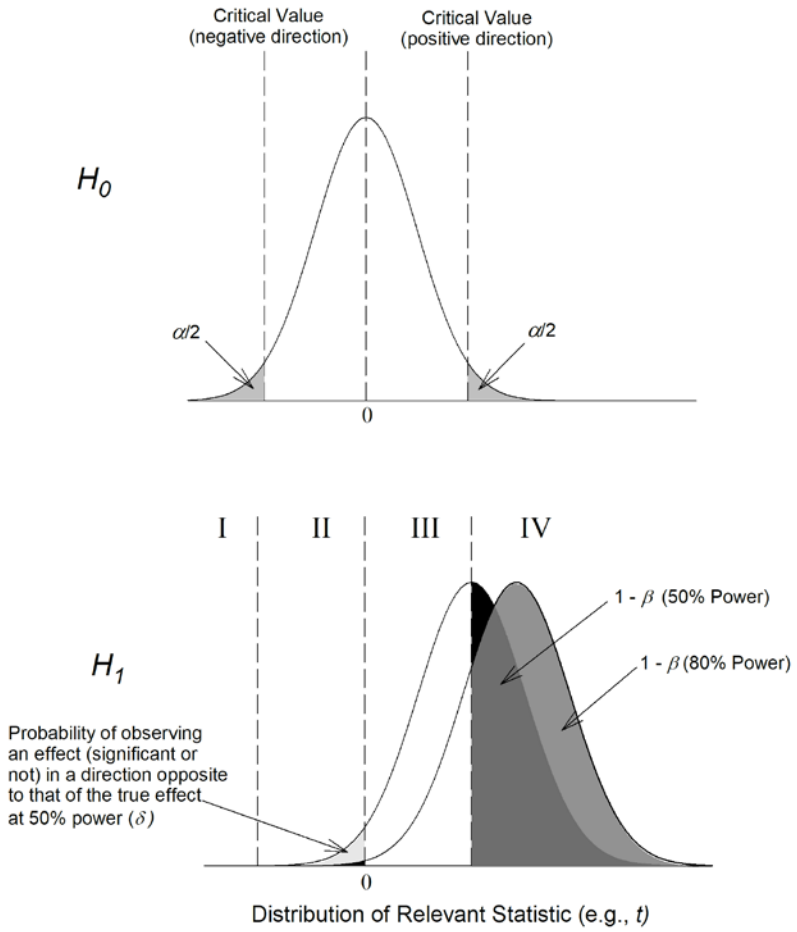
Schimmack , U., & Brunner , J. (2017, November 16). Z-Curve: A Method for Estimating Replicability Based on Test Statistics in Original Studies. Retrieved from https://replicationindex.wordpress.com/2017/11/16/preprint-z-curve-a-method-for-the-estimating-replicability-based-on-test-statistics-in-original-studies-schimmack-brunner-2017/

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-Curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666-681.

Szucs, D., & Ioannidis J. P. A. (2017) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, 15(3): e2000797. doi:10.1371/journal.pbio.2000797.

**In Detail Box 1 – Calculating *PPV***

*PPV* can be estimated by understanding the probability of different outcomes. $H_0$ shows a distribution of effects that are not real. It is centered on zero because any effects are due purely to noise. Half of the time effects observed when $H_0$ is true will be positive, and half of the time effects observed when $H_0$ is true will be negative. The area shaded in the upper and lower tails of $H_0$ represent the Type I error rate. $H_1$ shows two distributions of effects that are real, one for 50% power and the other for 80% power, where either 50% or 80% of the distribution, respectively, exceeds the critical value of the test statistic. If a replication experiment has 50% power, the probability of a real effect going in the opposite direction as the original experiment (denoted as $\delta$) is ~.02. If a replication experiment instead has 80% power, $\delta \approx .003$ (a probability so small that it is hard to see in the figure). Thus, the full equation for $\omega$ is $\omega = (1-\delta)PPV + .5(1-PPV)$ such that $PPV = (\omega - .5) / (.5 - \delta)$. Because $\delta$ is likely



negligible for the replication studies, we rely on the simpler equation, $PPV = 2\omega - 1$, as a close approximation. Johnson, Payne, Wang, Asher, and Mandal (2017) used a different approach to estimate *PPV* and obtained a similar estimate to the one obtained using our simple equation (see Supplementary Materials for more details).

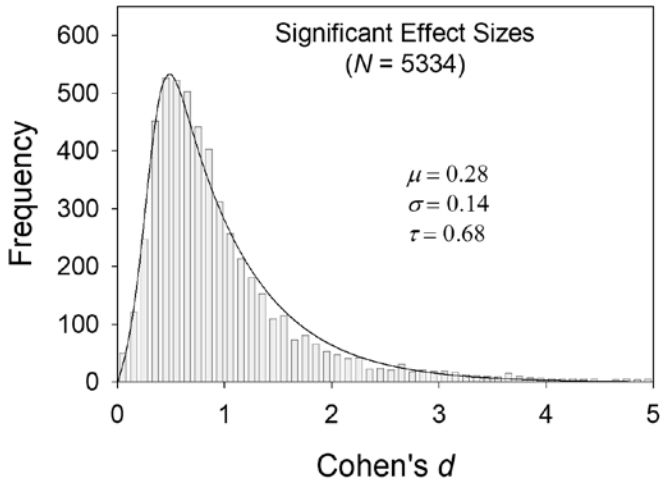
For $H_1$ the area in region I corresponds to the probability of getting a *significant* effect in the wrong direction if the effect is real. Clearly, it is extremely unlikely that a significant result reported in the original publications is in the wrong direction even when power is as low as 50%. Indeed, Gelman and Carlin (2014) showed that this kind of error (which they termed a "Type S error") becomes appreciable only when power drops well below 20%. We assume that even in the original studies, power was not well below 20%. If that assumption is correct, then significant Type S errors can be ignored in our analyses without appreciably affecting our estimates of *PPV*. The area in region II shows the probability of getting a nonsignificant effect in the wrong direction if an effect is real. The area in region III shows the probability of getting a nonsignificant result in the correct direction if an effect is real, and the area in region IV shows the probability of getting a significant result in the correct direction if an effect is real.

**In Detail Box 2 – Effect sizes are continuously distributed**

The analysis presented in the main text is based on binary logic (i.e., it assumes that effects are either real or not), which is a convenient assumption for thinking through these issues but is also unrealistic. Undoubtedly, effect sizes are continuously distributed, beginning with an effect size of zero and increasing in continuous fashion to effect sizes that are much greater than zero. Thus, for example, if one assumes that some hypotheses have effect sizes of 0, it would seem odd to suppose that there are not other hypotheses with an effect size of 0.001, still others with an effect size of 0.002, and so on. In this section, we consider the implications of the fact that effect size is likely a continuous variable. The main point we make in this section is that our conclusion about the prior odds of testing a real effect in cognitive vs. social psychology holds even if we assume that effect sizes are continuously distributed.

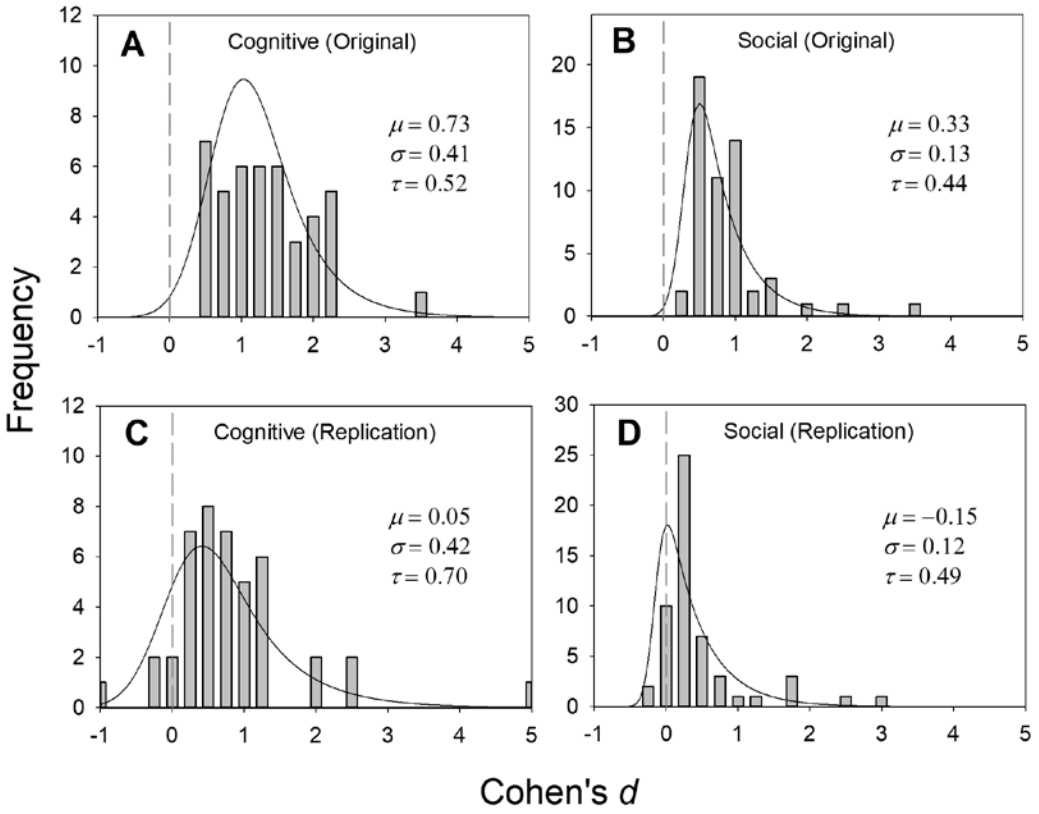
In some fields of research, such as in genome-wide association studies (e.g., Park et al., 2010), there are theoretical reasons to believe that effect sizes are exponentially distributed. If effect sizes happened to be exponentially distributed in psychology, then empirical effect sizes, which are measured with Gaussian error, would be distributed according to an ex-Gaussian (i.e., exponentially distributed effect sizes with Gaussian measurement error). Szucs and Ioannidis (2017) analyzed the distribution of published effect sizes from 3,801 recently published articles in psychology and cognitive neuroscience. From that dataset, we examined the distribution of significant effect sizes in psychology for values of Cohen's *d* less than 5.0 (a very small percentage of effects was larger than that). The adjacent figure

shows the results. The curve shows the maximum-likelihood fit of the 3-parameter ex-Gaussian distribution, which appears to provide a reasonable approximation to the large majority of effect sizes with $d < 5$. The mean of the exponential distribution is represented by $\tau$, and the mean and standard deviation of the Gaussian component are $\mu$ and $\sigma$, respectively. The overall average effect size is equal to $\mu + \tau$, or 0.28 + 0.68 = 0.96 for these data. Because these are maximum-likelihood estimates, the estimated average effect size of 0.96 is the same value one obtains by simply averaging the 5334 effect sizes.



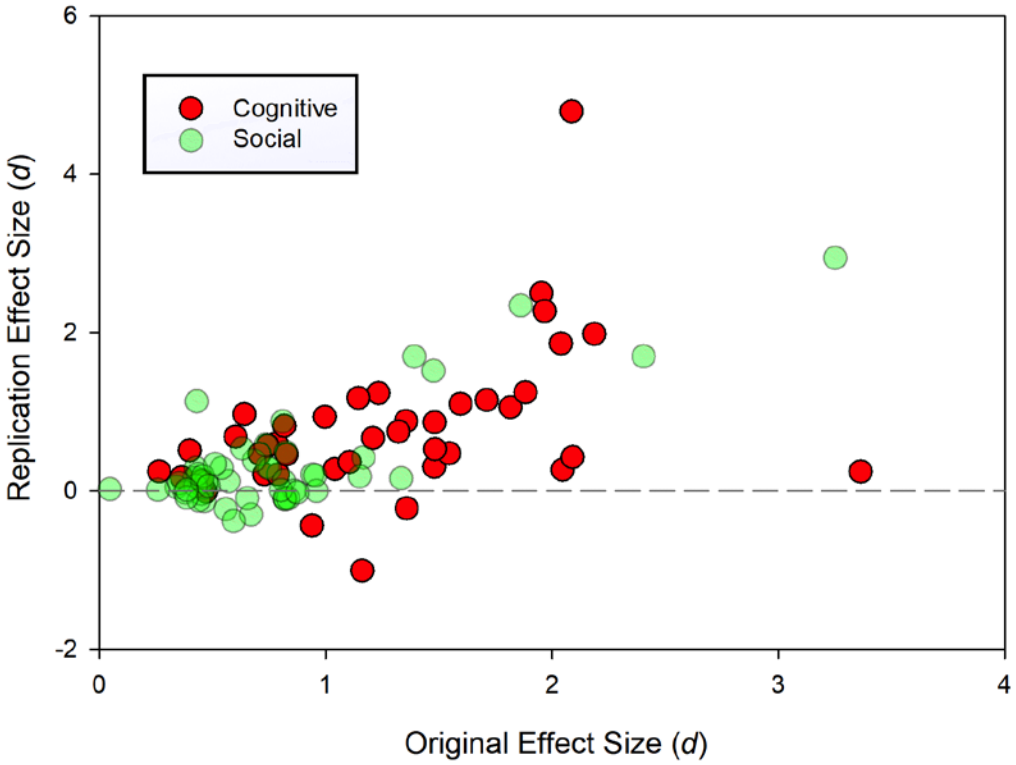

The next figure shows the frequency distributions of reported effect sizes for cognitive and social psychology studies from the Open Science Collaboration (2015) along with the best-fitting ex-Gaussian (fit using maximum likelihood estimation). The effect sizes reported in terms of *r* were converted to Cohen's *d* for this analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009). The data are quite variable, but the results are consistent with the notion that effect sizes are continuously distributed, more or less as an ex-Gaussian distribution. The distribution of published effect sizes for the original studies is almost certainly right shifted (i.e., published effect sizes are likely inflated relative to the true effect sizes). The replication distributions are not inflated, and they are consistent with an

exponential-like distribution of effect sizes (mode = effect size of 0, continuously increasing from there) with Gaussian error.




The next figure shows the distribution of replication effect sizes plotted against the corresponding distribution of original effect sizes, separately for cognitive and social psychology experiments. The average effect size (i.e., $\mu + \tau$) for the cognitive psychology studies is larger than for the social psychology studies, and this is true of both the original (cognitive mean = 1.25, social mean = 0.77) and the replication studies (cognitive mean = 0.75, social mean = 0.34). Are these apparent differences in average effect size reliable? To find out, we first conducted null hypothesis tests on the effect-size data and then performed a Bayesian analysis on

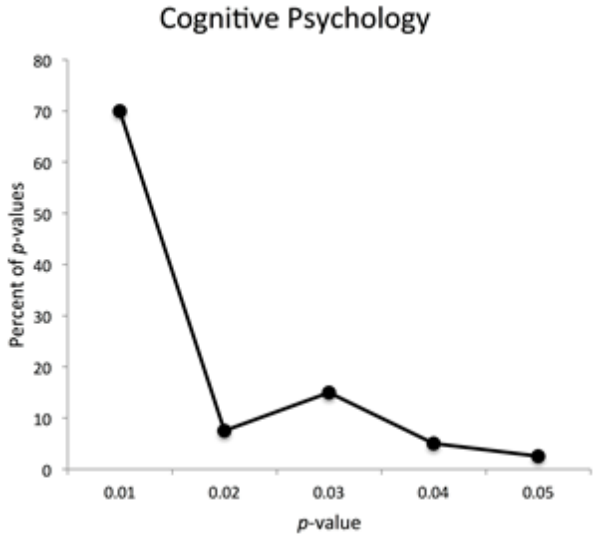
the same data. For the original studies, the effect size was significantly larger for cognitive psychology than social psychology, $t(92) = 3.90$, $p < .001$. For the replication studies, the effect size was again significantly larger for cognitive psychology than social psychology, $t(92) = 2.54$, $p = .013$. The overall decline in effect sizes from the original to the replication studies appears to be comparable for both fields. For the Bayesian analysis applied to the original data, we used a Cauchy uninformative prior (scale = 0.707), and the resulting Bayes factor was 130.0 (see Figure A1). For the replication data, we used a Gaussian informed prior (mean = 1.253 - 0.774 = 0.479, sigma = 0.25), and the resulting Bayes factor was 14.5 (see Figure A2). Thus, we conclude that the average effect size for the cognitive psychology studies was larger than the average effect size for the social psychology studies.




The key point of this section is that if effect sizes are continuously distributed (e.g., according to an exponential with Gaussian measurement error), which we believe they likely are, then we would need to conceptualize *R* in terms of the proportion of tested effects that fall in a region close to zero. In that case, a field with a lower average effect size would have a higher proportion of tested effects that are of negligible effect size no matter how "negligible" is defined (such as Cohen's $d < 0.10$). Thus, whether effect sizes are conceptualized in discrete or continuous terms, *R* is lower for the social psychology studies than the cognitive psychology studies in the Open Science Collaboration (2015).

*Figure 1.* Distributions of *p*-values for studies selected to be replicated in the Open Science Collaboration (2015). Panel A shows the *p*-values from the original cognitive psychology studies, and panel B shows the *p*-values from the original social psychology studies.

A.




B.

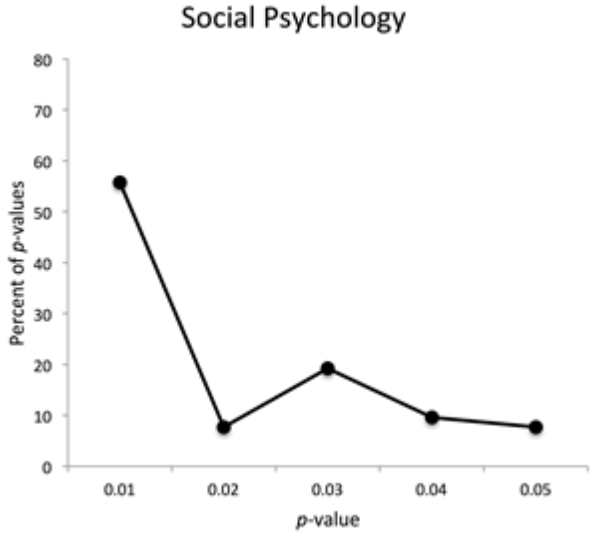
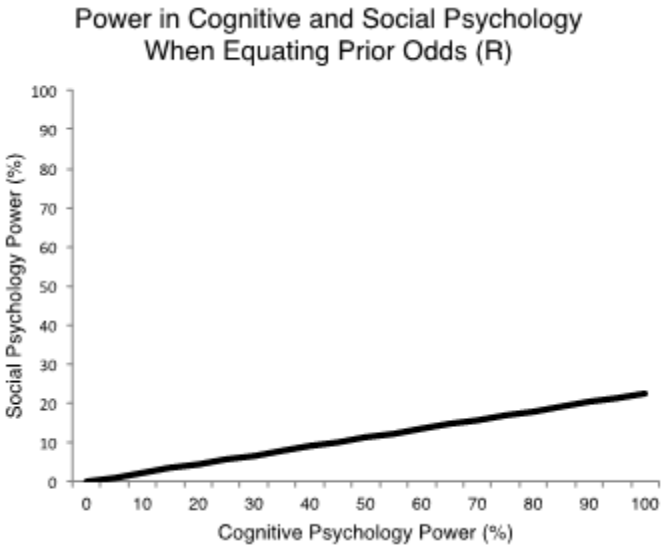
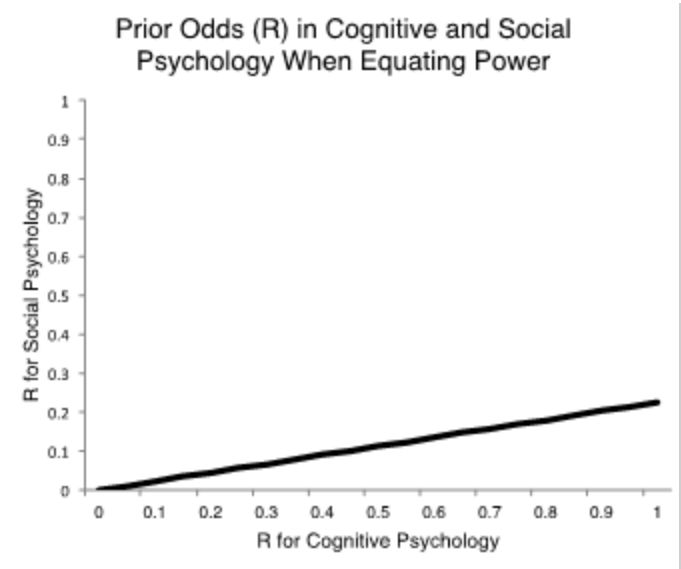


*Figure 2.* The reason social psychology experiments were less likely to replicate in the Open Science Collaboration (2015) could be due to either a lower base rate of true effects (*R*) in social than cognitive psychology or lower power in social than cognitive psychology. Panel A shows how much lower the power in social psychology would have to be than cognitive psychology in order for the prior odds of studying true effects to be equal. Panel B shows how much lower the prior odds of studying true effects would have to be in social relative to cognitive psychology for power to be equal.

A.




B.



# Appendix

Figure A1. Effect-size analysis (cognitive vs. social) for the original studies

**Bayesian Independent Samples T-Test**

Bayesian Independent Samples T-Test

| | $BF_{10}$ | error % |
|---|---|---|
| Cohens_d_O | 130.0 | 6.245e −6 |

**Inferential Plot**

**Cohens_d_O**

**Prior and Posterior**

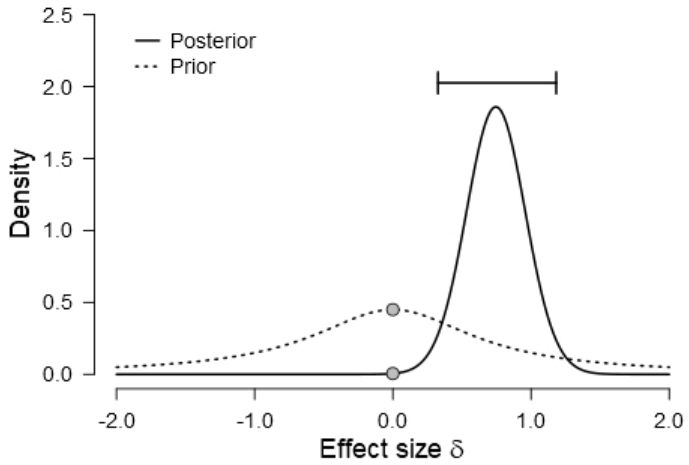


Figure A2. Effect-size analysis (cognitive vs. social) for the replication studies

**Bayesian Independent Samples T-Test**

Bayesian Independent Samples T-Test

| | $BF_{10}$ | error % |
|---|---|---|
| Cohens_d_R | 14.51 | |

**Inferential Plot**

**Cohens_d_R**

**Prior and Posterior**