Is Psychology Suffering From a Replication Crisis?

What Does "Failure to Replicate" Really Mean?

Scott E. Maxwell — *University of Notre Dame*
Michael Y. Lau — *Teachers College, Columbia University*
George S. Howard — *University of Notre Dame*

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Psychology has recently been viewed as facing a replication crisis because efforts to replicate past study findings frequently do not show the same result. Often, the first study showed a statistically significant result but the replication does not. Questions then arise about whether the first study results were false positives, and whether the replication study correctly indicates that there is truly no effect after all. This article suggests these so-called failures to replicate may not be failures at all, but rather are the result of low statistical power in single replication studies, and the result of failure to appreciate the need for multiple replications in order to have enough power to identify true effects. We provide examples of these power problems and suggest some solutions using Bayesian statistics and meta-analysis. Although the need for multiple replication studies may frustrate those who would prefer quick answers to psychology's alleged crisis, the large sample sizes typically needed to provide firm evidence will almost always require concerted efforts from multiple investigators. As a result, it remains to be seen how many of the recently claimed failures to replicate will be supported or instead may turn out to be artifacts of inadequate sample sizes and single study replications.

Keywords: false positive results, statistical power, meta-analysis, equivalence tests, Bayesian methods

Psychologists have recently become increasingly concerned about the likely overabundance of false positive results in the scientific literature. For example, Simmons, Nelson, and Simonsohn (2011) state that "In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not" (p. 1359). In a similar vein, Ioannidis (2005) concluded that for disciplines where statistical significance is a virtual prerequisite for publication, "most current published research findings are false" (p. 696). Such concerns led Pashler and Wagenmakers (2012) to conclude that there appears to be a "crisis of confidence in psychological science reflecting an unprecedented doubt among practitioners about the reliability of research findings in the field" (p. 528). Simmons et al. (2011) state that "a field known for publishing false positives loses its credibility" (p. 1359).

An initial reaction might be that psychology is immune to such concerns because published studies typically appear to control the probability of a Type I error (i.e., mistakenly reporting an effect when in reality no effect exists) at 5%. However, as a number of authors (e.g., Gelman & Loken, 2014; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011) have discussed, data analyses in psychology and other fields are often driven by the observed data. Data-driven analyses include but are not limited to noticing apparent patterns in the data and then testing them for significance, testing effects on multiple measures, testing effects on subgroups of participants, fitting multiple latent variable models, including or excluding various covariates, and stopping data collection once significant results have been obtained. Some of these practices may be entirely appropriate depending on the specific circumstances, but even at best the existence of such practices makes it difficult to evaluate the accuracy of a single published study because these practices typically increase the probability of obtaining a significant result. Gelman and Loken (2014) state that "Fisher offered the idea of p values as a means of protecting researchers from declaring truth based on patterns in noise. In an ironic twist, p values are now often manipulated to lend credence to noisy claims based on small samples" (p. 460).

The question of whether a pattern seemingly identified in an original study is in fact more than just noise can often best be addressed by testing whether the pattern can be replicated in a new study, which has led to increased attention to the role of replication in psychological research. Moonesinghe, Khoury, and Janssens (2007) have shown that successful replications can greatly lower the risk of inflated false positive results. Both Moonesinghe et al. (2007, p. 218) and Simons (2014, p. 76) maintain that replication is "the cornerstone of science" because only replication can adjudicate whether a single study reporting an original result represents a true finding or a false positive result. *Perspectives on Psychological Science* devoted a special section to replicability in 2012 (Pashler & Wagen-

Scott E. Maxwell, Department of Psychology, University of Notre Dame; Michael Y. Lau, Department of Counseling and Clinical Psychology, Teachers College, Columbia University; George S. Howard, Department of Psychology, University of Notre Dame.

We thank Melissa Maxwell Davis, David Funder, John Kruschke, Will Shadish, and Steve West for their valuable comments on an earlier version of this paper.

Correspondence concerning this article should be addressed to Scott E. Maxwell, Department of Psychology, University of Notre Dame, Notre Dame, IN 46556. E-mail: smaxwell@nd.edu

© 2015 American Psychological Association 0003-066X/15/$12.00
Vol. 70, No. 6, 487–498 http://dx.doi.org/10.1037/a0039400

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Scott E. Maxwell

makers, 2012). More recently, this journal has begun a new type of article, a Registered Replication Report (RRR). In an APS *Observer* column, Roediger (2012) stated that "By following the practice of both direct and systematic replication, of our own research and of others' work, we would avoid the greatest problems we are now witnessing" (para. 19). Along these lines, collaborative efforts such as the Reproducibility Project (Open Science Collaboration, 2012) and the psychfiledrawer.org website, which provides an archive of replication studies, reflect systematic efforts to assess the extent to which original findings published in the literature are replicable and can be trusted.

Several recent apparent replication failures have been widely publicized and have begun to cast doubt in some minds on the extent to which the field more broadly is beset with a preponderance of results that cannot be replicated. Most notably, various replication studies (e.g., Galak, LeBoeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012) apparently fail to confirm Bem's (2011) highly publicized findings regarding the existence of psi. Another highly publicized example is the apparent failure of Doyen, Klein, Pichon, and Cleeremans (2012) and Pashler, Coburn, and Harris (2012) to replicate Bargh's work on the influence of subtle priming on behavior. More generally, out of 14 replication attempts organized by Nosek and Lakens (2014), nine were interpreted as failing to replicate the original study and the other five were viewed as only partial replications. Such apparent replication failures are hardly unique to psychology. Scientists at the biotechnology firm Amgen attempted to replicate 53 landmark studies in hematology and oncology, but were able to confirm the original findings in only six cases, implying an apparent failure rate of 89% (Begley & Ellis, 2012). Although concerns about replication failures have arisen in several disciplines, much of the concern has focused on psychology. A 2012 article in *The Chronicle of Higher Education*, for example, raised the question, "Is Psychology About to Come Undone?" (Bartlett, 2012). More recently, a 2014 *Chronicle of Higher Education* article described the apparent crisis as "repligate" (Bartlett, 2014).

A particular replication may fail to confirm the results of an original study for a variety of reasons, some of which may include intentional differences in procedures, measures, or samples as in a conceptual replication (Cesario, 2014; Simons, 2014; Stroebe & Strack, 2014). Although conceptual replication studies can be very informative, they may not be able to identify false positive results in the published literature, because if the replication study fails to find an effect previously reported in a published study, any discrepancy in results may simply be due to procedural differences in the two studies. For this reason, there has been an increased emphasis recently on exact (or direct) replications. If exactly replicating the procedures of the original study fails to replicate the results, then it might seem reasonable to conclude that the results of the original study are in reality nothing more than a Type I error (i.e., mistakenly reporting an effect when, in reality, no effect exists). The primary purpose of our article is to explain why even an exact replication may fail to obtain findings consistent with the original study and yet the effect identified in the original study may very well be true despite these discrepant findings.

It might seem straightforward to decide whether a replication study is a success or a failure, at least from a narrow statistical perspective. Generally speaking, a published original study has in all likelihood demonstrated a statistically significant effect. In the current zeitgeist, a replication study is usually interpreted as successful if it also demonstrates a statistically significant effect. On the other hand, a replication study that fails to show statistical significance would typically be interpreted as a failure.[1]

An immediate limitation of this perspective is that the replication study may have failed to produce a statistically significant result because it may have been underpowered. There is always some probability that a nonsignificant result may be a Type II error (i.e., failing to reject the null hypothesis even though it is false). However, this limitation seems to have an immediate solution, namely to design the replication study so as to have adequate statistical power and thus minimal risk of a Type II error. As Simons (2014) states, "If an effect is real and robust, any competent researcher should be able to obtain it when using the same procedures with adequate statistical power" (p. 76). Unfortunately, in practice, things are rarely so simple. A major

[1] Because the recent debate about replication failures in psychology has been framed in terms of statistical significance, we will use significance testing as the framework for this article. However, it is important to note that replication can also be conceptualized in terms of effect sizes. In particular, a replication study might address the extent to which the sample value of an effect size reported in a published paper can or cannot be duplicated in a replication study.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Michael Y. Lau

goal of our paper is to explain why this approach is, at best, a partial solution toward designing a replication study and interpreting studies that fail to achieve statistical significance. The rest of the paper is separated into five sections. We first discuss complications in determining what is considered adequate power for any replication study. In particular, conventional practice fails to take into account sampling variability of the original effect size estimate. Next, even if adequate power is achieved, a statistically nonsignificant replication result does not unequivocally point to the failure of replication. We show how a statistically nonsignificant replication finding can sometimes be equally suggestive of a nonnull finding. Central to replication attempts is the desire to know if a nonsignificant finding argues for the acceptance of the null. The next section of the paper offers frequentist and Bayesian[2] solutions to assess more accurately the evidence against the findings of an original study. The ensuing section shows that one consequence of these alternative methods is that very large sample sizes will typically be needed if the goal of a study is to show that an effect is essentially zero. This realization leads to the final section, which emphasizes the need for more than a single replication study.

Difficulties in Adequately Powering a Replication Study

Several factors complicate designing an adequately powered replication study. First, a decision must be made as to how much power is in fact adequate. The most popular convention is to design a study so as to have statistical power of .80 (Cohen, 1988), which corresponds to an 80% chance of rejecting the null hypothesis if it is false (i.e., finding a true effect). Although any choice of power value is necessarily somewhat subjective, allowing a 20% chance of failing to reject the null hypothesis may be unwise for replication studies. Suppose 100 published studies have correctly rejected 100 different false null hypotheses. Should we be satisfied if attempts to replicate each of these studies suggest that 20 of these findings cannot be trusted? A strong argument can be made that replication studies should be designed to have power greater than .80. For reasons that will become clear later in the paper, we refrain from recommending a specific level of desired power, but in general we believe that values in the neighborhood of .90 and .95 may be more appropriate than .80.

Second, even after identifying the value for desired power, the appropriate sample size depends greatly on the presumed effect size. Designing a replication study would seem to have a decided advantage over designing an original study, because an effect size value can be obtained from the original study that is being replicated. Unfortunately, things are less straightforward than they first appear. One immediate problem is that effect sizes reported in the literature are generally biased estimates of true population effect sizes. Such factors as publication bias and selective reporting lead to published effect size values that are larger than their actual population counterparts (Greenwald, Gonzalez, Harris, & Guthrie, 1996; Lane & Dunlap, 1978; Maxwell, 2004; Schmidt, 1992). As a result, a replication study designed to have adequate power based on a published effect size will tend to be underpowered and thus will be too likely to fail to replicate the original finding. For those interested in pursuing this further, Gelman and Carlin (2014) demonstrate that misleading conclusions can be quite probable when sample effect size values are taken at face value from published studies, and describe an alternate approach to avoid this problem.

Third, basing sample size of a replication study on the effect size reported in an original study fails to take into account the sampling variability in the original sample effect size. Statisticians have distinguished between “conditional power” and “predictive power.” The former is the probability of rejecting the null hypothesis conditional on an effect size that is presumed to be known with certainty. In contrast, “predictive power” acknowledges that the effect size is typically not known with certainty but instead is at best an estimate. Predictive power averages the power over the plausible values of effect size (based on the estimated standard error of the estimated effect size) so as to obtain a point estimate of the power taking uncertainty into account. Dallow and Fina (2011) explain that “predictive power can lead to much larger sample sizes than either conditional power or standard sample size calculations when used with the same nominal value for

[2] There are two main schools of statistical inference, frequentist and Bayesian. Traditional null hypothesis significance testing is based on a frequentist conceptualization of probability. From this perspective, parameters are fixed, and the goal is to infer what would happen over repeated sampling. In contrast, Bayesians regard parameters as variables and calculate probabilities associated with different parameter values. See Kruschke (2014) for an excellent introduction to the Bayesian approach.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

George S. Howard

power" (p. 311).[3] Psychologists who use power analysis to design replication studies typically rely on conditional power calculations and thus implicitly assume a single value of the unknown population effect size. However, as Dallow and Fina state, this practice "almost certainly gives rise to undersized, underpowered studies" (p. 317). Thus, even if published effect sizes were unbiased, typical replication studies are unlikely to control Type II error rates at the desired level.

A hypothetical original study that compared the means of two independent groups using 40 participants per group (and thus a total sample size of 80) illustrates the impact of sampling variability. Suppose this study produced a statistically significant t value of 2.24. A researcher decides to replicate this study. How large does the sample need to be for conditional power to equal the current conventional standard of .80? The t value of 2.24 corresponds to a Cohen's d value of 0.50 (a "medium" effect size [Cohen, 1988]), which, in turn, implies that 64 participants per group (and thus a total sample size of 128) are needed to achieve a power of 0.80. However, the effect size value of 0.50 is only an estimate. The true population value could be either smaller or larger than 0.50. For example, a 50% confidence interval for the population value when $n = 40$ per group ranges from approximately 0.35 to 0.65. Thus, there is a 25% chance that the population value is less than 0.35, just as there is a 25% chance that the population value is larger than 0.65.[4] How powerful will the replication study with $n = 64$ per group be if the true effect size is only 0.35 or is actually as large as 0.65? If the true effect size is 0.65, the corresponding power will be .95, which naturally is larger than the anticipated value of .80. One might expect a similar drop in power if the effect size is smaller than 0.50. In reality, the power drops all the way to .50 if the true effect size is 0.35. From this perspective, there is a 25% chance that the true power of the replication study is at most .50 even if the replication exactly duplicates every detail of the original study except that the sample size is increased by over 50% in the hope of obtaining adequate power. Of course, the power is even less if the replication study is designed with the same sample size as the original study.

The practical implication is that basing sample size calculations for a replication study on the effect size obtained in an original study is less straightforward than it might first appear. A researcher who feels confident that his or her replication study has been designed with statistical power of .80 may in fact unknowingly be confronted with a much less powerful replication study. A primary contributing factor here is the nonlinear relationship between effect size and power. In the previous demonstration, the power for an effect size of 0.50 was .80. Even though effect sizes of 0.35 and 0.65 are equidistant from 0.50, their respective power values (i.e., .50 and .95) are *not* equidistant from the power of .80 for an effect size of 0.50. Instead, the power loss associated with an effect size of 0.35 is much greater than the power gain associated with an effect size of 0.65. Thus, failing to take sampling variability into account in planning a replication study can lead to greater risk of having an underpowered study than suggested by conventional practice. For those interested in a method to address this problem, Kruschke (2013) describes how Bayesian approaches to sample size planning can take sampling variability into account because "one uses an entire distribution of parameters instead of a single point value for the effect size" (p. 581).

Taylor and Muller (1996) developed a procedure for sample size determination that takes into account both effect size uncertainty and censoring that occurs as a result of publication bias. Unfortunately, practical implementation of the method will often be problematic in psychology because original studies are frequently underpowered (Button et al., 2013). In this case, Yuan and Maxwell (2005) show that any attempt to use the effect size from the original study in order to plan the sample size of a replication study is likely to be a wild guess. Both Taylor and Muller (1996) and Yuan and Maxwell (2005) suggest that researchers should use a confidence interval for the population effect size in order to plan future sample size instead of relying on a point estimate of the effect size. However, Yuan and Maxwell (2005) conclude that with a relatively small effect size, the sample size of the original study needs to be surprisingly large or else the lower limit of the

[3] Dallow and Fina's (2011) conceptualization of predictive power assumes that original data will be combined with the data to be collected, whereas we use the term predictive power to refer to the power that will occur based on only the new data that will be collected.

[4] Strictly speaking, this interpretation requires a Bayesian perspective. More will be said about Bayesian statistics later in the paper. Also, 50% confidence instead of the more typical 95% confidence level is used here to illustrate the effect sizes corresponding to the 25th and 75th percentiles instead of the 5th and 95th percentiles.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

confidence interval for power may not even exceed .05. Taylor and Muller's (1996) approach to sample size planning is an improvement over standard practice when the goal is to show that an effect exists, but as we will discuss later, other methods of sample size planning may be more appropriate when the goal is to show that an effect does not exist or is so small it is essentially zero.

McShane and Böckenholt (2014) emphasize that beyond the uncertainty associated with any sample effect size there is almost always variability in effect sizes even across seemingly equivalent studies. Meta-analyses of various psychological phenomena typically reveal substantial variability across studies even when those studies are designed to be exact replicates of one another. McShane and Böckenholt show that such variability leads to overly optimistic power calculations, thus making the actual probability of a successful replication less likely than it appears. For those interested in pursuing this further, McShane and Böckenholt have developed an approach to adjust power calculations for the degree of heterogeneity thought to be present for the effect being studied.

Difficulties Interpreting a Nonsignificant Result in a Replication Study

The previous section explained why designing a replication study to have adequate power is generally far from straightforward. Suppose, however, that it were possible to design a replication study with adequate power. Furthermore, suppose that the replication study, unlike the original study, fails to produce a statistically significant result. Common practice would imply that the replication failed. In particular, it would seem to follow that the results of the original study have been overturned. We now proceed to show that this interpretation is often premature.

Consider once again an original study that compared the means of two independent groups of 40 individuals per group (i.e., total $N = 80$) and reported a statistically significant t value of 2.24. A psychologist decides to replicate this study. The psychologist begins by calculating the effect size obtained in the original study, and finds that Cohen's d is 0.50. Following conventional practice, the psychologist chooses a sample size of 86 per group (total $N = 172$) for the replication study, because this will provide a power of .90 to detect a medium effect at an alpha level of .05 (two-tailed).[5] Notice that achieving a power of .90 requires more than twice as many participants per group as the original study, and even a power of .80 requires more than a 50% increase in sample size.

Suppose the replication study yields a t value of 1.50. The corresponding two-tailed p value is .14, so the replication study fails to obtain a statistically significant result. Conventional wisdom would state that the replication study failed, casting doubt on the validity of the original study. In particular, if the sample effect size in the original study is the true value of the population effect size, the fact that a sample size of 86 per group in the replication study yields power of .90 guarantees that there is only a 10% chance of failing to obtain a statistically significant result. Some participants in the replication debate would conclude that the replication study fails to support the original study, and thus the results of the original study should be regarded with skepticism. Beyond that, some individuals mistakenly infer that the null hypothesis here is likely to be true (or at least essentially true). In other words, the effect size in question is mistakenly interpreted to be zero (or so close to zero that it is reasonable to regard it as being essentially equal to zero). For example, psi does not exist, or subtle social cues do not actually have an effect on behavior. These are exactly the sorts of conclusions often reached when replication studies produce nonsignificant results. For example, when Pashler, Coburn, and Harris' (2012) attempts to replicate two of Williams and Bargh's (2008) priming studies yielded nonsignificant findings, they concluded that it is possible "that the Williams and Bargh results are simply not valid, representing, for example, Type I errors" (p. 6). Similarly, Ritchie, Wiseman, and French (2012) state that their failure to obtain significant results in attempting to replicate Bem (2011) "leads us to favor the 'experimental artifacts' explanation for Bem's original result" (p. 4).[6]

Nonsignificant replication results raise a fundamental question, namely, to what extent do nonsignificant results support the truth of the null hypothesis? To answer this question, we need to consider our hypothetical replication study in more detail. The observed t value of 1.50 implies that the sample value of Cohen's d for the replication study was 0.23. A 95% confidence interval for the population value of d based on this study yields an interval of -0.07 to 0.53. Because zero is contained in this interval, zero is a plausible value for the true population effect size. However, it is very different to conclude that zero is a plausible value than to conclude that the effect size is exactly zero. In fact, the results of this study in no way support a strong conclusion that the effect size is even close to zero. Based on the confidence interval, the results of the replication study leave open a conclusion that the population effect size could plausibly be "medium" (e.g., a value of 0.50). It is clearly inappropriate to conclude that no effect exists when it is plausible that the effect size could actually be medium.

Rosenthal and Rubin's (1994) "counternull" effect size offers a complementary perspective, by providing the effect size estimate that is as equally supported by the data as is a null finding. In other words, for any given sample estimate, Rosenthal and Rubin's counternull value is defined to be the effect size that is supported by the data exactly as much as the null value of zero is supported by

[5] The sample size of 86 per group provides conditional power of .90 if the population effect size is truly medium. As we have shown, the predictive power may be considerably less than .90, but we base our discussion on conditional power because this is the typical approach by which power analysis has been used to determine sample size.

[6] The authors of these studies emphasized that other interpretations are also possible, but many readers may have focused on the eventual conclusion that the phenomena in question may not exist.

the data. Consider our example of the replication study with a Cohen's *d* of 0.23. The counternull effect size here is approximately 0.46. This implies that a population effect size of 0.46 is as plausible as an effect size of zero, based on the replication study. Although it is tempting to conclude that the nonsignificant statistical test supports a conclusion that the true effect size is zero, the data equally support a conclusion that the true effect size is 0.46. From this perspective, it is equally likely that the unknown population effect size is essentially medium (i.e., a Cohen's *d* of 0.50) as it is that the population effect size is zero. Thus, concluding that the true effect size is zero or even very close to zero based on the nonsignificant test is misguided because it is just as likely that the effect size is medium.

Although the researcher who conducted the replication study followed all of the rules and designed the study very carefully, the end result is a study that neither confirms nor contradicts the original study. Although it is plausible that the population effect size is zero, it is essentially equally likely that the population effect size is in fact as large as the 0.50 value that was found in the original study. Thus, the nonsignificant result obtained in the replication study offers only weak evidence in support of a conclusion that the null hypothesis is true. Although the replication study failed to confirm the original study, it does not follow that the replication study has overturned the original study. Unfortunately, the replication study has produced equivocal results that neither confirm nor contradict the original study.

The Problem and Possible Solutions

The essential problem is how to justify a conclusion that the results of an original study are not trustworthy. The conventional answer is simple, namely that nonsignificant results in a replication study justify overturning the original study if the replication study was adequately powered. However, our major point so far is that this simple answer is often wrong. The fact that a replication study has resulted in a nonsignificant statistical test does not necessarily mean that the results of the original study should be discounted.

The statistical problem is that rejecting the results of the original study involves accepting the null hypothesis of the replication study. For example, a replication study might be interpreted as showing that individuals exposed to subtle social cues behave the same as individuals not exposed to such cues. However, failing to reject the null hypothesis is not the same as accepting it. Although we teach this message to our students, it seems to have been forgotten when we interpret the nonsignificant results obtained in a replication study. On a related note, Finch, Cumming, and Thomason (2001) found that 37% of studies with statistically nonsignificant results interpreted their results as providing evidence that the null hypothesis was true.

The dilemma is how to use the data from a replication study to decide whether to discard the results of an original study in favor of a conclusion that the effect in question is essentially null. Fortunately, statisticians have developed methods that can answer this question. One method derives from a frequentist perspective, while two methods are available from a Bayesian perspective. Excellent general introductions to these methods already exist (and we cite them throughout the following pages), so we focus on explaining the relevance of these methods for interpreting replication studies with nonsignificant statistical results.

For a frequentist, the fundamental question is whether the results of a replication study imply that the null hypothesis is essentially true by a nominal level of uncertainty. Answering this question requires establishing a region of equivalence, which represents a range of parameter values for which the null hypothesis is for all intents and purposes essentially true. This idea is similar in spirit to Serlin and Lapsley's (1985) concept of the "good-enough principle" that science sets standards to evaluate the extent to which experimental results are "good enough" to support an underlying scientific theory. This approach requires researchers to establish how close to a specific theoretical value (e.g., zero) an effect needs to be to conclude that the correspondence between the data and the theory is good enough that the data supports the theory. For example, researchers might decide that a Cohen's *d* value anywhere between -0.10 and 0.10 is good enough when a theory predicts a null effect. It is important to realize that this region is expressed in terms of population parameters, not sample observations. As a result, it does not suffice simply to see whether the sample value of Cohen's *d* falls within the specified interval. Instead, it is necessary to take sampling variability into account. This can be done either by performing a statistical test (Rogers, Howard, & Vessey, 1993) or by forming a confidence interval (Seaman & Serlin, 1998). Understanding the rationale for the method is easiest from the perspective of confidence intervals. From this perspective, the results of a replication study fall into one of three categories.

First, a confidence interval for the effect may fall entirely within the equivalence region. For example, the confidence interval for Cohen's *d* might be entirely contained within the bounds of ± 0.10. In this case, the effect is essentially zero, even at its highest and lowest values taking sampling variability into account. As a result, it is certain to a high degree that the true population effect size is no larger than 0.10 in absolute value, in which case a proper conclusion is that the effect size is essentially zero. Such a result provides strong statistical evidence contrary to an original study that had found a statistically significant effect.[7]

Second, the confidence interval for the effect size can fall entirely outside the equivalence region. This implies that the effect is almost certainly not close to zero. Not only is the effect statistically significant, but it also has some real theoretical or practical importance. Such a result un-

[7] Notice that even if the confidence interval for the effect size were to exclude zero, it is still appropriate to conclude that the effect is essentially zero when the entire confidence interval is contained within the equivalence region. In this situation the effect is statistically significant but is judged to be of no real theoretical or practical importance.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

equivocally supports an original study that had found a statistically significant effect.

Third, the confidence interval can overlap the equivalence region. Unfortunately, this implies that the effect could plausibly be trivial but could also plausibly be nontrivial. Such a result is equivocal, because it neither clearly contradicts nor supports an original study that had found a statistically significant effect.

To examine these three possible outcomes further, reconsider our hypothetical replication study that obtained a *t* value of 1.50 with 86 participants per condition. The corresponding 90% confidence interval[8] for the population value of *d* based on this study yields an interval of approximately −0.02 to 0.48. This interval overlaps the equivalence region, which means that the results of the replication study are equivocal. It is impossible to say whether the replication study supports or refutes the original study.

To understand how different this approach is from traditional hypothesis testing, consider a recent replication study conducted by Wortman, Donnellan, and Lucas (2014). These authors attempted to replicate a previous study by Bargh and Shalev (2012), which had shown that the experience of physical warmth leads to reduced reports of feeling lonely. The original study had reported a Cohen's *d* value of 0.61. The replication study obtained a Cohen's *d* value of 0.02, thus almost exactly zero. At first glance, this seems to represent overwhelming evidence against the original study if the replication study was adequately powered. The authors of the replication study went to considerable lengths to ensure that their study was in fact adequately powered, and ended up with 260 participants, as compared to the original study, which had 75 participants. The corresponding 90% confidence interval for Cohen's *d* for the replication study ranges from −0.18 to 0.22. Thus, it is plausible that the population effect is larger than Cohen's "small" effect size value of 0.20. This result would provide clear evidence against the existence of the effect found in the original study only if *d* values as large as 0.22 were regarded as essentially zero. Even if an apparently adequately powered replication study shows no effect whatsoever from a traditional hypothesis testing approach, it does not necessarily follow that the replication study strongly supports the lack of an effect, in which case the finding of a nonzero effect in the original study may still be plausible.

The Wortman et al. (2014) replication study produces equivocal results although from a conventional perspective it casts doubt on the original study. Unfortunately, the replication study has failed to provide a definitive answer to the question of whether the effect found in the original study is implausible. The problem is that the confidence interval is too wide to justify a clear conclusion. Although the replication study was designed to have adequate statistical power to reject the null hypothesis, it was not designed to have adequate statistical power for the test of equivalence. A later section of the paper addresses the question of sample size planning when the goal is to show that an effect is essentially zero.

Two Bayesian methods exist for assessing whether a replication study clearly fails to support the results of an original study. The first method involves a region of practical equivalence (ROPE, Kruschke, 2014). The general logic is similar to that of the frequentist equivalence approach, except for two potentially important differences. First, a Bayesian highest density interval, unlike a frequentist confidence interval, "actually includes the 95% of parameter values that are most credible" (Kruschke, 2013, p. 592), so "when the 95% HDI [highest density interval] falls within the ROPE, we can conclude that 95% of the credible parameter values are practically equivalent to the null value" (Kruschke, 2013, p. 592). Second, the Bayesian method incorporates a prior distribution for the effect size. Then a highest posterior density interval is formed for the effect size and this interval is compared to the ROPE as in the frequentist approach. An attractive feature of Bayesian methods is that they often facilitate robust estimation (Kruschke, 2013). However, this generally requires the availability of raw data, which is why we do not present a numerical example of the ROPE. Interested readers are referred to Kruschke (2013, 2014) for such examples.

The Bayesian perspective offers a second method that does not directly rely on a confidence interval for the effect size. The question of interest here could be stated as how probable is it that the null hypothesis is true, given the results of the replication study? This question is meaningless from a frequentist perspective, but can be answered from a Bayesian perspective (Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). In particular, the Bayesian approach can quantify the degree to which the data support either the null hypothesis or the alternative hypothesis. The important distinction here is that "from a Bayesian perspective, the fact that the null hypothesis is unlikely is not sufficient reason to reject it—the data may be even more unlikely under the alternative hypothesis" (Wagenmakers, 2007, p. 790).

Kass and Raftery (1995) pointed out that whereas estimation and testing are basically complementary in frequentist statistics, this is not necessarily true in Bayesian statistics. While the ROPE has its roots in estimation, Bayesian testing provides a different perspective on the question of evaluating the plausibility of the null hypothesis. As the previous quote from Wagenmakers indicates, in Bayesian testing it is necessary to consider not only how likely the null hypothesis is given the data, but also how likely the alternative hypothesis is given the data. This is commonly accomplished through the Bayes factor (BF).

The BF specifies the ratio of the probability of one model to an alternative model. For example, the BF comparing the null model and the alternative model determines the relative probabilities of these two models. For the

[8] Confidence intervals for establishing equivalence typically are based on 90% confidence instead of 95% because the underlying logic hinges on two one-tailed tests of significance. Each one-tailed test is performed with an alpha level of .05, but it is impossible for both null hypotheses to be false, so the overall alpha level is in fact .05 even though two tests have been performed, each with an alpha level of .05.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

hypothetical replication study with $t = 1.50$, the BF[9] for the null hypothesis relative to the alternative hypothesis is 2.14. This implies that the null hypothesis is 2.14 times more likely to be true than the alternative hypothesis. Thus, this result supports the plausibility of the null hypothesis, but does not provide overwhelming evidence in its favor. According to Jeffreys' (1961) classification scheme, the BF of 2.14 implies "weak" evidence in favor of the null hypothesis. Alternatively, the probability that the null hypothesis is true based on the data in the replication study is given by

$$\Pr(H_0 true) = \frac{B_{01}}{B_{01} + 1}, \tag{1}$$

where B_{01} is the BF for the null hypothesis. Substituting the value of 2.14 into Equation 1 shows that the probability that the null hypothesis is true based on the replication study data equals .68. Thus, although this analysis favors the null hypothesis, it is still quite plausible that the alternative hypothesis is true, despite a nonsignificant statistical test from the frequentist perspective. In contrast to the frequentist approach, the Bayesian approach directly addresses the probability of a null or alternative hypothesis. Additionally, in contrast to conventional approaches to understanding replication results, statistically nonsignificant findings may or may not be found to support the conclusion of a successful replication.

The BF is controversial even among Bayesians for two reasons. First, proponents as well as critics agree that results obtained with the BF approach can depend greatly on the choice of an alternative hypothesis. However, proponents such as Rouder, Speckman, Sun, Morey, and Iverson (2009) see this as a plus, whereas others such as Kruschke (2011) suggest it can sometimes be a disadvantage. Second, researchers who generally favor parameter estimation over model comparison express concerns such as, "The BF by itself can be misleading, for example in cases where the null hypothesis is favored despite huge uncertainty in the magnitude of the effect size" (Kruschke, 2013, p. 602), whereas supporters maintain that "As a rule of thumb, inference based on evaluating a null without comparison to alternatives tends to overstate the evidence against the null" (Rouder et al., 2009, p. 227). Additional information on the BF is available in Masson (2011); Rouder et al. (2009) and Wagenmakers (2007).

Sample Size Planning for Equivalence Studies

Statisticians have developed procedures for sample size planning to provide adequate statistical power for equivalence tests. In fact, such procedures are readily available in some statistical packages (e.g., SAS, R) for basic statistical tests such as comparing the means of two independent groups. Unfortunately, calculations will often reveal that very large samples are required to have adequate statistical power for equivalence tests. Of course, sample size depends on how narrow or wide the region of equivalence is defined to be, but for typical definitions of equivalence in psychology, the necessary sample size is likely to come as a shock to most researchers. An approximate formula for appropriate sample size is provided by Chow, Shao, and Wang (2003):

$$n = \frac{2\left(1.645 + z_{\beta/2}\right)^2}{\delta^2}, \tag{2}$$

where n is the sample size per group, β is the desired probability of a Type II error, and δ defines the value of Cohen's d that is deemed to be essentially null. It is important to note that this formula is similar to the standard formula for calculating sample size for a typical t test comparing the means of two independent groups (e.g., Gonzalez, 2009, pp. 122–124), but is not exactly the same because the test of equivalence is functionally two one-sided tests (hence the acronym TOST). It should also be noted that Equation 2 is developed from a frequentist perspective, but Kruschke (2013, 2014) describes a Bayesian sample size approach based on the ROPE.

Table 1 shows the necessary sample size (per group) to obtain specified levels of power for two definitions of equivalence in terms of Cohen's d. It is immediately obvious that for the definitions of "equivalence" considered in the table, the necessary sample size per group is much larger than the typical replication study. The sample sizes shown in the table are larger than might be expected because (a) it is important to distinguish very small effects in an equivalence study, and (b) the equivalence test is statistically significant (meaning that the results are consistent with an essentially null finding) only when the sample mean difference is small and also the corresponding confidence interval is narrow. As a result, multiple replication studies will often be necessary instead of only a single study if the goal is to show convincingly that an effect is for all practical purposes essentially zero, as anticipated by such authors as Bonett (2012); Kahneman (2012), and Nosek, Spies, and Motyl (2012), or even as far back as Hunter (2001), where such methods as cumulative meta-analysis and crowd sourcing of replication efforts (where multiple investigators join together to conduct replication studies, such as in the RRR) can be used to track the proper interpretation of multiple studies over time.

Using Multiple Studies to Address the Question of Equivalence

The previous sections of the paper implicitly assumed that only one replication study had been conducted, but after seeing Table 1 it is clear that multiple replication studies will almost always be necessary to establish that an effect

[9] The BF was calculated using the online web calculator at pcl.missouri.edu, using a scale r of $\sqrt{2}/2$ for the scaled JZS Bayes factor, which, here, is based on a Cauchy distribution centered at zero for the standardized effect size (Rouder et al., 2009). This approach follows a precedent established by Jeffreys (1961) and continued by Zellner and Siow (1980), hence the JZS acronym mentioned in the previous sentence. However, it is important to stress that the specific value of the BF can depend greatly on the presumed alternative.

Table 1
Sample Size (per Group) to Obtain Specified Power for Test of Equivalence

Equivalence region	Power	*n*
$-0.10 < d < .10$	.50	1,077
	.80	1,714
	.90	2,166
	.95	2,600
$-0.05 < d < .05$	.50	4,305
	.80	6,852
	.90	8,659
	.95	10,397

Note. This table is based on an independent groups comparison where the underlying population means are exactly the same, assuming an alpha level of .05 (two-tailed), normality, equal variances, and equal sample sizes.

is so small as to be considered nonexistent. This raises a question about how to analyze the data obtained from multiple studies. The natural answer is to use meta-analysis.

An argument could be made that fixed effects meta-analysis might be appropriate for a collection of direct replication studies. In reality, even studies designed as direct replications may differ from each other in unanticipated ways, suggesting that random effects meta-analysis may often provide a better model. Simons, Holcombe, and Spellman (2014) point out that "Even though all of the studies in an RRR adopt the same procedures, they might not be measuring exactly the same effect" (p. 554). More generally, a random effects model may be preferred because the ultimate goal is usually to draw inferences about effect parameters in a population of studies based on a random sample of studies (Hedges & Vevea, 1998). In addition, cumulative meta-analysis (Lau et al., 1992) can provide a valuable model by incorporating the information provided in the latest replication study into a previous meta-analysis conducted on prior replication studies.

Furthermore, meta-analysis can provide a valuable method for studying potential moderators explaining why different studies obtain different effects. In particular, meta-analysis provides an alternative to the flawed approach of inferring that studies differ if an original study reports a statistically significant effect but a replication study fails to find a significant effect. It is tempting to infer that the studies have produced truly different results and to begin the search for possible moderators, but as Gelman and Stern (2006) have pointed out, such an inference is often misguided. The fact that one study produces a significant effect while the other fails to produce a significant effect does not necessarily imply that the results of the two studies are significantly different from each other. For example, consider the Wortman et al. (2014) replication of Bargh and Shalev (2012). As mentioned earlier, Wortman et al. (2014) obtained a nonsignificant effect size estimate of 0.02, whereas Bargh and Shalev (2012) obtained a significant effect size estimate of 0.61. The sizable discrepancy in effect size estimates would seem to necessitate pursuing possible moderators in an effort to understand why the two studies produced such different results. However, Wortman et al. (2014) report that the difference between the estimated effect sizes of the two studies is itself not statistically significant at the .05 level. Thus, it is plausible that sampling error completely accounts for the difference between the effect sizes of the two studies. From a methodological perspective, it is important to realize that testing heterogeneity of effects in a meta-analysis should replace categorizing studies as either significant or nonsignificant, and then searching for reasons to explain why different studies are in different categories.

A previous section of the paper explained how a BF can be used to assess the extent to which a single replication study supports the null hypothesis of no effect. Kuiper, Buskens, Raub, and Hoijtink (2013) describe a Bayesian method that can be used to evaluate the plausibility of the null hypothesis by combining evidence from multiple studies. In particular, a BF as well as an accompanying posterior model probability can once again be calculated to evaluate the extent to which the available data support the existence or nonexistence of an effect. Alternatively, a Bayesian meta-analysis can be used to synthesize effect size estimates (e.g., Hedges, 1998; Higgins, Thompson, & Spiegelhalter, 2009).

Summary and Conclusions

Our main point is that the proper design and interpretation of replication studies is less straightforward than conventional practice would suggest. In particular, designing studies with adequate power encounters several important complications. Furthermore, interpreting a nonsignificant replication study becomes complicated even when the study appears to have been adequately powered according to currently accepted practices. Most importantly, the mere fact that a replication study yields a nonsignificant statistical result should not by itself lead to a conclusion that the corresponding original study was somehow deficient and should no longer be trusted, even if the replication study appears to have been adequately powered.

We suspect that researchers may well discover that designing appropriate replication studies frequently requires larger sample sizes than most researchers are accustomed to, or else the results of any single replication study are likely to be equivocal. A main reason for this is due to not attending to the sampling variability of effect size estimates in original studies. As a result, just as it may be unwise to consider a single original study as definitive, it may also be unwise to regard a single replication study as providing the final word. Instead, researchers should expect that multiple replication studies will often be needed to resolve apparent inconsistencies in the literature. Of course the value of multiple replication studies extends well beyond purely statistical considerations. Scientific psychology invariably involves establishing boundary conditions and moderators to ascertain the extent to which effects

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

generalize to other interventions, outcomes, persons, and settings. As Shadish, Cook, and Campbell (2002) have stated, "most of this knowledge about generalization is the product of multiple attempts at replication, or reflection on why successful or failed generalization might have occurred, and of empirical tests of which reasons are true" (p. 342).

The RRR recently initiated by *Perspectives on Psychological Science* exemplifies one type of approach we recommend. In particular, the first article of this type perfectly illustrates the perils of relying on a single replication study. Alogna et al. (2014) report the results of 31 labs that replicated the procedures originally described in Schooler and Engstler-Schooler's (1990) article on verbal overshadowing. In particular, each lab calculated a confidence interval for the effect of verbal overshadowing in order to see whether each replication study supported the significance of the effect. Interestingly, "all of the confidence intervals for the individual replications in RRR1 included zero" (Alogna et al., 2014, p, 570). Because the interval for each individual study contained zero, no study found a significant effect. Each study taken alone might then seem to refute the original finding that verbal overshadowing exists. However, instead of interpreting each replication separately, the authors used meta-analysis to obtain a cumulative confidence interval based on all of the replication studies combined. The meta-analytic confidence interval did not contain zero, and clearly supported the existence of a nonzero effect of verbal overshadowing. Alogna et al. (2014) conclude, "Had we simply tallied the number of studies providing clear evidence for an effect in RRR1, we would have concluded in favor of a robust failure to replicate—a misleading conclusion" (pp. 570–571). In other words, every single one of the individual replication studies failed to replicate the original finding in the sense that none of them obtained a significant result and yet the meta-analysis revealed a significant effect.

Hedges (1987) also illustrates the value of conducting multiple studies of the same phenomenon and using meta-analysis to interpret the combined results. Hedges states that "the notion that experiments in the social sciences produce relatively inconsistent (empirically noncumulative) results is not supported by these data" (p. 450). Surprisingly, Hedges also shows that single studies in the physical sciences cannot necessarily be relied upon because "The data from the physical sciences show that even research based on sound theories and strong methodology may not always yield results that are consistent in an absolute sense by a statistical criterion" (p. 450). In a similar vein, what could be more constant than a physical constant, such as the speed of light? Amazingly, Henrion and Fischoff (1986) show that estimates of physical constants such as the speed of light changed throughout the 20th century. In fact, examining results over time caused "deBray to suggest that the speed of light was not constant but decreasing by about 4 km/s/yr" (Henrion & Fischoff, 1986, p. 793). Eventually, it became clear that the speed of light is in fact a constant and has not been changing over time, but the intervals formed in earlier decades often did not overlap with intervals obtained in subsequent decades.

Tversky and Kahneman's (1971) representation hypothesis explains why "most psychologists have an exaggerated belief in the likelihood of successfully replicating an obtained finding" (p. 105). According to this hypothesis, "if we expect all samples to be very similar to one another, then almost all replications of a valid hypothesis should be statistically significant" (Tversky & Kahneman, 1971, p. 108). Alogna et al.'s (2014) RRR reinforces Tversky and Kahneman's point that the uncertainty inherent in individual studies tends to far exceed intuition, emphasizing the value of multiple studies.

Despite raising doubts about the extent to which apparent failures to replicate necessarily reveal that psychology is in crisis, we do not intend to dismiss concerns about documented methodological flaws in the field. Questionable research practices such as those identified by John et al. (2012) and Simmons et al. (2011) clearly need to be addressed because they produce inflated estimates of effect sizes and render *p* values largely uninterpretable. Similarly, we agree that the continuation of underpowered studies in many areas of psychology (e.g., Button et al., 2013) undermines scientific psychology. We support efforts such as those of Funder et al. (2014) to improve the quality of research in psychology. Furthermore, we support the increased emphasis on replication, and believe that replication plays an essential role in developing a cumulative science of psychology. Nevertheless, we also believe that psychologists need to be aware of the limitations of single replication attempts, especially when those attempts may seem to contradict original studies. Enormous sample sizes, much larger than those typical in psychology, are generally required for demonstrating that an effect is so small that it can essentially be regarded as null. Rarely will a single replication study by itself be able to show that an effect reported in an original study is no longer trustworthy.

Although we have focused on statistical issues in evaluating the extent to which replication studies indicate that an effect found in an original study may not truly exist, we want to emphasize that statistical considerations are only one aspect of what should always be a broader consideration. Cook, Gruder, Henningan, and Faly (1979), Greenwald (1975), and Wilson and Shadish (2006) all offer valuable perspectives on broader issues involved in evaluating the veracity of null effects. From the perspective of replication studies, Brandt et al. (2014) have developed helpful guidelines for conducting replication studies.

Finally, it may seem discouraging that the design and interpretation of replication studies is more complicated than current practice in the discipline implies. However, the potential silver lining is that some of the apparent replication failures currently plaguing the field may turn out not to be failures after all.

REFERENCES

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, Š., Birch, S., Birt, A. R., . . . Zwaan, R. A. (2014). Registered replication report: Schooler

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

and Engstler-Schooler (1990). *Perspectives on Psychological Science, 9,* 556–578. http://dx.doi.org/10.1177/1745691614545653

Bargh, J. A., & Shalev, I. (2012). The substitutability of physical and social warmth in daily life. *Emotion, 12,* 154–162. http://dx.doi.org/10.1037/a0023527

Bartlett, T. (2012). Is psychology about to come undone? *The Chronicle of Higher Education*. Retrieved November 23 2014, from http://chronicle.com/blogs/percolator/is-psychology-about-to-come-undone/29045

Bartlett, T. (2014). Replication crisis in psychology research turns ugly and odd. *The Chronicle of Higher Education*. Retrieved November 23 2014, from http://chronicle.com/article/Replication-Crisis-in/147301/

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483,* 531–533. http://dx.doi.org/10.1038/483531a

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100,* 407–425. http://dx.doi.org/10.1037/a0021524

Bonett, D. G. (2012). Replication-extension studies. *Current Directions in Psychological Science, 21,* 409–412. http://dx.doi.org/10.1177/0963721412459512

Brandt, M. J., IJzerman, H., Dijkstehuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217–224. http://dx.doi.org/10.1016/j.jesp.2013.10.005

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. http://dx.doi.org/10.1038/nrn3475

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9,* 40–48. http://dx.doi.org/10.1177/1745691613513470

Chow, S.-C., Shao, J., & Wang, H. (2003). *Sample size calculations in clinical research*. Boca Raton, FL: Taylor and Francis.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, T. D., Gruder, C. L., Henningan, K. M., & Faly, B. R. (1979). History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin, 86,* 662–679. http://dx.doi.org/10.1037/0033-2909.86.4.662

Dallow, N., & Fina, P. (2011). The perils with the misuse of predictive power. *Pharmaceutical Statistics, 10,* 311–317. http://dx.doi.org/10.1002/pst.467

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE, 7,* e29081. http://dx.doi.org/10.1371/journal.pone.0029081

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181–210. http://dx.doi.org/10.1177/00131640121971167

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review, 18,* 3–12. http://dx.doi.org/10.1177/1088868313507536

Galak, J., Leboeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate ψ. *Journal of Personality and Social Psychology, 103,* 933–948. http://dx.doi.org/10.1037/a0029709

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (Sign) and Type M (Magnitude) errors. *Perspectives on Psychological Science, 9,* 641–651. http://dx.doi.org/10.1177/1745691614551642

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102,* 460–465. http://dx.doi.org/10.1511/2014.111.460

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60,* 328–331. http://dx.doi.org/10.1198/000313006X152649

Gonzalez, R. (2009). *Data analysis for experimental design*. New York, NY: Guilford Press.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20. http://dx.doi.org/10.1037/h0076157

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175–183. http://dx.doi.org/10.1111/j.1469-8986.1996.tb02121.x

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42,* 443–455. http://dx.doi.org/10.1037/0003-066X.42.5.443

Hedges, L. V. (1998). Bayesian meta-analysis. In B. S. Everitt & G. Dunn (Eds.), *Statistical analysis of medical data: New developments* (pp. 251–275) New York, NY: Oxford University Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504. http://dx.doi.org/10.1037/1082-989X.3.4.486

Henrion, M., & Fischoff, B. (1986). Assessing uncertainty in physical constants. *American Journal of Physics, 54,* 791–798. http://dx.doi.org/10.1119/1.14447

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A, Statistics in Society, 172,* 137–159. http://dx.doi.org/10.1111/j.1467-985X.2008.00552.x

Hunter, J. E. (2001). The desperate need for replications. *Journal of Consumer Research, 28,* 149–158. http://dx.doi.org/10.1086/321953

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* e124. http://dx.doi.org/10.1371/journal.pmed.0020124

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). New York, NY: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524–532. http://dx.doi.org/10.1177/0956797611430953

Kahneman, D. (2012, September 26). A proposal to deal with questions about priming effects. [Letter emailed to social priming researchers]. Retrieved from http://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. http://dx.doi.org/10.1080/01621459.1995.10476572

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299–312. http://dx.doi.org/10.1177/1745691611406925

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142,* 573–603. http://dx.doi.org/10.1037/a0029146

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and BUGS* (2nd ed.). Burlington, MA: Elsevier.

Kuiper, R. M., Buskens, V., Raub, W., & Hoijtink, H. (2013). Combining statistical evidence from several studies: A method using Bayesian updating and an example from research on trust problems in social and economic exchange. *Sociological Methods & Research, 42,* 60–81. http://dx.doi.org/10.1177/0049124112464867

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31,* 107–112. http://dx.doi.org/10.1111/j.2044-8317.1978.tb00578.x

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *The New England Journal of Medicine, 327,* 248–254. http://dx.doi.org/10.1056/NEJM199207233270406

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43,* 679–690. http://dx.doi.org/10.3758/s13428-010–0049-5

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163. http://dx.doi.org/10.1037/1082-989X.9.2.147

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science, 9,* 612–625. http://dx.doi.org/10.1177/1745691614548513

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. (2007). Most published research findings are false—but a little replication goes a long way. *PLoS Medicine, 4,* e28. http://dx.doi.org/10.1371/journal.pmed.0040028

Nosek, B. A., & Lakens, D. (2014). A method to increase the credibility of published results. *Social Psychology, 45,* 137–141. http://dx.doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615–631. http://dx.doi.org/10.1177/1745691612459058

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7,* 657–660. http://dx.doi.org/10.1177/1745691612462588

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PLoS ONE, 7,* e42510. http://dx.doi.org/10.1371/journal.pone.0042510

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7,* 528–530. http://dx.doi.org/10.1177/1745691612465253

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's "retroactive facilitation of recall" effect. *PLoS ONE, 7,* e33423. http://dx.doi.org/10.1371/journal.pone.0033423

Roediger, H. L. (2012, February). Psychology's woes and a partial cure: The value of replication. *Observer, 25*. Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2012/february-12/psychologys-woes-and-a-partial-cure-the-value-of-replication.html

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553–565. http://dx.doi.org/10.1037/0033-2909.113.3.553

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science, 5,* 329–334. http://dx.doi.org/10.1111/j.1467-9280.1994.tb00281.x

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. http://dx.doi.org/10.3758/PBR.16.2.225

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181. http://dx.doi.org/10.1037/0003-066X.47.10.1173

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22,* 36–71. http://dx.doi.org/10.1016/0010-0285(90)90003-M

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3,* 403–411. http://dx.doi.org/10.1037/1082-989X.3.4.403

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83. http://dx.doi.org/10.1037/0003-066X.40.1.73

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9,* 76–80. http://dx.doi.org/10.1177/1745691613514755

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to Registered Replication Reports at Perspectives on Psychological Science. *Perspectives on Psychological Science, 9,* 552–555. http://dx.doi.org/10.1177/1745691614543974

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9,* 59–71. http://dx.doi.org/10.1177/1745691613514450

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics Theory and Methods, 25,* 1595–1610. http://dx.doi.org/10.1080/03610929608831787

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110. http://dx.doi.org/10.1037/h0031322

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804. http://dx.doi.org/10.3758/BF03194105

Wagenmakers, E-J., Lee, M. D., Lodewyckx, T., & Iverson, G. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Goelen (Eds.), *Bayesian Evaluation of Informative Hypotheses* (pp. 181–207). New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-09612-4_9

Williams, L. E., & Bargh, J. A. (2008). Keeping one's distance: The influence of spatial distance cues on affect and evaluation. *Psychological Science, 19,* 302–308. http://dx.doi.org/10.1111/j.1467-9280.2008.02084.x

Wilson, D. B., & Shadish, W. R. (2006). On blowing trumpets to the tulips: To prove or not to prove the null hypothesis—Comment on Bösch, Steinkamp, and Boller (2006). *Psychological Bulletin, 132,* 524–528. http://dx.doi.org/10.1037/0033-2909.132.4.524

Wortman, J., Donnellan, M. B., & Lucas, R. E. (2014). Can physical warmth (or coldness) predict trait loneliness? A replication of Bargh and Shalev (2012). *Archives of Scientific Psychology, 2,* 13–19. http://dx.doi.org/10.1037/arc0000007

Yuan, K.-H., & Maxwell, S. E. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30,* 141–167. http://dx.doi.org/10.3102/10769986030002141

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: Proceedings of the first international meeting* (pp. 585–603). Valencia, Spain: University of Valencia Press.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.