Psychological Methods
2016, Vol. 21, No. 1, 1–12

1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000051

There's More Than One Way to Conduct a Replication Study: Beyond Statistical Significance

Samantha F. Anderson and Scott E. Maxwell
University of Notre Dame

As the field of psychology struggles to trust published findings, replication research has begun to become more of a priority to both scientists and journals. With this increasing emphasis placed on reproducibility, it is essential that replication studies be capable of advancing the field. However, we argue that many researchers have been only narrowly interpreting the meaning of replication, with studies being designed with a simple statistically significant or nonsignificant results framework in mind. Although this interpretation may be desirable in some cases, we develop a variety of additional "replication goals" that researchers could consider when planning studies. Even if researchers are aware of these goals, we show that they are rarely used in practice—as results are typically analyzed in a manner only appropriate to a simple significance test. We discuss each goal conceptually, explain appropriate analysis procedures, and provide 1 or more examples to illustrate these analyses in practice. We hope that these various goals will allow researchers to develop a more nuanced understanding of replication that can be flexible enough to answer the various questions that researchers might seek to understand.

Keywords: replication, data analysis, confidence interval, effect size, equivalence test

Replication, a once largely ignored premise, has recently become a defining precept for the future of psychology. Reproducibility has been referred to as the "cornerstone" (Simons, 2014, p. 76) and "Supreme Court" (Collins, 1985, p. 19) of science, and as "the best and possibly the only believable evidence for the reliability of an effect" (Simons, 2014, p. 76). In fact, "findings that do not replicate are worse than fairy tales" (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012, p. 633).

The idea of replication is not new. Even prior to Sir Ronald Fisher and the advent of modern experimental design (circa 1935), the field of agriculture used replication to assess accuracy and reliability (Yates, 1964). In fact, Fisher himself emphasized the importance of replication, believing that experimental findings are only established if "a properly designed experiment rarely fails to give . . . significance" (Fisher, 1926, p. 504). In 1969, Tukey noted that "confirmation comes from repetition" and that ignoring the need for replication would "lend[s] itself to failure and more probably destruction" (Tukey, 1969, p. 84). However, replications were rarely conducted due to lack of incentive and rarely published due to lack of novelty (Nosek & Lakens, 2014). This lack of incentive gradually started to change when concerns about "the reliability of research findings in the field" began to emerge (Pashler & Wagenmakers, 2012, p. 528). The field has been amid "a crisis of confidence" (Pashler & Wagenmakers, 2012, p. 528) wherein published findings are regarded with a greater degree of skepticism in the wake of potentially too much flexibility in research practices (e.g., Simmons, Nelson, & Simonsohn, 2011).

Due to these growing concerns over the potential unreliability of reported results in psychology, researchers have begun to emphasize the importance of reproducing results and call for a greater focus on replication. Over the past decade, the number of articles focused on replication has grown steadily. A PsycINFO search for scholarly documents with replication or any of its derivatives in the title yields 82 articles in 2003, 121 articles in 2008, and 154 articles in 2013. Major journals have dedicated special sections and issues to the topic (e.g., *Perspectives on Psychological Science*, 2012; *Social Psychology*, 2014) in the hopes of creating incentives for researchers to engage in replication studies. The Center for Open Science (2012) has introduced a project aimed at assessing the bias present in the current psychological literature by inviting scientists to attempt to replicate findings from a sample of published findings from prominent journals in 2008.

Yet, despite increased appreciation for the role of replication and motivation to focus on replication, the current state of replication research remains seemingly incapable of truly advancing the field. The failure rate of replications is alarmingly high, as evidenced in a recent issue of *Social Psychology*. Out of 14 replication attempts arranged by Nosek and Lakens (2014), nine did not replicate the original study and another five were only partial replications, more nuanced manifestations of the effect (i.e., the effect only appeared in specific conditions), or had smaller effect sizes. This lack of replicability lends itself to questions regarding the reason that so many fail. In some cases, failure to replicate may be due to issues with the original study, including researcher degrees of freedom (Simmons et al., 2011). However, other replications may fail due to problems with the replication study itself. In addition to low power of the replication study, a number of other factors have limited the effectiveness of recent replications (Braver, Thoemmes, & Rosenthal, 2014). First, re-

This article was published Online First July 27, 2015.

Samantha F. Anderson and Scott E. Maxwell, Department of Psychology, University of Notre Dame.

Correspondence concerning this article should be addressed to Samantha F. Anderson, Department of Psychology, University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556. E-mail: sander10@nd.edu

searchers have often displayed a rather narrow perspective on replication, with study goals often being to replicate a statistically significant effect. Less often, the intention seems to be to show that a presumed effect does not exist. We assert that there are a number of additional worthy goals that researchers have rarely considered in planning replication studies. Further, even when pursuing noble goals, the analyses used to achieve these goals are often inadequate and often do not truly match the intended research question.

This article aims to offer readers an appreciation for a number of replication-related goals that may lead the field to a more nuanced understanding of the replicability of prior findings (see Table 1). Further, we intend to provide a conceptual and practical overview of recommended analysis strategies, each paired with illustrative examples.

General Considerations

We conducted a PsycINFO search for scholarly, peer-reviewed articles published in 2013 with replicat* in the title. Of the 154 results, we selected 50 to code. The other 104 studies were excluded based on the following properties: less relevant to general psychology (e.g., business journals, nursing journals), language other than English, qualitative-only results, replications of psychometric properties, and genome-wide association studies. This selection of 50 replications yields 44 that seem to decide the success of the replication based on a statistical test alone.[1] These studies generally interpreted the *p* value as either in line or divergent with the original study, based on whether both studies came to the same or different conclusions regarding statistical significance. Although this general strategy may at times be the most appropriate for the question at hand, we argue that authors may be considering replication in an overly narrow context. Along these lines, we invite authors to consider both the additional goals we outline and the analyses appropriate to those goals. The following section introduces six potential goals for replication. Later sections will be devoted to further developing those goals and associated analyses.

Replication of significance may indeed be a worthy goal to pursue, as replicating an effect in the same direction as the original study is often enlightening in its own right. This may be especially true if the original study resulted in unexpected or counterintuitive findings. For example, consider the seminal findings on thought-suppression (Wegner, Schneider, Carter, & White, 1987). Surprisingly, participants who were instructed to suppress thoughts of a white bear were rather ineffective at doing so, thinking about the bear once per minute, on average. The second finding was even more surprising. When the same participants repeated the experiment with instructions to think about the white bear, they had thoughts of the bear significantly more often than a control group who never received the thought-suppression instructions. This rebound effect was an unexpected phenomenon, and one could argue that its existence and the directionality of the effect are more important than the size of the effect. Replications of results of this nature may simply need to reproduce a statistically significant effect in the intended direction in order to lend support to the original findings. Unexpected results may have a special role in theory testing, underscoring the importance of hypothesis testing as opposed to effect-estimation when examining theoretical predictions (Morey, Rouder, Verhagen, & Wagenmakers, 2014).

However, significance-based replication (Goal 1) may not always be the most advantageous goal, and the overwhelming emphasis on this goal may be limiting the contribution that replication research can make to the field as a whole. Recent apparent nonreplications of controversial results, such as those of Bargh's subtle priming and Bem's ESP experiment, speak to the importance of detecting spurious findings (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012; Galak, LeBoeuf, Nelson, & Simmons, 2012). Indeed, there were a number (16) of authors in the 2013 PsycINFO sample of replication studies that reported a nonreplication of original findings. Although one of these studies explicitly referenced nonreplication as the goal, others were vaguer in their intentions. It is thus somewhat unclear what the goal truly was in some of these cases, but what is clear is the fact that the often reported claims of "null" results did not match the analyses performed. In fact, all of these 14 studies conducted analyses capable only of evidencing a statistically significant effect, but not a null effect. It is well known that failure to reject the null hypothesis does not necessarily constitute evidence that the null hypothesis should be accepted, but our review of replication studies showed authors regularly making this mistake. Several studies published in the 2014 replication special issue of *Social Psychology* also fell victim to this mismatch between reported nonreplication and the inappropriate analyses used to support that conclusion.

Thus, whether or not authors are aware of the utility of intentions to show a null effect (nonreplication; Goal 2), they are most often using analysis strategies that fail to support the interpretation given for the results and fail to answer what may have been the real question of interest. There are two separate cautions here. First, authors who expect a statistically significant result (replication) are absolutely justified in conducting analyses in line with this goal. However, these authors must be careful in interpreting nonsignificant findings as direct evidence of a failure to replicate. More notably, authors who desire to evidence that there is no effect have often failed to utilize an analysis that can substantiate this goal. Again, though nonsignificant findings indicate a failure to reject the null hypothesis, many researchers claim that *p* values greater than .05 are evidence in favor of the null hypothesis or are a metric from which to determine the probability that the null is true. In fact, *p* values greater than .05 by themselves reveal little about the probability that the null hypothesis is true. We recommend that authors make use of equivalence tests (frequentist) or Bayesian methods in order to adequately support the claim of a null replication effect, which will be described in greater detail later. We note here that although Bayesian methods can be used for other hypothesis and interval-based situations presented in this paper, we limit our presentation of Bayesian methods to Goal 2, as these methods are especially helpful for answering questions regarding the lack of an effect. Readers interested in Bayesian methods more generally may consult Kruschke (2014) and Gelman et al. (2013).

[1] This estimate of 44 studies is likely conservative. Three additional studies based their main analysis on surface level comparisons (correlations, stepwise regression models, and sensitivity/specificity), without testing whether these parameters statistically differed. Thus, these studies did not decide success directly based on a statistically significant–nonsignificant distinction, but rather an even simpler visual inspection of the estimates in question.

Table 1
Six Replication Goals and Descriptions

No.	Goal	Recommended analysis	Success criterion
1	To inter the existence of a replication effect	Repeat analysis of original study	$p < .05$
2	To infer a null replication effect	Equivalence test	Confidence interval falls completely inside region of equivalence
3	To precisely estimate the replication effect size	AIPE, construct confidence interval for effect size	Effect size estimated with desired level of precision
4	To combine replication sample data with original results	Construct confidence interval for the average effect size of replication and original studies	Building on prior knowledge; more precise estimate of the effect of interest
5	To assess whether replication is clearly inconsistent with original	Construct confidence interval for the difference in effect sizes	Confidence interval for difference in effect sizes does not include 0
6	To assess whether replication is clearly consistent with original	Equivalence test, using confidence interval for the difference in effect sizes	Confidence interval for difference in effect sizes falls completely inside region of equivalence

In addition to showing the existence or nonexistence of a previously published finding, we believe that there are other potential goals of replication that have largely been overlooked in the recent push toward reproducibility. For example, researchers may have reason to question the size of the effect reported in the original study. Research has shown that published effect sizes are likely to be upwardly biased, which may motivate researchers to attempt to better estimate the true population effect size (Lane & Dunlap, 1978; Maxwell, 2004). Thus, it may be worthwhile to estimate the size of the effect of the original study, providing evidence that it is indeed as sizable as the original authors claimed (Biesanz & Schrager, 2010; Goal 3). This goal warrants the formation of a confidence interval around the replication effect size. Many researchers seem to be unaware of this goal, as only 23 studies from our PsycINFO sample provided an effect size, only four provided a confidence interval around it, and none discussed or interpreted these confidence intervals.

Another goal may be to replicate the original study by combining it with a new sample of participants in something akin to a small meta-analysis (Goal 4). Although two studies in the 2013 sample had access to the original study raw data, this access is often not possible. This goal allows comparison between a replication and original study without this requirement. However, no studies in our sample followed this goal. Authors may also want to show that a replication effect is clearly inconsistent with the original study's effect through more than simply direction/significance alone, meriting a test of the difference in effect sizes (Goal 5). This goal is an extension of Goal 3, wherein the replication effect size must be significantly distinct from the original to support nonreplication. Authors who declare their study a nonreplication in response to finding a smaller effect size in their sample, without testing its disparity from the original effect size, seem to be unaware of this goal, or of the proper analyses to accomplish the goal (three studies from our PsycINFO sample). Only one study from our 2013 sample used an analysis in line with Goal 5. Conversely, it may also be enlightening to show that a replication is clearly consistent with the original study through an analysis such as an equivalence test of the difference in effect sizes (Goal 6; no studies from our PsycINFO sample). Similarly to Goal 5, authors who declare their study a replication in response to finding a similar effect size, without testing its equivalence to the original effect size, seem to be unaware of this goal and associated analyses (three studies from our PsycINFO sample.)

It is important to note some caveats regarding direct (exact) versus conceptual replications. While direct replications were once avoided for lack of originality, authors have recently urged the field to take note of the benefits and importance of direct replication. According to Simons (2014), this type of replication is "the only way to verify the reliability of an effect" (p. 76). With respect to this recent emphasis, the current article will assume direct replication. However, despite the push toward direct replication, some have still touted the benefits of conceptual replication (Stroebe & Strack, 2014). Importantly, many of the points and analyses suggested in this paper may translate well to conceptual replication. However, readers should be cautioned that there are exceptions to this, as in replication studies with multiple dependent variables. Further, the interpretation of results may not be as straightforward in conceptual replications, as nonsignificant or disparate findings could be due to a host of uncontrolled factors, such as differences in, conditions, measurement tools, and participants.

Further, this article will mainly limit its focus to single replications of original studies. However, as others have noted, a single replication is usually insufficient to accept or refute a published effect with absolute confidence. We echo the importance of multiple replications which can then lend themselves to a future meta-analysis and we will broaden our discussion to these cases when possible (Hunter, 2001). Nevertheless, it is also important for researchers to be aware of various questions that may be addressed with a single replication study, as well as knowing the most appropriate analytic methods for each type of question. We note that just as in meta-analysis, access to the raw data is not required for any of our proposed methods.

Finally, the issue of sample size planning for replication studies is beyond the scope of this article. However, we emphasize that many, if not all, of the goals may require much larger sample sizes than are commonly seen in the literature. Replication research often suffers from low power due to the uncertainty and bias inherent in the sample effect sizes (from the original study) that inform the replication's planned sample size (Maxwell, Lau, & Howard, in press). Thus, even replication studies that claim to have power greater than .8 may have actual power that is much lower.

We urge researchers to attend to the nuances of their proposed analyses in making sample size decisions. We recommend Taylor and Muller (1996) for a power analysis method that handles both publication bias and the distribution inherent in sample effect size estimates. Finally, it is important to note that the sample size of the original study plays an important role anytime the goal involves either comparing or combining the results of the original study and the replication study.

We will discuss each of the aforementioned potential goals in more detail later in this article. We emphasize that these goals are not mutually exclusive and often may be combined when appropriate based on the questions at hand. We caution that goals should be decided on *a priori*, before conducting analyses. Performing replication studies with attention to a wider variety of definitions that could constitute replication may provide more illuminating answers as to the validity of purported effects in the literature.

Goal 1: To Infer the Existence (and Direction) of an Effect

As described previously, the goal of replicating the statistical significance of an effect seems to be the most common purpose described in recent replication studies. This is not surprising, given that psychologists often have "an exaggerated belief in the likelihood of successfully replicating an obtained finding" (Tversky & Kahneman, 1971, p. 105). If reproducibility is indeed the gold standard of science, it makes sense to attempt to replicate the statistical significance (and in many cases replicate the directionality) of previously reported effects. After selecting an appropriate sample size, the statistical methods chosen should typically mirror those in the original study. This may be a regression, an ANOVA, or something more complex. We provide a two group example below, although we acknowledge that this is only representative of some replications of interest to researchers. Of course, there are many ways an original study could have been performed, each with its own standard analysis. Researchers should be sensitive to the context of the original study in planning the most appropriate way to conduct the replication analyses.

Suppose a researcher is interested in replicating the scope-severity paradox. The original study on the topic found a surprising series of results that stood in contrast to the common sense view of the time. Specifically, participants randomly assigned to conditions judged equivalent crimes *less* severely when more people had been victimized by the crime and recommended *more* punishment for crimes of equal magnitude when fewer people were victimized (Nordgren & McDonnell, 2011). For simplicity, suppose the researcher's replication will focus only on the perceived severity of the crime, where in the original study, small scope vignettes were judged with more severity ($M = 6.37$, $SD = 1.67$) than large scope vignettes ($M = 5.51$, $SD = 1.33$), $F(1, 59) = 4.88$, $p = .03$.[2] The corresponding original sample effect size was $d = 0.57$ (approximately Cohen's medium effect). Notice that the original study found a significant effect of the scope of the crime on perceived severity. The researcher may first want to replicate the statistical significance of the effect, without attention to its size. In this case, it is important to the researcher to say that participants indeed perceived crimes to be more severe when fewer people were affected by them, but not whether that effect is large enough to be of clinical or practical importance. In other words, the fact that such a surprising effect exists is noteworthy, while its size is less vital to the theory. Consequently, analyses should proceed as in the original study. In this case, the researcher would perform an independent samples t test on two groups randomly assigned to scope conditions. A p value of less than .05 would indicate that the replication attempt was successful, while a larger p value would indicate that varying the scope of a crime did not have a statistically significant effect on perceived severity, though not that the influence of scope was essentially zero.

We argue that in these cases, it is often not only the statistical significance of the effect, but also the directionality inherent in the statistically significant finding that is important to convey. For a successful replication of a two-group study, the replication effect must not only have a p value less than .05, but also reproduce the direction of the mean difference found in the original study. In the example above, a replication finding that participants judge crimes *more* severely when they victimize more people would likely be considered unsuccessful, even if the mean difference was statistically significant in both cases. We acknowledge, however, that there may be a few situations where even direction does not matter. Although many studies involving three or more groups eventually involve analyzing contrasts, where direction is of interest to the theory, some theories may simply contrast *any* difference between means with *no* difference between means. For example, a seminal study found that infants preferred to look at faces over scrambled faces and blank screens (Goren, Sarty, & Wu, 1975). We argue that in this case, however, a replication of any visual preference may still be considered successful by some, if the contrasted theories are thought to be no visual preference versus any visual preference (indicating that the infant visual system is more developed than had been previously thought).

Goal 2: To Infer a Null Effect

In 2011, an uproar ensued over a controversial study published in the *Journal of Personality and Social Psychology* (JPSP; Bem, 2011). The article, through nine experiments, claimed that undergraduates successfully displayed retroactive influence of future events on current responses, an indication of the existence of psi, with a mean sample effect size of $d = 0.22$. Skeptics attempted to reproduce Bem's findings and failed multiple times. But what truly constitutes a nonreplication? A study using methods akin to the original, but failing to produce a statistically significant result may be viewed by many as a failure to replicate. In fact, a highly publicized replication attempt of Bem's study made essentially these conclusions based on nonsignificant results (Ritchie, Wiseman, & French, 2012). Although other replications went on to apply more appropriate analyses as evidence for nonreplication, these and other similar conclusions are a sign of a general lack of understanding of what nonsignificant findings actually reveal.

If one is skeptical of an original study's results, a goal may be to infer a null effect. In this case, it is necessary to show evidence in favor of the null hypothesis, rather than simply a failure to reject the null. As discussed earlier, a failure to reject the null hypothesis is not necessarily informative about the likelihood that the true

[2] The authors report 1 and 59 degrees of freedom for a 1 way ANOVA with 60 participants. If the description is accurate, correct *df* would be 1 and 58.

effect is zero or does not exist, even when the study is seemingly adequately powered. In fact, when the goal is to infer a null effect, the alternative hypothesis should be the default hypothesis, and it should take sufficient evidence to overturn the alternative in favor of the null, so the two hypotheses effectively play opposite roles from their usual role in traditional hypothesis testing (Walker & Nowacki, 2011). In light of this, we do not recommend using the traditional statistical methods of the original study. Three approaches are capable of satisfying the goal of being able to conclude that an effect is null or essentially null in many common psychological designs.

Frequentist Method

The method most accessible to psychology researchers is the equivalence test (or two one-sided tests; TOST), because it is derived from the traditional frequentist perspective familiar to those conducting hypothesis tests. The first step is to establish what is known as a region of equivalence or region of indifference. This is an interval of values of which the researcher believes to be so small as to be essentially zero. Notice that this interval is based entirely on theory and must be specified prior to collecting replication data. The logic of this is consistent with the "good enough principle," which acknowledges that in strict terms, the null hypothesis may never be exactly true (Serlin & Lapsley, 1985). The authors encourage forming "a good-enough belt width of delta" in the null prediction (p. 79). Following the traditional analyses, the second step is to form a $(1 - 2\alpha) \times 100\%$ confidence interval around the estimate of the effect. For an α level of .05, a 90% confidence interval should be computed. Although 95% confidence intervals are more common in traditional null hypothesis testing, the equivalence test corresponds to two one-tailed tests, each at $\alpha = .05$ (Walker & Nowacki, 2011). The logic of TOST is that if the confidence interval of the estimate falls entirely within the region of equivalence, the null hypothesis can be claimed to be functionally true with a low amount of uncertainty.

Continuing with the scope-severity paradox example introduced in Goal 1, suppose a skeptic believes the original study was flawed and thus would like to show that the number of individuals affected by a crime essentially does not impact perceived severity at all. Based on theory in the domain of crime severity, the skeptic determines that a mean difference of half of a point on the 10-point severity measure used in the original study is too small to constitute any real effect. The region of equivalence would extend from −0.5 to 0.5. After conducting the replication study with 40 participants per group, imagine the skeptic finds a mean difference of 0.2 points on the scale between small scope crimes ($M = 6.05$, $SD = 1.30$) and large scope crimes ($M = 5.85$, $SD = 1.15$), $t(78) = 0.729$, $p = .468$. Although this estimate falls within the region of equivalence, the important question is whether a 90% confidence interval around 0.2 falls entirely inside the region. In this scenario, the researcher would find a 90% confidence interval that extends from −0.25 to 0.65. In this case, then, the skeptic determines that the replication was not able to demonstrate equivalence, as the upper limit of the observed confidence interval extends beyond the region of equivalence.

Had the skeptic's replication sample yielded a more precise, narrow confidence interval (e.g., from −0.05 to 0.45), there would have been convincing evidence that the scope of the crime had a negligible impact on severity. This brings up an important caveat about power and sample size for studies attempting to claim no effect. It should not be surprising that large sample sizes may be necessary, as it requires more precision to establish equivalence than it does to simply fail to reject the null.

One limitation of equivalence tests is that they typically pertain to situations involving a unidimensional confidence interval. Even so, the method can often be applied in multivariate settings by reexpressing effects in terms of summary measures of proportion of variance accounted for, such as R squared. Alternatively, Wellek (2010) describes the union-intersection principle as an approach for establishing multidimensional equivalence.

Bayesian Methods

In addition to the frequentist equivalence testing, Bayesian methods also allow for conclusions regarding the relative evidence for the null hypothesis. The region of practical equivalence method (ROPE; Kruschke, 2011) is similar to equivalence testing, but incorporates prior information regarding the effect size and consequently has a somewhat different interpretation. A second Bayesian method specifically quantifies the degree of support for the null hypothesis. The Bayes factor can reflect the ratio of the probability of the null hypothesis over that of the alternative hypothesis (see Rouder & Morey, 2012, and Rouder, Morey, Speckman, & Province, 2012 for reviews on the use of Bayes factors for regression and ANOVA models, respectively). The general equation for the Bayes factor is as follows,

$$B_{01} = \frac{pr(\mathbf{D}|H_0)}{pr(\mathbf{D}|H_1)}, \quad (1)$$

where $\mathbf{D}$ is the observed data and H_0 and H_1 designate the two competing hypotheses (Kass & Raftery, 1995). For example, a Bayes factor of 2 would indicate that the null hypothesis is two times more likely to be true than the alternative. Although the null is more likely than the alternative, the evidence is not overwhelming in this case. The Bayes factor can be further used to specify the probability that the null hypothesis is true, given the following formula, assuming equal prior probabilities for the null and alternative hypothesis:

$$Pr(H_o) = \frac{B_{01}}{B_{01} + 1}, \quad (2)$$

where B_{01} is the Bayes factor. In the preceding example with a Bayes factor of 2, the probability the null is true would be two thirds, or about 66%. Jeffreys (1961) and Kass and Raftery (1995) have developed classification schemes that may be used as rough guidelines for interpreting the strength of evidence suggested by the Bayes factor. However, the practical meaning of the Bayes factor may depend on personal judgment as well as what is regarded as appropriate in a researcher's particular subdiscipline. For example, Rouder, Morey, Speckman, and Province (2012) noted that a Bayes factor of 40 supporting the existence of Bem's ESP findings may still not be high enough to indicate overwhelming support, given that ESP stands in contrast to basic scientific principles.

Continuing our scope-severity example from a Bayesian perspective, remember that the previous replication researcher conducted an equivalence test on a replication attempt with 40 par-

ticipants per group yielding a nonsignificant difference between small scope crimes ($M = 6.05$, $SD = 1.30$) and large scope crimes ($M = 5.85$, $SD = 1.15$), $t(78) = 0.729$, $p = .468$. This researcher instead could alternatively consider a Bayesian perspective. To determine the strength of evidence that this nonsignificant result suggests for the null hypothesis, the skeptic can calculate a Bayes factor using an online calculator such as Rouder, Speckman, Sun, Morey, and Iverson's (2009; http://pcl.missouri.edu/bayesfactor), which requires inputting the sample size per group, the *t*-value, and the scale *r*, or prior size.[3] The default *r* is medium ($\sqrt{2}/2$), which says that 50% of the prior effect sizes are within the interval [−0.7071, 0.7071]. Entering the appropriate information into the calculator results in a Bayes factor of 3.42 in favor of the null, which from Equation 2 yields a 77% probability that the null is true.

A possible limitation of the typical Bayes factor approach is that as typically formulated it assesses the strength of evidence in favor of the point null hypothesis, that is, a hypothesis that the effect in question is precisely zero (some authors refer to this as a nil hypothesis). The plausibility of any effect being literally exactly zero is questionable, leading some researchers to question the value of such evidence. Thus, instead of testing a hypothesis of exact equality, it may be more appropriate to test a hypothesis of approximate equality, such as

$$H_0^{\varepsilon}: |\mu_O - \mu_R| \leq \varepsilon, \tag{3}$$

where ε reflects the largest effect deemed to be equivalent to zero. However, Berger and Delampady (1987) showed that the probability associated with the precise null hypothesis tends to be very close to the probability associated with an approximate null hypothesis as long as ε is small and the sample size is at most moderately large. As a result, in many circumstances the usual Bayes factor will be functionally equivalent to the technically more appropriate Bayes factor for the approximate null hypothesis. Even so, there may be situations where it is preferable to calculate the Bayes factor associated with an approximate null hypothesis. Morey and Rouder (2011) describe such a method for the specific case of a single mean or paired means and also provide an online calculator along with an R package. Hoijtink and colleagues have also described how to calculate Bayes factors for approximate hypotheses in a variety of situations. For example, Hoijtink and Klugkist provide a conceptual comparison of Bayes factors for nil hypotheses with Bayes factors for "non-sharp null models" (Hoijtink & Klugkist, 2007, p. 83). van de Schoot et al. (2011) provide an especially accessible example of testing a hypothesis of approximate equality. Hoijtink (2011) provides comprehensive coverage of calculating Bayes factors for informative hypotheses for a wide variety of designs and analyses.

Goal 3: To Quantify the Size of an Effect

In many cases, it is not only important to replicate the existence of an effect, but also to replicate its size. As mentioned earlier, in addition to carrying with them uncertainty and variability, published effect sizes tend to be positively biased (Lane & Dunlap, 1978; Maxwell, 2004; Maxwell, Kelley, & Rausch, 2008). More clearly, the requirement of statistical significance for publication can act as a censoring mechanism to make reported effect sizes systematically too large (Taylor & Muller, 1996). However, appropriate sample size planning in a replication study, with attention paid to accurately estimating effect size, can result in less biased effect sizes for replication. Specifically, as replications more often are published even with nonsignificant results, the replication may not suffer the same fate of bias as the original. A researcher may thus be suspicious that the true population effect size is not as large as claimed in the original study or may simply desire more confidence in the actual size of the effect. This goal, accuracy in parameter estimation (AIPE), is intertwined with sample size planning, as it represents an alternative approach to traditional statistical power. In contrast to traditional power, which is defined as the probability of obtaining a statistically significant result for a true alternative hypothesis, AIPE allows for sample size to be based on the ability to estimate the size of an effect and to do so with a certain amount of precision (Maxwell, Kelley, & Rausch, 2008). Once the proper sample size is determined for the desired amount of precision, the actual analysis can proceed as in the original study with one small change: the author should provide a confidence interval around his or her specified effect size estimate, which can serve as a metric for determining the potential boundaries of the effect estimate.

To illustrate this goal in practice, let us develop a more applied example, in which the size of the effect may be an important contributor in whether study findings can be implemented in a real world context. Suppose a researcher is studying the effectiveness of an intervention program on adolescent self-esteem. A prior study using the same intervention found that those randomly assigned to the intervention group reported significantly higher self-esteem than those in a control group, with an effect size of $d = 0.5$ (Cohen's medium effect, $n = 50$ per group). However, the 95% confidence interval around this effect size, extending from 0.10 to 0.90, indicates substantial imprecision in the estimate. Assume that the funding agency will not allow the program to be used on a large scale in the school system until the researcher can hone in on the size of the effect with a greater degree of precision. It will no longer be sufficient to only evidence a statistically significant intervention effect. Perhaps the researcher determines that a half-width of 0.15 will allow him or her to estimate the effect of the intervention with adequate certainty. Proper sample size planning will provide him or her with the correct number of participants to guarantee this precision (here, approximately 350 individuals per group). After conducting the *t* test, as in Goal 1, and computing an effect size, suppose the sample Cohen's *d* is 0.35. A 95% confidence interval can then be computed either by hand or via statistical software. Packages such as MBESS in R contain user friendly functions for estimating confidence intervals around effect sizes (Kelley & Lai, 2012). Additionally, Bonett (2008, 2009) provides computational examples for calculating confidence intervals around the standardized mean difference (SMD) for a variety of research designs.

In the current example, the researcher calculates a 95% confidence interval of [0.20, 0.50]. In a case such as this, the researcher is able to estimate the effect with a higher degree of certainty and may be better equipped to convince the funding agency with this

[3] The Bayes factor reported is the JZS Bayes factor, which is based on a Cauchy prior (Rouder et al., 2009; based on work by Jeffreys, 1961 and Zellner & Siow, 1980).

more accurate estimate. Here, the success of the replication can be measured, in a sense, by the degree of precision surrounding the estimated effect size, without mention of the similarity between the new and original effect sizes. This approach does not directly compare the two effect sizes in question and thus does not evidence a statistical difference between the results of the original and replication study. Thus, we reiterate that the goals in this paper are not mutually exclusive and can be used in combination when appropriate.

Confidence intervals can be formed around a variety of effect size estimates, in addition to the SMD. Just as in Goal 1, the basic set of analyses applied by the researcher will mirror those of the original study. For example, a researcher could replicate a regression study via the original methods and additionally provide a confidence interval surrounding the replication's sample regression coefficient for the variable of interest. Package MBESS has functionality for confidence intervals around the regression coefficient, multiple correlation (R), multiple correlation squared (R^2), standardized contrast, standardized mean, and the SMD (Kelley & Lai, 2012). The importance of confidence intervals has not gone unnoticed in research areas beyond replication. Although the field largely ignored Rozeboom's, 1960 recommendation and even Wilkinson's, 1999 recommendation to provide confidence intervals for parameter estimates, researchers have recently been urged to heed this recommendation, given the uncertainty in sample effect sizes and the availability of simple ways to calculate confidence intervals (Bonett, 2008; Stukas & Cumming, 2014). We argue further that the confidence interval has a special place, too, in replication research.

Goal 4: To Infer an Effect Based on a Combination of the Original and Replication Studies

If reproducibility is the gold standard of science, then meta-analysis may be considered the gold-standard of reproducibility. Statistically combining the results of multiple studies not only increases power, but also pulls from a more influential evidence base in drawing conclusions (Cohn & Becker, 2003). In the case of a single replication, one may want to infer an effect utilizing the resources of the original study in addition to the new sample, especially if there is little reason to suspect the original was flawed. Bonett (2009) emphasized the benefits of statistically combining a replication attempt with prior studies to define a more precise estimate of the effect size in question. This small meta-analysis may then serve as the initial step in the continuously cumulating meta-analysis (CCMA) technique (Braver et al., 2014; Lau et al., 1992; Mullen, Muellerleile, & Bryant, 2001; Rosenthal, 1990) Each replication attempt is not viewed as a standalone piece of evidence for or against the null hypothesis, but rather is statistically combined with the prior studies (Bonett, 2009). This can happen incrementally. For example, the first replication attempt could be combined with the original study, the second replication attempt could be combined with both the original and first replication, and so on. Following an initial replication attempt, the resulting effect size would be a pooled estimate of the replication and original study, benefitting from the data and resources of both studies (Bonett, 2009). This method is thus able to begin to correct some of the issues amounting from the low levels of power often inherent in research.

As discussed above, nonsignificant findings in replication attempts are often viewed as failures and diminish the validity of a purported effect. However, the results of CCMA may sometimes be in contrast to this traditional belief and may be able to reverse this perception. For example, when meta-analytically combined with an original study, a nonsignificant replication may show a statistically significant effect. Further, the combined effect may even be more significant than the original effect. That is, a nonsignificant replication attempt may actually lead to stronger evidence against the null hypothesis than the original statistically significant findings. Braver, Thoemmes, and Rosenthal (2014) were able to numerically illustrate this paradox via simulation studies. In their example, the original study had a p value of .033, the replication had a p value of .198, and the combined meta-analytic p value was .016. Although conceptually counterintuitive, this emphasizes the power advantage of meta-analysis, even when only two studies are involved.

Although significance-based methods exist for converting the p values from each study into an average meta-analytic p value (see Braver et al., 2014), these methods do not take into account sample size differences between studies and may be biased in cases of effect size heterogeneity (Bonett, 2009; Braver et al., 2014). Confidence interval methods have also been developed. Here, it is important to note the distinction between fixed and random effects meta-analysis. In fixed effect meta-analysis, the estimates are assumed to generalize only to the exact studies involved in the meta-analysis. Meta-analytic confidence intervals have been proposed (see Bond, Wiitala, & Richard, 2003; Hedges & Vevea, 1998), but these share the inadequacies of the significance test method (Bonett, 2009). Random effects meta-analytic confidence intervals assume that the studies used are random selections from a specified larger population were introduced as a potential solution (see Bond et al., 2003; Hedges & Vevea, 1998), but this assumption is often faulty in practice (Bonett, 2009). To combat these issues, Bonett (2008 [correlations], 2009 [standardized and unstandardized mean differences]; Bonett & Price, 2014 [proportions]) proposed an alternative set of fixed effect meta-analytic confidence intervals that are robust to even large degrees of heterogeneity of variance and non-normality with sample sizes larger than 30 per group. The confidence interval formula for the SMD in the case of two studies is as follows:

$$\bar{d} \pm z_{\alpha/2}\left[\frac{b_1^2 var(\hat{d}_1) + b_2^2 var(\hat{d}_2)}{4}\right]^{1/2}, \tag{4}$$

where $z_{\alpha/2}$ is a two-tailed critical value and $\bar{d}$ is obtained from the following formula:

$$\bar{d} = \frac{b_1\hat{d}_1 + b_2\hat{d}_2}{2} \tag{5}$$

b_i is a bias adjustment obtained from:

$$b_i = 1 - \frac{3}{4(n_{i1} + n_{i2}) - 9} \tag{6}$$

for studies with two groups, where i represents study i. Finally, the variance of $\hat{d}$ is obtained from:

$$var(\hat{d}_i) = \hat{d}_i^2\left(\frac{\hat{\sigma}_{i1}^4}{df_{i1}} + \frac{\hat{\sigma}_{i2}^4}{df_{i2}}\right)\Big/ 8\hat{\sigma}_i^4 + \left(\frac{\hat{\sigma}_{i1}^2}{df_{i1}} + \frac{\hat{\sigma}_{i2}^2}{df_{i2}}\right)\Big/ \hat{\sigma}_i^2, \tag{7}$$

where $df_{ij} = n_{ij} - 1$ and

$$\hat{\sigma}_i = \left(\frac{\hat{\sigma}_{i1}^2 + \hat{\sigma}_{i2}^2}{2}\right)^{1/2}. \quad (8)$$

For more general forms of these equations that can accommodate multiple replications, see Bonett (2009).

As an example, let us return to the scope-severity example from above. Recall that the original study found a significant effect of the scope of the crime on perceived severity, $F(1, 59) = 4.88$, $p = .03$, $d = 0.57$, such that small scope vignettes were judged with more severity ($M = 6.37$, $SD = 1.67$) than large scope vignettes ($M = 5.51$, $SD = 1.33$). Suppose that this new researcher's replication attempt with 40 participants per group found that small scope crimes ($M = 6.20$, $SD = 1.50$) were not judged significantly differently than large scope crimes ($M = 5.75$, $SD = 1.20$), $t(78) = 1.48$, $p = .143$, $d = 0.33$. Using Bonett's (2009) formula for the average standardized effect size, the researcher would obtain a meta-analytic confidence interval of [0.10, 0.79]. Again, note that in this case, the average effect size ($d = 0.44$) is significant, despite a nonsignificant replication when viewed alone. Further, the researcher has the benefit of a more precise indicator of the effect of scope severity (the replication study alone had a wider confidence interval of [−0.11, 0.77]), as well as a contribution that is able to progressively build upon the knowledge base of the relevant domain.

Despite the conceptual and methodological advantages to adopting CCMA, we caution readers that this approach does not ameliorate the issue of publication bias or the upward bias often inherent in published effect size estimates (Braver et al., 2014). However, the approach as geared toward a single replication is still rather early in its development and application, and merits continued attention. Further, although the emphasis of the present article is on direct replication, approaches have recently been developed that use meta-analytic methods to provide interesting comparisons in conceptual replications and extensions. Finally, although robust to moderate non-normality, other approaches are preferred when more extreme levels of non-normality are present or when normality is unknown. The interested reader should consult Bonett (2009).

Goal 5: To Assess Whether Replication is Clearly Inconsistent With Original Study

Although considering a replication unsuccessful when the results are essentially zero (as in Goal 2) has many merits, investigators could alternatively consider a more nuanced definition of replication. In fact, critics of traditional significance testing have convincingly emphasized that the difference between a significant finding and a nonsignificant finding is often not statistically significant (Gelman & Stern, 2006). Rather than simply evidencing a nonsignificant effect (or null effect, as in Goal 2) as support for a conclusion that an effect is not real, more refined support could be that the effect size of the replication is significantly in contrast to that reported in the original. This goal, essentially a test of heterogeneity of effect sizes, involves more of a direct comparison between the two studies (e.g., Hedges & Vevea, 1998). Although it adds a layer of complexity, the analysis strategy we recommend incorporates the fundamental concepts we have already introduced—effect sizes and confidence intervals. After conducting the replication study in keeping with the analyses reported in the original study, we suggest that investigators should form a confidence interval for the difference in effect sizes (Bonett, 2009). One may wonder whether this analysis could be simplified by computing a confidence interval around each sample effect size and looking for overlap between the confidence intervals. Although intuitively appealing and interpreted by many researchers (as demonstrated by Higgins, Thompson, Deeks, & Altman, 2003), this is incorrect (Schenker & Gentleman, 2001). Although two nonoverlapping confidence intervals are indeed significantly different, the converse is not always true. In fact, two confidence intervals can overlap, but their parameter estimates can still be significantly different from each other (Schenker & Gentleman, 2001). Thus, the method is too conservative when the null hypothesis is true and lacks power when it is false. One might also wonder whether it is sufficient to claim that two studies differ because one had $p < .05$ and the other $p > .05$. Again, this is inadequate. As noted earlier, the difference between significant and nonsignificant effects may not be statistically significant (Gelman & Stern, 2006). In fact, p values often tell little about how distinct one finding is from another. For example, $p = .049$ and $p = .06$ may tell essentially the same story if the studies are similar, but the field has continued to cling to the notion that there is something magical about .05 (Rosenthal & Rubin, 1979). Thus, it is necessary to calculate a single confidence interval for the difference between the original and replication effect sizes.

The general notion of directly comparing effect sizes was originally developed about three decades ago, first with p values, then with effect size estimates (Hedges, 1982; Hsu, 1980; Rosenthal & Rubin, 1979, 1982). The original Rosenthal and Rubin (1979) method for comparing p values involved converting each p value to a Z score, dividing their difference by $\sqrt{2}$, and converting the resulting Z back to a p value, although they proposed that this method could also be used with effect sizes. However, this method was criticized for its restriction to studies of similar size when directly comparing effect sizes, resulting in too many Type I or Type II errors, depending on which study had a larger n (Hsu, 1980). Hsu instead proposed a method more robust to differences in sample size, which was criticized for producing upwardly biased values of Z (Hedges, 1982). Hedges (1982) developed a parallel method using an unbiased effect size estimator and Rosenthal and Rubin (1982) refined their method. Despite initial interest in this area, it appears that researchers have failed to see the importance of these types of comparisons for replication studies.

Although these early approaches laid the groundwork for tests of heterogeneity of effect sizes, the methods all consist of a significance test. More recently, authors have cautioned against the use of significance test approaches to effect size heterogeneity (Bonett, 2008, 2009). Statistical tests are often misused: Failing to find a significant difference in effect sizes does not necessarily imply they are homogeneous and statistical significance does not imply a meaningful difference (Bonett & Wright, 2007). We argue that providing a confidence interval for the difference in effect sizes is able to convey additional information over a significance test alone. This can be accomplished through methods outlined in Bonett (2009). Specifically, one can define a linear contrast of effect sizes and compute a confidence interval for that contrast. For the case of a comparison between the SMD of an original study to that of a direct replication, the contrast could be coded as +1, −1, in the following form: $d_O - d_R$, where subscripts O and R indicate

the original and replication studies. The confidence interval for this comparison would then be:

$$d_O - d_R \pm z_{\alpha/2}[var(d_O) + var(d_R)]^{1/2}, \quad (9)$$

where $z_{\alpha/2}$ is the two-tailed critical value (Bonett, 2009).

Adapting our self-esteem intervention example for this goal, recall that the hypothetical original study found that the intervention ($n = 50$) had a positive impact on self-esteem when compared with a control group ($n = 50$), with an effect size of $d = 0.5$. Because group variances are needed for this analysis, assume both the control and intervention groups have a standard deviation of 2.5 on self-esteem. Let us now suppose that a new investigator, skeptical of the original, attempts to replicate the original study using a larger sample ($n = 75$ per group, $SD = 2.25$ in both groups). After obtaining a replication effect size of $d = 0.15$, the skeptic's initial reaction might be that the replication results undermine the original study because the new effect size estimate is both nonsignificant and much smaller than the original estimate. However, before arriving at this conclusion, the skeptic should define a contrast $d_O - d_R$, with contrast coefficients $c_O = +1$ and $c_R = -1$. The skeptic can obtain a confidence interval for this contrast by calculating the variance of each sample $\tilde{d}$ (from Equation 7) and substituting the results into Equation 9. The skeptic finds his resulting confidence interval to be $[-0.17, 0.87]$. Because this confidence interval contains zero, the skeptic cannot conclude with certainty that the replication study effect size is distinct from the original effect size. At first glance, this may seem counterintuitive: A replication with a larger sample size that is nonsignificant and seemingly has a much weaker effect is nevertheless not significantly different from the significant original study. Yet, the width of this confidence interval highlights the difficulty of finding a significant difference between two studies unless both studies have large enough sample sizes to yield a precise interval.

This method can be adapted for other measures of effect. Let us then consider a second example involving a different type of effect size. Suppose a researcher is attempting to replicate a piece of correlational research. In our hypothetical example, an original study with 100 individuals found the correlation between Graduate Record Examination (GRE) scores and academic grades to be .8, even larger than a large relationship as defined by Cohen. However, the researcher thinks that performance in classes is now assessed with more diverse indicators of intelligence, so that the GRE and school grades may be more divergent from one another than when the original study was conducted. To this end, the researcher conducts a replication study in a similar sample of 200 individuals and finds the correlation to be .6. Zou (2007) provides the following formula for a confidence interval for the difference between independent correlations, which can be adapted for replication with the following:

$$upperlimit = r_O - r_R - [(r_O - L_O)^2 + (U_R - r_R)^2]^{1/2} \quad (10)$$

and

$$lowerlimit = r_O - r_R + [(U_O - r_O)^2 + (r_R - L_R)^2]^{1/2} \quad (11)$$

Again, subscripts O and R represent the original and replication studies, respectively. The coefficients L and U represent the lower and upper bounds of the confidence interval for the sample correlation of each study, which can be obtained with the formula:

$$tanh(tanh^{-1}(r) \pm z_{\alpha/2} var[tanh^{-1}(r)]^{1/2}) \quad (12)$$

In the case of a single correlation, r is simply the sample correlation estimate. In these situations, $var[tanh^{-1}(r)]$ can be easily obtained from the following:

$$var[tanh^{-1}(r)] = \frac{[(1 - r_i^2)^2/(n_i - 3)]}{(1 - r_i^2)^2}. \quad (13)$$

$Tanh(x)$, the hyperbolic tangent of x, and $tanh^{-1}(x)$, the inverse hyperbolic tangent of x- often designated arctanh- can be simply calculated in many computer programs.[4] After applying these formulas, the researcher finds that the confidence interval for the difference in correlations is [.08, .31]. Because zero is not included, the researcher can claim that the replication effect size is distinct from that of the original study.

In cases of multiple replications, when a confidence interval is not applicable, it is possible to perform a test of effect size heterogeneity or report a descriptive measure. However, these meta-analytic methods, including the Q statistic (Cochran, 1954) and I^2 (percent of variation due to heterogeneity rather than chance; Higgins et al., 2003), have been criticized for dependency on sample size and imprecision (Ioannidis, Patsopoulos, & Evangelou, 2007).

Goal 6: To Assess Whether Replication is Clearly Consistent With Original Study

The final goal that we recommend researchers consider in planning replication studies is similar to Goal 5, but with the intention of declaring two studies have essentially identical effect sizes. Rather than simply declaring equivalence when the analyses presented in Goal 5 yield a confidence interval containing zero, we emphasize the importance of conducting an equivalence test to indicate that the effect sizes in question are the same, beyond a reasonable doubt. We mirror the logic presented under Goal 2, in that showing that two studies are not significantly different from each other is not the same as showing that they are indeed equivalent. Even recent methodological reports on comparing studies erroneously suggest that a nonsignificant difference in sample effect sizes is evidence that the phenomenon as originally described is real (e.g., Braver et al., 2014).

As may be anticipated, our recommended analysis is an equivalence test on the difference in effect sizes between the original and replication study. This is a simple extension of the analyses of Goals 2 and 5. As in Goal 2, we advise using theory and past research to define a confidence interval for a difference in effect sizes that would be viewed as essentially zero. As in Goal 5, the next step is to construct a confidence interval for the difference in effect sizes between the two studies. We emphasize again that this confidence interval corresponds to two one-sided tests, and thus should be a 90% confidence interval if α of .05 is used. If the confidence interval calculated in Step 2 rests fully inside the confidence interval defined in Step 1, one can conclude that the two effect sizes are homogeneous. In other words, the sample effect

[4] If calculating without the aid of a computer package, $tanh(x)$ has the following formula: $\frac{e^{2x}-1}{e^{2x}+1}$. The formula for $tanh^{-1}(x)$ is: $\frac{1}{2}[ln(1 + x) - ln(1 - x)]$.

sizes are essentially the same. Partially overlapping confidence intervals would not indicate conclusive evidence for homogeneity.

Returning to our self-esteem example one last time, suppose that the researcher instead finds a replication effect size of $d = 0.48$ ($SD = 0.5$ in both groups), which seems potentially indistinct from the original effect size of $d = 0.5$. Prior to conducting the replication study, the researcher determined that a difference of effect sizes less than or equal to 0.15 in either direction would be justifiable as evidence of equivalence. The researcher then calculates the confidence interval for the difference in effect sizes, using linear contrasts and the formula outlined in Goal 5 (assume that the replication sample size and variances are equivalent to those outlined in Goal 5). This results in a confidence interval of [−0.50, 0.54]. Because this confidence interval does not completely fall inside the a priori interval [−0.15, 0.15], the researcher cannot claim that the two effect sizes are the same even though they differ by only 0.02.[5] The large width of the researcher's confidence interval emphasizes the typical requirement of very large sample sizes to have power to establish equivalence, especially when the original study has a small sample size. If this was the researcher's specific goal, the new study would not be successful in meeting that goal.

Conclusion

Clearly, replication is not as simple as it once may have seemed. Part of the reason for the crisis of confidence may be that the field is viewing replication as a black and white dichotomy, wherein statistically significant findings based on analyses mirroring the original study are deemed successful, whereas nonsignificant findings are deemed failures. Instead, replication should be viewed as a construct that can be amenable to varying purposes and flexible in answering the questions that are most beneficial to moving the field forward in the domain of interest.

We posit that there are two major issues preventing the full nature of replication from being explored in psychology. First, researchers often seem to be heavily relying on Goals 1 and 2, without appreciation of what the other goals can offer in terms of replication. However, it can be difficult to determine what authors' motivations are in many cases, so a more objective strategy may be to look at what they did. From this angle, it is clear that authors are very often relying on a single analysis strategy—the significance test—to determine the success of a replication. In other words, authors are failing to utilize more nuanced types of analysis and instead relying on the suggested analysis for Goal 1. Thus, even if authors are aware of Goals 2–6, they are often failing to use the analyses that can provide the answers they seek.

To address this problem, we outlined six goals that can provide varying paths to the dichotomous choice of replication/nonreplication, providing examples for each goal. In the preceding discussion, we hope that it became clear that these goals are not mutually exclusive. We invite researchers to flexibly use these goals to conform to their criterion of interest and to combine them when appropriate. For example, it may be logical to report a confidence interval around the replication effect size (consistent with Goal 3) and to then compare that effect size with a purportedly distinct original effect size (consistent with Goal 5).

Truly, if replication is indeed the "cornerstone of science," then the field must treat both the outcome and the process of replication as such. Attention to the variety of ways in which replication can be defined as well as the proper analyses to address these goals may serve to provide the information necessary to pull the field out of crisis.

[5] Some readers may consider an interval of ± 0.15 to be rather large in defining a difference in effect sizes to be essentially zero; however, as is evident from the width of the observed confidence interval in this example, much larger sample sizes are needed to establish stricter definitions of equivalence.

References

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

Berger, J., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science, 2,* 317–335. http://dx.doi.org/10.1214/ss/1177013238

Biesanz, J. C., & Schrager, S. M. (2010). *Sample size planning with effect size estimates*. Manuscript under review. Vancouver, Canada: University of British Columbia School of Psychology.

Bond, C. F., Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8,* 406–418. http://dx.doi.org/10.1037/1082-989X.8.4.406

Bonett, D. G. (2008). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods, 13,* 173–181. http://dx.doi.org/10.1037/a0012868

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods, 14,* 225–238. http://dx.doi.org/10.1037/a0016619

Bonett, D. G., & Price, R. M. (2014). Meta-analysis methods for risk differences. *British Journal of Mathematical and Statistical Psychology, 67,* 371–387. http://dx.doi.org/10.1111/bmsp.12024

Bonett, D. G., & Wright, T. A. (2007). Comments and recommendations regarding the hypothesis testing controversy. *Journal of Organizational Behavior, 28,* 647–659. http://dx.doi.org/10.1002/job.448

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9,* 333–342. http://dx.doi.org/10.1177/1745691614529796

Center for Open Science. (2012). *Reproducibility project*. Retrieved from https://osf.io/ezcuj/wiki/home/

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10,* 101–129. http://dx.doi.org/10.2307/3001666

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8,* 243–253. http://dx.doi.org/10.1037/1082-989X.8.3.243

Collins, H. M. (1985). *Changing order*. London, UK: SAGE.

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

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33,* 503–513.

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. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

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

Goren, C. C., Sarty, M., & Wu, P. Y. K. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics, 56,* 544–549.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499. http://dx.doi.org/10.1037/0033-2909.92.2.490

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–501. http://dx.doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557–560. http://dx.doi.org/10.1136/bmj.327.7414.557

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton, FL: CRC Press. http://dx.doi.org/10.1201/b11158

Hoijtink, H., & Klugkist, I. (2007). Comparison of hypothesis testing and Bayesian model selection. *Quality & Quantity: International Journal of Methodology, 41,* 73–91. http://dx.doi.org/10.1007/s11135-005-6224-6

Hsu, L. M. (1980). Tests of differences in p levels as tests of differences in effect sizes. *Psychological Bulletin, 88,* 705–708. http://dx.doi.org/10.1037/0033-2909.88.3.705

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., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *British Medical Journal, 335,* 914–916. http://dx.doi.org/10.1136/bmj.39343.408449.80

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

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

Kelley, K., & Lai, K. (2012). *Package "MBESS."* Retrieved from http://cran.r-project.org/web/packages/MBESS/MBESS.pdf

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. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). San Diego, CA: Academic Press.

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

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

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–563. http://dx.doi.org/10.1146/annurev.psych.59.103006.093735

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (in press). Is psychology suffering from a replication crisis? *American Psychologist*.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16,* 406–419. http://dx.doi.org/10.1037/a0024377

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E. J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science, 25,* 1289–1290. http://dx.doi.org/10.1177/0956797614525969

Mullen, B., Muellerleile, P., & Bryant, B. (2001). Cumulative meta-analysis: A consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin, 27,* 1450–1462. http://dx.doi.org/10.1177/01461672012711006

Nordgren, L. F., & McDonnell, M. H. M. (2011). The scope-severity paradox: Why doing more harm is judged to be less harmful. *Social Psychological & Personality Science, 2,* 97–102. http://dx.doi.org/10.1177/1948550610382308

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

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

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior and Personality, 5,* 1–30.

Rosenthal, R., & Rubin, D. B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin, 86,* 1165–1168. http://dx.doi.org/10.1037/0033-2909.86.5.1165

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500–504. http://dx.doi.org/10.1037/0033-2909.92.2.500

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research, 47,* 877–903. http://dx.doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. N. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56,* 356–374. http://dx.doi.org/10.1016/j.jmp.2012.08.001

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

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428. http://dx.doi.org/10.1037/h0042040

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55,* 182–186. http://dx.doi.org/10.1198/000313001317097960

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

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

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

Stukas, A. A., & Cumming, G. (2014). Interpreting effect sizes: Toward a quantitative cumulative social psychology. *European Journal of Social Psychology, 44,* 711–722. http://dx.doi.org/10.1002/ejsp.2019

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,* 1–13.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work. *American Psychologist, 24,* 83–91. http://dx.doi.org/10.1037/h0027108

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

van de Schoot, R., Mulder, J., Hoijtink, H., Van Aken, M. A. G., Dubas, J. S., de Castro, B. O., . . . Romeijn, J. W. (2011). An introduction to Bayesian model selection for evaluating informative hypotheses. *European Journal of Developmental Psychology, 8,* 713–729. http://dx.doi.org/10.1080/17405629.2011.621799

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 632–638. http://dx.doi.org/10.1177/1745691612463078

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine, 26,* 192–196. http://dx.doi.org/10.1007/s11606-010-1513-8

Wegner, D. M., Schneider, D. J., Carter, S. R., III, & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology, 53,* 5–13. http://dx.doi.org/10.1037/0022-3514.53.1.5

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC Press. http://dx.doi.org/10.1201/EBK1439808184

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. http://dx.doi.org/10.1037/0003-066X.54.8.594

Yates, F. (1964). Sir Ronald Fisher and the design of experiments. *Biometrics, 20,* 307–321. http://dx.doi.org/10.2307/2528399

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.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods, 12,* 399–413. http://dx.doi.org/10.1037/1082-989X.12.4.399

Received November 26, 2014
Revision received June 10, 2015
Accepted June 15, 2015 ■