Psychological Science OnlineFirst, published on March 23, 2015 as doi:10.1177/0956797614567341

General Article

Psychological Science
1–11
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797614567341
pss.sagepub.com

Small Telescopes: Detectability and the Evaluation of Replication Results

Uri Simonsohn
The Wharton School, University of Pennsylvania

Abstract

This article introduces a new approach for evaluating replication results. It combines effect-size estimation with hypothesis testing, assessing the extent to which the replication results are consistent with an effect size big enough to have been detectable in the original study. The approach is demonstrated by examining replications of three well-known findings. Its benefits include the following: (a) differentiating "unsuccessful" replication attempts (i.e., studies yielding $p > .05$) that are too noisy from those that actively indicate the effect is undetectably different from zero, (b) "protecting" true findings from underpowered replications, and (c) arriving at intuitively compelling inferences in general and for the revisited replications in particular.

Keywords

replication, statistical power, hypothesis testing, open materials

Received 10/14/14; Revision accepted 12/16/14

Replication studies provide new data. These can be used to answer at least three different questions:

> *Question 1:* When we combine data from the original and replication studies, what is our best guess as to the magnitude of the effect of interest?
>
> *Question 2:* Is the effect of interest in the replication study detectably different in magnitude from what the original study found?
>
> *Question 3:* Does the replication study suggest that the effect of interest is undetectably different from zero?

This article focuses on Question 3 for two main reasons: First, we already have widely accepted meta-analytic tools to answer Questions 1 and 2, and second, it is Question 3 that quite often motivates replication attempts in the first place.

Assessing replicability, rather than increasing precision, is what places replications at the core of the scientific method (Fisher, 1926; Popper, 1935/2005).[1] Assessing the "overall rate of reproducibility [in psychology]" is among the key motivators of the large-scale Reproducibility Project (Open Science Collaboration, 2012), and the desire to identify "findings that are robust, replicable, and generalizable" (Association for Psychological Science, 2014) is one of the principles behind the recent Registered Replication Results initiative in the journal *Perspectives on Psychological Science*.

Only once we are past asking whether a phenomenon exists at all and we come to accept it as qualitatively correct may we become concerned with estimating its magnitude more precisely. Before lines of inquiry arrive at the privileged position of having identified a phenomenon that is generally accepted as qualitatively correct, researchers require tools to help them distinguish between those that are and are not likely to get there.

If an original study found that, under prescribed conditions, people can levitate 9 in. on average, but an attempted replication found 0 in. of levitation, few researchers would be interested in the approximately 4.5-in. meta-analytic average, or in whether the new estimate is significantly different from 9 in.; most would want to know, instead, if the replication convincingly contradicts the qualitative conclusions of the original study that levitation is a detectable phenomenon. Concluding that levitation does not happen requires accepting the

Corresponding Author:
Uri Simonsohn, University of Pennsylvania–The Wharton School, 3730 Walnut St., 500 Huntsman Hall, Philadelphia, PA 19104
E-mail: uws@wharton.upenn.edu

Downloaded from pss.sagepub.com at UNIV OF PENNSYLVANIA on March 24, 2015

hypothesis of no effect. This is not something we can do when we engage in traditional significance testing using zero as the null hypothesis; failing to reject zero is not the same as accepting zero.

An old solution to this problem consists of testing the null hypothesis that the effect is "small," rather than zero, and concluding that it is (basically) zero if we obtain a result that is statistically significantly smaller than that small effect (see, e.g., Cohen, 1988, pp. 16–17; Greenwald, 1975, pp. 16–19; Hodges & Lehmann, 1954; Kraemer, 1983; Rindskopf, 1997; Serlin & Lapsley, 1985, 1993). There are at least two problems with this old solution. First, it requires determining an effect size that is small enough for it to no longer support the theory. Psychological theories are almost exclusively qualitative rather than quantitative—predicting the sign rather than the magnitude of effects—and hence are not well equipped to help us identify when an effect is too small to be of theoretical interest. One way to think of this problem is that because every confidence interval, no matter how tight, includes nonzero values, confidence intervals obtained in all studies end up including values consistent with theoretically interesting effects, whether the studies examine real or nonreal effects.[2]

The second problem with testing the null hypothesis of a small effect is that for such a test to be properly powered, it requires a sample size typically not feasible for psychology experiments. For example, to have 80% power to conclude that $|d|$ is less than 0.1 when the true d is 0 requires about 3,000 observations. These two problems may explain why this 60-year-old idea of testing small effects has not been implemented.

This article introduces a new approach to "accepting zero" that bypasses both problems: It does not require determining an effect size that is too small to be theoretically interesting, nor does it require collecting implausibly large samples of data. In particular, instead of asking if the effect obtained in the replication study is close enough to zero for it to lack theoretical value, we ask if it is close enough to zero that the original study would have been unable to meaningfully study an effect that small.

Imagine an astronomer claiming to have found a new planet with a telescope. Another astronomer tries to replicate the discovery using a larger telescope and finds nothing. Although this does not prove that the planet does not exist, it does nevertheless contradict the original findings, because planets that are observable with the smaller telescope should also be observable with the larger one.

It is generally very difficult to prove that something does not exist; it is considerably easier to show that a tool is inadequate for studying that something. With a small-telescopes approach, instead of arriving at the conclusion that a theoretically interesting effect does not seem to exist, we arrive at the conclusion that the original evidence suggesting a theoretically interesting effect exists does not seem to be adequate.

An attractive feature of the small-telescopes approach is that it combines traditional null-hypothesis significance testing with considerations of effect size, alleviating important limitations of the former (see, e.g., Cohen, 1994; Cumming, 2014). For instance, when we take into account detectability, not all nonsignificant replication findings are created equal. Some arise from effect-size estimates that are small enough and precise enough to be at odds with the claim that an effect is detectable, whereas some do not.

In what follows, I first use results from replications of well-known studies to demonstrate the problems with the two standard approaches for interpreting replication results—that is, with asking whether the effect obtained in the replication study is (a) significantly different from zero or (b) significantly different from the original effect-size estimate. I then introduce the new detectability standard and show how our interpretation of these replication results changes when we use this standard.

Problems With Asking Only Whether the Effect Obtained in the Replication Is Statistically Significant

Despite the notable shortcomings of this approach (see, e.g., Asendorpf et al., 2013; Cumming, 2008; Valentine et al., 2011), effects obtained in replication studies are currently evaluated almost exclusively on the basis of whether or not they are significantly different from zero. As of early 2013, for example, this approach was used by 9 of the 10 "most-seen" replication studies in the PsychFileDrawer Web site, and by all 10 of the most cited psychology articles with "Failure to Replicate" in their title.[3]

Nonsignificant results are indeed expected (with 95% chance) when the effect of interest does not exist, but they are also expected when the effect does exist and its estimate is imprecise. For this reason, we seldom interpret $p > .05$ in an original study as suggesting that an effect does not exist.

Evaluating replications by asking whether the effect obtained is statistically significant can easily lead to inferences opposite to what the evidence warrants. First, underpowered replication attempts of (possibly) true findings *predictably* reduce our confidence in those findings. For example, in Study 1 by Zhong and Liljenquist (2006), subjects who were asked to recall an unethical deed generated more cleanliness-related words (e.g., completing "S–P" as "SOAP") than those who were asked to recall an ethical deed. This original study, with

Downloaded from pss.sagepub.com at UNIV OF PENNSYLVANIA on March 24, 2015

60 subjects, obtained a difference across conditions of $\hat{d} = 0.53$ (i.e., the means differed by 0.53 *SD*); if that were the true effect size, a replication would need about 110 total subjects to have 80% power. Gámez, Díaz, and Marrero (2011) attempted to replicate this finding with 47 subjects; for the original effect size, this experiment had only 44% power. Even if Zhong and Liljenquist (2006) were exactly right not only about the existence of the effect but also about its size, our confidence in their finding would typically drop if we evaluated such an underpowered replication by asking only if the obtained effect was significantly different from zero.

Second, if a replication obtains an estimated effect size with a magnitude similar to that of the original study's, but the effect is sufficiently imprecisely estimated to be nonsignificant, the replication would—despite being entirely consistent with the original—reduce our confidence in the effect. Consider, for example, the tendency of people to ask for more money to sell something than they are willing to pay for it (Knetsch & Sinden, 1984). The most cited demonstration of this *endowment effect* for valuations indicated that median selling prices are about 2.5 times higher than median buying prices (Kahneman, Knetsch, & Thaler, 1990). The most cited "failure" to obtain the endowment effect, the study by Coursey, Hovis, and Schulze (1987), in turn, indicated that median selling prices are about 2.6 times higher than median buying prices.[4]

Third, when a large-sample study "successfully replicates" a small-sample one (i.e., it also obtains an effect at $p < .05$), effect-size differences may be colossal enough that treating the effects in the original and replication studies as consistent with one another can be untenable. For example, in Schwarz and Clore's (1983) classic study (Study 2), a research assistant called college students in Urbana-Champaign, Illinois, to ask about their life satisfaction, which was found to be higher on sunny than on rainy days. In a replication study, Feddersen, Metcalfe, and Wooden (2012) analyzed archival data from an Australian survey that included life-satisfaction questions (N = 96,472) to assess the impact of the weather on the day people were called. They wrote, "despite the difference in magnitude, we do confirm Schwarz and Clore's (1983) finding that cloudiness matters" (p. 6). However, the difference in magnitude that they refer to is quite substantial. Schwarz and Clore's study, with 14 respondents per cell, obtained an effect more than 100 times larger than the effect Feddersen et al. obtained (for calculations, see Supplement 2 in the Supplemental Material available online). For a study with a sample as small as Schwarz and Clore's, an effect as small as that documented by Feddersen et al. is indistinguishable from 0.

Problems With Comparing Effect Sizes

A natural alternative to asking if the effect in a replication study is significantly different from zero is to ask if it is significantly different from the original estimate. This approach has several problems. First, it answers whether the effect of interest is smaller than previously documented (Question 2 in the introduction) rather than whether a detectable effect exists (Question 3). Imagine again that study documenting 9 in. of levitation. It is interesting and possibly important to assess whether a replication in which only 7 in. of levitation was observed obtained an estimate that is significantly different from 9 in., but this is distinct from assessing if levitation is a detectable phenomenon.

A second problem with comparing point estimates is more pragmatic: This standard is simply not very good at identifying false-positive findings. False-positive results tend to be barely significant (e.g., $p > .025$ rather than $p < .025$) and hence have confidence intervals that nearly touch zero (Simonsohn, Nelson, & Simmons, 2014b). The wider the confidence interval around an estimate is, the less likely another estimate is to be statistically significantly different from it. Indeed, if the original finding is a false positive, a replication with the same sample size has less than a 34% chance of obtaining a statistically significantly different estimate (see Supplement 3). The intuition behind this low probability of identifying false positives is that the replication needs to obtain a precise and *opposite-sign* point estimate to differ significantly from the false-positive original, but this does not happen very often if the true effect is zero. It follows that most replications of most false-positive findings will obtain results that are not significantly closer to zero.

One could fix this problem by treating the original estimate as a point null (ignoring its confidence interval). However, this approach further removes us from the question of interest. We act as if the original estimate, which usually contains considerable noise, is a magnitude of intrinsic interest. In addition, because of publication bias, most true effects are overestimated in the published literature (Hedges, 1984; Ioannidis, 2008; Lane & Dunlap, 1978), so most sufficiently powered replication studies of true effects will obtain statistically significantly smaller estimates than the originals.

Detectability: A New Approach to Accepting Zero

The most widely used metric for detectability is statistical power: the probability that a study will obtain a statistically significant effect given a particular combination of effect size and sample size. In Simonsohn et al. (2014b),

Downloaded from pss.sagepub.com at UNIV OF PENNSYLVANIA on March 24, 2015

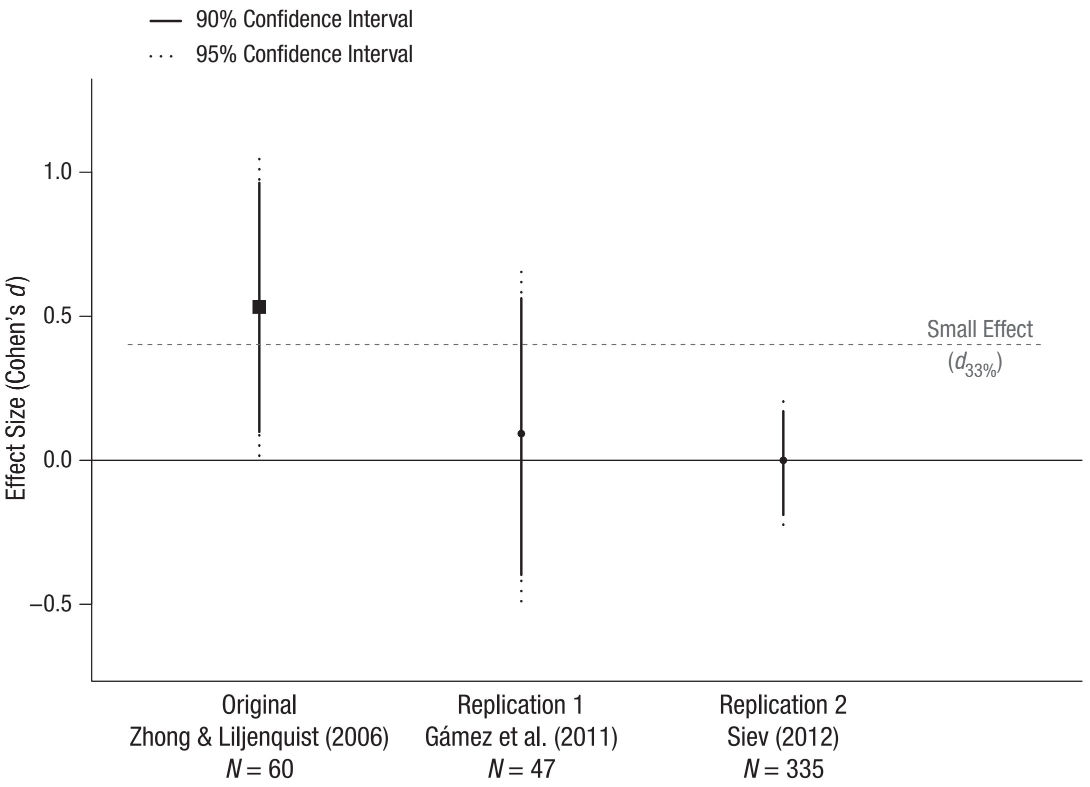

Fig. 1. Results from Zhong and Liljenquist's (2006) Study 1 and two replication studies (Gámez, Díaz, & Marrero, 2011; Siev, 2012). The markers indicate effect-size estimates, and the vertical bars their confidence intervals. The dashed line indicates the effect size ($d_{33\%}$ = 0.401) that would give the original study, with a sample size of 30 per cell, 33% power. See Supplement 1 in the Supplemental Material for the calculations behind this figure.

we considered studies powered below 33%, those in which the odds are at least 2:1 against obtaining a statistically significant effect, as being severely underpowered. For expository purposes, I use that same reference point here and define a small effect as one that would give 33% power to the original study; I refer to this effect size as $d_{33\%}$. A replication that obtains an effect size that is statistically significantly smaller than $d_{33\%}$ is inconsistent with the notion that the studied effect is large enough to have been detectable with the original sample size.[5]

The approach works as follows. Consider an original study with a two-cell between-subjects design and 20 observations per cell. We first ask: What effect size would give this study 33% power? The answer is, an effect size of 0.5, so $d_{33\%}$ = 0.5. We then use the estimate from the replication, $\hat{d}$, to test the null hypothesis that d = 0.5 against the one-sided alternative that d < 0.5. If we reject the null hypothesis, we conclude that the original telescope was too small.

Although d = 0.5 is generally considered a moderate, rather than small, effect size (Cohen, 1988) and may be large from a theoretical perspective for most domains, an effect of this size is not easily detectable with samples of 20 per cell, and an effect of d < 0.5 is even less detectable. So we may very well care about an effect with d = 0.5, but the point here is that the original study could not have meaningfully examined such an effect.

We can apply the detectability approach without relying on $d_{33\%}$ in particular or hypothesis testing more generally. Instead of testing whether the effect size obtained in the replication is significantly smaller than $d_{33\%}$, we can express the confidence interval for that effect size in terms of the statistical power it confers to the original study. For instance, imagine an original study with 20 observations per cell. If a replication study obtains a point estimate $\hat{d}$ = 0.1 with confidence interval [−0.1, 0.3], the point estimate implies that the original study was powered to a meager 6.1%, and the confidence interval for its power goes up only to 15%.

How Things Change When We Consider Detectability

It seems likely that many replications that have been considered failures in the past, because they did not reject the null hypothesis of a zero effect, involved uninformatively imprecise confidence intervals for the underlying effect sizes. For example, Figure 1 shows that the

Downloaded from pss.sagepub.com at UNIV OF PENNSYLVANIA on March 24, 2015

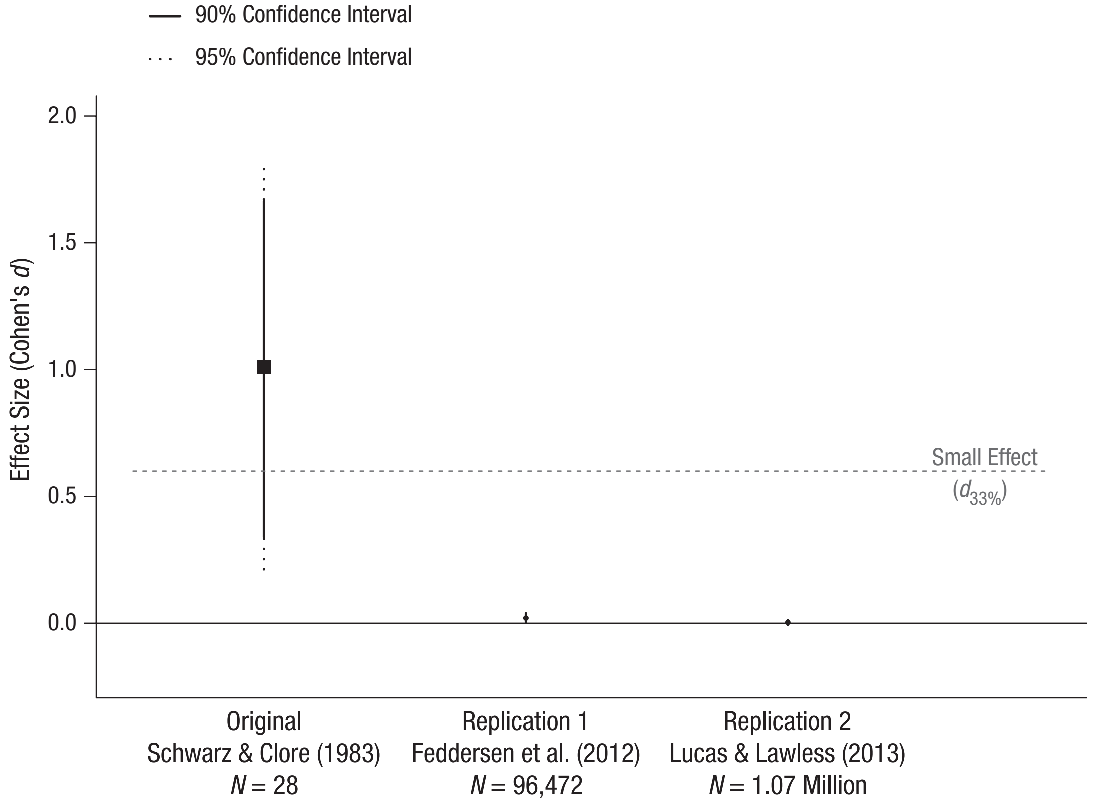

Fig. 2. Results from Schwarz and Clore's (1983) Study 2 and two replications of it (Feddersen, Metcalfe, & Wooden, 2012; Lucas & Lawless, 2013). The markers indicate effect-size estimates, and the vertical bars their confidence intervals. The dashed line indicates the effect size that would give the original study, with a sample size of 14 per cell, 33% power. See Supplement 2 in the Supplemental Material for the calculations behind this figure.

confidence interval in the replication of Zhong and Liljenquist's (2006) Study 1 by Gámez et al. (2011) included not only 0 but also $d_{33\%}$. The one-sided test for the effect being smaller than $d_{33\%}$ is not rejected, $p = .14$. The 90% confidence interval for d in this replication had 0.562 as its upper end; an effect this big confers the original study 57% power. (See Supplement 1 for calculations.) The "failed" replication, in other words, is consistent with an effect size that would have been detectably different from zero with the original sample.

Figure 1 contrasts this uninformatively imprecise replication result with the small and precisely estimated effect size obtained by Siev (2012). Its confidence interval included 0 but not $d_{33\%}$ (rejecting the null hypothesis that the true effect is $d_{33\%}$ with $p < .0001$). The biggest effect size within the 90% confidence interval, 0.204, confers the original study just 9.9% power. (See Supplement 1 for calculations.)

An interesting case is when a replication $\hat{d}$ is significantly different from 0 in the same direction as in the original study, but also significantly smaller than $d_{33\%}$. The sign of the effect has been replicated, but the notion that the original study meaningfully informs our understanding of the phenomenon of interest is inconsistent with the new data.

For example, Figure 2 shows that although Feddersen et al. (2012) obtained a statistically significant effect of sunshine on reported life satisfaction (though only after controlling for, among other things, respondent fixed effects), the effect was so small that the hypothesis that an effect was remotely detectable in the original study (Schwarz & Clore, 1983), with 14 respondents per cell, is firmly rejected. The planet Feddersen et al. reported seeing is not a planet Schwarz and Clore could have seen. Figure 2 also shows that Lucas and Lawless (2013) obtained an informative failure to replicate the effect of sunshine on life satisfaction in a professional survey of Americans with a sample size about 50,000 times the original. (See Supplement 2 for calculations.) Table 1 summarizes the contrasts among the different approaches for evaluating replication results.

How Many Observations for a Replication?

The status quo

Currently, replication studies are often powered on the basis of the effect size obtained in the original study. For example, researchers may set the sample size for a

Downloaded from pss.sagepub.com at UNIV OF PENNSYLVANIA on March 24, 2015

Table 1. Comparison of Approaches for Evaluating Replication Results

Question that the existing approach focuses on	What happens when the approach is used	What happens when the detectability (small-telescopes) approach is used
Is the estimated effect in the replication statistically significant, $p < .05$?	1. Underpowered replications fail to replicate original results and reduce confidence in true findings. 2. A replication obtaining the same effect size as the original study can be deemed a replication failure. 3. A replication obtaining an effect size colossally smaller than the original effect size is deemed a successful replication.	1. Underpowered replications are deemed uninformative and do not influence confidence in true findings. 2. A replication obtaining the same effect size as the original study is never deemed a replication failure. 3. If a replication obtains an effect size colossally smaller than the original effect size, the original study is deemed inadequate.
Is the estimated effect in the replication significantly smaller than the original finding?	1. The replication does not inform the question of interest (i.e., whether a detectable effect exists). 2. Most false-positive findings survive most replication attempts.	1. The replication does inform the question of interest. 2. Replications of studies with false-positive findings have an 80% chance of being informative replication failures.
Is the estimated effect in the replication significantly smaller than the original finding if we ignore its sampling error?	1. The replication does not inform the question of interest (i.e., whether a detectable effect exists). 2. Large-sample replications fail to replicate most true findings.	1. The replication does inform the question of interest. 2. Large-sample replications successfully replicate most true findings.

replication to have 80% power to detect an effect as big as that originally documented. There are at least two serious problems with this approach.

First, as mentioned earlier, because of publication bias, published effect-size estimates are greatly inflated (see, e.g., Fig. 2 in Simonsohn, Nelson, & Simmons, 2014a). This means that even when original findings are true, powering replications on the basis of published effect sizes leads to much lower power than intended.

Figure 3 reports some basic calibrations of the statistical power that replications actually obtain when sample sizes are set on the basis of observed effect sizes, assuming that original research is published only if the predicted effect reaches $p < .05$. The lower the power of the

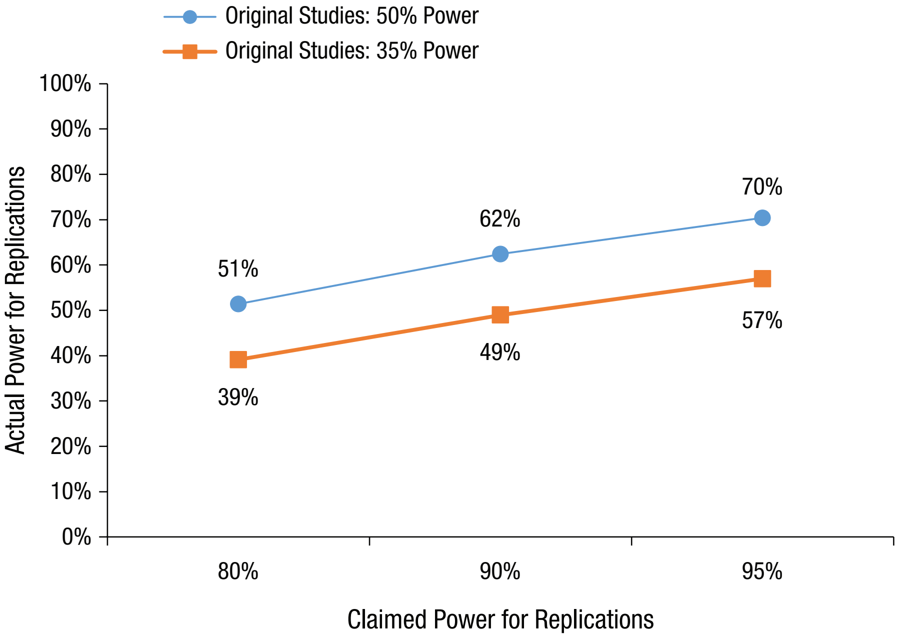

Fig. 3. Actual versus claimed power in replications when sample size is based on the effect size observed in the original study. The figure reports statistical power in replications with sample sizes that, given the effect size observed in the original study, would obtain 80%, 90%, and 95% power, when original results are observable only if $p < .05$ and hence overestimated on average. See Supplement 5 in the Supplemental Material for the calculations behind this figure.

Downloaded from pss.sagepub.com at UNIV OF PENNSYLVANIA on March 24, 2015

original research, the more this subset of results overestimates effect size, and hence the more the claimed power of replications is inflated. If the original study is powered to 50%, an optimistic scenario (Button et al., 2013; Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1989), the resulting upward bias in effect size would be severe enough that replications with samples sizes set to obtain 80% power, given the reported effects, would on average have achieved only 51% power.

Second, if a replication is powered on the basis of the observed effect size, only the power of the "is the replication significant?" test has been considered, and as I have shown, this test leads to untenable inferences. If we power replications this way, too many will obtain confidence intervals that are too wide, including both zero and detectable effect sizes. A power calculation based on a reported effect size, then, lacks face validity and is too likely to lead to uninformative results.

Replication sample size = 2.5 × original sample size

The sample size for original research is supposed to be set so that it leads to a reasonable level of statistical power, typically proposed to be around 80%. To determine the power associated with a given sample size for a study, we need to determine what question it is that we want to answer with the study—what statistical test we plan on conducting. For example, a 2 × 2 study needs at least twice as many subjects per cell if we want to test an attenuated interaction as opposed to a simple effect (Simonsohn, 2014).

When setting the sample size for a replication, we could focus on the statistical power for at least two different tests. We could set the sample size to have enough power to reject the traditional null hypothesis of zero effect, assuming that the true effect is larger than zero. Alternatively, we could set the sample size to have enough power to reject the null hypothesis of a detectable effect (e.g., $d_{33\%}$), assuming that the true effect is zero.

The former objective, detecting an effect, is analogous to the objective of original research; hence, whatever approach researchers take to set sample size for original research would apply to replications with this objective. For the latter objective, accepting the null hypothesis, we ask the following question: "If the true effect size is 0, how many observations do we need to have an 80% chance of concluding that the effect is undetectably small?"

The answer to this question turns out to be very simple and does not require consulting power tables or software like gPower. Whether the statistical test is based on the normal, Student, F, or χ^2 distribution; whether it is conducted with a between- or within-subjects design; and whether the data are from an experiment or an observational study, if the true effect is zero, a replication needs 2.5 times as many observations as the original study to have about 80% power to reject $d_{33\%}$. For example, if the true effect is zero, and an original study had 20 observations per cell, a replication with 50 observations per cell would have about 80% power to reject the hypothesis that the effect is $d_{33\%}$ (see Fig. 4).

What about very large samples?

Sometimes studies have very large sample sizes, for example, a few hundred thousand Implicit Association Test takers (Nosek, Banaji, & Greenwald, 2002), a few million Danes (Dahl, Dezső, & Ross, 2012), or several million Facebook users (Bond et al., 2012). Would one really need 2.5 times those sample sizes to properly power replication attempts?

First, it is important to establish that large-sample experiments are rather rare in psychology. The median sample size for studies published in *Psychological Science* between 2003 and 2010 was about 20 per cell, and about 90% of these studies had a per-cell sample size smaller than 150.[6]

Second, whether we need 2.5 times the original sample size or not depends on the question we wish to answer. If we are interested in testing whether the effect size is smaller than $d_{33\%}$, then, yes, we need about 2.5 times the original sample size no matter how big that original sample was. When samples are very large, however, that may not be the question of interest. If the original study had 100,000 subjects per cell, 33% power is obtained with an exceedingly small effect size—$d_{33\%}$ = 0.007. Under these circumstances, we may be able to accept the null hypothesis in the "old-fashioned" way—by showing that the effect is small enough to be lacking in theoretical interest.

For example, if the researcher running the replication deems effects with $|d| < 0.1$ as negligible, to have 80% power to reject $d = 0.1$ when $d = 0$, the researcher needs about 3,000 observations total. No matter how big the original study was, if as replicators we are content accepting the null when we conclude that $d < .1$, we need "only" an n of 1,500 per cell. To be clear, in those cases, the replication could have fewer observations than the original study.

To further examine the impact of sample size on the interpretation of replications, consider Figure 5, which depicts results of hypothetical original studies, all obtaining effect-size estimates of $\hat{d} = 0.6$, with different sample sizes. The figure also depicts two hypothetical replication results.[7] Let us begin with the first replication, which obtained a somewhat imprecisely estimated zero effect.

Downloaded from pss.sagepub.com at UNIV OF PENNSYLVANIA on March 24, 2015

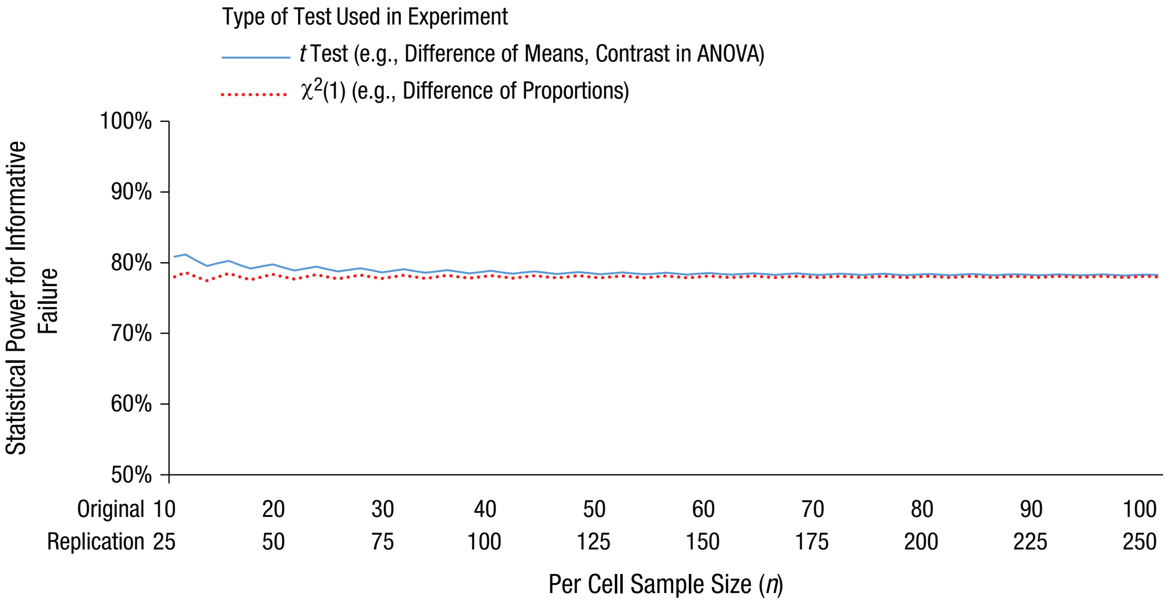

Fig. 4. Power of a replication to reject the hypothesis of a detectable effect ($d_{33\%}$), given that the true effect is 0, when the sample size of the replication is 2.5 times the sample size of the original study. Power results are shown for both *t* tests and χ^2 tests, for original samples ranging from 10 to 100. See Supplement 4 in the Supplemental Material for the calculations behind this figure. ANOVA = analysis of variance.

This replication rejects the null hypothesis of $d_{33\%}$ for the original studies with *n*s of 20, 50, and 100, deeming such samples small telescopes. The same replication, however, does not reject $d_{33\%}$ for the original studies with larger samples. This highlights that (a) the detectability approach evaluates the original study's *design* rather than its result—the telescope rather than the planet—and (b) when samples are large enough, it is informative to accompany the small-telescopes test with a simple test comparing the effect sizes obtained in the original and

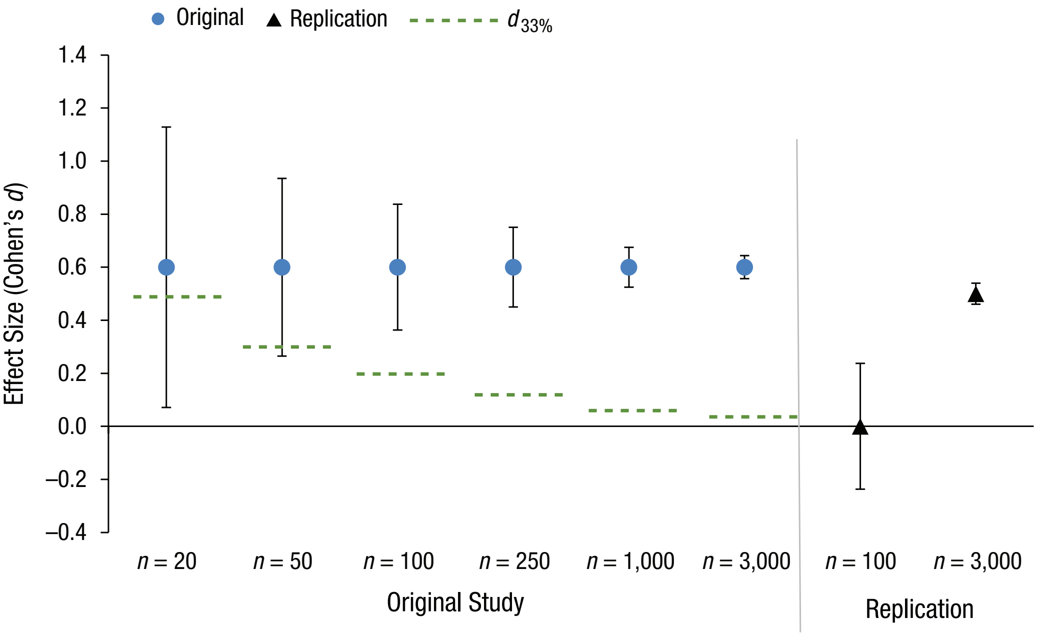

Fig. 5. Relations among sample size, confidence intervals, and $d_{33\%}$. All studies are hypothetical. Vertical lines correspond to 90% confidence intervals.

Downloaded from pss.sagepub.com at UNIV OF PENNSYLVANIA on March 24, 2015

replication studies. The second replication depicted in the figure, which obtained a precisely estimated $\hat{d} = 0.5$, reminds us why this simple comparison of effect sizes alone, without considerations of detectability, can be misleading. The replication's $\hat{d} = 0.5$ with n of 3,000 is significantly smaller than the original $\hat{d} = 0.6$ with n of 3,000, but for almost any psychological theory, these results are entirely consistent with one another.

What About Bayesian Analyses?

Bayesian statistics offer two paths to accepting the null hypothesis. One involves accepting it when the (Bayesian) confidence interval excludes small effects. When applied to simple designs (e.g., comparing means across conditions), using flat priors, Bayesian confidence intervals are numerically equivalent to traditional ones, so this approach is equivalent to the old frequentist approach for accepting the null. Although the interpretation of the results is different, the results themselves are not (Lindley, 1965, pp. 76–79; Rouanet, 1996, pp. 150–151).

The other path is Bayesian hypothesis testing (Jeffreys, 1961), which assesses whether the observed data are more consistent with an effect of zero or with some alternative hypothesis. The default Bayesian t test (Rouder, Speckman, Sun, Morey, & Iverson, 2009) sets as the alternative hypothesis that the effect has the standard normal distribution. When a default Bayesian test favors the null hypothesis, the correct interpretation of the result is that the data favor the null hypothesis more than that one specific alternative hypothesis. The Bayesian test could conclude *against* the same null hypothesis, using the same data, if a different alternative hypothesis were used, say, that the effect is distributed normal but with variance of 0.5 instead of 1, or that the distribution is skewed or has some other mean value.

Verhagen and Wagenmakers (2014) proposed using the posterior from the original study (starting from uniform priors) as the alternative hypothesis for replications. This is conceptually and mathematically similar to testing whether the effect size obtained in the replication is closer to the effect size in the original study than it is to zero.

What About Meta-Analyses?

The detectability approach proposed here for assessing replications complements rather than replaces meta-analysis. First, it answers a different question (Question 3 instead of Question 1). When conducting meta-analyses, we take all results at face value and combine them. When conducting a replication, in contrast, we are typically asking *whether* we should take the original study at face value.

Second, meta-analysis necessitates the accumulation of individual studies, and there will often be an interest in assessing each of them as they are completed. Each replication has a replicator who wonders, upon analyzing the data, if it would be warranted to accept the null hypothesis of zero effect.

Third, meta-analytic results can be interpreted by focusing on detectability. For example, Alogna et al. (2014) meta-analyzed 31 simultaneous replication attempts of Study 4 by Schooler and Engstler-Schooler (1990), obtaining an overall effect-size estimate of −4%, with a confidence interval of [−7%, −1%]. The original Study 4 had 75 subjects, and therefore too small a telescope to meaningfully inform conclusions regarding an effect as small as the replications suggest exists. Even the end of the confidence interval, a 7% difference, indicates that the original sample had less than 10% power.[8] As a whole, the replications were consistent, then, with the theoretical development of the original report, but not with its original design having been adequate. The theory was right, but the evidence was almost pure noise.

Interpreting Failures to Replicate Original Effects

When a replication indicates that for the original sample size, an effect is undetectably different from zero, the original finding has not been annulled, nor shown to be a false positive, p-hacked, or fraudulently obtained. What has been determined is that sampling error alone is an unlikely explanation for why the replication resulted in such a categorically smaller effect size. An effect detectable by the original study should have resulted in a larger estimate in the replication.

A replication failure may arise because the true effect studied in the replication is different from the true effect studied in the original study. Just as it is not possible to step twice into the same river, it is not possible to run the same study twice. Differences in materials, populations, and measures may lead to differences in the true effect under study.

Statistical techniques help us identify situations in which something other than chance has occurred. Human judgment, ingenuity, and expertise are needed to know what has occurred instead.

Author Contributions

U. Simonsohn is the sole author of this article and is responsible for its content.

Acknowledgments

Joe Simmons and Leif Nelson provided invaluable suggestions and encouragement throughout the execution of this project.

Downloaded from pss.sagepub.com at UNIV OF PENNSYLVANIA on March 24, 2015

The telescope analogy for sample size, and the levitation analogy for qualitative effects studied in psychology, are due to Leif Nelson as well. Geoff Cumming, Brian Nosek, and E.-J. Wagenmakers provided extraordinarily valuable signed reviews. Danny Kahneman, Ken Kelley, Deborah Mayo, Hal Pashler, and Norbert Schwarz provided valuable suggestions.

Declaration of Conflicting Interests

The author declared that he had no conflicts of interest with respect to his authorship or the publication of this article.

Supplemental Material

Additional supporting information can be found at http://pss.sagepub.com/content/by/supplemental-data and also at https://osf.io/adweh/files

Open Practices

Data files and R programs used to generate all figures in this article have been made publicly available via Open Science Framework and can be accessed at https://osf.io/adweh/files/. The complete Open Practices Disclosure for this article can be found at http://pss.sagepub.com/content/by/supplemental-data. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at https://osf.io/tvyxz/wiki/view/ and http://pss.sagepub.com/content/25/1/3.full.

Notes

[1] Popper (1935/2005) wrote, "The scientifically significant physical effect may be defined as that which can be regularly reproduced by anyone who carries out the appropriate experiment in the way prescribed" (pp. 23–24). Fisher (1926) wrote, "A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance [referring to $p < .05$]" (p. 504).

[2] This reasoning assumes that the true effect is zero or of the predicted sign. When effects predicted by a theory and true effects have opposite signs, we do reliably obtain confidence intervals that reject theoretically interesting values.

[3] I searched for the term "Failure to Replicate" in the titles of articles indexed on the Web-of-Science (https://webofknowledge.com), sorted these articles by number of citations, and then examined the first 10 articles listed to see how the conclusions that the effect was not replicated were substantiated.

[4] This ratio of median valuations was not reported in the published article on this study, but Kahneman, Knetsch, and Thaler (1990, p. 1336, footnote 5) indicated that an unpublished version of the article did report this result.

[5] Geoff Cumming, when providing feedback on an earlier version of this article, intuited that for the normal distribution, $d_{33\%}$ corresponds to 61% of the length of one arm of the 95% confidence interval. I verified that this is indeed the case and that for sample sizes with 10 or more observations per cell, this is a very close approximation for the Student distribution also. This rule of thumb can prove useful for evaluating replications for which confidence intervals, but not $d_{33\%}$s, are reported.

[6] These numbers are based on degrees of freedom reported for t tests (N = 4,275; see Supplement 6 in the Supplemental Material for details).

[7] Figure 5 is based on a suggestion by E.-J. Wagenmakers.

[8] In the original Study 4, on average across the control and verbal conditions, 59.85% of subjects shown a lineup correctly identified the suspect they had seen in a video. To compute the statistical power implied by a 7% difference between conditions, I split this difference evenly between the two cells. The resulting accuracies of 56.35% versus 63.35% give a two-sample proportions test (with $n_1 = 38$ and $n_2 = 37$) 9.4% power.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., . . . Zwaan, R. A. (2014). Registered Replication Report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*, 556–578.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119.

Association for Psychological Science. (2014). *Registered Replication Reports*. Retrieved from http://web.archive.org/web/20140623042346/http://www.psychologicalscience.org/index.php/replication

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*, 295–298.

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

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Coursey, D., Hovis, J., & Schulze, W. (1987). The disparity between willingness to accept and willingness to pay measures of value. *Quarterly Journal of Economics*, *102*, 679–690.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Dahl, M. S., Dezső, C. L., & Ross, D. G. (2012). Fatherhood and managerial style: How a male CEO's children affect the wages of his employees. *Administrative Science Quarterly*, *57*, 669–693.

Feddersen, J., Metcalfe, R., & Wooden, M. (2012). *Subjective well-being: Weather matters; climate doesn't* (Melbourne

Downloaded from pss.sagepub.com at UNIV OF PENNSYLVANIA on March 24, 2015

Institute Working Paper Series, 25/2012). Melbourne, Victoria, Australia: University of Melbourne. Retrieved from http://web.archive.org/web/20150107020727/http://melbourneinstitute.com/downloads/working_paper_series/wp2012n25.pdf

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

Gámez, E., Díaz, J. M., & Marrero, H. (2011). The uncertain universality of the Macbeth effect with a Spanish sample. *The Spanish Journal of Psychology*, *14*, 156–162.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics*, *9*, 61–85.

Hodges, J., & Lehmann, E. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *16*, 261–268.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–646.

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

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, *98*, 1325–1348.

Knetsch, J. L., & Sinden, J. A. (1984). Willingness to pay and compensation demanded: Experimental evidence of an unexpected disparity in measures of value. *Quarterly Journal of Economics*, *99*, 507–521.

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational and Behavioral Statistics*, *8*, 93–101.

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.

Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian viewpoint* (Vol. 2). Cambridge, England: Cambridge University Press.

Lucas, R. E., & Lawless, N. M. (2013). Does life seem better on a sunny day? Examining the association between daily weather conditions and life satisfaction judgments. *Journal of Personality and Social Psychology*, *104*, 872–884.

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*, 101–115.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.

Popper, K. R. (2005). *The logic of scientific discovery* [Taylor & Francis e-Library edition]. Retrieved from http://web.archive.org/web/20150218163435/http://strangebeautiful.com/other-texts/popper-logic-scientific-discovery.pdf (Original work published 1935)

Rindskopf, D. M. (1997). Testing 'small', not null, hypotheses: Classical and Bayesian approaches. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests* (pp. 319–332). Mahwah, NJ: Erlbaum.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*, 646–656.

Rouanet, H. (1996). Bayesian methods for assessing importance of effects. *Psychological Bulletin*, *119*, 149–158.

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.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*, 36–71.

Schwarz, N., & Clore, G. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*, 513–523.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 73–83.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Mahwah, NJ: Erlbaum.

Siev, J. (2012). [Attempt to replicate Zhong & Liljenquist]. Unpublished raw data.

Simonsohn, U. (2014). *[17] No-way interactions*. Retrieved from http://web.archive.org/web/20150206205257/http://datacolada.org/2014/03/12/17-no-way-interactions-2/

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

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

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., . . . Schinke, S. P. (2011). Replication in prevention science. *Prevention Science*, *12*, 103–117.

Verhagen, J., & Wagenmakers, E.-J. (2014). A Bayesian test to quantify the success or failure of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.

Zhong, C.-B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*, 1451–1452.

Downloaded from pss.sagepub.com at UNIV OF PENNSYLVANIA on March 24, 2015