aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2014, Vol. 9(3) 305–318
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691614528518
pps.sagepub.com
SAGE

Expectations for Replications: Are Yours Realistic?

David J. Stanley and Jeffrey R. Spence
University of Guelph

Abstract

Failures to replicate published psychological research findings have contributed to a "crisis of confidence." Several reasons for these failures have been proposed, the most notable being questionable research practices and data fraud. We examine replication from a different perspective and illustrate that current intuitive expectations for replication are unreasonable. We used computer simulations to create thousands of ideal replications, with the same participants, wherein the only difference across replications was random measurement error. In the first set of simulations, study results differed substantially across replications as a result of measurement error alone. This raises questions about how researchers should interpret failed replication attempts, given the large impact that even modest amounts of measurement error can have on observed associations. In the second set of simulations, we illustrated the difficulties that researchers face when trying to interpret and replicate a published finding. We also assessed the relative importance of both sampling error and measurement error in producing variability in replications. Conventionally, replication attempts are viewed through the lens of verifying or falsifying published findings. We suggest that this is a flawed perspective and that researchers should adjust their expectations concerning replications and shift to a meta-analytic mind-set.

Keywords

replication, reliability, meta-analysis, individual differences, methodology, scientific

How should failed replication attempts be interpreted? Imagine a scenario in which a researcher publishes a finding that eating fruit is associated with increased sleep quality. A second researcher tries to replicate the published result and finds no such relation. Undeterred, the second researcher tries again, but still fails to find an effect. What are we to conclude about the published relation between eating fruit and sleep quality? Many researchers intuitively expect that an observed relation can be verified by a subsequent study. As a result of a failed replication attempt, one could conclude that (a) the original researcher got it wrong, (b) the original researcher was fraudulent or incompetent, or (c) the replicating researcher was wrong twice. The correct interpretation of replication failures is crucial for the legitimacy of research because the ability to consistently reproduce research findings is central to credible science. Of late, the credibility of research has been questioned in response to a growing awareness of an inability to reproduce published findings in several disciplines, including medical research (Ioannidis, 2005), neuroscience (Button et al., 2013), and psychology (Asendorpf et al., 2013; Yong, 2012).

The Replication Crisis

In psychology, recent concerns regarding replication are linked to what has been referred to as a "crisis of confidence" (Pashler & Wagenmakers, 2012). Although the origin of the "crisis" is multifaceted, it was largely caused by controversies such as the Diederick Stapel, Dirk Smeesters, and Lawrence Sanna fraud cases (Funder & Task Force on Publication and Research Practices, Society for Personality and Social Psychology, 2014; Pashler & Wagenmakers, 2012; Stroebe, Postmes, & Spears, 2012); suspiciously high rates of hypothesis confirmation (Fanelli, 2012); and widespread questionable research practices (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). Moreover, recent data indicate that replications are infrequently published and that, when they are, they have success rates a little over 1%

Corresponding Author:
David J. Stanley, Department of Psychology, University of Guelph,
50 Stone Rd. East, Guelph, Ontario N1G 2W1, Canada
E-mail: dstanley@uoguelph.ca

(Makel, Plucker, & Hegarty, 2012). One hypothesis that has emerged from these recent circumstances is that, if there is chicanery in research, a symptom of empirical trickery should be an inability to replicate findings. Or, viewed from another perspective, if we see that there is an inability to replicate research, there may be a problem with the credibility of the researchers, their research practices, or both.

Psychologists are not taking these concerns lightly, and recent initiatives such as the Reproducibility Project are specifically aimed at investigating replicability in psychological research. On the main page of its Web site, the Reproducibility Project is identified as a "crowdsourced empirical effort to estimate the reproducibility of a sample of studies from scientific literature" (Reproducibility Project, 2013). The general aims of the project are to identify the rate of replication and factors that contribute to the success or failure of replications in select psychology journals.

The Reproducibility Project is an attempt to accumulate multiple, well-powered replications and to draw inferences about the results in the aggregate. However, as the issue of direct replications has come to the forefront of the current conversation about best practices, researchers are increasingly trying to replicate their own and other people's results using small numbers of studies or, sometimes, single studies. It is therefore important to understand both what we can and what we cannot learn from small-scale replications so researchers can make replication studies maximally informative.

Intuitively, many researchers may think that two identical studies should produce very similar results. However, both researchers and consumers of research are faced with an inescapable challenge when interpreting replication attempts: There is no such thing as the perfect study (Hunter & Schmidt, 2004). That is to say, all studies contain sources of error, or *artifacts*, such as sampling error, measurement error, range restriction, artificial dichotomization of continuous variables, and imperfect validity, which can create alternative explanations for results (Hunter & Schmidt, 2004). In any given study, these artifacts are expected to produce fluctuations in data that are not due to the phenomena under investigation. As a result, researchers cannot be certain if an effect that is observed, or not observed, is a reflection of artifacts or the underlying truth. Complicating the situation further, these fluctuations are separate from instances in which phenomena or relations change across time, contexts, and samples as a result of factors such as culture or history effects (e.g., Bonanno & Jost, 2006; Markus & Kitayama, 1991; Schwartz & Rubel, 2005).

In the context of sampling error and measurement error, the assumption that methodologically identical studies will produce consistent results if an effect is "real" is not accurate. Sampling error and measurement error are two artifacts that are present in *every* study, and both can produce fluctuations in estimates (Hunter & Schmidt, 2004). Of these two artifacts, sampling error is the most publicized in the context of replication (e.g., Asendorpf et al., 2013; Button et al., 2013; Funder et al., 2014; Tressoldi, 2012). Of interest to us in the current article is the influence of the lesser-known and often ignored artifact known as measurement error (or its inverse, score reliability). Measurement error is universal in the social sciences and can be characterized as random variation in study observations (Nunnally & Bernstein, 1994). Measurement error is typically indexed with its inverse, reliability, which indicates how much stability or consistency is present in data. Given that replication is ostensibly the repeatability of research findings and that reliability is conceptualized as being akin to consistency and stability (Nunnally & Bernstein, 1994), the connection between replication and score reliability would appear to be natural. In fact, replication discussions often characterize individual studies or research as a whole as being "reliable" or "unreliable" (e.g., Button et al., 2013; Tressoldi, 2012). However, misconceptions about the effects of unreliability on observed associations can make the importance of reliability in replications unclear. As Cunningham (1986) stated, "Despite its being relatively easy to compute, there are few subjects in the field of measurement more difficult to understand than reliability" (p. 101).

Overview

In the current article, our goal is to highlight the importance of measurement error in replication discussions by isolating its effect on replication attempts. Drawing on recent work on the effect of unreliability on observed relations (Charles, 2005; Zimmerman, 2007), we present a series of data simulations to illustrate that unreliability can have large and irregular effects on observed relations that can greatly influence the success of replication attempts. In doing so, we illustrate that low rates of replication success may reasonably be attributed to measurement error alone. By extension, we suggest that questions surrounding researcher integrity and poor research practices may not be warranted in the context of measurement error, a documented, ubiquitous, and expected characteristic of data.

We examine the influence of measurement error on replications from three perspectives. First, we take an *omniscient perspective*, in which the *true correlation* (i.e., the correlation between variables measured without error) is known. Second, we take a *realistic perspective*, in which the true correlation is unknown and a researcher attempts to replicate a correlation from a published study in the presence of measurement error. And third, we extend the realistic perspective and the difficulties associated with interpreting and replicating a published finding by including the effects of both measurement error and sampling error.

In the examples we present below, we use the correlation coefficient as our index of effect size. Two common measures of effect size are the correlation coefficient *r* and Cohen's *d*. Although the correlation coefficient tends to be used with correlational studies and Cohen's *d* tends to be used with experimental studies, this is mostly a matter of convention. Coefficient *r* and Cohen's *d* simply use different units to express effect size. This is illustrated by the fact that a formula can be used to convert from one to the other (see Borenstein, Hedges, Higgins, & Rothstein, 2009). Consequently, the results and conclusions we present can be extended to experimental studies and are not limited only to observational studies. For those who are more acquainted with Cohen's *d* than coefficient *r*, the following translation will be helpful for interpreting our examples: $d = .63$ is equivalent to $r = .30$ (using equations in Borenstein et al., 2009).

Omniscient Perspective on Replication: The True Correlation Is Known

Consider a scenario in which a researcher, Jane, is interested in the relation between age (*X*) and psychological well-being (*Y*).[1] Jane is an upstanding researcher and does not engage in any questionable or unethical research practices. To conduct her study, Jane collects data from 40 participants and computes a correlation between age and well-being (i.e., an *observed correlation*). From our omniscient perch, we know that the true correlation between age and psychological well-being is $r = .30$.[2] However, because Jane lives in the real world, she is unaware of the true correlation and must estimate it from the data she collects. Additionally, because Jane lives in the real world, she must also contend with measurement error. Given the objective nature of age, Jane is able to measure it without error, by using birth certificates to indicate date of birth. In contrast, psychological well-being is measured with a self-report scale, and scores contain random measurement error. In this example, the well-being data has a reliability of .70, a value considered by many to be an acceptable reliability cut-off (see Lance, Butts, & Michels, 2006, for a review).[3]

Given these parameters—a sample size of 40, age measured with perfect reliability, psychological well-being measured with a reliability of .70, and a true correlation of $r_{\text{true}} = .30$—what correlation is Jane expected to observe?

A "best-guess" perspective on Jane's scenario

One approach to determining the correlation that Jane should observe is to use the Spearman attenuation formula (Spearman, 1904, 1910). This formula provides an estimate of the observed correlation based on a true correlation and known levels of reliability (Spearman, 1904, 1910). An application of this formula, where we know the true correlation, indicates that Jane should observe a correlation of $r = .25$. The formula is presented below in Equation 1:

$$\begin{aligned}\text{observed } r_{xy} &= \text{true } r_{xy} \times \sqrt{r_{yy}}\\ &= .30 \times \sqrt{.70}\\ &= .25\end{aligned}$$

Here, *true* r_{xy} is the true correlation of .30, r_{yy} is the reliability of psychological-well-being scores (.70), and *observed* r_{xy} is the expected correlation between age and psychological well-being. A common interpretation of this calculation is that Jane should observe a correlation of .25. However, the observed $r_{xy} = .25$ value provided by the formula is not the only value that Jane can observe in her study. It represents a best guess of the correlation Jane will observe in her study, which means that $r = .25$ may not be the value she actually obtains. Why?

Understanding why the "best-guess" perspective is still a guess

In this section, we will address a few common misunderstandings of the Spearman attenuation formula presented above.

Expected values are not exact values. A frequent misunderstanding of the Spearman attenuation formula (Equation 1) is that it indicates the exact correlation that Jane will observe under the specified conditions (i.e., a true correlation of .30 and reliability of .70). Unfortunately, it is not possible to predict the exact correlation that Jane will observe. In a single study, the presence of measurement error can produce a wide range of correlations. In Jane's study, measurement error is introduced by her psychological-well-being data.

In truth, the value provided by the Spearman formula is simply a best guess of the correlation that is expected to be observed by Jane in the long run—in the same way that a mean can be the best guess of the value for any given individual in a sample. Indeed, the value provided by the Spearman formula is an average of all the possible correlations that could be observed by Jane when the reliability of the criterion scores is .70. In mathematical terms, the value provided by the Spearman formula is an expected value. We encourage methodologically minded individuals to see the more detailed explanation of this issue in our Supplemental Material.

Attenuation and nuisance correlations. Another common misunderstanding of the Spearman formula is that observed correlations will always be attenuated when data contain measurement error. That is, when the reliability of scores (r_{yy}) is less than 1.00, one can mistakenly presume that the observed correlation (observed r_{xy}) will always be less than the true correlation (true r_{xy}). In Jane's case, this would mean that she would always observe a correlation less than .30. In practice, when there is measurement error, Jane could observe a range of correlations, with the average of these observed correlations falling below the true correlation. However, the variability in observed correlations can also produce observed correlations that are *larger* than the true correlation. We illustrate this potential below.

Score reliability is based on the premise that an observed value (Y_{observed}) contains not only the true value (Y_{true}) that the researcher intends to measure but also random error (Y_{error}). These components are formally indicated below in Equation 2:

$$Y_{\text{observed}} = Y_{\text{true}} + Y_{\text{error}}$$

Score reliability is the proportion of variance in observed scores that is due to true scores.[4] Thus, when score reliability is .70, it indicates that 70% of the variance in observed scores is due to true variability. A key assumption of classical test theory is that errors are uncorrelated with true scores. Following this, researchers typically assume that random error is uncorrelated with true scores in their data. In Jane's situation, this would mean that random measurement errors in psychological well-being (Y_{error}) are unrelated to psychological-well-being true scores (Y_{true}) or age true scores (X_{true}).

In practice, the assumption of classical test theory that random measurement errors are uncorrelated with true scores (on the criterion or predictor) is rarely met (Zimmerman, 2007; Zimmerman & Williams, 1977). Indeed, it is likely to hold only in extraordinarily large sample sizes. In real-world scenarios, the correlation between random errors and true scores is not zero. These nonzero correlations between true scores and random errors are referred to as *nuisance correlations* (Charles, 2005; Zimmerman, 2007). Our Supplemental Material provides a more in-depth description of nuisance correlations.

Making the connection with the Spearman attenuation formula discussed earlier, Equation 1 can be viewed as providing an estimate of the observed correlation that would occur in a situation with no nuisance correlations. In a realistic scenario, a range of observed correlations are possible because of the presence of these nuisance correlations. To summarize and emphasize, under everyday circumstances, the Spearman attenuation formula provides a best guess of the correlation a researcher will observe for a given true correlation and reliability. This best guess is not the only possible value.

We will use Jane's scenario to illustrate how observed correlations vary because of nuisance correlations, which are a product of measurement error. Recall that in Jane's scenario, the true correlation between age and well-being is $r_{\text{true}} = .30$, and the reliability of the criterion (well-being) scores is .70 (age is measured without error).

To illustrate the impact of unreliability on observed correlations, we consider several thousand *ideal replications* of Jane's study. To achieve these ideal replications, we constructed the most perfect replication scenario we could imagine: one in which Jane's study is repeated several thousand times, with the same materials, *on the same participants*, who have no memory of the previous replication attempts. Because the same participants are used in each replication, the same predictor and criterion true scores are used and the true correlation does not change. Moreover, because the same participants are used in each replication, the sampling error associated with selecting different participants from the population is not present. The only factor that changes across replications is random measurement error (study participants are held constant). Moreover, relative proportions of true scores and random errors are also held constant across replication attempts. The results of these ideal replications, generated through a computer simulation using R (R Core Team, 2013), are presented in Figure 1a.

Figure 1a is a histogram of the results of 100,000 replications of Jane's study. Each replication results in an observed correlation that we refer to as a *replication correlation*. The average replication correlation between age and well-being, over the 100,000 studies, is $r = .25$, and the middle 99% of replication correlations fall between $r = .03$ to $r = .46$. The vertical line on the figure denotes the true correlation of $r_{\text{true}} = .30$. Recall, the average of .25 is precisely what the Spearman formula predicts. The range of values that are observed over the course of the 100,000 replications is important to note. Specifically, in 29% of the replication studies, the correlation was over .30. This means that in more than one quarter of the replication studies, the replication correlation was higher than the true correlation and well above what is expected as a result of attenuation.[5] Moreover, many replications

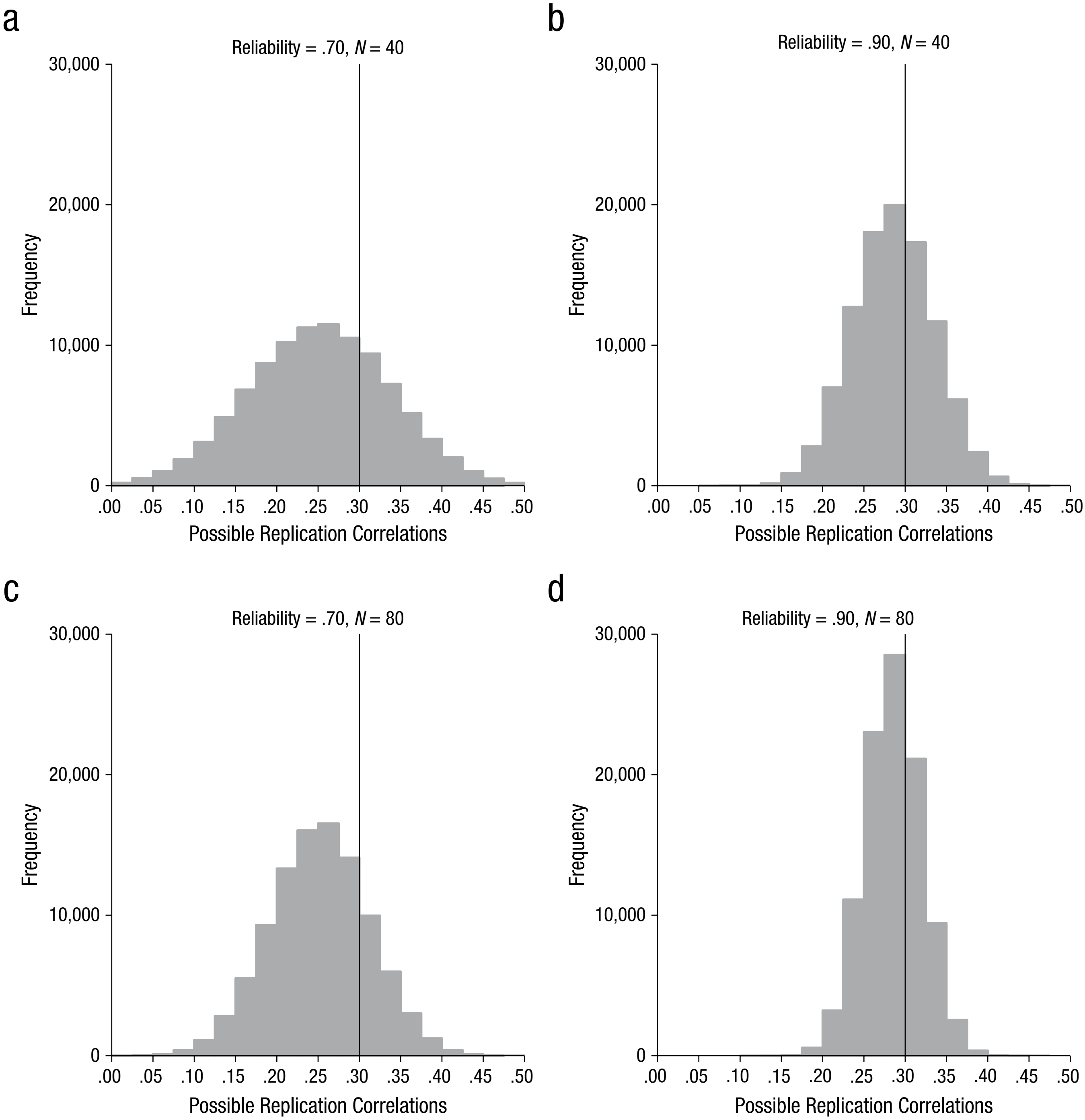

Fig. 1. Histograms showing the range of possible correlations obtained from ideal replications with various reliability and sample-size conditions. The vertical line on each graph indicates the true correlation (.30). Each histogram illustrates the results of 100,000 replications with the same participants. The variability in replication correlations illustrated here is due only to the effects of measurement error—specifically, the nonzero nuisance correlations between true scores and random measurement errors.

produced correlations that were less than .25, with some (beyond the middle 99%) even producing negative correlations. As a result, we can see that, even under ideal replication conditions, the range of results one can expect from replications is quite large.

Indeed, for a given underlying reality in which the correlation is .30 (i.e., a true correlation of .30), any correlation between r = .03 and r = .46 may be considered a replication that was not influenced by any unethical or unsavory research practices, just measurement error. This range of correlations may be unsettling to many researchers and is sufficiently wide to raise the question of what researchers should reasonably expect to find when attempting to replicate the results of a given study.

What if Jane had observed a correlation of $r = .46$ in her first study attempt? A colleague from the same university, Richard, might attempt to replicate Jane's finding under an ideal replication scenario (i.e., using the same participants). On the basis of the above results, Richard could conceivably observe a correlation of $r = .03$. The low correlation obtained by Richard might lead him to wonder if Jane engaged in unethical research practices or made a mistake—especially given that he and Jane used the same participants yet obtained a substantially different result. However, in reality, the only reason Richard's findings differed from Jane's was the random measurement error associated with the criterion.

Variability in replications using different parameters

To further illustrate the effect of reliability on replications, we ran our simulations using other reliability and sample-size parameters. We adjusted the parameters in a variety of optimistic ways. The results presented in Figure 1b illustrate a decreased variability in replication correlations when the reliability was increased from .70 to .90. The results presented in Figure 1c illustrate a decreased variability in replication correlations when the sample size was increased from 40 to 80. Finally, the results presented in Figure 1d illustrate the substantial decrease in the variability of replication correlations that occurred when the sample size was doubled to 80 and the reliability was increased to .90. In other words, investing in measures that produce reliable data as well as increasing the sample size of the original study and replication attempts can substantially increase reproducibility.

Additionally, we examined the replication correlations that are possible as a result of measurement error across a range of effect sizes (holding sample size at 40 and reliability at .70). These results are presented in Figure 2 and illustrate that there is slightly more variability in replication correlations when the underlying correlation is smaller ($r = .10$) than when it is larger ($r = .50$).

Summary. These simulations illustrate that researchers can effectively decrease the expected variability in replication correlations by using larger sample sizes and by having data with high reliability. Notably, though, when larger sample sizes are used with high levels of reliability, there is still considerable variability in the correlations that may be observed, even under ideal circumstances (i.e., using the same participants). Finally, there was slightly more variability in replication correlations when the true correlation was small as opposed to large—although the size of the underlying effect is outside the control of researchers.

Because the true correlation in the above examples is known, the results from Jane's scenario are effective for demonstration purposes. However, in practice, the true correlation is never known prior to research being conducted. As a result, it is not possible to generate the range of possible replication results a priori. In our next section, we discuss the challenge of a researcher observing a correlation in a single study and then trying to ascertain the extent to which it can be established as a replicable effect.

A Realistic Perspective on Replication With Measurement Error: The True Correlation Is Unknown

Imagine the following scenario. A published study reports a correlation of $r = .30$ between age (measured without error) and well-being (measured with a reliability of .70). Mary is a researcher who is interested in replicating this published study. To isolate and highlight the effect of measurement error, let us again assume that Mary can conduct a theoretically ideal replication with the same amnesic participants who have no memory of participating in the published study and are not affected by having done so. Under these ideal conditions, what results can Mary expect?

We propose that even under these ideal replication conditions (with the same participants), determining the correlation (or range of correlations) Mary might observe when replicating the study is exceedingly difficult. As a result, trying to ascertain the extent to which Mary's replication attempt is "successful" is a tenuous proposition. Mary's difficulty in determining what to expect from her replication stems from the fact that she does not know the true correlation that produced the published correlation. Indeed, an unfortunate consequence of the variability in correlations that is caused by measurement error (as illustrated above) is that for any observed correlation, there are a multitude of true correlations that could have produced the observed correlation.

We illustrate this problem using another simulation, in which the published correlation that Mary wants to replicate could be due to any one of several true correlations. To do this, we repeated the simulation described in the previous section but did so four times, using different true correlations: $r_{\text{true}} = .10$, .20, .30, and .40. As a result, each simulation represents one possible underlying reality. That is, there were only four sets of true scores, one set corresponding to each underlying reality ($r_{\text{true}} = .10$, .20, .30, and .40). The results are displayed in Figures 3a,

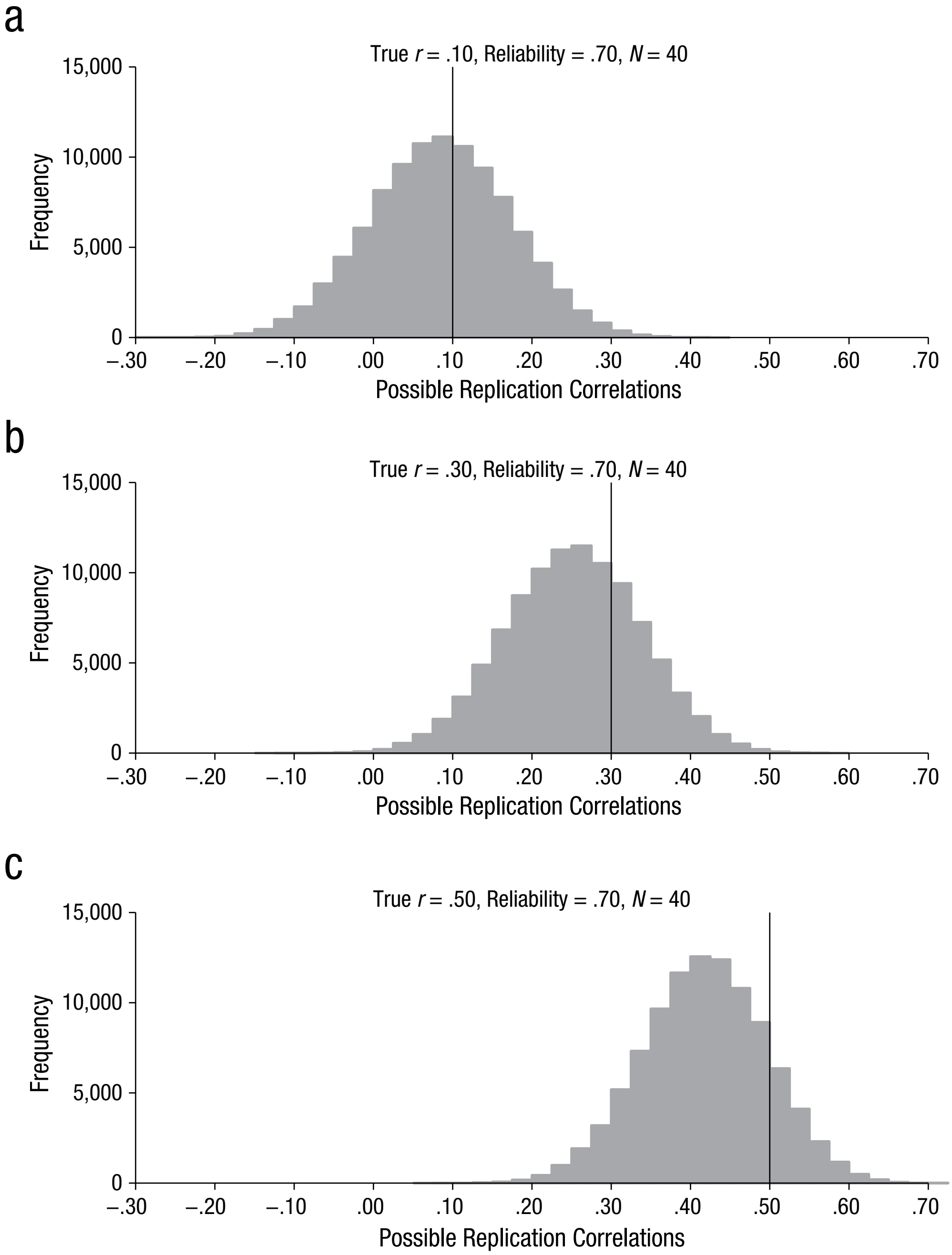

Fig. 2. Histograms showing the range of possible correlations obtained from ideal replications with various true correlations. Each histogram illustrates the results of 100,000 replications with the same participants. The vertical line on each graph indicates the true correlation (.10, .30, and .50, respectively). The variability in replication correlations illustrated here is due only to the effects of measurement error—specifically, the nonzero nuisance correlations between true scores and random measurement errors.

3b, 3c, and 3d, respectively. The dotted vertical lines at .30 on the x-axes illustrate that a published correlation of .30 is possible in each underlying reality. The solid lines denote the true correlations. The variability of replication correlations in these graphs is due solely to measurement error and not the error associated with sampling different individuals from a population (i.e., participants were held constant in each graph).

Figures 3a and 3b illustrate underlying realities of r_{true} = .10 and r_{true} = .20, respectively. For both, the published r = .30 correlation Mary wants to replicate is an overestimate of the true correlation. Figure 3c illustrates

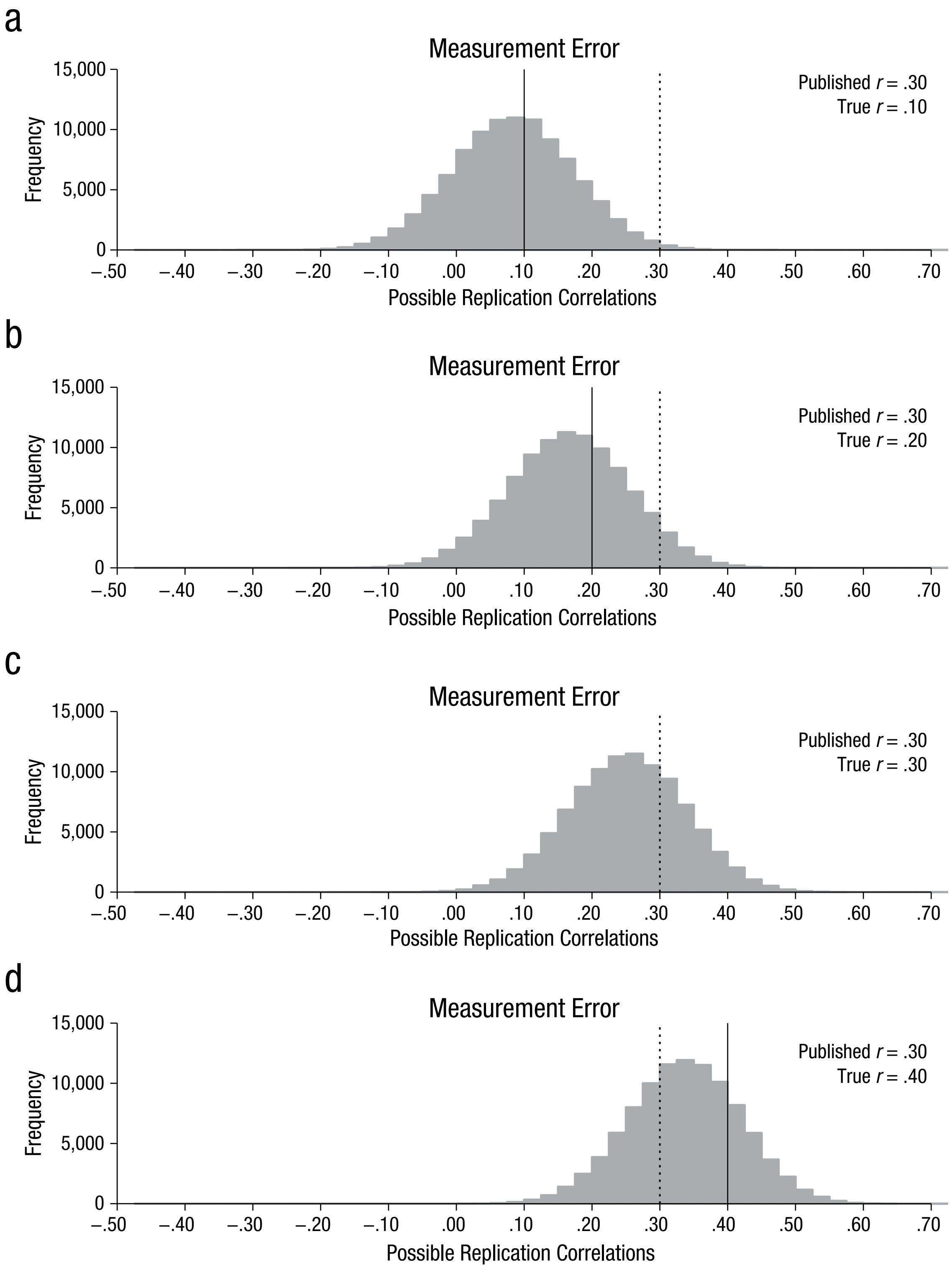

Fig. 3. Histogram showing that, as a result of measurement error, multiple true correlations can produce the same published correlation. A single published correlation of .30 (represented by the dotted vertical lines) may be due to any one of a number of true correlations (represented by solid vertical lines). Each histogram illustrates the results of 100,000 replications with the same 40 participants where the reliability is .70 (r_{yy} = .70). The variability in replication correlations illustrated here is due only to the effects of measurement error—specifically, the nonzero nuisance correlations between true scores and random measurement errors.

an underlying reality of r_{true} = .30. In this case, the published r = .30 correlation is a perfect estimate of the r_{true} = .30 true correlation. Figure 3d illustrates an underlying reality of r_{true} = .40. In this case, the published r = .30 correlation underestimates the r_{true} = .40 true correlation.

Collectively, Figures 3a through 3d illustrate that any one of several true correlations could have produced the published correlation of r = .30. Additionally, the simulations illustrate that without omniscient knowledge of the true correlation that produced the published correlation, it can be extremely difficult to estimate the correlation that one might expect when attempting to replicate the study. With a single published correlation of r = .30 between two variables that have reliabilities of 1.00 and .70, our results suggest that the true correlation could be r_{true} = .10, .20, .30, or .40. The results of these four simulations, taken together, suggest that the correlation observed in a replication could be as low as r = −.14 and as high as r = .54. In this example, we call the range of effect sizes that could be observed in a replication the conceivable *replication interval*.

Where did the replication interval of −.14 to .54 come from? If the published correlation of r = .30 was due to a true correlation of r_{true} = .10, then 99% of replication correlations would fall between −.14 and .31 (see Figure 3a). Thus, a replication correlation could be as low as −.14. Likewise, if the published correlation of r = .30 was due to a true correlation of r_{true} = .40, then 99% of replication correlations would fall between .12 and .54 (see Figure 3d). Thus, a replication correlation could potentially be as high as r = .54. Given that the "correct" underlying reality is unknown, we need to incorporate many possible underlying realities when constructing the replication interval for a single correlation. Consequently, when Mary attempts to replicate a published correlation of r = .30 (with criterion reliability of .70), she may obtain a correlation anywhere between r = −.14 and r = .54 simply because of random measurement error. This is a very large range—especially for a replication that used the same participants as the published study. It is important to note that we examined only four underlying realities; the replication interval is actually larger when all possible underlying realities are considered.

We can see that with only the results from a single study, or even a handful of studies, it is not possible to draw firm conclusions about the underlying reality (i.e., the true correlation). As a result, it is not clear how to interpret results of studies on the same phenomenon that do not agree. What would Mary or any trained, competent, level-headed researcher think if she or he were to obtain a correlation of r = −.14 in a study attempting to replicate a published correlation of r = .30?

As noted above, although our simulations represent a large range of possible true correlations, they are conservative. Specifically, they are not comprehensive, and the real range of possible true correlations is actually larger (and the replication interval correspondingly wider). Additionally, real-world replications would involve sources of error beyond just measurement error in the criterion. In many circumstances, both the predictor and criterion would be contaminated by measurement error. Furthermore, the error associated with sampling different sets of participants from the population would also be a source of error, one that is not reflected in these simulations. We consider the combined effect of sampling error and measurement error below.

A Realistic Perspective on Replication With Measurement Error and Sampling Error: The True Correlation Is Unknown

Until now, we have concentrated on demonstrating the unique influence of measurement error on replication attempts by isolating measurement error. To do so, we conducted ideal replications on the same participants in order to eliminate sampling error. In practice, such ideal replications are not possible, and different participants are used. As a result, researchers must contend with error that is due to both measurement error and sampling error (Hunter & Schmidt, 2004). In this section, we include sampling error alongside measurement error to examine the relative contribution of both sources of error in the variability of replication attempts.

We begin by revisiting Mary's attempt to replicate a published correlation of r = .30 between age and well-being where well-being has a reliability of .70. This time, we include the effects of both measurement error and sampling error. Because we now include the effects of sampling error, the term *true correlation* represents the *population correlation* in this section.

We created four scenarios with true correlations of r_{true} = .10, .20, .30, and .40. For each of these scenarios, we created a population (N = 1,000,000) and drew a random sample of participants (N = 40) for every replication. Then, we calculated the correlation (influenced by measurement error) in each sample. The results are displayed in Figures 4a, 4b, 4c, and 4d, respectively. A dotted vertical line at .30 on the x-axis illustrates how a published correlation of r = .30 is possible for all four underlying realities (i.e., all four true correlations). On these figures, we also report the relative influence of measurement error and sampling error on replication correlations; these are the *mean-change-due-to-sampling-error* and *mean-change-due-to-measurement-error* statistics.[6] The

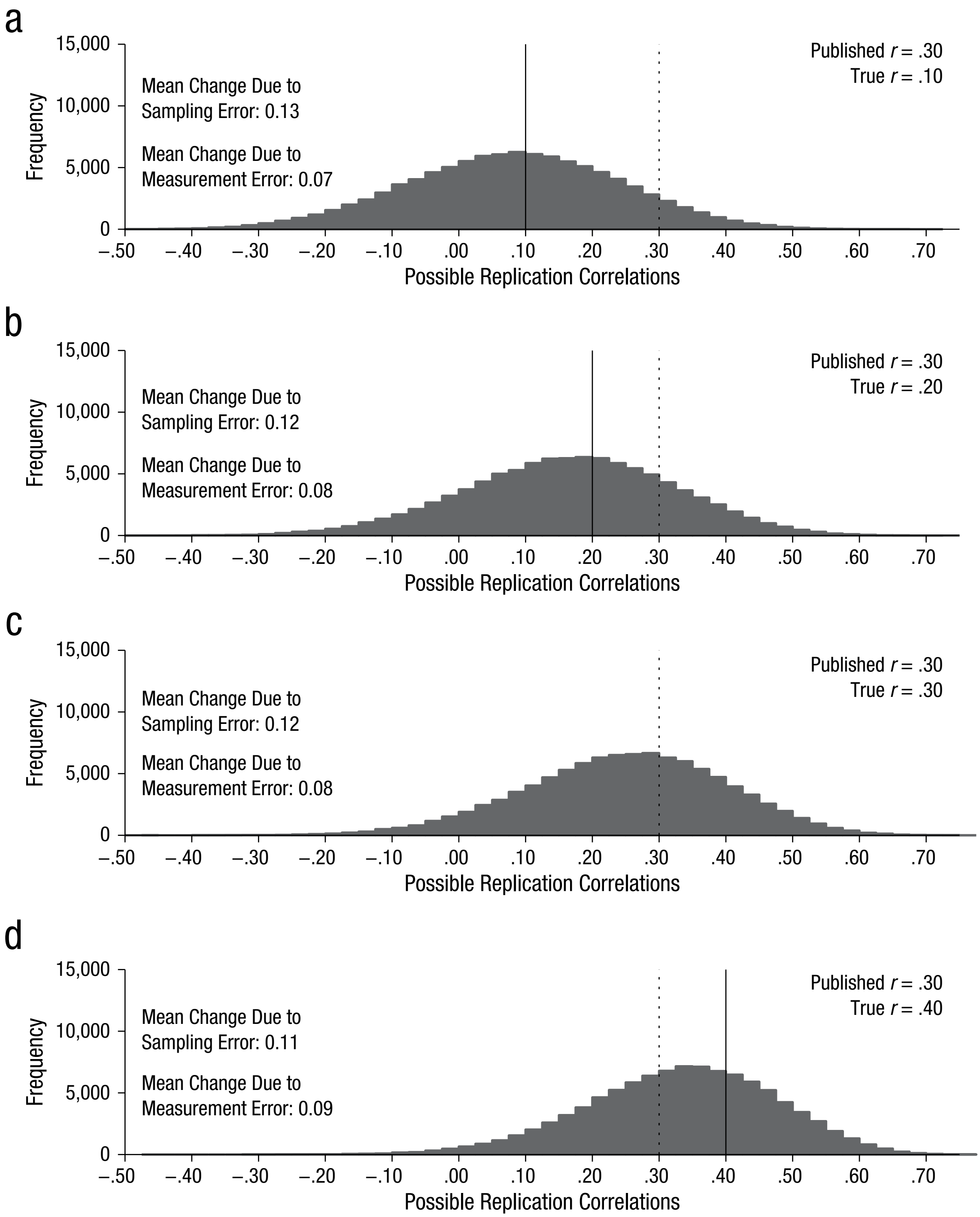

Fig. 4. Histogram showing that, as a result of sampling and measurement error, multiple true correlations can produce the same published correlation. A single published correlation of .30 (represented by the dotted vertical lines) may be due to any one of a number of true correlations (represented by solid vertical lines). Each histogram illustrates the results of 100,000 replications with samples of 40 participants each randomly drawn from a single population. In each sample, the reliability is .70 (r_{yy} = .70). The variability in replication correlations is due to random sampling error and nonzero nuisance correlations between true scores and random measurement errors.

mean-change values are reported in absolute correlation units, so the values can be interpreted as the average amount a replication correlation would move up or down as a result of the specified source of error.

Notably, the distributions in Figure 4 (from simulations including both measurement and sampling error) are substantially wider than those in Figure 3 (from simulations including measurement error alone). For example, in Figure 4a, where the true correlation was r_{true} = .10 (reliability of .70), 99% of replication correlations (influenced by measurement and sampling error) fell between r = −.33 and r = .47. In contrast, in Figure 3a, where only measurement error was considered, the range of possible replication correlations was a narrower r = −.14 to r = .31. Likewise, in Figure 4d, where the true correlation was r_{true} = .40 (with reliability of .70 in the criterion), 99% of replication correlations (influenced by measurement and sampling error) fell between r = −.06 to r = .65. In contrast, in Figure 3d, where only measurement error was present, it was a narrower r = .12 to r = .54. Thus, unsurprisingly, when both sampling error and measurement error were taken into account, the range of replication correlations was wider than when only measurement error was taken into account.

Examining the mean-change-due-to-sampling-error and mean-change-due-to-measurement-error statistics, we see that both sources of error were nontrivial. Measurement error ranged from .07 to .09 correlation units, and sampling error ranged from .11 to .13 correlation units. Across the different scenarios, the relative influence of sampling error was consistently larger than the effect of measurement error, with measurement error ranging from 54% to 82% of the size of sampling error.

In these more realistic examples including measurement and sampling error, if the published r = .30 correlation was due to a true correlation of r_{true} = .10, Mary might legitimately obtain a correlation as low as r = −.33 in her replication attempt. In contrast, if the published r = .30 correlation was due to a true correlation of r_{true} = .40, Mary might legitimately obtain a correlation as high as r = .65 in her replication attempt. Consequently, the replication interval is −.33 to .65. Once again, because we examined only four possible underlying realities (r_{true} = .10, .20, .30, and .40), the replication interval reported here is an underestimate of the true replication interval. This is because there are more extreme true correlations that we did not include in the simulation that could also have created the published correlation of .30. For example, a true correlation of r = −.10 could also produce a published correlation of r = .30 (when the true correlation is r = −.10, 99% of replication correlations fall between r = −.47 and r = .33). Figure S4 in the Supplemental Material illustrates that interpretational difficulties, although somewhat reduced, are still present when the sample size is doubled to 80. Overall, these simulations effectively showcase the futility associated with trying to draw conclusions about the size of an underlying effect from a single study. To put it plainly, under these very reasonable and common parameters, a researcher could expect anywhere from a medium-sized negative correlation to a large positive correlation when attempting to replicate a positive correlation of .30.

Moving Forward: New Expectations for Replications

The scenarios and results we have presented above illustrate that substantial variability in replication correlations can be expected as a result of measurement error alone. We also showed that the range of possible replication correlations increased when both measurement error and sampling error were present. Our data and examples showcased the inherent difficulties associated with trying to interpret replication attempts as verifying or falsifying primary studies. Specifically, it is not possible to know the true correlation that produced a particular published correlation. As a result, without omniscient knowledge of the true correlation, the range of correlations one might expect during a replication study can be incredibly large. From our results, we can see that researchers can reduce the range of replication correlations by increasing sample size and reliability. Moreover, although not under researchers' control, larger effect sizes show smaller ranges of variability than smaller ones.

Consider a scenario in which a researcher conducts Study 1 and obtains a significant result. That same researcher then conducts Study 2, using the same materials and sample size, and fails to replicate the findings from Study 1. One reaction might be for the researcher to abandon the program of research because the effect that he or she is pursing is not reliable. Alternatively, the researcher might put Study 2 in a file drawer and conduct a new "Study 2." We suggest that both courses of action are inappropriate. Abandoning a line of research on the basis of one or two studies is premature, given the scenarios we have illustrated. Similarly, failing to publish a nonsignificant finding is detrimental to knowledge generation in a field; even nonsignificant studies contain valid effect-size estimates. Of course, in this scenario, the failure to publish the nonsignificant finding would be due in part to the reluctance of reviewers and editors to acknowledge the variability in study results that occurs as a result of random sampling and measurement error (see Maner, 2014, this issue).

Power analysis

One approach to addressing the replication crisis is increasing the use of power analysis. Researchers advancing solutions to the replication crisis have correctly emphasized the importance of running adequately powered studies in order to increase the reliability of research (e.g., Asendorpf et al., 2013; Button et al., 2013; Tressoldi, 2012). This increased emphasis on power analysis is extremely valuable because it substantially increases the probability that replication attempts will correctly reflect the existence of a nonzero effect. Indeed, power analysis provides a statistically grounded expectation about what replication studies should show in the long run.

A practical challenge when conducting power analyses is that the calculations depend on knowledge of the true effect size (e.g., the true/population correlation). Because the true effect size is unknown, estimates of the true effect size must be used in power calculations. These estimates are often acquired from a published study or a pilot study. Our simulations illustrate that an effect-size estimate from a published study may deviate substantially from the true effect size as a result of measurement error and sampling error (see also Perugini, Gallucci, & Costantini, 2014, this issue). Indeed, even well-powered studies can produce effect-size estimates that will vary across studies (Hunter & Schmidt, 2004).

Meta-analysis

An important implication of our results is that the replication process should be thought of differently. Researchers currently attempt to replicate the results of previous research in order to verify those results. We suggest that this mind-set is based on the flawed premise that there is substantive meaning that can be obtained from the results of a single study. Our simulations (which considered only two of many sources of random error) clearly illustrate that this premise is flawed: A single study often provides little information about an underlying true effect. Consequently, we suggest moving from a mind-set focused on *verification* of individual studies to one that is based on *estimation*. Researchers must shift their mind-set from thinking that individual studies provide definitive insight into the validity of a research hypothesis to a mind-set in which the results of a single study are viewed as a mere estimate of an underlying reality. The estimation mind-set implies that multiple approximations need to be averaged to determine the true underlying reality. Fortunately, this approach to knowledge generation already exists—it is known as meta-analysis.

Shifting to a meta-analytic mind-set emphasizes that a single study provides information, but that this information is limited (see Chan & Arvey, 2012). A study may be best viewed as a data point in a broader data set to be examined via meta-analysis (see also Braver, Thoemmes, & Rosenthal, 2014, this issue). This is a familiar mind-set and is analogous to how a researcher may view the contribution of an individual participant to his or her study. That is, a researcher would never interpret individual participant data through a lens of verification, whereby Participant 2's data is viewed as replicating or failing to replicate the results for Participant 1. Likewise, the data from one study should not be interpreted as replicating or failing to replicate the results of another study.

A variety of approaches to meta-analysis have been proposed over the years (c.f. Borenstein et al., 2009; Cooper, Hedges, & Valentine, 2009; Glass, 1976; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter & Schmidt, 2004; Schmidt & Hunter, 1977). Most types of meta-analysis focus on eliminating the error associated with sampling different sets of individuals from a population by averaging across several samples. The standard meta-analytic process provides an attenuated estimate of the population correlation (i.e., the one indicated by the Spearman attenuation formula). Fortunately, one type of meta-analysis known as psychometric meta-analysis (e.g., Hunter & Schmidt, 2004) also corrects measurement error (among other sources of error) and provides an accurate estimate of the true population correlation. Increasing reliance on this meta-analytic approach (or on similar ones) allows for accurate estimation of the relations among variables. Meta-analysis has been around for decades and is not a new methodology—we are not suggesting that it is. Instead, we suggest that the replication crisis perhaps exists only for those who do not view research through the lens of meta-analysis.

Another way of thinking about this mind-set shift is that researchers must move away from attempting to replicate results and move toward replicating methodology. That is, by carefully replicating the methodology of previous research (assuming it is of high quality), researchers are providing high-quality data points for a future meta-analysis. Moreover, this approach is consistent with the editorial policy of some journals that emphasize methodological rigor over particular findings (e.g., *PLOS ONE*). Within this mind-set, replication studies are not only desirable for knowledge generation—they are essential.

A natural problem with shifting to this mind-set is that, as researchers, we tend to give undue weight to single studies—even when we try not to do so. This problem is illustrated particularly well in the context of journal reviewers and editors who accept the first few studies in a particular area, but not subsequent ones. Of course, the first studies in a particular area may have value beyond subsequent ones in that the first studies may encourage research in a new area. There is, however, no reason to suspect that the first studies published in an area provide a more accurate estimate of a relation than later ones. Indeed, meta-analytic procedures can account for this bias and have steps to address issues such as the "file-drawer problem" with statistical adjustments to estimate portions of the distribution that are not known because of publication bias. Consequently, given the ability of meta-analysis to ameliorate the effects of both sampling error and publication bias, meta-analytic estimates are ideally suited as effect-size estimates for power analyses in a mature research area.

A shift to this meta-analysis mind-set cannot inoculate research against concerns regarding questionable research practices and data fraud. A meta-analysis is only as good as the studies it is based on; therefore, any corruption of the meta-analytic data by questionable research practices can influence the meta-analytic estimate. Consequently, we propose that a meta-analytic perspective and interventions to address questionable research practices are complementary and that both are required for the advancement of psychological research.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Notes

1. The variables used in this example were chosen for illustration purposes only. The concepts we illustrate in this scenario apply to all variables, not just age and well-being.

2. The parameters used for Jane's scenario, sample size (N = 40) and true correlation (r = .30), were chosen to be representative values in psychological research. To choose the true correlation, we searched the literature to determine an effect size that generalized to many studies. We obtained an estimate based on a meta-analysis that compiled the average effect size across 100 years of social psychology, in 18 topic areas, based on 25,000 studies (Richard, Bond, & Stokes-Zoota, 2003). This investigation revealed that the average effect size was r = .21. We recognized, however, that this was an observed correlation and estimated that the true correlation might be closer to .30 (if the underlying published studies had reliabilities of approximately .70 on average). Likewise, we based the sample size in our simulation on an investigation of sample sizes across four areas of psychology (abnormal, applied, developmental, and experimental; Marszalek, Barber, Kohlhart, & Holmes, 2011), which revealed that the interquartile range for sample size in 2006 was N = 18 to N = 136, with a median of N = 40 (which we used in our simulation). Consequently, we believe that the particular conditions under which we ran our replication simulations are representative of the conditions under which a substantial amount of psychological research takes place.

3. We use error only on the criterion to simplify our discussion. Moreover, this approach has the advantage of ensuring that our illustrations are applicable to experimental research in which there is a manipulation used as the "predictor" and measurement error (i.e., reliability) influences only the dependent variable.

4. Reliability is defined as $r_{yy} = \frac{\sigma^2_{Y_{\text{true}}}}{\sigma^2_{Y_{\text{true}}} + \sigma^2_{Y_{\text{error}}}}$. This definition is applied throughout this article.

5. Because of publication bias, studies like these may be the most likely to be published.

6. To compute the mean-change values, we followed a two-step process. First, we examined the distance between the true correlation (i.e., the population correlation) and the sample correlation (calculated without measurement error). We then averaged these distances across all 100,000 replications in a scenario to obtain the mean change due to sampling error. Second, we examined the sample correlation calculated without measurement error and the sample correlation calculated with measurement error. We then averaged the distances between these two numbers across all 100,000 replications to obtain the mean change due to measurement error.

References

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

Bonanno, G. A., & Jost, J. T. (2006). Conservative shift among high-exposure survivors of the September 11th terrorist attacks. *Basic and Applied Social Psychology*, *28*, 311–323.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*, 333–342.

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

Chan, M. E., & Arvey, R. D. (2012). Meta-analysis and the development of knowledge. *Perspectives on Psychological Science*, *7*, 79–92.

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods*, *10*, 206–226.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Cunningham, G. K. (1986). *Educational and psychological measurement*. New York, NY: Macmillan.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.

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

Funder, D. C., & Task Force on Publication and Research Practices, Society for Personality and Social Psychology (2014). Notice: PSPB articles by authors with retracted articles at PSPB or other journals: Stapel, Smeesters, and Sanna. *Personality and Social Psychology Bulletin*, *40*, 132–135. doi:10.1177/0146167213508152

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, *5*, 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. Retrieved from http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0020124

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

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*, 202–220.

Makel, M., Plucker, J., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*, 537–542.

Maner, J. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science*, *9*, 343–351.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224–253.

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual & Motor Skills*, *112*, 331–348.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

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.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*, 319–332.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org/

Reproducibility Project. (2013). Retrieved from https://osf.io/ezcuj/wiki/home/

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331-363. doi:10.1037/1089-2680.7.4.331

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529–540.

Schwartz, S. H., & Rubel, T. (2005). Sex differences in value priorities: Cross-cultural and multimethod studies. *Journal of Personality and Social Psychology*, *89*, 1010–1028.

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.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, *3*, 271–295.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, *7*, 670–688.

Tressoldi, P. E. (2012). Replication unreliability in psychology: Elusive phenomena or "elusive" statistical power. *Frontiers in Psychology*, *3*, 1–5.

Yong, E. (2012, May 16). Replication studies: Bad copy. *Nature*. Retrieved from http://www.nature.com/news/replication-studies-bad-copy-1.10634

Zimmerman, D. W. (2007). Correction for attenuation with biased reliability estimates and correlated errors in populations and samples. *Educational and Psychological Measurement*, *67*, 920–939.

Zimmerman, D. W., & Williams, R. H. (1977). Validity coefficients and correlated errors in test theory. *Journal of Experimental Education*, *45*, 4–9.