aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2016, Vol. 11(4) 539–544

DOI: 10.1177/1745691616646366
pps.sagepub.com
SAGE

What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science

Prasad Patil, Roger D. Peng, and Jeffrey T. Leek
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Abstract

A recent study of the replicability of key psychological findings is a major contribution toward understanding the human side of the scientific process. Despite the careful and nuanced analysis reported, the simple narrative disseminated by the mass, social, and scientific media was that in only 36% of the studies were the original results replicated. In the current study, however, we showed that 77% of the replication effect sizes reported were within a 95% prediction interval calculated using the original effect size. Our analysis suggests two critical issues in understanding replication of psychological studies. First, researchers' intuitive expectations for what a replication should show do not always match with statistical estimates of replication. Second, when the results of original studies are very imprecise, they create wide prediction intervals—and a broad range of replication effects that are consistent with the original estimates. This may lead to effects that replicate successfully, in that replication results are consistent with statistical expectations, but do not provide much information about the size (or existence) of the true effect. In this light, the results of the *Reproducibility Project: Psychology* can be viewed as statistically consistent with what one might expect when performing a large-scale replication experiment.

Keywords

replication, prediction intervals, *p* values, reproducibility, *Reproducibility Project: Psychology*

It is natural to hope that when two scientific experiments are conducted in the same way, they will lead to identical conclusions. This is the intuition behind the recent tour-de-force replication of 100 psychological studies by the Open Science Collaboration's *Reproducibility Project: Psychology* (Open Science Collaboration, 2015). At incredible expense and with painstaking effort, the researchers attempted to replicate the exact conditions for each experiment, collect the data, and analyze them just as they were in the original study.

The original analysis considered both subjective and quantitative measures of whether the results of the original study were replicated in each case. They compared average effect sizes, compared effect sizes with confidence intervals, and measured subjective and qualitative assessments of replication. Despite the measured tone of the manuscript, the resulting mass, social, and scientific media coverage of the article fixated on the statement that in only 36% of the studies were the original results replicated (Patil & Leek, 2015).

Although one may hope that a properly replicated study will provide the same result as the original, statistical principles suggest that this may not be the case. The *Reproducibility Project: Psychology* study coincided with extensive discussion on what it means for a study to be reproducible and how to account for different sources of variation when replicating studies (Ledgerwood, 2014). Stanley and Spence (2014) showed through simulation how sampling and measurement variation interact with the size and reliability of an effect to produce wide distributions of replication effect sizes. These examinations

Corresponding Author:
Jeffrey T. Leek, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD 21205
E-mail: jtleek@gmail.com

were accompanied by discussions of adequate study power (Maxwell, 2004; McShane & Böckenholt, 2014), sample size (Gelman & Carlin, 2014; Schönbrodt & Perugini, 2013), and how meta-analysis may address the consequences of inadequate power or sample size (Braver, Thoemmes, & Rosenthal, 2014). Anderson and Maxwell (2015) furthered these concepts by categorizing the different goals of replicating a study and recommending appropriate analyses and equivalence tests specific to each goal. In sum, the sources of variability that make replicating the result of a particular study so difficult were well documented while the *Reproducibility Project: Psychology* study was under way.

In the current article, we present a view of replication based on prediction intervals—a statistical technique for predicting the range of effects that researchers should expect in a replication study. This technique respects both the variability in the original study and the variability in the replication study to come to a global view of whether the results of the two are consistent. The statistical analysis shows that researchers' intuitive understanding of replication can be flawed. The key point is that variability exists in both the original study and the replication study. When the original study is small or poorly designed, there will be a large range of potential replication estimates that are consistent with the original estimate. Larger, more carefully designed studies will have a narrower range of consistent replication estimates. Consequently, many smaller studies will show statistically consistent replications even if they provide very little information about the quantity of interest. In other words, the replication may be statistically successful but may carry little information about the true effects being studied.

We re-emphasize the importance of well-designed studies that are run with sufficient sample sizes for drawing informative conclusions. We also suggest that replicating studies with small original sample sizes may be relatively uninformative—the replication estimates will be statistically consistent even when the estimates change signs or are quite different from the original study.

Defining and Quantifying Replication Using *p* Values

In the original article describing the *Reproducibility Project: Psychology* (Open Science Collaboration, 2015), a number of approaches to quantifying reproducibility were considered. The widely publicized 36% figure refers only to the percentage of study pairs that reported a statistically significant ($p < .05$) result in both the original and replication studies. The relatively low number of results that were statistically significant in both studies was the focus of extreme headlines (e.g., "Over half of psychology studies fail reproducibility test"; Baker, 2015) and played into the prevailing narrative that science is in crisis (Gelman & Loken, 2014).

The most widely disseminated information from the original article was based on a misinterpretation of reproducibility and replicability. Reproducibility is defined informally as the ability to recompute data-analytic results given an observed data set and knowledge of the statistical pipeline used to calculate them (Peng, 2011; Peng, Dominici, & Zeger, 2006). The expectation is that a reproducible study is one in which the exact same numbers will be produced from the same code and data every time. "The replicability of a study is the chance that a new experiment targeting the same scientific question will produce a consistent result" (Leek & Peng, 2015, p. 1645; see also Asendorpf et al., 2013; Ioannidis, 2005). When a study is replicated, it is not expected that the same numbers will result; this is true for many reasons, including both natural variability and changes in the sample population, methods, or analysis techniques (Leek & Peng, 2015).

Researchers therefore should not expect to get the same answer even if a perfect replication is performed. Defining replication as consecutive results with $p < .05$ does square with the intuitive idea that replication studies should arrive at similar conclusions, so it makes sense that despite the many reported metrics in the original article (Open Science Collaboration, 2015), the media have chosen to focus on this number. However, this definition is flawed because there is variation in both the original study and in the replication study, as has been much studied in the psychology community of late. Even if you performed 10,000 perfect studies and 10,000 perfect replications of those studies and tallied the number of times both the original study and the replication yielded significant results, and you repeated that again and again, you would expect that number to vary from one set of 10,000 studies to the next.

In real studies, researchers do not know the truth—what the real effect size is or whether the study found it. An alternative is to generate simulated data for which the effect size and variability are already known, then apply statistical methods to see what characteristics these data show. We conducted a small simulation that was based on the effect sizes presented in the original article. In the original study, the authors applied transformations to 73 of the 100 studies whose effects were reported via test statistics other than the correlation coefficient (e.g., *t* statistics, *F* statistics). We simulated 10,000 perfect replications of each of these 73 studies using 1-degree-of-freedom tests. These 10,000 simulations represented error-free, perfect versions of the studies in the Reproducibility Project. In each case, the percentage of *p* values that was less than .05 ranged from 73% to 91% (i.e., first to third quartile; maximum = 100%, minimum = 6%) with a high degree of variability (see Fig. S1 in the Supplemental Material available online).

Prediction Intervals

If replication is defined by a *p*-value cutoff, sampling variation alone may contribute to a failure to replicate results. Instead, we considered a more direct approach by asking the question, "What effect would we expect to see in the replication study once we have seen the original effect?" This expectation depends on many variables about how the experiments are performed (Goodman, 1992). We assumed that the replication experiment was indeed a true replication—a not-unreasonable assumption in light of the effort expended to replicate these experiments accurately.

One statistical quantity that incorporates what researchers can reasonably expect from subsequent samples is the prediction interval. A traditional 95% confidence interval describes the uncertainty about a population parameter of interest. An odds ratio might be reported as "OR = 1.6, 95% confidence interval (CI) = [1.2, 2.0]," where 1.6 is the best estimate of the true population odds ratio based on the observed data. The confidence limits 1.2 and 2.0 define the 95% CI constructed from this study. If one were able to observe 100 samples and construct a 95% confidence interval for each sample, 95 of the 100 samples would contain the true population odds ratio.

A prediction interval makes an analogous claim about an individual future observation given what has already been observed. In this context, given the observed original correlation and some distributional assumptions (described in detail in the Supplemental Material), one could construct a 95% prediction interval and state that if the exact same study were replicated 100 times, 95 of the observed replication correlations would fall within the corresponding prediction interval.

Using Prediction Intervals to Assess Replication

Assuming the replication is true and using the derived correlations from the original manuscript, we applied Fisher's z transformation (Fisher, 1915) to calculate a pointwise 95% prediction interval for the replication effect size given the original effect. The 95% prediction interval is

$$\hat{r}_{\text{orig}} \pm z_{0.975}\sqrt{\frac{1}{n_{\text{orig}}-3}+\frac{1}{n_{\text{rep}}-3}},$$

where $\hat{r}_{\text{orig}}$ is the correlation estimate in the original study, n_{orig} and n_{rep} are the sample sizes in the original and replication studies, respectively; and $z_{0.975}$ is the 97.5% quantile of the normal distribution (see the Supplemental Material). The prediction interval accounts for variation in both the original study and the replication study through the sample sizes incorporated in the expression of the standard error.

We observed that for the 92 studies in which a replication correlation effect size could be calculated, 69 (or 75%) were covered by the 95% prediction interval based on the original correlation effect size (Fig. 1). In two cases, the replication effect was actually larger than the upper bound of the 95% prediction interval. Considering the asymmetric nature of the comparison, one might consider these effects as having *replicated with clear effect*. We then estimated that 71 of 92 (77%) of replication effects are in or above the 95% prediction interval based on the original effect. Some of the effects that changed signs on replication still fell within the 95% prediction intervals calculated from the original effects. This in unsurprising in light of the relatively modest sample sizes and effects in both the original and replication studies (see Fig. S2 in the Supplemental Material).

We noted that of the 69 replication effect sizes that were covered by the 95% prediction interval, two replications showed a slightly negative correlation (−0.005 and −0.034) rather than a positive correlation as in the original study (0.22 and 0.31, respectively). In the first study, the original and replication sample sizes were 110 and 222, respectively; in the second study, they were 53 and 72, respectively. We would classify these two studies as *replicated with ambiguous effect* as opposed to replicated with clear effect because of the change in direction of the effect, although both are very close to zero. All other negative replication effects did not fall into the 95% prediction intervals, and hence these studies were considered *did not replicate*.

In 51 of 73 (70%) studies that the authors reported to be based on 1-degree-of-freedom tests, the replication effect was within the 95% prediction interval. The two cases in which the replication effect exceeded the 95% prediction interval were in this set, leaving us with the estimate that 53 of 73 (73%) of these studies had replication effects that were consistent with the original effects.

On the basis of the prediction-interval theory, we expected about 2.5% of the replication effects to be above the prediction interval bounds and 2.5% of the replication effects to be below the prediction interval bounds. About 23% were below the bounds, which suggests that not all effects were replicated or that there were important sources of heterogeneity among the studies that were not accounted for. The key message is that replication data—even for studies that should replicate—is subject to natural sampling variation in addition to a host of other confounding factors.

It is noteworthy that almost all of the replication-study effect sizes were smaller than the original-study effect

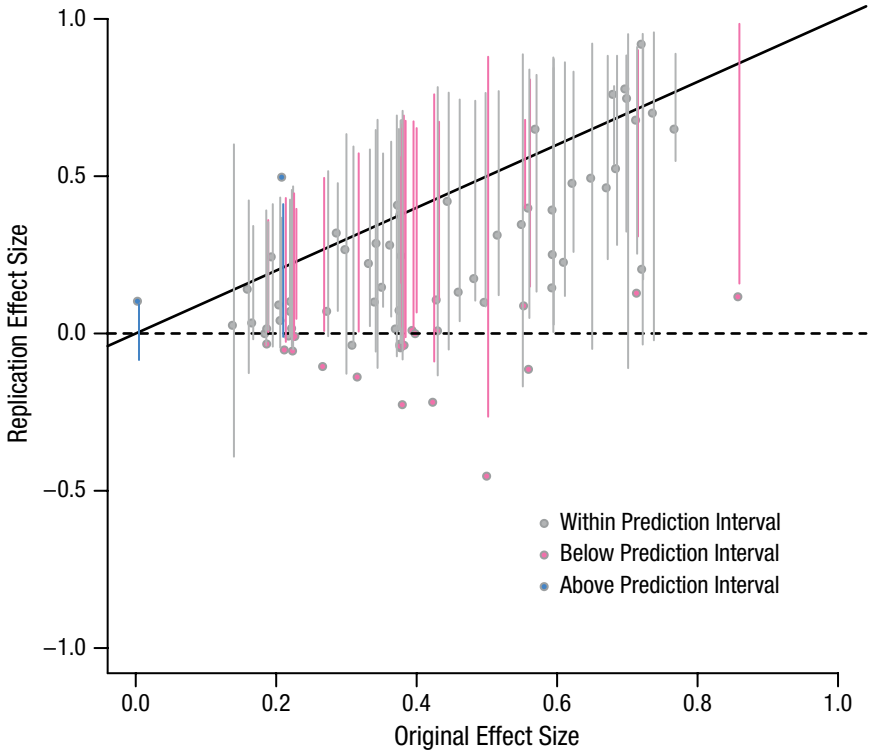

Fig. 1. Ninety-five percent prediction intervals and original and replication effect sizes. The scatterplot (with best-fitting regression line) shows the relationship between original effect sizes (correlation coefficient; *x*-axis) and replication effect size (correlation coefficient; *y*-axis). Each vertical line is the 95% prediction interval based on the original effect size. The dashed line indicates a replication effect of zero.

sizes, regardless of whether they fell inside the 95% prediction interval. In the original set of 92 studies, of those for which the replication effect fell within the 95% prediction interval (69 studies), 55 of 69 (80%) had a replication effect size that was smaller than the original effect size.

There is almost certainly some level of publication bias in the original estimates' $\hat{r}_{orig}$. This bias means that there would be an expectation of a nonzero difference between the reported original effect and the true correlation. If we make the reasonable assumption that people usually report larger effects, then the bias in the quantity will be positive. Given the calculations in the Supplemental Material, prediction intervals are less likely to cover the true value when bias exists in the original studies. This is likely the reason for some of the discrepancy between the observed coverage and the expected coverage of prediction intervals.

This finding speaks to the notion that many biases are likely to pervade the original study, pertaining mostly to the desire to report a statistically significant effect—even if that effect is small or unlikely to be replicable (Gelman & Weakliem, 2009). In this sense, our analysis complemented the findings of the Open Science Collaboration, (2015) and simultaneously provided some additional perspective on the expectation of replicability.

Conclusion

Researchers need a new definition for replication that acknowledges variation in both the original study and in the replication study. Specifically, a study replicates if the data collected from the replication are drawn from the same distribution as the data from the original experiment. Multiple independent replications of the same study will be needed to definitively evaluate replication. This view is consistent with the long-standing idea that a claim will be settled only by a scientific process, not a single definitive scientific study. We support Registered Replication Reports (Simons, Holcombe, & Spellman, 2014) and other such policies that incentivize researcher contribution to these efforts.

The *Reproducibility Project: Psychology* study highlights the fact that effects may be exaggerated and that replicating a study perfectly is challenging. We were caught off guard by the immediate and strong sentiment that psychology and other sciences may be in crisis (Gelman & Loken, 2014). However, many effects fell

within the predicted ranges despite the long interval between the original and replication studies, the complicated nature of some of the experiments, and the differences in populations and investigators performing the studies; these are all reasons for some guarded optimism about the scientific process. It is also in line with estimates we have previously made about the rate of false discoveries in the medical literature (Jager & Leek, 2014).

However, our analysis also allows us to make two general points about studying replication in psychological science. First, replication should consider the variability in both the original study and the replication study. When both original and replication variability are considered, studies may replicate statistically in ways that are unintuitive. For example, replication effects with opposite signs may still be statistically consistent with the original study.

Second, our work highlights the critical importance of good study design and sufficient sample sizes both when performing original research and when deciding which studies to replicate. Our work shows that studies with small sample sizes—like many in the *Reproducibility Project: Psychology*—will produce wide prediction intervals.

Although this may mean that the replication estimates will be statistically consistent with the original estimates, they may not be very informative. Replication of studies that are poorly designed or insufficiently powered may not provide much information about replication. But if the replication is well designed and well powered, it may impart something about whether the effect appears to be there at all.

We stress that the approach outlined in the current article is easily applied when the result of interest in a study can be summarized by one value that one can assume comes from a certain distribution. In reality, most scientific studies are more complex than one value can convey, dealing in multiple stimuli (Westfall, Kenny, & Judd, 2014), adaptation over time and circumstance (Berry, 2011), and complicated data sources (Cardon & Bell, 2001), just to name a very few. Our suggestion of 95% prediction intervals to help assess replication is meant to establish a conceptual framework and motivate researchers to begin considering what constitutes a reasonable expectation for a replicated effect. Extending these concepts to modern study designs is the next step in understanding the replicability of scientific 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.

Supplemental Material

Additional supporting information can be found at http://pps.sagepub.com/content/by/supplemental-data

References

Anderson, S. F., & Maxwell, S. E. (2015). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*, 1–12. doi:10.1037/met0000051

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. doi:10.1002/per.1919

Baker, M. (2015, August 27). Over half of psychology studies fail reproducibility test. *Nature News and Comment*. doi:10.1038/nature.2015.18248

Berry, D. A. (2011). Adaptive clinical trials: The promise and the caution. *Journal of Clinical Oncology*, *29*, 606–609.

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

Cardon, L. R., & Bell, J. I. (2001). Association study designs for complex diseases. *Nature Reviews Genetics*, *2*, 91–99.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*, 507–521.

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641–651.

Gelman, A., & Loken, E. (2014, November-December). The statistical crisis in science. *American Scientist*, *102*, 460. doi:10.1511/2014.111.460

Gelman, A., & Weakliem, D. (2009, July-August). Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, *97*, 310–316. doi:10.1511/2009.79.310

Goodman, S. N. (1992). A comment on replication, *P*-values and evidence. *Statistics in Medicine*, *11*, 875–879. doi:10.1002/sim.4780110705

Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, *294*, 218–228.

Jager, L. R., & Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, *15*, 1–12.

Ledgerwood, A. (2014). Introduction to the special section on advancing our methods and practices. *Perspectives on Psychological Science*, *9*, 275–277.

Leek, J. T., & Peng, R. D. (2015). Statistics: *P* values are just the tip of the iceberg. *Nature*, *520*, 612.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice when power analyses are optimistic. *Perspectives on Psychological Science*, *9*, 612–625.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6521). doi:10.1126/science.aac4716

Patil, P., & Leek, J. T. (2015). *Reports of the '36% value' in the media*. Retrieved from https://github.com/jtleek/replication_paper/blob/gh-pages/in_the_media.md

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*, 1226–1227. doi:10.1126/science.1213847

Peng, R. D., Dominici, F., & Zeger, S. L. (2006). Reproducible epidemiologic research. *American Journal of Epidemiology*, *163*, 783–789.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at *Perspectives on Psychological Science*. *Perspectives on Psychological Science*, *9*, 552–555.

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*, 305–318.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*, 2020–2045. doi:10.1037/xge0000014