Published in final edited form as:
Perspect Psychol Sci. 2016 July ; 11(4): 539–544. doi:10.1177/1745691616646366.

What should we expect when we replicate? 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 in the paper, mass, social, and scientific media adhered to the simple narrative that only 36% of the studies replicated their original results. Here we show that 77% of the replication effect sizes reported were within a 95% prediction interval based on the original effect size. Our analysis suggests two critical issues in understanding replication of psychological studies. First, our 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 consistent replication effects. This may lead to effects that replicate successfully, in that replication results are consistent with statistical expectations, but that do not provide much information about the size (or existence) of the true effect. In this light, the results of *Reproducibility Project: Psychology* can be viewed as statistically consistent with what you would expect when performing a large scale replication experiment.

Introduction

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, *Reproducibility Project: Psychology* (Collaboration et al., 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 identically to 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 to 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 paper fixated on a statement that only 36% of the studies replicated the original result (Patil & Leek, 2015).

Although we 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

(Ledgerwood, 2014). Stanley and Spence (Stanley & Spence, 2014) showed through simulation how sampling and measurement variation interplay with the size and reliability of an effect to produce wide distributions of replication effect sizes. These examinations 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 (Anderson & 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 when the *Reproducibility Project: Psychology* study was underway.

Here we present a view of replication based on prediction intervals - a statistical technique for predicting the range of effects we would 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 our intuitive understanding of replication can be flawed. The key point is that there is variability both in the original study and the replication study. When the original study is small or poorly designed, this means that the range of potential replication estimates consistent with the original estimate will be large. Larger, more carefully designed studies will have a narrower range of consistent replication estimates. With this view 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 still may carry little information about the true effects being studied.

Our analysis re-emphasizes the importance of well designed studies that are run with sufficient sample sizes for drawing informative conclusions. It also suggests that replicating studies with small original sample sizes may be relatively uninformative - the replication estimates will be statistically consistent even in cases where the estimates change sign or are quite different from the original study.

Defining and Quantifying Replication Using P-values

In the original paper describing the *Reproducibility Project: Psychology*, 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 ($P <$ 0.05) 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 like “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 report from this paper is based on a misinterpretation of reproducibility and replicability. Reproducibility is defined informally as the ability to recompute data analytic results conditional on an observed data set and knowledge of the statistical pipeline used to calculate them (Peng, 2011; Peng, Dominici, & Zeger, 2006). The expectation for a study to be reproducible is that the exact same numbers will be produced from the same code and data every time. Replicability of a study is the chance that a new experiment targeting the same scientific question will produce a consistent result (Asendorpf et al., 2013; Ioannidis, 2005). When a study is replicated, it is not expected that the same numbers will result for a host of reasons including both natural variability and changes in the sample population, methods, or analysis techniques (Leek & Peng, 2015).

We therefore do not expect to get the same answer even if a perfect replication is performed. Defining replication as consecutive results with $P < 0.05$ squares 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 paper, the media has chosen to focus on this number. However, this definition is flawed since there is variation in both the original study and in the replication study, as has been much-studied in the psychology community to date. Even if you performed 10,000 perfect studies and 10,000 perfect replications of those studies, you would expect the number of times both P-values are less than 0.05 to vary.

In real studies we don’t know the truth - what the real effect size is or whether the study found it. An alternative is to generate simulated data where the effect size and variability are already known, then apply statistical methods to see what characteristics these methods show. We conducted a small simulation 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 these 73 studies based on one degree of freedom tests. Each of these 10,000 simulations represents a perfect version of the Reproducibility Project with no errors. In each case, we calculated the percentage of P-values less than 0.05. The percentage of P-values less than 0.05 ranged from 73% to 91% (1^{st} to 3^{rd} quartile; high: 100%; low: 6%) with a high degree of variability (Figure S1).

Prediction Intervals

Sampling variation alone may contribute to “un-replicated” results if you define replication by a P-value cutoff. We instead consider 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). Here we assume the replication experiment is 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 we can reasonably expect from subsequent samples is the prediction interval. A traditional 95% confidence interval describes our uncertainty about a population parameter of interest. We may see an odds ratio reported in a paper as 1.6 [1.2, 2.0]. Here, 1.6 is our best estimate of the true population odds ratio based on the observed data. The range [1.2, 2.0] is our 95% confidence interval constructed from this study. If we were able to observe 100 samples and construct a 95% confidence interval for each sample, 95 of the 100 would contain the true population odds ratio.

A prediction interval makes an analogous claim about an individual future observation given what we have already observed. In our context, given the observed original correlation and some distributional assumptions (described in detail in the Supplementary section), we can construct a 95% prediction interval and state that if we were to replicate the exact same study 100 times, 95 of our observed replication correlations will 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 $\hat{r}_{\mathrm{orig}} \pm z_{0.975}\sqrt{\frac{1}{n_{\mathrm{orig}}-3}+\frac{1}{n_{\mathrm{rep}}-3}}$, where $\widehat{r_{orig}}$ is the correlation estimate in the original study; n_{orig}, n_{rep} are the sample sizes in the original and replication studies; and $z_{0.975}$ is the 97.5% quantile of the normal distribution (Supplementary Methods). The prediction interval accounts for variation in both the original study and in the replication study through the sample sizes incorporated in the expression of the standard error.

We observe that for the 92 studies where 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 (Figure 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 effect clear". We then estimate that 71/92 (or 77%) of replication effects are in or above the 95% prediction interval based on the original effect. Some of the effects that changed signs upon replication still fell within the 95% prediction intervals calculated based on the original effects. This in unsurprising in light of the relatively modest sample sizes and effects in both the original and replication studies (Figure S2).

We note here that of the 69 replication effect sizes that were covered by the 95% prediction interval, two replications showed a slightly negative correlation (−0.005, −0.034) as compared to a positive correlation in the original study (0.22, 0.31, respectively). In the first study, the original and replication sample sizes were 110 and 222; in the second study, they were 53 and 72. We would classify these two studies as "replicated with ambiguous effect" as opposed to "replicated with effect clear" due to 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 were considered "did not replicate".

We also considered the 73 studies the authors reported to be based on one degree of freedom tests. In 51 of these 73 studies (70%), the replication effect was within the 95% prediction interval. The same two cases where the replication effect exceeded the 95% prediction interval were in this set, leaving us with an estimate of 53/73 (73%) of these studies had replication effects consistent with the original effects.

Based on the theory of the prediction interval we expect about 2.5% of the replication effects to be above and 2.5% of the replication effects to be below the prediction interval bounds. Since about 23% were below the bounds, this suggests that not all effects replicate or that there were important sources of heterogeneity between 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 notable that almost all of the replication study effect sizes were smaller than the original study effect sizes, whether or not they fell inside the 95% prediction interval. In the original set of 92 studies, of those where the replication effect falls within the 95% prediction interval (69 studies), 55/69 (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 $\hat{\beta_{orig}}$. This bias means that the difference will have a non-zero expectation. If we make the reasonable assumption that people usually report larger effects then the bias in the quantity will be positive. Based on the calculation (in Supplementary Methods) our 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 our observed and expected coverage of prediction intervals.

This speaks to the notion that there are likely a host of biases that pervade the original study, pertaining mostly to the desire of reporting a statistically significant effect - even if it is small or unlikely to replicate (Gelman & Weakliem, 2009). In this sense, our analysis complements the finding of the Open Science Collaboration while simultaneously providing some additional perspective on the expectation of replicability.

Conclusion

We 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. To definitively evaluate replication we will need multiple independent replications of the same study. This view is consistent with the long-standing idea that a claim will only be settled by a scientific process rather than a single definitive scientific paper. 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). The fact that many effects fall within the predicted ranges despite the long interval between original and replication study, the complicated nature of some of the experiments, and the differences in populations and investigators performing the studies is a reason 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 makes two more general points about studying replication in psychological science. First, that replication should consider both the variability in 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. This means that replication of studies that are poorly designed or insufficiently powered may not tell us much about replication. But if the replication is well designed and powered, it may tell us something about whether the effect appears to be there at all.

We stress that the approach outlined here is easily applied when the result of interest in a study can be summarized by one value upon which we can ascribe distributional assumptions. In reality, most scientific studies are more complex, 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 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 is 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Anderson SF, Maxwell SE. There's more than one way to conduct a replication study: Beyond statistical significance. Psychological Methods. 2015 http://dx.doi.org/10.1037/met0000051.

Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJ, Fiedler K, et al. Recommendations for increasing replicability in psychology. European Journal of Personality. 2013; 27(2):108–119.

Baker, M. Over half of psychology studies fail reproducibility test. 2015 Aug. (http://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248 [Online; posted 27-August-2015])

Berry DA. Adaptive clinical trials: the promise and the caution. Journal of Clinical Oncology. 2011; 29(6):606–609. [PubMed: 21172875]

Braver SL, Thoemmes FJ, Rosenthal R. Continuously cumulating meta-analysis and replicability. Perspectives on Psychological Science. 2014; 9(3):333–342. [PubMed: 26173268]

Cardon LR, Bell JI. Association study designs for complex diseases. Nature Reviews Genetics. 2001; 2(2):91–99.

Collaboration OS, et al. Estimating the reproducibility of psychological science. Science. 2015; 349(6251):aac4716. [PubMed: 26315443]

Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika. 1915:507–521.

Gelman A, Carlin J. Beyond power calculations assessing type s (sign) and type m (magnitude) errors. Perspectives on Psychological Science. 2014; 9(6):641–651. [PubMed: 26186114]

Gelman A, Loken E. The statistical crisis in science. American Scientist. 2014; 102(6):460.

Gelman A, Weakliem D. Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. American Scientist. 2009:310–316.

Goodman SN. A comment on replication, p-values and evidence. Statistics in medicine. 1992; 11(7): 875–879. [PubMed: 1604067]

Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. Jama. 2005; 294(2):218–228. [PubMed: 16014596]

Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics. 2014; 15(1):1–12. [PubMed: 24068246]

Ledgerwood A. Introduction to the special section on advancing our methods and practices. Perspectives on Psychological Science. 2014; 9(3):275–277. [PubMed: 26173263]

Leek JT, Peng RD. Statistics: P values are just the tip of the iceberg. Nature. 2015; 520(7549):612–612. [PubMed: 25925460]

Maxwell SE. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. Psychological methods. 2004; 9(2):147. [PubMed: 15137886]

McShane BB, Böckenholt U. You cannot step into the same river twice when power analyses are optimistic. Perspectives on Psychological Science. 2014; 9(6):612–625. [PubMed: 26186112]

Patil, P.; Leek, JT. Reporting of 36% of studies replicate in the media. 2015 Sep. (https://github.com/jtleek/replication_paper/blob/gh-pages/in_the_media.md [Online; updated 16-September-2015])

Peng RD. Reproducible research in computational science. Science (New York, Ny). 2011; 334(6060): 1226.

Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. American journal of epidemiology. 2006; 163(9):783–789. [PubMed: 16510544]

Schönbrodt FD, Perugini M. At what sample size do correlations stabilize? Journal of Research in Personality. 2013; 47(5):609–612.

Simons DJ, Holcombe AO, Spellman BA. An introduction to registered replication reports at perspectives on psychological science. Perspectives on Psychological Science. 2014; 9(5):552–555. [PubMed: 26186757]

Stanley DJ, Spence JR. Expectations for replications are yours realistic? Perspectives on Psychological Science. 2014; 9(3):305–318. [PubMed: 26173266]

Westfall J, Kenny DA, Judd CM. Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. Journal of Experimental Psychology: General. 2014; 143(5):2020. [PubMed: 25111580]

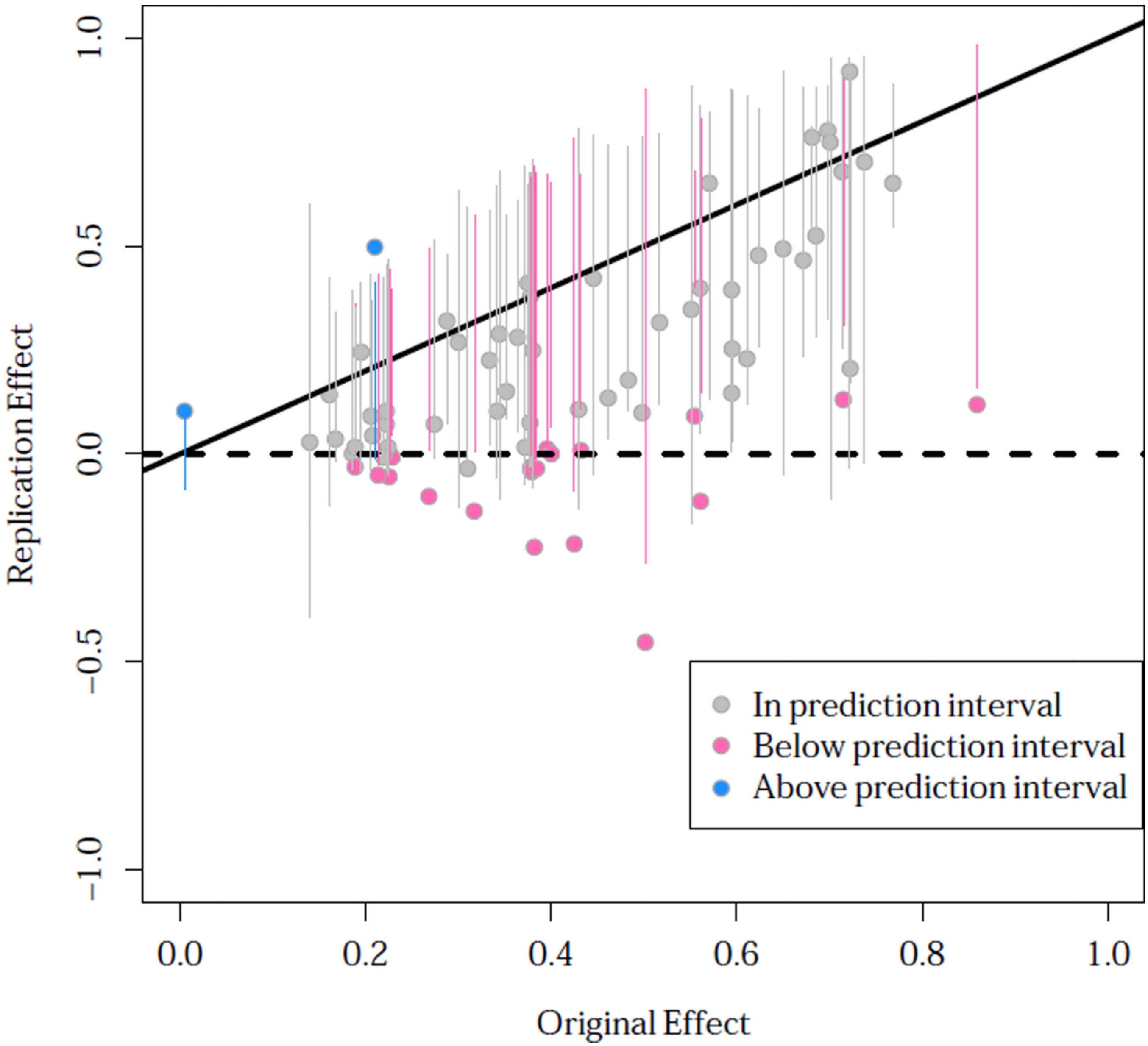

Figure 1. 95% prediction intervals suggest most replication effects fall in the expected range

A plot of original effects on the correlation scale (x-axis) and replication effects (y-axis). Each vertical line is the 95% prediction interval based on the original effect size. Replication effects could either be below (pink), inside (grey), or above (blue) the 95% prediction interval.