Authors of target article: Rolf A. Zwaan, Alexander Etz, Richard E., Lucas, and M. Brent Donnellan

Abstract word count: 60

Main text word count: 999

References word count: 132

Total word count: 1191

Title: Don't characterize replications as successes or failures

Author: Andrew Gelman

Institution: Columbia University

Mailing address: Department of Statistics, Columbia University, New York, N.Y. 10027

Telephone number: 212-851-2142

Email: gelman@stat.columbia.edu

Url: http://www.stat.columbia.edu/~gelman

Abstract:

No replication is truly direct, and I recommend moving away from the classification of replications as "direct" or "conceptual" to a framework in which we accept that treatment effects vary across conditions. Relatedly, we should stop labeling replications as successes or failures and instead use continuous measures to compare different studies, again using meta-analysis of raw data where possible.

Main text:

I agree wholeheartedly that replication, or the potential of replication, is central to experimental science, and I also agree that various concerns about the difficulty of replication should, in fact, be interpreted as arguments in *favor* of replication. For example, if effects can vary by context, this provides more reason why replication is necessary for scientific progress. I also agree with the target article it is an error when, following a disappointing replication result, proponents of the original published studies "irrationally privilege the chronological order of studies over the objective characteristics of those studies when evaluating claims about quality and scientific rigor." As a remedy to this fallacy I have proposed a "time-reversal heuristic" (Gelman, 2016): the thought experiment of imagining the large, preregistered replication study coming first, followed by the original, uncontrolled study.

It may well make sense to assign lower value to replications than to original studies, when considered as *intellectual products*, as we can assume the replication requires less creative effort. When considered as *scientific evidence,* however, the results from a replication can well be better than those of the original study, in that the replication can have more control in its design, measurement, and analysis.

It is also good to present and analyze all the data from an experiment. Selection, forking paths, and researcher degrees of freedom have led us into the replication crisis—but these problems are all much reduced with analyses that use all the data. Conversely, if we don't have access to raw data, many published results are close to useless, and when there is a high-quality preregistered replication, I'd be inclined to pretty much ignore the original paper, rather than, say, to assume the truth lies somewhere in between the original and replication results.

Beyond this, I would like to add two points from a statistician's perspective.

First, the idea of replication is central not just to scientific practice but also to formal statistics, even though this has not always been recognized. Frequentist statistics relies on the reference set of repeated experiments, and Bayesian statistics relies on the prior distribution which represents the population of effects—and in the analysis of replication studies it is important for the model to allow effects to vary across scenarios.

My second point is that in the analysis of replication studies I recommend continuous analysis and multilevel modeling (meta-analysis), in contrast to the target article which recommends binary decision rules which which I think are contrary to the spirit of inquiry that motivates replication in the first place.

The target article follows the conventional statistical language in which a study is a "false positive" if it claims to find an effect where none exists. But in the human sciences, just about all the effects we are trying to study are real; there are no zeros. See Gelman (2013) and McShane et al. (2017) for further discussion of this point. Effects can be hard to detect, though, because they can be highly variable and measured inaccurately and with bias. Instead of talking about false positives and false negatives, we prefer to speak of type M (magnitude) and type S (sign) errors (Gelman and Carlin, 2014). Related is the use of expressions such as "failed replication." I've used such phrases myself but they get us into trouble with their implication that there is some criterion under which a replication can be said to succeed or fail. Do we just check whether $p < 0.05$? That would be a very noisy rule, and I think we would all be better off simply reporting the results from the old and new studies (as in the graph in Simmons and Simonsohn, 2015). If there is a need to count replications in a larger study of studies such as the Open Science Collaboration, I'd prefer to do so using continuous measures rather than threshold-based replication rates.

The authors write, "if there is no theoretical reason to assume that an effect that was produced with a sample of college students in Michigan will not produce a similar effect

in Florida, or in the UK, or Japan, for that matter, then a replication carried out with these samples would be considered direct." The difficulty here is that theories are often so flexible that all these sorts of differences *can* be cited as reasons for a replication failure. For example, Michigan is colder than Florida, and outdoor air temperature was used as an alibi for a replication failure of a well-publicized finding in evolutionary psychology (Tracy and Beall, 2014). And there is no end to the differences between the UK and Japan that could be used to explain away a disappointing replication result in social psychology. The point is that any of these could be considered a "direct replication" if that interpretation is desired, or a mere "extension" or "conceptual replication" if the results do not come out as planned. In social psychology, at least, it could be argued that no replication is truly direct: society, and social expectations, change over time. The authors recognize this, citing Schmidt (2009) and also in their discussion of why contextual variation does not invalidate the utility of replications; given this, I think the authors could improve their framework by abandoning the concept of "direct replication" entirely, instead moving to a meta-analytic approach in which it is accepted ahead of time that the underlying treatment effects will vary between studies. Rather than trying to evaluate "whether a study is a direct or conceptual" replication, we can express the difference between old and new studies in terms of the expected variation in the treatment effect between conditions.

That said, if the measurements in the original study are indirect and noisy (as is often the case) and it is impossible or inconvenient to reanalyze the raw data, the question is moot, and it can make sense to just take the results from the replication or extension studies as our new starting point.

References:

Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, Jennifer L. Tackett (2017). Abandon statistical significance. https://arxiv.org/abs/1709.07588

Andrew Gelman (2013). I'm negative on the expression "false positives." http://andrewgelman.com/2013/11/07/nix-expression-false-positives/

Andrew Gelman (2016). The time-reversal heuristic—a new way to think about a published finding that is followed up by a large, preregistered replication (in context of Amy Cuddy's claims about power pose). http://andrewgelman.com/2016/01/26/more-power-posing/

Andrew Gelman and John B. Carlin (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. Perspectives on Psychological Science 9, 641-651.

Joseph Simmons and Uri Simonsohn (2015). Power posing: Reassessing the evidence behind the most popular TED talk. http://datacolada.org/37

Jessica L. Tracy and Alec T. Beall (2014). The impact of weather on women's tendency to wear red or pink when at high risk for conception. Plos-One 9(2), e88852.