Commentary

Psychological Science
1–3

DOI: 10.1177/0956797616665220
pss.sagepub.com

Each Reader Decides if a Replication Counts: Reply to Schwarz and Clore (2016)

Uri Simonsohn
The Wharton School, University of Pennsylvania

Received 5/25/16; Revision accepted 7/31/16

To learn from data, we need to ask two questions: What happened? And what does it mean? Asking *whether* a replication failed is an example of the first question; asking *why* it failed is an example of the second. I recently proposed the *small-telescopes* approach to answering the first question (Simonsohn, 2015b). It calls a replication a failure when the replication rejects effects big enough to have been detectable by the original study. It combines effect-size estimation with hypothesis testing, and treats underpowered nonsignificant replications as inconclusive rather than as failures.

The small-telescopes approach, however, does not address the second question. Schwarz and Clore (2016) help make this clearer. They highlight the importance of assessing the similarity between original and replication studies. For instance, "A replication failure may arise because the true effect studied in the replication is different from the true effect studied in the original study. . . . Differences in materials, populations, and measures may lead to differences in the true effect under study."

Actually, they did not write that. I did. It is in the closing section of my "Small Telescopes" article (Simonsohn, 2015b, p. 567). My point is that I do not disagree with Schwarz and Clore on the importance of paying attention to differences between original and replication studies—but I do disagree on how to go about it.

Disagreement 1: Manipulation Checks

In "Small Telescopes," I briefly discussed two failures to replicate (Feddersen, Metcalfe, & Wooden, 2012; Lucas & Lawless, 2013) the classic finding that people report being less happy with their lives on rainy days (Schwarz & Clore, 1983). Schwarz and Clore (2016) point out that the large-sample replication studies differed in several ways from their study, and that the former did not measure mood, a variable of key theoretical importance in their study. They propose that "just as original studies do, replications need to ensure that the theoretically specified variables are realized" (p. XXX). Concretely, they argue that replications must include (successful) manipulation checks for the small-telescopes test to make sense. I disagree. Manipulation checks are useful, but not necessary, to diagnose replication failures.

First, empirical findings can be of interest independent of the theory motivating them. Life-satisfaction researchers may be indifferent to the mood-as-information hypothesis motivating the original experiment, and yet be interested in the replicability of the finding that trivial factors, such as the weather, have extremely large effects on measured life satisfaction. (The effect of rain on life satisfaction reported by Schwarz & Clore, 1983, is larger than the documented difference in happiness between people who have recently gotten married and those who have recently been widowed.)[1]

Second, if a failed replication includes a manipulation check, we can better identify where the failure originates, but we do not need such a check to realize that the failure has occurred. The original finding was that weather affects mood, which, in turn, affects life satisfaction. If weather does not affect life satisfaction in a replication study, the original finding has not been replicated. Is it because weather does not really affect mood? Is it because current mood does not really affect self-reported life satisfaction? These are interesting questions, but answering them is not necessary to conclude that the replication has failed. Finkel (in press), for example, rightly treated a Registered Replication Report (Cheung et al., in press) that failed to replicate the manipulation check of an earlier study of his as a failure to replicate the study as a whole.

Third, what if the original manipulation check is a false positive? Many nonreplicable findings presumably include nonreplicable manipulation checks. Consider a

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

different and extreme case: If we cannot replicate Larry Sanna's manipulation checks, should we refrain from concluding that his faked findings cannot be replicated?

Disagreement 2: Stating Versus Testing Hypotheses

Schwarz and Clore propose qualitatively comparing the original and replication studies and leaving our beliefs about the original unchanged if we subjectively decide that the replication is sufficiently different. This allows our motivated minds to find excuses not to update beliefs we do not wish to update (Lord, Ross, & Lepper, 1979; Mahoney, 1977). Koehler (1993) provided direct evidence on this problem. He found that researchers judge studies that contradict their prior beliefs to be of lower quality than studies that confirm their prior beliefs.

Presumably, original authors believe their findings are replicable, so they will find studies failing to replicate the original effect to be of lower quality than those successfully replicating it. Relying purely on original authors' subjective assessment is akin to determining the outcome of sports matches by asking home fans which team they feel played better.

Motivated reasoning is reduced if we test the predictions that follow from the hypotheses we generate to explain why a replication failed. For example, Schwarz and Clore hypothesize that the large samples in the replications of their study were more diverse than the original sample, and hence the data in the replications are noisier. If this hypothesis were correct, the standard deviation of life satisfaction should have been greater in the replication by Feddersen et al. (2012), which used the same 11-point scale, than in the original study. But it was smaller (1.52 vs. 1.69).[2] This hypothesis also predicts that reducing sample variability should bring back the effect, but Feddersen et al. estimated a model with respondent fixed effects (a dramatically less noisy within-subjects analysis), and the impact of weather was still not replicated.

Similarly, Schwarz and Clore hypothesize that weather fluctuation may have been milder in the large-sample studies than in the original study, and thus would have produced smaller effects. In that case, the effect should have been replicated if analysis focused on the subset of days with large weather swings, but Lucas and Lawless (2013) did not replicate the effect even in such analysis. (For a more detailed discussion, see Simonsohn, 2015a.)

Who Decides if the Replication Counts? Each Reader Does

A referee reviewing an earlier version of this manuscript asked a question I think is on many people's minds: Where should the burden of proof lie in deciding whether a replication study is sufficiently similar to the original? Do the original authors need to prove that it is different, or do the replicators need to prove that it is similar?

I would not frame the problem this way, as it positions replications as moderated two-party debates in which preset rules are used to determine the victor. Scientific communications are between authors and *all* readers. Readers' assessments of how compelling authors' arguments are determine "winners."

The burden of proof is on Schwarz and Clore not because they are the original authors, but because they have raised specific post hoc auxiliary hypotheses that are critical to their argument and that lead to falsifiable predictions testable with available data. The need to be empirically compelling is dictated not by their role in the debate, but by their role in society: scientists. Moreover, when proof is so easily accessible (e.g., look up the standard deviations in the relevant tables), proof is not a burden, it is an advantage. Original authors and replicators should make compelling evidence-based arguments. That is what we get paid to do.

Another reservation I have with the "who has the burden of proof?" frame is that it positions replications as historical exercises aimed at understanding single past studies, rather than as scientific exercises aimed at better understanding the world around us. Replications may differ from original studies in ways that make them more informative about the phenomenon of interest. If a replication study eliminates a confound or uses clearer instructions, it is meaningful to ask, "Is the effect replicated once problems in the original study are addressed?" And it is meaningful to use the small-telescopes approach to answer that question.

Original authors sometimes are perceived as moving the goalposts of what constitutes a valid replication, adding post hoc hypotheses for failed replications as needed. But so what? Fields do not move forward when original authors update their beliefs; they move forward when a substantial share of readers do. Readers respond to facts more than to talk.

Replicators Should Identify Differences

For readers to decide on the importance of design differences, they must be aware of them. It is the replicator's responsibility to ensure that this occurs. Unfortunately, the set of differences between any two studies is arguably infinite. A heuristic I now use is to ask: "Would a reader be surprised to find out that X differed between the original and replication study and yet the replicator did not mention X?"

I developed this heuristic after reading the concerns Gilbert, King, Pettigrew, and Wilson (2016) raised with the Open Science Collaboration (2015) not having told

readers about differences in design between original and replication studies—differences that, at least when evaluated outside the full context of the underlying studies, seemed quite surprising (e.g., a scenario involving a honeymoon was used in a replication instead of the original scenario involving military service). The point is not that the presence of these or any surprising difference precludes interpreting a study as a replication; rather, the point is that the presence of surprising differences should be explicitly discussed by the researchers who are interpreting a study as a replication (Simonsohn, 2016).

Going back to the "Small Telescopes" article, I imagine that readers of that article who were unfamiliar with Schwarz and Clore's (1983) original report might not have realized that mood was key for the hypothesis of interest to them, and that it was only the replicators who were intrinsically interested in the effect of weather. I should have mentioned that. Moreover, after publishing my article, I wrote a blog post discussing additional differences across studies and analyzing their potential importance in the failures to replicate (Simonsohn, 2015a). If I were writing my article today, I would mention the differences in the main text and would include the analyses reported in the blog post as a supplement. I would help readers decide rather than decide for them.

Action Editor

D. Stephen Lindsay served as action editor for this article.

Author Contributions

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

Declaration of Conflicting Interests

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

Notes

1. The reported effect of rainy versus sunny day was 1.7 on an 11-point life-satisfaction scale. Lucas (2007) reported a life-satisfaction difference of about 1.5, on the same scale, between people who got married and those who were widowed within the last year (see his Fig. 1).

2. The standard deviation of 1.69 corresponds to the pooled standard deviation, that is, the average of the within-cell standard deviations. (Schwarz & Clore, 1983, did not report standard deviations. I computed them off the reported *t* tests; see Supplement 2 in Simonsohn, 2015b.)

References

Cheung, I., Campbell, L., LeBel, E., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., . . . Yong, J. C. (in press). Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*.

Feddersen, J., Metcalfe, R., & Wooden, M. (2012). *Subjective well-being: Weather matters; climate doesn't* (Melbourne Institute Working Paper Series, Working Paper No. 25/12). Retrieved from http://web.archive.org/web/20160707030801/http://melbourneinstitute.com/downloads/working_paper_series/wp2012n25.pdf

Finkel, E. J. (in press). Reflections on the commitment-forgiveness Registered Replication Report. *Perspectives on Psychological Science*.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*, 1037.

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, *56*, 28–55.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.

Lucas, R. E. (2007). Adaptation and the set-point model of subjective well-being: Does happiness change after major life events? *Current Directions in Psychological Science*, *16*, 75–79.

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

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*, 161–175.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 943. doi:10.1126/science.aac4716

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

Schwarz, N., & Clore, G. L. (2016). Evaluating psychological research requires more than attention to the *N*: A comment on Simonsohn's (2015) "small telescopes." *Psychological Science, 26*, XXX–XXX.

Simonsohn, U. (2015a, November 16). Rain & happiness: Why didn't Schwarz & Clore (1983) 'replicate'? [Web log post]. Retrieved from http://datacolada.org/43

Simonsohn, U. (2015b). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569. doi:10.1177/0956797614567341

Simonsohn, U. (2016, March 3). Evaluating replications: 40% full ≠ 60% empty [Web log post]. Retrieved from http://www.datacolada.org/47