Abstract.
I agree with Schwarz & Clore (pdf) on the importance of considering differences between original and replication studies when interpreting replication failures. I disagree on the proposition that without manipulation checks replications cannot be statistically analyzed as such, and disagree on their approach to considering hypotheses for why a replication failed. They stop after generating hypotheses, I show the benefits of also testing them. I propose a heuristic for deciding which design differences replicators should highlight in their articles. A unifying theme is that replicators are communicating with all readers, not just the original authors; the goal is informing the former rather than persuading the latter.

Word count: 1539
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 of the second. I recently proposed the “Small Telescopes” approach
to answering the first question (Simonsohn, 2015b). It calls a replication a failure when it
rejects effects big enough to be detectable by the original study. It combines effect size
estimation with hypothesis testing, and treats underpowered non-significant 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 didn’t write that. I did. It is the closing paragraph of the “Small
Telescopes” paper. 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,
Metcalfè, & Wooden, 2012; Lucas & Lawless, 2012) 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 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, and propose that “Just as original studies, replications need to ensure that the theoretically specified variables are realized.” (p.xx). Concretely, they argue 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 life-satisfaction measures (The effect of rain on life-satisfaction reported by Schwarz & Clore (1983) is larger than the documented effects of having recently widowed and having recently gotten married, combined).¹

Second, if a failed replication includes a manipulation check, we can better identify where the failure originates, but we do not need it to realize it occurred. The finding is that weather affects mood, which, in turn, affects life-satisfaction. If weather does not affect life-satisfaction, 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 the replication failed.

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

¹ The reported effect of rainy vs sunny day is 1.7 in an 11 point life-satisfaction scale. Lucas (2007) reports a difference of about 1.5 between people who got married vs widowed within the last year; see his Figure 1.
different and extreme case: If we cannot replicate Larry Sanna’s manipulation checks, should we refrain from concluding his faked studies do not replicate?

**Disagreement 2. Stating vs testing hypotheses**

Schwarz & Clore propose qualitatively comparing the original and replication studies and leaving our beliefs about the original unchanged if we subjectively decide 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) provides direct evidence on this problem. He finds that researchers judge studies that contradict their prior beliefs to be of lower quality.

Presumably original authors believe their findings are replicable, so they will find failures to be of lower quality than successes. 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 the large sample replications were more diverse and hence noisier. This predicts the standard deviation of life satisfaction will be greater in the replication. But it is smaller, SD=1.52 vs SD=1.69.\(^2\) It also predicts reducing sample variability should bring back the effect; but with respondent fixed effects (a dramatically less noisy within-subject analysis) the impact of weather was still not replicated.

Similarly, they hypothesized that weather fluctuation may have been milder in the large sample studies leading to smaller effects. This predicts the effect should have re-

\(^2\) SD=1.69 corresponds to the pooled SD, which averages SDs *within* cells (SDs are not reported in the original paper; I computed them off F-stats, see Supplement 2 in Simonsohn, 2015b).
Each Reader

appeared focusing on large-weather contrast days, but it did not. (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 when deciding if a replication is sufficiently similar to the original. Do the original authors need to prove it is different, or the replicators prove that it is similar?

I wouldn’t frame the problem this way, it positions replications as moderated debates between two parties where pre-set 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 (2016) not because they are the original authors, but because they 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 not dictated by their role in the debate, but by their role in society: scientists. Moreover, when proof is so easily accessible (e.g., lookup the SD in the relevant tables) it is not a burden, it is an advantage. Original authors and replicators should make compelling evidence-based arguments, that’s 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 a historical exercise attempting to understand that one past study, rather than as a scientific exercise attempting to better understand 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 eliminates a confound or uses clearer instructions, it is meaningful to ask “Does the effect replicate once problems in the original are addressed?” And it is meaningful to use “Small Telescopes” to answer that question.

Original authors sometimes are perceived as moving the goals-posts of what constitutes a valid replication, adding post-hoc hypotheses for failed replications as needed. But so what? Fields don’t 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 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 X differed between 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 to replicate a study with a 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, that the
presence of surprising differences should be explicitly discussed by those interpreting it as a replication (Simonsohn, 2016).

Going back to the “Small Telescopes” article, I imagine readers unfamiliar with Schwarz and Clore (1983) may have been surprised to learn that mood was key for the hypothesis of interest to them, and that it was only the replicators who were intrinsically interested in the impact of weather. I should have mentioned that. Moreover, after publishing the paper, I wrote a blog post discussing additional differences across studies and analyzing their potential importance in the failures to replicate (http://DataColada.org/43). If I were writing the paper today, I would mention the differences in the main text and would include the analyses reported in the blog post as a supplement. Helping readers decide rather than deciding for them.
References