Authors of target article: Zwaan, R.A., Etz, A., Lucas, R.E., & Donnellan, M. B.

Word counts:
ABSTRACT: 91
MAIN TEXT: 904
REFERENCES: 22
ENTIRE TEXT (TOTAL + ADDRESSES etc.): 1078

Title:
The meaning of a claim is its reproducibility

Full name and institution:
Jan P. de Ruiter
Tufts University
Medford, Massachusetts
United States of America
Email: jp.deruiter@tufts.edu
Telephone: +1-617-627-2531

Abstract

A scientific claim is a *generalization* based on a reported statistically significant effect. The reproducibility of that claim is its scientific meaning. Anything not explicitly mentioned in a scientific claim as a limitation of the claim's scope means that it implicitly generalizes over these unmentioned aspects. Hence, so-called "conceptual" replications that differ in these unmentioned aspects from the original study are legitimate, and necessary to test the generalization implied by the original study's claim.

I commend the authors for carefully addressing some of the canards that have emerged in the recent attempts to downplay the crucial role of replication in psychology. However, they fail to avoid a widespread conceptual confusion that has substantially contributed to the declining reputation of replication in psychology.

In the target article, the words "finding", "result", "effect" and "claim" are used interchangeably, probably for reasons of stylistic variation. That's fine, but it also obscures a number of relevant distinctions. Independent of the words we use, it is important in the context of replication to distinguish between the following concepts: *data*, the raw recordings of the dependent measure(s), *difference*, the descriptive difference between aggregated values of the data for the relevant conditions, and *significant difference*, or *effect*, which is the statistical generalization of an observed difference, demonstrating that the difference cannot be explained by chance alone. The presence of an *effect* is necessary (but not sufficient) for making a *claim*, which is an effect believed to be generalizable to the population and context of interest. Claims either support or undermine theories, which is why we make them in the first place.

Every one of these concepts is an avenue for replication. If we replicate *data*, we essentially double-check for measurement errors or fraud. If we replicate a *difference* found in the data, we double-check for the way the data were aggregated, e.g. to avoid Simpson's paradox (Simpson, 1951). If we replicate an *effect*, we reproduce the statistical procedure that was used to make sure that the difference was not due to chance alone. Finally, and most importantly for the

present discussion, if we do a replication study about a *claim*, we check the generalizability of the effect over the population, task, and other aspects of the context that the claim was about. In that sense, the *reproducibility of a claim is its scientific meaning*. If we establish in a controlled experiment at Tufts University that undergrads in Computer Science perform better at a math test when they've had coffee than when they haven't, the fact that those specific undergrads performed better at that specific math test on that specific day after having that specific amount of that specific type of coffee is not particularly interesting. It is the underlying claim (which one hopes is clearly specified in the study) that people, young adults, or students perform better at math, analytic problem solving, or whatever the claim says, when they've had coffee, or caffeine, or whatever the claim says.

So the discussion of direct vs. conceptual replication, as well as the assessment of the value of a "conceptual replication" can be elegantly addressed once we realize that all the replication of a claim does is explore the generalizability of that claim.

The more general the claim the finding is held to support, the more "conceptual" the replication of the supporting findings can (and should) be. Suppose we have an effect E that we report in order to claim evidence for scientific claim C. Then, if C is *identical* to E, such that C is a claim of the type "The participants in our experiment did X at time T in location L performing task X", it is impossible to replicate that claim because the exact circumstances under which E was found were unique and therefore by definition irreproducible. But in this case (that C = E), C obviously has no generality at all, and is therefore scientifically irrelevant. If, on the other hand, C is more general than E, the level of detail that is provided in the claim should be sufficient to enable readers to attempt to replicate the claim, allowing for variation that the authors do not consider important. If the authors remark that the effect arises under condition A, but acknowledge that it might not arise under condition B (for instance, with participants who are aged 21-24 rather than 18-21), then clearly a follow-up experiment under condition B isn't a valid replication. But if their claim doesn't specify the age for which the claim should hold, then a follow-up study involving condition B is a perfectly legitimate replication. The failure to specify any particular limitation of the claim might reasonably be considered an implicit statement that the claim is so general that changing this aspect in a replication study should not matter.

So assuming the data are accurate and the statistical generalization is solid, if we just use the rule: "whatever isn't specified in the claim is something the claim is generalizing over", we accomplish three (good) things. First, we create an incentive for authors to be more careful in specifying the generalizability of their claims. Second, we make it easier to replicate studies to assess the validity of their claims. And third, we avoid the possible cop-out for authors of nonreplicated studies that the study didn't replicate because of "unknown moderator variables". If these variables were not excluded in the original study by limiting the generality of the claim, they can't be invoked to discredit a failed replication.

A possible argument against the proposed rule is that it becomes much harder to make claims that hold up under replication. My response to that argument is that this is not a bug but a feature. Finding general effects in psychology *is* very difficult, and it would be a good first step to address our replication crisis if we stopped pretending it is not.

REFERENCES

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological), 13*(2), 238-241.