Commentary on Zwaan, Etz, Lucas and Donnellan target article, Behavioral and Brain Sciences (submitted)

17 January, 2017

Word counts
Abstract: 60
Main text: 995
References: 217
Entire text: 1366

Three Strong Moves to Improve Research and Replications Alike

Roger Giner-Sorolla, University of Kent
Email: rsg@kent.ac.uk Web: https://www.kent.ac.uk/psychology/people/ginerr/

David M. Amodio, University of Amsterdam, New York University
Email: david.amodio@gmail.com Web: http://amodiolab.org/

Gerben A. van Kleef, University of Amsterdam
Email: G.A.vanKleef@uva.nl Web: http://home.staff.uva.nl/g.a.vankleef

Corresponding author contact:

Professor Roger Giner-Sorolla
School of Psychology (Keynes E3.6)
University of Kent
Canterbury, Kent CT2 7NP
United Kingdom
tel. +44 (0)1227 823085

Abstract

We suggest three additional improvements to replication practices. First, original research should include concrete checks on validity, encouraged by editorial standards. Second, the reasons for replicating a particular study should be more transparent, and balance systematic positive reasons with selective negative ones. Third, methodological validity should also be factored into evaluating replications, with methodologically inconclusive replications not counted as non-replications.

Commentary Text

Although we largely agree with Zwaan, Etz, Lucas & Donnellan's (this issue; henceforth ZELD) analysis, we want to add to it, drawing on our experiences with replications as authors and editors. Over the past years in psychology, successful reforms have been based on concrete suggestions with visible incentives. We suggest three such moves that ZELD might not have considered.

Anticipate Replication in Design

In answering concerns about context variability, ZELD suggest that original authors' reports should be more detailed and acknowledge limitations. But these suggestions miss what lets us meaningfully compare two studies across contexts: calibration of methods, independent from the hypothesis test.

Often, suspicions arise that a replication is not measuring or manipulating the same thing as the original. For example, the Reproducibility Project (Open Science Collaboration, 2015) was criticized for substituting an Israeli vignette's mention of military service with an activity more common to the replication's US participants (Gilbert, King, Pettigrew & Wilson, 2016). All the methods reporting in the world cannot resolve this kind of debate. Instead, we need to

know whether both scenarios successfully affected the independent variable. Whether researchers have the skill to carry out a complex or socially subtle procedure is also underspecified in most original and replication research., surfacing only as a doubt when replications fail.

Unfortunately, much original research doesn't include procedures to check that manipulations affected the independent variable, or to validate original measures. Such steps can be costly, especially if participant awareness concerns require a separate study for checking. Nevertheless, the highest standard of research methodology should include validation that lets us interpret both positive and negative results (Giner-Sorolla, 2016; Le Bel & Peters, 2011). Although the rules of replication should allow replicators to add checks on methods, such checks should also be a part of original research. Specifically, by adopting the Registered Report publication format (Chambers, Dienes, et al., 2015), evaluation of methods precedes data collection, so that planning to interpret negative results is essential. More generally, publication decisions should openly favor studies that take the effort to validate their methods.

Discuss and Balance Reasons to Replicate

Providing a rationale for studying a particular relationship is pivotal to any scientific enterprise, but there are no clear guidelines for choosing a study to replicate. One criterion might be importance: theoretical weight, societal implications, influence through citations or textbooks, mass appeal. Alternatively, replications may be driven by doubt in the robustness of the effect. Currently, most large-scale replication efforts (e.g., Ebersole, Atherton, et al. 2016; Klein, Ratliff, et al., 2014; Open Science Collaboration, 2015) have chosen their studies either arbitrarily (e.g., by journal dates) or by an unsystematic and opaque process.

Without well-justified reasons and methods for selection, it is easy to imagine doubt motivating any replication. Speculatively, many individual replications seem to be attracted by a profile of “surprising results, weak theory and methods.” But if replications hunt the weak by choice, conclusions about the robustness of a science will skew negative. This problem is compounded by the psychological reality that findings that refute the status quo (such as failed replications) attract more attention than findings that reinforce the status quo (such as successful replications).

Replicators (like original researchers) should provide strong justification for their choice of topic. When replication is driven by perceptions of faulty theory or implausibly large effects, this should be stated openly. Most importantly, replications should also draw on selection criteria *a priori* based on positive traits, such as theoretical importance, or diffusion in the academic and popular literature. Indeed, we are aware of one attempt to codify some of these traits, but it has not yet been finalized or published (Lakens, 2016).

Although non-replication of shaky effects can be valuable, encouragement is also needed to replicate studies that are meaningful to psychological theory and literature. Importance could be one criterion of evaluation for single replication articles. Special issues and large-scale replication projects could be planned around principled selection of important effects to replicate. The CREP student replication project (CREP, 2018), for example, chooses studies for replication based on a priori citation criteria.

Evaluate Replication Outcomes More Accurately

The replication movement also suffers from an underdeveloped process for evaluating the validity of its findings. Currently, replication results are reported and publicized as a success or failure. But “failure” really represents two categories: valid non-replications and invalid (i.e., inconclusive) research. In original research, a null result could reflect a true lack of

effect, or problems with validity (a manipulation or measure not being operationalized precisely and effectively). Validity is best established through pilot testing, manipulation checks, and the consideration of context, sample, and experimental design, and evaluated through peer review. If validity is inadequate, then the results are inconclusive, not negative.

Indeed, most replication attempts try hard to avoid inconclusive statistical outcomes, often allotting themselves stronger power than the original study. But there has not been as much attention to identifying inconclusive methodological outcomes, such as when a replication's manipulation check fails, or a method is changed in a way that casts doubts upon the findings. One hindrance is the attitude, sometimes seen, that direct replications do not need to meet the same standards of external peer review as original research. For example, the methods of the individual replications in Open Science Collaboration (2015) were reviewed only by one or two project members and an original study author, pre-data collection.

Conclusion and Recommendations

Reasons for replicating a particular effect should be made transparent, with positive, systematic methods encouraged. Replication reports and original research alike should include evidence of the validity of measures and manipulations, with standards set before data collection. Methods should be externally peer reviewed for validity by experts, with clear consequences (revision, rejection) if they are judged inadequate. And, when outcomes of replication are simplified into "box scores", they should be sorted into three categories: replication, non-replication, and inconclusive. By improving the validity of replication reports, we will strengthen our science, while offering a more accurate portrayal of its state.

References

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: realigning incentives in scientific publishing. *Cortex*, *66*, A1-A2.

CREP (2018). Current study list and selection methods. Retrieved from https://osf.io/flaue/wiki/home/.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68-82.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037-1037.

Giner-Sorolla, R. (2016). Approaching a fair deal for significance and other concerns. *Journal of Experimental Social Psychology*, *65*, 1-6.

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., ... & Cemalcilar, Z. (2014). Data from investigating variation in replicability: A "Many Labs" Replication Project. *Journal of Open Psychology Data*, *2*(1).

Lakens, D. (2016). The replication value: What should be replicated? Retrieved from http://daniellakens.blogspot.co.uk/2016/01/the-replication-value-what-should-be.html

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*(4), 371-379.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.