replication), but instead focuses on a seemingly minor point that was mentioned in the target article (i.e., the outlet in which the debate takes place). However, I am convinced that the outlet is of critical relevance here.

First and most important, I believe that most mainstream scientists still read scientific journals more frequently and more intensely than they follow social media. Thus, it is simply more efficient to publish fresh ideas in journals to gain optimal access to "the silent majority" whom authors would like to convince. A perfect example here is the success of the "False-Positive Psychology" article published in *Psychological Science* (Simmons et al. 2011; see also Simmons et al. 2018). A few additional examples that readily come to mind are the publication of the results of the "Replication Project: Psychology" in *Science* (Open Science Collaboration 2015), the – regrettably renamed – "Voodoo Correlations" paper in *Perspectives on Psychological Science* (Vul et al. 2009), the "Scientific Utopia" article in *Psychological Inquiry* (Nosek & Bar-Anan 2012), and the mind-boggling "Political Diversity" paper in *Behavioral and Brain Sciences* (Duarte et al. 2015).

Of course, it is certainly difficult and all too often very frustrating to try to publish innovative ideas or critiques of established theories in journals because the thorny peer-review process sometimes seems to be abused by established scholars in their roles of reviewers and editors in efforts to block innovations and criticism. By contrast, all ideas can quickly and without filtering be published in blogs, and there have been several additional clever arguments put forward in favor of blogs over journals (e.g., open data, code, and materials, open reviews, no eminence filter, better error correction, and open access; Lakens 2017). On the other hand, established scholars sometimes complain about, for example, a lack of reflection, a lack of peer advice, impulsivity, personalized debates, and personal accusations triggered by the features of social media. Although I believe that the "tone debate" has been largely exaggerated – "Don't dish it out if you can't take it" – there is some evidence that intellectual opponents and especially third parties might be more efficiently convinced if the arguments are presented in a friendly tone. Thus, the more formal and down-to-earth tone used in scientific journals might in fact be helpful for convincing others. Similarly, mainstream journals are, in general, still more highly respected than most social media outlets. Thus, especially more conservative scholars will trust arguments exchanged in journals more than those that come from debates fought out in blogs.

This should by no means be interpreted to mean that blogs and social media do not have their merits in the replication debate and beyond. To the contrary: They are fast, they are subjective, they are mostly short and to-the-point, they may be provocative, and so forth. My argument is instead that the important debates in our discipline (e.g., whether and how to replicate) should not be restricted to these media but should also be published in established mainstream journals. Although such journals are necessarily somewhat slower, they offer another form and style and can potentially present a more elaborated form of the argument. If one mainstream journal rejects your paper, please try another (and so on). There are also newly founded – not yet so well-established – journals such as *Collabra*, *Metapsychology*, or *Advances in Methods, and Practices in Psychological Science* (to name just a few) that might be alternatives in the face of repeated publication failure in more traditional journals.

Taken together, the formal publication of well-crafted and clever articles (e.g., this one on replication in BBS) seems to offer the best and most efficient way to reach a maximal audience and especially to convince as yet undecided individuals to, for example, join the replication movement in order to make replication mainstream, thereby providing one contribution (out of many possible ones) to psychology's renaissance (Nelson et al. 2018).

## A pragmatist philosophy of psychological science and its implications for replication

doi:10.1017/S0140525X18000626, e127

Ana Gantman, Robin Gomila, Joel E. Martinez, J. Nathan Matias, Elizabeth Levy Paluck, Jordan Starck, Sherry Wu, and Nechumi Yaffe

*Department of Psychology, Princeton University, Princeton, NJ 08544.*
**agantman@princeton.edu rgomila@princeton.edu joelem@princeton.edu jmatias@princeton.edu epaluck@princeton.edu jstarck@princeton.edu jueyuw@princeton.edu myaffe@princeton.edu**
**anagantman.com**
**www.robingomila.com**
**http://socialbyselection.wordpress.com/**
**https://twitter.com/natematias**
**www.betsylevypaluck.com**
**www.sherryjwu.com**

**Abstract:** A pragmatist philosophy of psychological science offers to the direct replication debate concrete recommendations and novel benefits that are not discussed in Zwaan et al. This philosophy guides our work as field experimentalists interested in behavioral measurement. Furthermore, all psychologists can relate to its ultimate aim set out by William James: to study mental processes that provide explanations for why people behave as they do in the world.

A pragmatist philosophy of psychological science offers to the direct replication debate concrete recommendations and novel benefits that are not discussed in Zwaan et al. Pragmatism starts from the premise that "thinking is for doing" (Fiske 1992). In other words, pragmatic psychological theories investigate the mental processes that predict observable behavior within the "rich thicket of reality" (James 1907, p. 68). This philosophy guides our work as field experimentalists interested in behavioral measurement. Furthermore, all psychologists can relate to its ultimate aim set out by William James: to study mental processes that provide explanations for why people behave as they do in the world.

***Recommendations.*** A pragmatist philosophy of science urges scientists to observe what behaviors emerge in the complexity of real life; it encourages active theorizing about individuals' contexts and the way that individuals construe or interpret them. Specifically, direct replications should research the context of the planned replication site (i.e., James's "thicket of reality") to determine when it is appropriate to use the precise materials of previous experiments and when researchers should translate materials at the new site so that they will replicate the original participants' construal (Paluck & Shafir 2017). Some methods for documenting context and adapting studies include well-designed manipulation checks, pretesting, reporting on the phenomenological experience of participants in any intervention, and collaboration with those who have actually implemented previous studies. An additional recommendation we propose is statistical: Investigators should statistically characterize the field, meaning that every study should report the amount of explained *and* unexplained variance of the treatment effect. In this way, replications and original findings can be explicitly situated by both the effect size and the amount of "noise" (e.g., from measurement error or unmeasured construal, context, and individual differences) that might help identify the source of differences across studies (Martinez et al. 2018).

***Benefits.*** A pragmatist approach draws out the creativity and rigor of replication research. For example, when conducting a replication of a field experiment at a new site, the question of whether to use the same materials or to create translated (construal-preserving) materials arises. Field replications create the most obvious opportunities to develop rigorous standards that describe and compare research settings. These standards could be adopted

by researchers working in many settings. Researchers can break new ground by developing these methodological standards, as opposed to basing replication decisions on unstated assumptions about context similarity. Theorizing the context of a proposed replication also entails creative theoretical integration in our highly differentiated field; specifically, the integration of theories that pertain to context (to situation, identity, culture, and perception) with the focal theory that is to be tested with the replication. Additionally, reporting the total unexplained and explained variance from a study is an explicitly cumulative exercise aimed at meta-analysis. Emphasizing measurement as a point of comparison between studies also addresses the chronology problem (Zwaan et al., sect. 5.1.1) in which studies that are "first" to ask a particular question are prioritized over replications.

Field researchers, who regularly face the challenge of theorizing a broader context, may have a larger leadership role in developing conventions of direct replication than implied by Zwaan et al., who predict fewer replications of field versus laboratory studies. For example, in the digital space, replications of marketing and media experiments proceed at a scale that vastly outstrips normal academic research. These studies represent enormous opportunities to examine the impact of context on causal relationships (Kevic et al. 2017). In the policy world, Campbell's vision for the experimenting society (Campbell 1969; 1991) lays out steps for cost-efficient and politically feasible replication of studies across real-world settings. Such experiments feature contextual variation of deep theoretical importance, including differing levels of economic inequality, demographic diversity, and political contestation (for an example, see Dunning et al., in press). Finally, articles based on field experimental replications can be models of compelling scientific writing, combating claims that replication research is rote and boring, because field studies lend themselves to a rich description of place, participants, history, and more generally the psychological and behavioral equilibrium into which a social scientist intervenes (Lewin 1943/1997).

## Don't characterize replications as successes or failures

doi:10.1017/S0140525X18000638, e128

Andrew Gelman

*Department of Statistics, Columbia University, New York, NY 10027*

gelman@stat.columbia.edu

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.

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 that it is an error when, following a disappointing replication result, proponents of the original published studies "irrationally privilege the chronological order of studies more than the objective characteristics of those studies when evaluating claims about quality and scientific rigor" (sect. 5.1.1, para. 3). As a remedy to this fallacy I have proposed a "time-reversal heuristic" (Gelman 2016b): the thought experiment of imagining the large, pre-registered 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 could 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 of 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 of the data. Conversely, if we do not have access to raw data, many published results are close to useless, and when there is a high-quality pre-registered replication, I would be inclined to pretty much ignore the original paper, rather than, say, to assume the truth lies somewhere 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 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 of 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 & Carlin 2014). Related is the use of expressions such as "failed replication." I have 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 $p<.05$? 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 & Simonsohn 2015). If there is a need to count replications in a larger study of studies such as the Open Science Collaboration, I would 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 United Kingdom or Japan, for that matter, then a replication carried out with these samples would be considered direct" (sect. 4, para. 3). 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 & Beall 2014). Also there is no end to the differences between the United Kingdom 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 in citing Schmidt (2009)