Published in final edited form as:

J Exp Soc Psychol. 2016 September ; 66: 145–147. doi:10.1016/j.jesp.2016.01.006.

How to Publish Rigorous Experiments in the 21st Century

Susan T. Fiske

Princeton University

Abstract

Crises provide an opportunity for the field to take stock, as do the articles in this special issue. Constructive advice for 21st century publication standards includes appropriate theory, internal validity, and external validity. First, well-grounded theory can produce a priori plausibility, testable logic, and a focus on the ideas involved, all cumulatively informed by meta-analysis across studies. Second, internal validity benefits from both exploratory work and confirmatory analyses on well-powered samples that require systematic detection and principled decisions about data quality. Inferences benefit from manipulated mediation analysis and from careful interpretation without over-claiming. Finally, external validity profits from a variety of exact and conceptual replications, best evaluated by meta-analysis.

Keywords

experiments; replicability; internal validity; external validity; theory

Social psychology has too many flashy demonstrations of counterintuitive and cute, but unreliable, phenomena: Sound familiar? Social psychology's previous crisis worried about issues related to the current crisis. The earlier (1970s) crisis focused on scientific standards, generalizability, and real-world relevance. The current one also focuses on scientific standards, but mostly replicability. In both cases, the benefits of the critiques, symposia, journal special issues, and task forces were and are to focus us on specifying psychological mechanisms, reducing experimenter bias, reporting effect sizes and power analysis, considering context as a generalizability concern, and testing our ideas in applied settings. Although panic, dismay, and despair were and are not uncommon reactions (several people in my graduate cohort apparently left the field for related reasons—and I worry about the same exodus now), the net effect was and will be constructive. Such crises allow the field to self-correct, as this *JESP* special issue on experimental rigor suggests. But the previous events hold a further lesson for the current ones: Crises do not require the field to adopt rigid rules or conduct inquisitorial investigations.

This essay explores the special issue's proposed, peer-reviewed solutions, with advice to current social psychological scientists about how to improve their work's theory and

sfiske@princeton.edu.

conceptualization, internal validity, and external validity. Along the way, the essay describes the emerging norms for how to publish—and get cited favorably—in today's context.

Theory and Conceptualization

Novices often arrive at a professor's office, armed with an intriguing method but no concept. Careful discussion and additional reading usually reveal an implicit theory that can be informed by the literature and guided by extant theory. Better still is the graduate student whose exposure to the literature reveals a gap to explore and test. Ideas matter first and foremost. Without concepts, even our best methods would just randomly sample reality, not a good route to cumulative science.

Across the many sources of ideas (Fiske, 2004, 2014), systematic theory provides the framework on which we hang our studies. Together, the studies inform the viability of the theory. As Stroebe (2016) argues, theory also predicts (informs the prior probability of) particular results. Thus, social psychology is not a-theoretical, observational research. The critique of epidemiology that "most published findings are false" does not likely hold true for theory-based experimental social psychology (except perhaps for our limited to one-shot and most surprising results). The distinction between a data- versus theory-driven field matters to experimental reliability because theory informs expected results.

Theory also suggests assembling cumulative results in meta-analyses to test the theory across studies. (This increased leverage augments advice to use sufficient power in the original individual studies.) Stroebe also suggests that mass replication projects will be less than informative because the chosen studies are not representative of the field (often being simple scenario studies). Also, he argues, mass replication does not allow conclusions about the validity of social psychology because of contextual and individual moderators. The implications for an individual investigator are to work from a priori theory, replicate before publishing, and meta-analyze cumulative results from one's own and preferably others' relevant studies.

As Crandall and Sherman (2016) note, the unit of analysis is the idea, not the single effect. We want to know whether the idea is generalizable, not as much whether the specific operationalization is, though that is useful too. For testing theory, they emphasize the role of conceptual replication (shifts in specific operations): If converging operations all support the idea, that is even more convincing than if repeated exact replications do, as Campbell and another Fiske noted decades ago (1959) regarding the strength of multiple methods. Conceptual replication serves as a corrective to method-specific effects and an obsessive focus on direct or exact replication. To be sure, the collective process is robust, with tolerance for both approaches. For an individual author, this suggests that papers have multiple studies, including both exact and certainly conceptual replications.

Starting from theory requires conceptual rigor in analyzing exactly what to test. As Schaller (2016) explains, conceptual clarity and precision can come from constructing the series of if-then steps logically implied by the theory and testing each in turn. Making the logic explicit a priori decreases an experimenter's later wiggle room for reinterpreting unruly results.

Generating detailed hypotheses as if-then steps also helps to depersonalize the theory from a protected personal possession to a set of collectively testable propositions. Experimenters will be more convincing if they follow this advice.

Internal Validity

Theory testing includes initial descriptive mapping of an area, becoming familiar with the territory, encountering relevant phenomena, investigating likely venues. Pilot testing and pre-testing have a role to play, before plunging into fully powered formal tests, according to Sakaluk (2016)'s scheme to "explore small, confirm big." (He argues that null-hypothesis significance testing can be used at the exploratory stage but has limited utility without confidence intervals—unless it is required by a focus on dichotomous outcomes.) Instead, he favors small-scale exploratory work, followed by large-N confirmatory work with pre-registered hypotheses. He illustrates with social-psychology-relevant Big Data: replication projects, large on-line samples, international collaborations, Google Correlate queries, social-media analyses, as well as meta-reanalysis coding for study characteristics that moderate effect sizes. Implications for investigators are clear: Exploratory work is useful, but further large-scale confirmation will convince even more of the audience.

Confirming big is not without hazards. One source of large Ns, online participants require quality controls to monitor careless or insufficient-effort responding (Curran, 2016). Various detection strategies include using response time to detect too-rapid (or too-slow) completion, analyzing long strings of identical responses, identifying univariate and multivariate outliers, assessing individual participant consistency (e.g., by comparison of odd and even responses, psychometric consistency, semantic consistency, individual standard deviations), and attention checks. Each strategy should specify a priori criteria for exclusion, to minimize ad hoc decisions influenced by confirmation bias.

Besides monitoring shirkers, perhaps one could motivate them. Unfortunately, incentives seem ineffective to improve response efforts (Berinsly, Margolis, & Sances, 2016). Inducements improve just responses to attention-check items, but not the main results. Training, incentivizing, pleading, and thanking do not have effects beyond the specific item. The researcher thus faces a dilemma: Excluding shirkers threatens external validity. And inducements by themselves increase attrition, also threatening external validity. (Some evidence suggests that low-effort responders are disproportionately lower income or minority.) At the same time, keeping shirkers in the dataset threatens internal validity because they are not attending sufficiently to stimuli and measures to give meaningful responses. Choosing between these alternatives entails being explicit about the tradeoffs.

Assuming good-enough data quality, internal validity benefits by specifying psychological mechanisms. Social psychologists' preferred strategy is mediation analysis. However, as Pirlott and MacKinnon (2016) remind us, these methods are correlational, so the causal role of the mediator remains ambiguous. As one alternative, one can manipulate the mediators, along with the main independent variables. A variety of strategies (block/enhance; encourage/discourage; one-study/separate-study designs) all can benefit programmatic research.

Given a set of results, inference follows. Whether the conclusions follow from the data is the last step in internal validity. Hales (2016) points out that, in the effort to report effects that are both significant and interesting, researchers may go beyond what the data allow. Over-claiming takes forms beyond the familiar Type I (false positive) and Type II (false negatives) errors. A proposed Type III error describes reaching an accurate conclusion but by flawed methods (e.g., confirmation bias, hypothesizing after results are known, discarding data). A proposed Type IV error describes reaching an accurate conclusion based on faulty evidence (insufficient power, invalid measures). Remedies include tracking one's own questionable research practices (e.g., ad hoc stopping, nondisclosure of failed replications, exploration reported as confirmation) or calculating the plausibility of one's data (e.g., checking for experimenter bias during analysis). Preregistration and transparency are encouraged.

Also focused on the internal validity issue of invalid interpretation, another approach to deal with over over-claiming trusts the researcher's motives and judgment much less than others here do. Jussim, Crawford, Anglin, Stevens, and Duarte (2016) disagree that science is self-correcting and propose that those who share their perspective work to unmask other researchers' hidden phenomena that conflict with an author's preferred narrative. "Masking" implies authorial malignant intent and suggests that researchers cannot be trusted to uphold standards. The proposed solution is for enterprising others (apparently beyond the peer-review process) to police and call-out perceived violations, as Jussim et al.'s article does. This approach sees the scientific glass as half empty. It ignores legitimate disagreement about which is a focal finding. In itself, this approach is also vulnerable to motivated reasoning. For authors, the implication is to be as transparent as possible. For those who want to join the unmasking approach, transparency suggests checking with authors before publishing an attack, which could be mistaken. Another suggestion for would-be unmaskers: A constructive approach would be to go beyond just criticism to one's own empirical tests and claims about the supposed core phenomenon of interest. Transparency is good advice for everyone.

External Validity

If internal validity is suspect, an improved replication adds clarity, as well as external validity. If generalizability is suspect, conceptual replication tests that. Much of the impetus for this special issue is a perceived crisis in replicability. We have seen that both theoretical and internal-validity concerns can contribute. Here we see various ways to address external validity.

Huffmeier, Mazei, and Schultze (2106) offer a replication typology: exact, close, constructive, conceptual (lab and field). Each has a role in progressively moving from repeating the same study to broader implications, from alternative explanations to moderators to real-world applications. An exact replication, conducted by the original authors, strives to repeat the original as much as possible, guarding against false positives. A close replication, conducted by independent researchers, strives to replicate the original, but will likely deviate in more respects than the exact replication, guarding against false positives that result from researchers owning the original study. A constructive replication aims to replicate the original but adds new element(s) for theoretical or methodological

reasons; these studies not only guard against false positives but also strive to discover moderators or mechanisms. Conceptual replications in the lab adhere to the same theory and hypotheses but use at least some new operationalizations. Finally, conceptual replications in the field generalize the most and often move from experimental to correlation methods, but with the advantage of larger sample sizes. A comprehensive, convincing program of research should consider each progressive type of replication.

Unpacking further the distinction between exact and conceptual replications, Fabrigar and Wegener (2016) note that successful conceptual replication helps build construct validity. But the evidentiary value of conceptual-replication failure is ambiguous precisely because the operations differ. Even with exact replications and certainly with conceptual ones, the evidence is only as good as the psychometric equivalence to the original. For example, factor structures and reliabilities need to be comparable. Moreover, even the statistical meaning of replication is ambiguous. Evaluating a replication does not require a necessarily significant result in the same direction as the original; even a nonsignificant same-directional result can add evidentiary value in a meta-analysis. Using meta-analysis to evaluate credibility allows cumulative estimates of an effect size. Individual researchers can do this even for their own research program.

At the far end of the replication continuum is field replication, recommended by Maner (2016). Field research tests external validity the most. Usually, participants are more diverse; extraneous variables are less controlled, and measures often include actual behavior. Maner argues that field research also reduces researcher degrees of freedom and publication bias because of usually having fewer measures, less daily control, and more investment. Arguably, these same factors could motivate field researchers to search for some result, any result to report. Field trials typically are not published if they yield no results. Nevertheless, the value of field replication and extension is unique.

A clearly distinctive external-validity proposal is prepublication close replication conduced simultaneously in multiple labs, as illustrated by Uhlmann’s massively collaborative pipeline project (Schweinsberg et al., 2016). In their view, a main advantage is limiting researcher degrees of freedom by pre-commitment regarding the main theoretical focus, to keep the logic from becoming slippery. Not every research program can do this. Acknowledging limits to the type of study that can be crowd-sourced, of course, would eliminate time-consuming or high-expertise research. Researchers might also worry about being scooped, but there would be a lot of witnesses to one’s claim to priority. This innovative method is not for the faint-hearted.

Conclusion

Norms about acceptable research methods change by social influence, not by regulation. As social psychology tells us, people internalize change when they trust and respect the source. A punishing, feared source elicits at best compliance and at worst reactance, not to mention the source’s own reputational damage. For the most part, the proposals in this special issue are persuasive communications, not threats. And all are peer-reviewed, not mere blog posts. And they are mostly reasoned advisory proposals, not targeted bullying. As such, they

appropriately treat other researchers as colleagues, not miscreants. This respectful discourse moves the field forward better than vigilantism. The authors and editors are to be commended for constructive contributions, and other researchers would do well to heed most of them.

References

Berinsky A, Margolis M, Sances M. Can we turn shirkers into workers? Journal of Experimental Social Psychology. 2016

Campbell DT Fiske DW Convergent and discriminant validation by the multitrait-multimethod matrix Psychological Bulletin 1959 56 2 81–105 http://dx.doi.org/10.1037/h0046016 [PubMed: 13634291]

Crandall CS, Sherman JW. On the scientific superiority of conceptual replications. Journal of Experimental Social Psychology. 2016

Curran PG. Methods for the detection of carelessly invalid responses in survey data. Journal of Experimental Social Psychology. 2016

Fabrigar LR, Wegener DT. Conceptualizing and evaluating the replication of research results. Journal of Experimental Social Psychology. 2016

Fiske ST Mind the gap: In praise of informal sources of formal theory Personality and Social Psychology Review 2004 8 132–137 [PubMed: 15223512]

Fiske ST Reis HT Judd CM Scratch an itch with a brick: Why we do research Handbook of research methods in social and personality psychology 2014 1–7 Cambridge University Press New York 2nd ed

Hales A. Does the conclusion follow from the evidence? Recommendations for improving research. Journal of Experimental Social Psychology. 2016

Huffmeier J, Mazei J, Schultze T. Reconceptualizing replication as a sequence of different studies: A replication typology. Journal of Experimental Social Psychology. 2106

Jussim L, Crawford JT, Anglin SM, Stevens ST, Duarte JL. Interpretations and methods: Towards a more effectively self-correcting social psychology. Journal of Experimental Social Psychology. 2016

Maner JK. Into the wild: Field studies can increase both replicability and real-world impact. Journal of Experimental Social Psychology. 2016

Pirlott A, MacKinnon DP. Design approaches to experimental mediation. Journal of Experimental Social Psychology. 2016

Sakaluk JK. Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research. Journal of Experimental Social Psychology. 2016

Schaller M. The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). Journal of Experimental Social Psychology. 2016

Schweinsberg M. et al. The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. Journal of Experimental Social Psychology. 2016

Stroebe W. Are most published social psychological findings false? Journal of Experimental Social Psychology. 2016