failure to replicate will ruin their reputation, it is imperative that we do not treat failures to replicate (outside of extreme circumstances) as having reputational consequences.

Although it is perhaps not feasible for every scientist to fully buy in to the "you are not your data" mantra, it is nonetheless important to increase its aggregate influence. To this end, scientists who demonstrate a willingness to divorce themselves from their data should be celebrated (see https://lossofconfidence.com/ for a group focused on this very thing). Awards and accolades should go to scientists who, beyond having a significant influence on their respective fields, can also provide evidence of identifying with the *process* of science and the pursuit of truth (e.g., via dedication to open science or revision of a previous stance based on new data; see Nosek et al. [2012]). Prestigious academic positions should be given to those who do research that is both impactful *and* sound (a notion that seems sufficiently obvious, but that does not necessarily correspond to the selection of individuals who have successfully created a "brand"). Finally, the significance of valuing the truth should be emphasized to graduate students and future generations of scientists, particularly in cases when the relaxing of scientific values is expedient. Ultimately, making replications mainstream will be easier if scientific incentive structures begin to align with a separation of identity and data.

ACKNOWLEDGMENTS

For comments on an earlier version, I thank Nathaniel Barr, Adam Bear, Shadi Beshai, Michal Bialek, Justin Feeney, Jonathan Fugelsang, Gordon Kraft-Todd, Srdan Medimorec, Sandeep Mishra, David Rand, Paul Seli, Nick Stagnaro, and Valerie Thompson.

The importance of exact conceptual replications

doi:10.1017/S0140525X18000821, e146

Richard E. Petty

Department of Psychology, Ohio State University, Columbus, OH 43210.
petty.1@osu.edu
https://richardepetty.com/

Abstract: Although Zwaan et al. argue that original researchers should provide a replication recipe that provides great specificity about the operational details of one's study, I argue that it may be as important to provide a recipe that allows replicators to conduct a study that matches the original in as many conceptual details as possible (i.e., an *exact conceptual replication*).

Zwaan et al. make the classic distinction between *exact replications* (using the same operations as in an original study) and *conceptual replications* (using different materials to instantiate the independent variables [IVs] and/or dependent variables [DVs]). They argue that exact replications are superior and therefore original authors should provide a "replication recipe" providing considerable detail about the specific operations used so others can duplicate one's study. Furthermore, Zwaan et al. claim that a finding is "not scientifically meaningful until it can be replicated *with the same procedures* that produced it in the first place" (sect. 6, para. 1). Instead, I argue that for much theoretical work in psychology, use of the same *operations* is not what is critical, but rather instantiation of the same *concepts*. Thus, theory testing researchers should emphasize conducting *exact conceptual replications* (ECRs) where the goal is to repeat as closely as possible not the precise methods of the original study, but to instantiate the same conceptual meaning of the original variables in the same conceptual context (Petty 2015).

In the physical sciences, the emphasis on carefully replicating operations is often reasonable. For instance, when mixing hydrogen and oxygen to create water, the choice of operations to represent the hydrogen and oxygen concepts is constrained because there is a tight link between concepts and operations (i.e., the operations and concepts are basically the same). Furthermore, the operations chosen are likely to represent the concepts across virtually all contexts. Thus, if you reasonably do the same thing, you should get the same result. In contrast, in many theory testing psychology studies, the choice of operations to represent concepts is vast and the link between the two is not assured. Thus, conducting a replication that is as close as possible to the original study will not necessarily help with replicability because the meaning of the original IVs and DVs in the new context may have changed.

Consider a psychologist mixing a credible source with a persuasive message to produce favorable attitudes toward some proposal. When Hovland and Weiss (1951) did this, Robert Oppenheimer was used as a credible source, and the Russian newspaper, *Pravda*, was the low credible source on the topic of building atomic submarines. Oppenheimer produced more favorable attitudes than *Pravda*. It seems unlikely that the same operations would produce the same result today. Does this render the original study scientifically meaningless? No. The initial result is meaningless only if you cannot conduct an ECR. ECRs are important because what we ultimately want to know is not whether Oppenheimer produces more favorable attitudes toward submarines than *Pravda*, but whether credibility affects persuasion.

The initial credibility study results would be meaningful if the study can be replicated in an ECR. Original authors can specify the criteria any replication study should meet. Namely, provide the *conceptual recipe*. This differs from the *operational recipe* that Zwaan et al. favor. Thus, if manipulating credibility, instead of only articulating operational details like replicators must have people see an 8 X 10 picture of the source with an 18 word description, original authors could also indicate that the high credibility manipulation should produce a rated level of credibility of 7 on an 11 point credibility scale and the low credibility condition should be at 4. But, it is not sufficient for replicators to produce a successful manipulation check. If the original study had high and low credibility means of 7 and 4 but the replication study had means of 1 and 4, the manipulation check in the replication study would seem "successful" (and the effect size of the manipulation check might be comparable to the original), but, the placement of the manipulation along the credibility continuum would be quite different and thus inappropriate for an ECR. In addition to providing information about the statistical properties of the IV manipulation check, original authors should specify what constructs the IV should *not* vary. Thus, original authors should not only provide the IV information just noted, but also what concepts should be assessed to ensure they are not confounded (e.g., source attractiveness and power).

Critically, similar arguments apply to the DV. In the chemistry example, the dependent variable (water) is easily assessed. However, there are multiple ways to assess favorable evaluations (e.g., explicit vs. implicit measures). Now consider a different original study in which investigators are examining the frustration to aggression link. These researchers should indicate how to determine if the dependent measure taps aggression. The original study might have measured how many teaspoons of hot sauce were administered, but in a replication attempt in Mexico, giving hot sauce may not signal aggressiveness. Thus, specifying what criteria the DV should meet (to gauge its conceptual meaning) is as important as specifying this for the IVs. For example, participants can rate how aggressive it is to give hot sauce.

Finally, the overall level of the DV on the aggression continuum in the new context is important. This is because unlike the chemistry example where there is only one way to produce water, psychology DVs are often multiply determined. There are many ways to produce aggression and there may be factors in the replication context that affect the hot sauce DV that were not present in the original research. Some of these may be alternative causes of

aggression (e.g., hot temperature), but others may influence giving out hot sauce for other reasons (e.g., its popularity or price in the culture). Each can be problematic and lead to replication failure. Thus, a replication recipe should focus on describing contextual factors that are plausibly linked to the DV. Most simply, one can report the mean level of the operational (amount of hot sauce) and conceptual (link to aggressiveness) DV in a control condition in which none of the critical IVs are varied. This is needed to assure that relevant background variables in the replication study that affect the DV are set at a similar level to the original study.

In sum, conceptually driven psychology research is different from the physical sciences, and our replication recipes should reflect this.

The replicability revolution

doi:10.1017/S0140525X18000833, e147

Ulrich Schimmack

Department of Psychology, University of Toronto, Mississauga, ON L5L 1C6, Canada.

Ulrich.schimmack@utoronto.ca

https://replicationindex.wordpress.com/

Abstract: Psychology is in the middle of a replicability revolution. High-profile replication studies have produced a large number of replication failures. The main reason why replication studies in psychology often fail is that original studies were selected for significance. If all studies were reported, original studies would fail to produce significant results as often as replication studies. Replications would be less contentious if original results were not selected for significance.

The history of psychology is characterized by revolutions. This decade is marked by the replicability revolution. One prominent feature of the replicability revolution is the publication of replication studies with nonsignificant results. The publication of several high-profile replication failures has triggered a confidence crisis.

Zwaan et al. have been active participants in the replicability revolution. Their target article addresses criticisms of direct replication studies.

One concern is the difficulty of re-creating original studies, which may explain replication failures, particularly in social psychology. This argument fails on three counts. First, it does not explain why published studies have an apparent success rate greater than 90%. If social psychological studies were difficult to replicate, the success rate should be lower. Second, it is not clear why it would be easier to conduct conceptual replication studies that vary crucial aspects of a successful original study. If social priming effects were, indeed, highly sensitive to contextual variations, conceptual replication studies would be even more likely to fail than direct replication studies; however, miraculously they always seem to work. The third problem with this argument is that it ignores selection for significance. It treats successful conceptual replication studies as credible evidence, but bias tests reveal that these studies have been selected for significance and that many original studies that failed are simply not reported (Schimmack 2017; Schimmack et al. 2017).

A second concern about direct replications is that they are less informative than conceptual replications (Crandall & Sherman 2016). This argument is misguided because it assumes a successful outcome. If a conceptual replication study is successful, it increases the probability that the original finding was true and it expands the range of conditions under which an effect can be observed. However, the advantage of a conceptual replication study becomes a disadvantage when a study fails. For example, if the original study showed that eating green jelly beans increases happiness and a conceptual replication study with red jelly beans does not show this effect, it remains unclear whether green jelly beans make people happier or not. Even the nonsignificant finding with red jelly beans is inconclusive because the result could be a false negative. Meanwhile, a failure to replicate the green jelly bean effect in a direct replication study is informative because it casts doubt on the original finding. In fact, a meta-analysis of the original and replication study might produce a nonsignificant result and reverse the initial inference that green jelly beans make people happy. Crandall and Sherman's argument rests on the false assumption that only significant studies are informative. This assumption is flawed because selection for significance renders significance uninformative (Sterling 1959).

A third argument against direct replication studies is that there are multiple ways to compare the results of original and replication studies. I believe the discussion of this point also benefits from taking publication bias into account. Selection for significance explained why the reproducibility project obtained only 36% significant results in direct replications of original studies with significant results (Open Science Collaboration 2015). As a result, the significant results of original studies are less credible than the nonsignificant results in direct replication studies. This generalizes to all comparisons of original studies and direct replication studies. Once there is suspicion or evidence that selection for significance occurred, the results of original studies are less credible, and more weight should be given to replication studies that are not biased by selection for significance. Without selection for significance, there is no reason why replication studies should be more likely to fail than original studies. If replication studies correct mistakes in original studies and use larger samples, they are actually more likely to produce a significant result than original studies.

Selection for significance also explains why replication failures are damaging to the reputation of researchers. The reputation of researchers is based on their publication record, and this record is biased in favor of successful studies. Thus, researchers' reputations are inflated by selection for significance. Once an unbiased replication produces a nonsignificant result, the unblemished record is tainted, and it is apparent that a perfect published record is illusory and not the result of research excellence (a.k.a flair). Thus, unbiased failed replication studies not only provide new evidence; they also undermine the credibility of existing studies. Although positive illusions may be beneficial for researchers' eminence, they have no place in science. It is therefore inevitable that the ongoing correction of the scientific record damages the reputation of researchers, if this reputation was earned by selective publishing of significant results. In this way direct replication studies complement statistical tools that can reveal selective publishing of significant results with statistical tests of original studies (Schimmack 2012; 2014; Schimmack & Brunner submitted for publication).

Constraints on generality statements are needed to define direct replication

doi:10.1017/S0140525X18000845, e148

Daniel J. Simons,[a] Yuichi Shoda,[b] and D. Stephen Lindsay[c]

[a]*Department of Psychology, University of Illinois, Champaign, IL 61820;* [b]*Department of Psychology, University of Washington, Seattle, WA 98195;* [c]*Department of Psychology, University of Victoria, Victoria, BC V8W 2Y2, Canada.*

dsimons@illinois.edu yshoda@uw.edu slindsay@uvic.ca

www.dansimons.com

http://www.psych.uw.edu/psych.php?p=358&person_id=2569

https://www.uvic.ca/socialsciences/psychology/people/faculty-directory/lindsaysteve.php

Abstract: Whether or not a replication attempt counts as "direct" often cannot be determined definitively after the fact as a result of flexibility in how procedural differences are interpreted. Specifying constraints on generality in original articles can eliminate ambiguity in advance, thereby leading to a more cumulative science.