Original Article

A Causal Replication Framework for Designing and Assessing Replication Efforts

Peter M. Steiner[1], Vivian C. Wong[2], and Kylie Anglin[3]

[1]Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, USA
[2]Department of Leadership, Foundations, and Policy, University of Virginia, Charlottesville, VA, USA
[3]Curry School of Education and Human Development, University of Virginia, Charlottesville, VA, USA

Abstract: Replication has long been a cornerstone for establishing trustworthy scientific results, but there remains considerable disagreement about what constitutes a replication, how results from these studies should be interpreted, and whether direct replication of results is even possible. This article addresses these concerns by presenting the methodological foundations for a replication science. It provides an introduction to the causal replication framework, which defines "replication" as a research design that tests whether two (or more) studies produce the same causal effect within the limits of sampling error. The framework formalizes the conditions under which replication success can be expected, and allows for the causal interpretation of replication failures. Through two applied examples, the article demonstrates how the causal replication framework may be utilized to plan prospective replication designs, as well as to interpret results from existing replication efforts.

Keywords: causal replication framework, replication crisis, replication assumptions, causal inference, potential outcomes

Efforts to promote evidence-based practices in psychology and in other disciplines assume that scientific findings are valid and reliable but also generalizable enough to warrant their use in decision-making. Replication has long been a cornerstone for establishing trustworthy scientific results. At its core is the belief that scientific knowledge should not be based on chance occurrences. Rather, reliable scientific knowledge should be cumulatively established through multiple systematic and transparent studies with findings that are generalizable to at least some current or future target population of interest (Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015).

Given the central role of replication in the accumulation of scientific knowledge, researchers have reevaluated the replicability of seemingly well-established findings. Results from these efforts have not been promising. The Open Science Collaboration (OSC) replicated 100 experimental and correlational studies published in high impact psychology journals and found that only 36% of these efforts produced results with the same statistical significance pattern as the original study (Open Science Collaboration, 2015). The findings prompted the OSC authors to conclude that replication rates in psychology were low, but not inconsistent with what has been found in other domains of science. For example, Ioannidis (2005) suggests that most findings published in the biomedical sciences were likely false. His review of more than 1,000 medical publications found that only 44% of replication efforts produced results that corresponded with the original findings (Ioannidis, 2008). Combined, these results and others contribute to a growing sense of a "replication crisis" occurring in multiple domains of science, including marketing (Madden, Easley, & Dunn, 1995), economics (Dewald, Thursby, & Anderson, 1986; Duvendack, Palmer-Jones, & Reed, 2017), education (Makel & Plucker, 2014), and prevention science (Valentine et al., 2011).

Despite consensus on the need to promote the replicability of results, there remains considerable disagreement about what constitutes as replication, how replication studies should be implemented, and how replication results should be interpreted. For example, Gilbert, King, Pettigrew, and Wilson (2016) argued that OSC's conclusions about replicability rates in psychology were overly pessimistic. They showed that besides sampling error and weak statistical power in the original studies, the replication efforts themselves may have been biased. Only 69% of their study protocols were endorsed by the original authors – suggesting substantial deviations in study factors across the original and replication efforts.

This article addresses these concerns by presenting the methodological foundations for a "replication science." We introduce the *causal replication framework*, which

defines "replication" as a research design that tests whether two (or more) studies produce the same causal effect within the limits of sampling error (Wong & Steiner, 2018b). The causal replication framework uses potential outcomes notation (Rubin, 1974) to specify a causal estimand of interest, as well as five assumptions under which replication success can be expected. Here, a causal estimand is defined as the unknown causal effect of a well-defined treatment-control contrast on a clearly specified outcome for a specific target population and setting. The replication assumptions include: treatment and outcome stability, equivalence of causal estimands, identification of estimands, unbiased estimation of effects, and correct reporting of effects. Under the causal replication framework, the purpose of direct replication is to replicate an identical causal estimand, which has been derived from subject-matter theory. While prior conceptualizations of replication emphasize repetition of methods and procedures, we show that repeating all methods and procedures may not be required – or even desired – to achieve replication success. What matters is the extent to which replication assumptions are met in field settings.

An important implication of the causal replication framework is that an effect will not replicate if *any* of the five replication assumptions are violated. *Replication failure* occurs when studies fail to produce the same effect estimate (within the limits of sampling error), or when studies fail to draw the same conclusions about the direction of an effect. A finding might not replicate if there are even small differences in the causal estimands across studies. This may be because of differences in treatment and control conditions, in outcomes, and in population and setting characteristics; it may also be because one or both studies fail to correctly identify, estimate, and report the same effect of interest. However, as we will show, replication failure is not inherently a problem *as long as the researcher is able to understand why the result was not reproduced*. In fact, identifying the source(s) of replication failure is crucial for understanding effect heterogeneity, and for generalizing an effect. Currently, there are no standards for how researchers should characterize the causal estimand of interest for replication. Rather, it must usually be inferred based on methods and procedures of the study. The causal replication framework provides a theoretical basis for more systematic reporting of the causal estimand of interest, which is essential for understanding why replication failure occurs.

In this article, we will also emphasize the importance of *prospective replication designs* for uncovering systematic sources of effect heterogeneity. In prospective replication designs, the researcher plans a series of replication studies that may occur simultaneously or at different times. This type of design allows the researcher to evaluate replication assumptions by systematically relaxing a single replication assumption while ensuring that all others are met. If replication failure is observed, then the researcher has high confidence that she has identified the source of the effect heterogeneity. In this way, the causal replication framework may be understood as "causal" in two ways – first, it identifies and estimates the causal effect for a well-defined treatment-control contrast in each study; second, it can be used to uncover the causal reasons for *why* a result does not replicate. Currently, *post hoc replication designs* are more common. Here, a researcher attempts to replicate a finding from an earlier study, addressing replication assumptions based only on what can be inferred through methods and procedures from the original study. The challenge here is that there may not be sufficient information from the original study to assess the extent to which replication assumptions are violated. In cases where multiple replication assumptions have been violated simultaneously, it is often difficult to understand why replication failure was observed.

The article proceeds as follows. We begin by introducing the causal replication framework, its required design assumptions, and implications of the theory. Although replication assumptions may seem stringent, our goal here is not to argue that replication designs are infeasible in field settings. Rather, it is to demonstrate that researchers must carefully attend to replication assumptions for studies to produce interpretable results. In the second section of this article, we show through an example how researchers may address replication assumptions using a prospective research design. We also discuss an example of a *post hoc* replication design, using the causal replication framework to highlight the challenges of interpreting results from this type of approach. We conclude by discussing how the causal replication framework may be applied more generally for planning replication studies to uncover effect heterogeneities.

What Is Replication?

Over the years, researchers have sought to clarify what is meant by "replication." Most definitions have focused on repeating methods and procedures from an original study (Schmidt, 2009). For instance, Brandt et al. (2014) define direct replications as studies "that are based on methods and procedures *as close as possible* to the original study" (p. 218, emphasis in original). Zwaan, Etz, Lucas, and Donnellan (2018) describe replication as "studies intended to evaluate the ability of a particular method to produce the same results upon repetition" (p. 5). Nosek and Errington (2017) define replication as "independently repeating the methodology of a previous study and obtaining the same results" (p. 1). Others have also emphasized the need for replication procedures to be carried out by independent researchers (Simons, 2014) on independent samples of participants (Lykken, 1968).

"Procedure-based" approaches to replication focus on ensuring that the same methods and tools are used in both the original and replication studies. Thus, the quality of the replication is judged by how closely the replication study is able to repeat methods and procedures from the original study (Brandt et al., 2014; Kahneman, 2014). Despite this seemingly straight-forward approach to replication, procedure-based approaches to replication pose multiple challenges for implementation in field settings. First, the original study may fail to report all relevant and necessary methods and procedures, rendering direct replications difficult if not impossible to achieve (Hansen, 2011). Second, this approach to replication privileges methods and procedures in the original study, but the original study itself may not have been well implemented. Thus, replicating flawed methods and procedures from the original study may not be feasible or even desired. Third, procedure-based approaches are inherently challenging because the repetition of methods itself is rarely the primary goal of any replication effort. More important is whether the intervention under consideration has a reliable and replicable *causal effect* on the outcome of interest. In procedure-based approaches, the causal effect of interest is usually not well-defined by the researcher, it can only be guessed from an accurate description of procedures and methods used. As a result, it may be challenging to assess whether the replicated procedures and methods are similar enough to the original study, and whether they are appropriate for replicating the same causal effect.

We argue that instead of focusing on repeating methods and procedures, the goal of a replication study is to systematically address replication assumptions via study design features such that the causal estimand of interest is the same across both studies. In the following section, we discuss the causal replication framework and required assumptions for a *direct replication* of results. The assumptions are set up from an idealistic point of view, where two studies are expected to focus on the exact same causal estimand, and differences in effect estimates arise only due to sampling or randomization uncertainty. In later sections, we show that the causal replication framework may be extended beyond the case of "direct replication" to "conceptual replications," where the goal may be to evaluate different causal estimands of interest in order to uncover sources of effect heterogeneity.

Replication Under the Causal Replication Framework

Schmidt (2009) describes direct replication as a "methodological tool based on a repetition procedure," but adds that its purpose is for "*establishing a fact, truth or piece of knowledge*" (p. 91, emphasis in original). In the causal replication framework, we begin with the premise that replication is for establishing a "fact, truth or piece of knowledge." Here, the piece of knowledge can be described as a causal estimand, which is the target of inference across Study 1 and Study 2. The causal estimand of interest is derived from subject-matter theory. That is, it is guided by theory that defines and operationalizes (a) a treatment-control contrast of interest, (b) an outcome measure that is meaningful, (c) the target population of relevance, and (d) a setting in which support and inhibitory factors may or may not be controlled. Thus, the causal estimand represents the "true" but unknown causal effect on an outcome Y for a well-specified replication population R, which might be a subpopulation of one or both studies' target population, and a specific setting S. Because the causal replication framework does not prioritize the repetition of an original study's methods and procedures, we now refer to the original and replication study as Studies 1 and 2.

Using potential outcomes notation (Imbens & Rubin, 2015), we define $Y(0)$ to be the potential control outcome (under the control condition $T = 0$) and $Y(1)$ to be the potential treatment outcome (under the treatment condition $T = 1$). These are the outcomes we would observe if a subject were assigned to the control or treatment condition, respectively. Though in practice, we observe only one of the two potential outcomes, the potential outcomes notation allows us to clearly define the causal estimand of interest. For example, the Average Treatment Effect (ATE) for the inference population R and setting S is defined as $\text{ATE}_{R,S} = E_R[Y(1) - Y(0)|S]$, the average difference in treatment and control potential outcomes. This is the average of individual treatment effects for all subjects in the replication population R (for instance, the ATE for all 8-year-old female students) under setting S. Other examples of causal estimands are the Average Treatment Effect for the Treated ($\text{ATT}_{R,S}$), the Intent-To-Treat effect ($\text{ITT}_{R,S}$), the Complier Average Treatment Effect ($\text{CATE}_{R,S}$), or the corresponding effects for any other replication (sub) population R (e.g., Morgan & Winship, 2015). These estimands are also relevant for Randomized Controlled Trials (RCTs) whenever attrition or one- or two-sided noncompliance with treatment assignment is present (Imbens & Rubin, 2015; Steiner, Kim, Hall, & Su, 2017). Attrition and noncompliance issues need to be explicitly addressed in analyzing RCTs, otherwise the causal estimands will neither be identified nor identical across studies.

The causal replication framework suggests two important insights about replication approaches (Wong & Steiner, 2018b). First, successful replication of the causal effect (within the limits of sampling error) can be expected only if the causal estimands of Studies 1 and 2 are identical –

both studies must focus on the same causal estimand, that is, the same treatment-control contrast on the same outcome Y for the same inference population R and setting S. Second, multiple assumptions about the identification, estimation, and reporting of the causal effect are required for a successful replication of the same causal effect. Overall, this means that the causal estimand of interest in both studies must be both identifiable and estimable without bias.

The causal estimand is *identifiable* if it can be *nonparametrically* estimated without bias from the hypothetically *infinite* replication population (Hernán & Robins, 2020). Here, "identification" refers to conditions needed for yielding replicable causal effects without systematic biases due to confounding, attrition, or measurement, for instance. Identification is not concerned with random fluctuations in estimates due to sampling or randomization error. Nonparametric identification allows us ignore issues related to parametric model specification, which is about estimation. *Estimation* addresses the question of whether an identified causal effect can be uniquely estimated from a *finite* sample without bias. Identification and estimation assumptions depend on the study design. For instance, the $ATE_{R,S}$ for participants in an RCT is identified if randomization is perfectly implemented and the Stable-Unit-Treatment-Value Assumption (SUTVA) is met (Imbens & Rubin, 2015). The identified $ATE_{R,S}$ can be estimated without bias if an unbiased estimator (e.g., regression estimator) is used and all technical assumptions are met (e.g., sufficient degrees of freedom or absence of perfect collinearity).

Direct replication of a causal effect (within the limits of sampling or randomization uncertainty) can be expected when all five identification and estimation assumptions are met, which we summarize below (for a more rigorous formalization and proof using potential outcomes notation, see Wong & Steiner, 2018b).

Assumption A1: Treatment and Outcome Stability Across Studies

A1.1 *No hidden variation in treatment and control conditions*. This assumption requires that the treatment and control conditions are clearly defined and identical in both studies, that is, there is no (unobserved) variation in the implementation of the treatment-control contrast across studies.

A1.2 *No variation in outcome measures*. Both studies measure exactly the same outcome construct. This can be achieved by holding measures, instruments, test setting and timing constant across studies. In particular, the outcome needs to be measured in both studies exactly the same time after the implementation of the treatment.

A1.3 *No mode-of-study-selection effects*. Selection into the two studies has no effect on the potential outcomes. For instance, it does not matter whether participants are randomly sampled or assigned to one of the two studies or whether they volunteer or self-select into one of the studies. Recruitment strategies like incentives for study participation must not affect the potential outcomes either.

A1.4 *No peer, spillover, or carry-over effects*. The potential outcomes of participants in Study 1 are unaffected by researchers, participants, and characteristics of Study 2, and vice versa.

Assumption A2: Equivalence of Causal Estimands

A2.1 *Same causal quantity of interest*. Both studies need to aim at the same causal quantity. For instance, the ATE or the ITT (in the presence of treatment non-compliance in an RCT).

A2.2 *Identical effect-generating processes*. If Studies 1 and 2 are implemented at different sites or times, the real-world process that generates the causal effects must be identical for both studies, that is, the effect-generating process does not vary across sites and time. This implies that all effect moderators have the same impact in both studies. This assumption presumes that nature (including humans) behaves lawfully and uniformly; otherwise, replication hardly establishes stable knowledge (for a discussion, see Schmidt, 2009).

A2.3 *Identical distribution of population characteristics*. The target populations of the two studies must be identical with respect to the joint distribution of individual characteristics. This does not imply that both studies have to focus on the same overall target populations, but it does suggest that the replication population R, for which we want to replicate the causal effect, must be covered by both studies. Matching or reweighting may be needed in one or both studies to achieve equivalence in population characteristics. Although the assumption does not need distributional equivalence in all population characteristics, it does require equivalence on all characteristics that *moderate* the causal effect. Since determination of moderating variables requires reliable subject-matter knowledge, the assumption is more likely met if researchers aim at the very same target populations (i.e., equivalence on all observed and unobserved characteristics) for both studies. This assumption would be violated, for instance, if the portion of male and female participants differs across studies and if gender is an effect moderator. Without the use of weighting or matching adjustments to equate the gender distribution across studies the causal estimands would differ.

A2.4 *Identical distribution of setting variables*. Both studies must have the same joint distribution of setting variables that *moderate* the causal effect. If the setting variables do not vary within studies, the two studies must be implemented in the same setting S, guaranteeing that all the factors that bring about or inhibit the causal effect must be absent or present to the same extent in both studies (Cartwright & Hardie, 2012).

Assumption A3: Identification of Causal Estimands

In both studies, the causal estimand (e.g., $ATE_{R,S}$) must be identified. For instance, if Study 1 uses an RCT, then it must be perfectly implemented (with no attrition, noncompliance, spill-over or peer effects) and cover the replication population R. If Study 2 is based on observational data, then in addition to the absence of spill-over and peer-effects, strong ignorability must be met with respect to $ATE_{R,S}$. That is, the set of observed covariates is able to remove any confounding bias. In short, this assumption requires the valid implementation of an experimental or quasi-experimental design. Discussions of the identification assumptions for different research designs can be found in Imbens and Rubin (2015), Morgan and Winship (2015), Steiner et al. (2017), Kim and Steiner (2019), or Wong, Wing, Steiner, Wong, and Cook (2012).

Assumption A4: Unbiased Estimation of Causal Estimands

In both studies, the causal estimand is estimable without bias. This requires the use of an unbiased or at least consistent estimator (provided sample sizes are sufficiently large). For example, when parametric models are used, the models need to be correctly specified and the technical assumptions must be met (e.g., no perfect collinearity and sufficient degrees of freedom).

Assumption A5: Correct Reporting of Estimands, Estimators, and Estimates

For both studies, the estimands, estimators, and estimates (including standard errors) are correctly reported. Mistakes in reporting may result in incorrect conclusions about whether the two studies actually aim at the same causal estimand or whether the results successfully replicate. In addition to the correct reporting, all the identification and estimation assumptions required to meet A3 and A4 need to be credibly defended, ideally based on strong subject-matter knowledge about the data-generating process and with empirical evidence to rule out most plausible validity threats.

When assumptions A1 and A2 hold, both Studies 1 and 2 focus on the same *causal estimand* (e.g., $ATE_{R,S}$), that is, the same treatment-control contrast for target population R in setting S. Assumptions A3, A4, and A5 then ensure that the causal estimand of interest is identified, estimated without bias, and correctly reported in each of the two studies.

The replication assumptions highlight the difference between traditional, procedure-based approaches to replication, and the causal replication framework. In procedure-based approaches, the goal and purpose of replication is repetition of methods. In the causal replication framework, the goal is that both studies identify and estimate the same causal estimand of interest. Importantly, repeating methods and procedures does not guarantee that all or even most replication assumptions are automatically met (see also Stroebe & Strack, 2014). The two studies may still identify and estimate quite different causal estimands. For example, using the same methods and procedures may yield different causal estimands if participants failed to comply with their treatment assignment status, or if setting characteristics changed in ways that amplified or dampened the effect, or if the outcome measures do not represent the same underlying construct (because of cultural differences or changes over time).

On the other hand, while replicating the same methods and procedures will often help meet assumptions A1 through A4, it is not always required that they are implemented in identical ways across studies. For example, two studies may have different research designs *as long as they identify and estimate the same well-defined causal estimand of interest*. Study 1 may use an RCT while Study 2 uses an observational design with self-selection. Or, two studies may have different but valid measurement instruments of the same underlying outcome construct. Thus, under the causal replication framework, the quality of the replication effort is judged by the extent to which the five causal replication assumptions are met in field settings rather than the successful replication of methods and procedures. Strong subject-matter theory about the causal effects under investigation and the data-generating processes underlying the two studies is indispensable for deriving testable implications and probing assumptions. Violations of *any* of the five assumptions likely result in a direct replication failure.

Research Designs for Causal Replication

The causal replication framework yields two important insights for practice. First, although assumptions for the

direct replication of results are stringent, it is possible for researchers to address or probe these assumptions through the thoughtful use of research designs and empirical diagnostic tests. Second, researchers may identify sources of effect heterogeneity by systematically inducing potential violations of one or multiple replication assumptions. In this case, a *prospective replication approach* may be used to ensure that all design assumptions are met with the exception of the one that is under investigation. If results fail to replicate, the researcher will know why there was a difference in effects. *Post hoc approaches* have the advantage of reflecting more natural sources of variation across studies, but may be more difficult to interpret in cases where replication failure is observed due to violations of multiple replication assumptions. Below, we discuss two example replication studies, and the extent to which design assumptions were met under the causal replication framework.

Example 1: A Prospective "Design-Replication" Study

In prospective replication designs, researchers may systematically test effect heterogeneities by examining whether crucial design assumptions have been met across studies. Although prospective designs for replication are relatively novel in the general replication literature, they have been established in a related but mostly independent literature called design-replication studies (also called within-study comparison designs). In a design-replication study, the researcher evaluates the performance of an observational study (i.e., a non- or quasi-experiment) by testing whether the observational study is able to replicate the treatment effect from a benchmark RCT with the same target population and setting. Difference in treatment effect estimates between the RCT benchmark and observational studies is interpreted as failure in the observational study to correctly identify the causal estimand of interest (A3). Although this approach has been used to evaluate the identification assumptions in replication designs, it may also be applied to systematically test differences in effects due to variations in treatment-control conditions and in population and setting characteristics across study arms.

Wong and Steiner (2018a) discuss research designs for "design-replication" studies, and highlight an example of a *prospective design-replication* study introduced by Shadish, Clark, and Steiner (2008). In this approach, researchers randomly assigned students to one of two study arms: Study 1 or Study 2. This ensured that students in both study arms were equivalent on the distribution of all covariates. Students who were assigned to study 1 were *randomly assigned again* into 1 of 2 treatment conditions– either a short vocabulary or a short math training. Students who were randomly assigned to Study 2, however, were allowed to *self-select* into their preferred training session (vocabulary or math) which introduced confounding bias. Participants in both study arms underwent treatment and control conditions simultaneously in the same setting (e.g., a college university classroom), and their outcomes were assessed on the same measures (post-intervention math and vocabulary tests) in the same time frame. To address selection into training conditions in Study 2, the researchers applied propensity score techniques to establish equivalence between groups. Once groups were matched, treatment effects were estimated using the same ANCOVA models as for the RCT (Study 1) to ensure there were no differences due to estimation techniques (A4). Observational treatment effects from Study 2 were then compared to the RCT benchmark results from Study 1. Any difference in estimated treatment effects (within the limits of sampling error) was interpreted as failure to replicate due to "non-experimental bias" in Study 2 (i.e., a violation of assumption A3).

The Shadish et al. (2008) study demonstrates how research design features were used to control for all other replication assumptions (A1, A2, A4, and A5), except for the one being tested (A3, causal identification of ATE in study 2). The assumptions are summarized in Table 1.

- The treatment and control conditions were well defined and implemented *simultaneously* under lab conditions in both studies (assumption A1.1).
- Outcomes were measured in the same way across treatment conditions and study arms, and administered at the same time (A1.2).
- Because the intervention was short and students were tested immediately after the intervention, there was no opportunity for spillover or peer effects within and across study arms (A1.4).
- In both study arms, the researchers aimed at the same causal quantity, the ATE (A2.1).
- Randomization into Studies 1 and 2 ensured that the target populations in both study arms were statistically equivalent (A2.3).
- Since the intervention and outcome measures were implemented at the same location and time in both study arms, variations due to changes in the effect-generating mechanism (A2.2) and due to study setting differences (A2.4) were also ruled out.
- The interventions were implemented in tightly-controlled, laboratory-like conditions, which resulted in a high-quality RCT with no differential attrition, noncompliance, or other issues. Thus, the RCT provided a valid benchmark for identifying the ATE (A3 for Study 1).
- Treatment effects were consistently estimated the same way in both studies (with the exception of using propensity score adjustments in Study 2), ensuring that

Table 1. Shadish et al. (2008) viewed through the causal replication framework

Assumption	Study 1: RCT	Study 2: Observational study
A1 Treatment and outcome stability		
A1.1	✔ Treatment and control conditions: identical	✔ Treatment and control conditions: identical
A1.2	✔ Outcome measure, instruments, and timing: identical	✔ Outcome measure, instruments, and timing: identical
A1.3	✔ No mode-of-study-selection effects	**?** No mode-of-study-selection effects
A1.4	✔ No peer-, spillover-, or carry-over effects	✔ No peer-, spillover-, or carry-over effects
A2 Equivalence of causal estimands		
A2.1	✔ ATE	✔ ATE
A2.2	✔ Effect-generating process: identical	✔ Effect-generating process: identical
A2.3	✔ Target population: identical	✔ Target population: identical
A2.4	✔ Setting: identical	✔ Setting: identical
A3 Identification	✔ ATE is identified (under RCT assumptions)	? ATE is identified if + All confounders are reliably measured + No preference effects are present
A4 Estimation	✔ Unbiased (mean difference)	✔ Unbiased/consistent (matching estimator)
A5 Reporting	✔ Correct reporting	✔ Correct reporting

Note. ✔ indicates assumptions that are likely met, **?** indicates potential violations of assumptions (intended or unintended).

there were no differences in estimation procedures across study arms (A4).
- Subsequent reanalysis of the original data by independent investigators found no reporting errors of results (A5).

Assuming the absence of any mode-of-study selection effects (A1.3) and preference effects in the Study 2 (part of A3), any difference in effect estimates may be credibly interpreted as failure in the observational method to identify valid causal effects in Study 2 (i.e., a violation of A3).

The prospective research design in Shadish et al. (2008) allowed researchers to address replication assumptions, and to identify potential sources of replication failure. Here, the researchers concluded that despite using different research designs for identifying and estimating effects in Studies 1 and 2, the studies were able to replicate the same causal effect. This finding implied that in this specific context at least, replication assumption A3 was met. Interestingly, the results of this design-replication study have been successfully reproduced in a conceptual replication (Pohl et al., 2009).

At first glance, the design-replication study may seem like a narrow application of replication. However, this example highlights several important features for considering and planning replication studies more generally. First, it demonstrates that it is possible to attend to replication assumptions in field settings, and that it is most convincing when research design features (such as randomization) are used to address assumptions. Second, Shadish et al. (2008) show that by systematically relaxing one or two replication assumptions, it is possible to learn about sources of effect heterogeneity through replication studies. The strength of the replication design rests on whether other design assumptions (that are not being evaluated) are met. Finally, although randomizing units into study arms may not be feasible in many field settings, there are other research design approaches that may be applied for addressing assumptions, such as ensuring equivalence in population characteristics across studies (A2.3). For example, the researcher may match or reweight participants from Studies 1 and 2 so that participant characteristics are the same across studies. Or, a researcher may apply a "switching replication" design (Shadish, Cook, & Campbell, 2002), where participants are randomized into groups and receive treatment and control conditions in an alternating, replicating sequence. This ensures that participant characteristics remain equivalent across replication cycles while other assumptions are tested. For example, the researcher may systematically introduce differences in contexts, settings, and timing for each replication cycle. The point here is that, by shifting the focus from replicating methods and procedures to addressing replication assumptions, it is possible to conceptualize replication through a wider array of research designs that can be used to systematically uncover sources of effect heterogeneity across multiple studies.

Example 2: A Post Hoc Direct Replication Study

Most replication studies are post hoc designs where the replication study is planned and designed only *after* the (statistically significant) results of the original study have

been published. As a result, if there are flaws or implementation challenges in the original study (e.g., the sample size is too small, the causal estimand it not explicitly defined and not well-identified due attrition or noncompliance issues), they cannot be changed or addressed by the replicator. Thus, there may be multiple differences across study arms in how treatment and control conditions are implemented and in how treatment effects are identified, estimated, and reported. While in prospective designs, the researcher may introduce systematic sources of variation across the two studies, in post hoc approaches, multiple study differences occur naturally and may not be researcher controlled.

To address replication assumptions, the researcher may attempt post hoc matching of characteristics of the original and replication study. The characteristics are related to the similarity of treatment and control conditions (A1), units and settings (A2), and methodology (A3 and A4) across both studies. A successful replication of results can be expected only if all assumptions (A1 through A5) are met. However, the researcher will often lack sufficient knowledge about whether even close repetition of methods and procedures succeeds in addressing all five replication assumptions. For example, it may not be clear which population and study factors moderate treatment effects (A2), or whether differences in the timing of treatment and measurement implementations produce differences in results (A1).

Klein et al.'s (2014) "Many Labs" study is an example of a post hoc replication study. This was a collaborative effort of 36 independent research teams from 12 countries that sought to examine the replicability of 13 well-known effects in psychology, and the robustness of these effects across samples, settings, and cultures. The research team selected the 13 effects to be replicated from 12 original studies based on several criteria. First, treatments had to be delivered in a standardized format online or in person. This helped maintain the integrity of the original treatment conditions under investigation. Second, the study designs had to be short and straight-forward for independent investigators to implement. This was to allow for multiple treatments to be evaluated in a single testing session. Third, with the exception of a single correlational study, treatments were evaluated using simple, two group experimental designs. Fourth, the 13 effects were selected to represent variations in topics, time frames since the original study was conducted, and certainty of their replicability. Each of the 36 research teams replicated all 13 effects in a single sample of participants. Labs delivered near identical scripts, translating and adapting the language as necessary. They documented key information about the sampling frame, recruitment process, achieved sample, and other factors related to the local context. Deviations from the original study protocol were also recorded.

The Klein et al. (2014) Many Labs study demonstrates that thoughtfully designed post hoc studies can limit, but rarely eliminate confounders between study arms. The Many Labs study was designed to examine the variation in replicability across 36 samples and settings through deliberate variations in populations and settings (i.e., potential violations of A2.3 and A2.4). With the exception of populations and settings, the authors intended to replicate the procedures of the original study, especially the treatment conditions, as closely as possible. Table 2 summarizes potential violations of the replication assumptions.

Table 2. Klein et al. (2014)viewed through the causal replication framework

Assumption	Original studies	Replication studies
A1 Treatment and outcome stability		
A1.1	✔ Treatment and control conditions clearly defined	**?** Slight variations in treatment conditions (e.g., due to translations)
A1.2	✔ Outcome measures, instruments, and timing	**?** Slight variations in outcome measures, instruments and timing (translations, online)
A1.3	✔ Mode-of-study-selection	**?** Mode-of-study-selection effects (incentives)
A1.4	✔ No peer-, spillover-, or carry-over effects	✔ No peer-, spillover-, or carry-over effects
A2 Equivalence of causal estimands		
A2.1	✔ ATE	✔ ATE
A2.2	✔ Effect-generating process	**?** Variations in effect-generating process
A2.3	✔ Target population	**?** Different target populations
A2.4	✔ Setting	**?** Different setting
A3 Identification	✔ ATE is identified (under RCT assumptions)	✔ ATE is identified (under RCT assumptions) but different populations/settings/treatments
A4 Estimation	✔ Unbiased (mean difference)	✔ Unbiased (mean difference)
A5 Reporting	✔ Correct reporting	✔ Correct reporting

Note. ✔indicates assumptions that are likely met, **?** indicates potential violations of assumptions (intended or unintended). All but one of the many labs replications featured a simple two-condition experiment. This table excludes the correlational study.

- The treatment and control conditions were well-defined and documented across the study arms. However, some small changes in the treatment conditions were necessary to account for differences in populations and time (e.g., treatment materials were translated into the relevant language) or for intervention timing. Despite these efforts, the culture- and language-specific constructs of the treatment conditions might nonetheless vary across studies (A1.1).
- Whenever possible, outcome measures in the original and replication studies used the exact same wording, translated if needed. However, the translations of questions and verbalized response scales could have resulted in slight differences in the underlying outcome construct. Moreover, the original studies measured the outcomes using pencil and paper assessments, while all replication studies used an online medium. Thus, it is questionable whether the exact same outcome was measured across both studies (A1.2).
- Both the original and replication studies relied on university participant pools and incentives to recruit participants. Recruitment strategies were well-documented in the replication studies, but not always documented in the original studies. Further, it was unclear at the outset whether variations in incentives and recruitment strategies would impact potential outcomes (A1.3).
- Peer, spillover, or carry-over effects between the original and replication studies are unlikely because most of the original studies were implemented decades ago (A1.4).
- Most of the causal quantities were ATE generated through experimental variation (A2.1).
- The Many Labs study was designed under the assumption that the original and replication studies had stable effect-generating processes. Given that the replication studies were implemented decades later and at different sites, a violation of assumption A2.2 is possible.
- By design, the replication studies deliberately varied target populations and study settings, such that assumptions A2.3 and A2.4 likely did not hold. For instance, variations in lab or online settings might affect the potential outcomes and change the causal estimand. Or, all replication studies tested each participant with respect to all 13 effects, while the original studies independently assessed only a single effect. Thus, context and order effects may have influenced the assessment of each single effect.
- Since most studies relied on experimental settings, the ATE was likely identified for all studies (A3), provided the absence of nonresponse, attrition, or noncompliance. However, the ATEs very likely referred to different target populations and treatment contrasts (due to potential violations of A1 and A2) such that different causal estimands were identified.
- Since treatment effects were estimated in the same way in both the original and replication studies, biases due to estimation procedures were less likely, unless there was differential missingness or systematic measurement error (A4).
- There was no evidence of incorrect reporting (A5), though 3 of the 12 original studies failed to report their sample size.

Overall, although the Many Labs replication study had the goal of assessing the replicability of effects across different populations and settings (A2.3 and A2.4), it is likely that other assumptions (A1 and A2) were also violated (variations in treatments, outcomes, populations, settings, effect-generating process). As such, the causal estimands were likely different across the original and replication studies. Nonetheless, the researchers concluded that 10 of the 13 results replicated when looking at the direction and significance patterns of effects (but they did not test the direct replication of the magnitude of effects). The potential violation of assumptions (intended or unintended) highlight the challenge of post hoc designs for even "close" replication studies. Replication bias may occur if any one of the design assumptions is violated and, as noted above there were many opportunities for violations in the Klein et al. study. The challenge here is that without research design elements to control these factors systematically, it is hard to interpret the source of effect variations in replication designs. However, post hoc designs are often needed when the replication of an important finding has yet to be established, and there is interest in assessing the robustness of results across different treatments, populations, settings, and outcomes. In these cases, investigators should conduct empirical diagnostics to probe and discuss each design assumption systematically.

Prospective Versus Post Hoc Replication Designs

The above case studies provide just two examples of research design variants that may be used to evaluate the replicability of results. Prospective replication designs are akin to the design of RCTs – they have the advantage of offering strong causal interpretations of results, especially in cases when results do not replicate and there is a strong need to know the source of treatment effect variation. However, prospective designs are limited because of the extensive resources needed to plan for these studies in advance, and may fail to reveal key sources of variation in effects. Post hoc designs have the potential to allow researchers

to assess the replicability and robustness of results over more natural and realistic sources of variation, but results from these studies may be challenging to interpret when multiple violations of replication assumptions occur simultaneously or when the results of the original study are a false-positive finding due to publication bias. Our approach is to recommend that multiple research designs for replication are needed for establishing robust scientific results, with the acknowledgment that each method has its relative strengths and weaknesses. Regardless of the research design, the causal replication framework will help researchers in planning replication studies and in systematically assessing and learning from replication failure.

Discussion

To promote high quality replication efforts in psychology and elsewhere, a clear understanding of replication as its own scientific method is needed. The causal replication framework shows that replication may be understood as a causal research design, with stringent assumptions for producing interpretable results. Just as a randomized experiment attempts to test a hypothesis by minimizing differences between treatment and control groups except for the clearly defined treatment-control contrast, the causal replication framework asks researchers to minimize differences between studies except for the assumptions of interest. The causal replication framework requires researchers to *explicitly* define the causal estimand of interest based on subject-matter theory and the research design chosen. This is in contrast to procedural replication approaches where the causal estimands are often only *implicitly* derivable from the description of methods and procedures used. Importantly, the framework suggests that replication approaches may be improved through the thoughtful use of research design features and diagnostic tests for systematically addressing and testing replication assumptions. A high quality replication effort is characterized by a replication design that is able to convincingly rule out most plausible validity threats (assumptions A1–A5, with the exception of those that are violated due to intended variations in the replication) in order to systematically identify sources of treatment effect heterogeneity.

As our review of Shadish et al. indicates, prospectively planned replication designs allow researchers to implement replication studies with design elements for probing the crucial replication assumptions. In post hoc replication efforts, however, the researchers implementing the replication study need to demonstrate that the same causal estimand is identified and estimated as in the original study, a task which is only possible if the original study provides sufficient information about the causal estimand and about potential identification and estimation issues (assumptions A3–A5). Our review of the Many Labs' replications demonstrates how researchers may address replication assumptions in field settings and the common challenges that may arise.

The Causal Replication Framework and Alternative Conceptualizations of Replication

Given recent interest in promoting replication efforts, researchers have suggested various topographies for categorizing and understanding different types of replication. For example, Schmidt (2009) introduced *direct* and *conceptual* replications. Direct replication requires the repetition of an experimental procedure, while conceptual replication involves the repetition of a hypothesis test or result using different methods or procedures. Somewhat similarly, Clemens (2017) proposed *verification* and *robustness tests*. Verification tests evaluate whether results are replicable using the same study protocol on the same data – or on new data that are resampled from the same underlying sampling distribution. Robustness tests examine the replicability of results when samples are drawn from a different sampling distribution, or when there is some variation in method or procedure from the original study (e.g., in the analysis code).

Although the causal replication framework focuses on design assumptions for the direct replication of a causal estimand, it is fully compatible with prior conceptualizations of replication types. The verification test in the Clemens approach (2017) is akin to replication efforts that are meant to address all five replication assumptions under the causal replication framework. Schmidt's (2009) definition of "direct" or "statistical" replication (Valentine et al., 2011) are also examples in which all five replication assumptions are expected to be met. Other types of replication designs, however, test whether potential violations to replication assumptions occurred by assessing the replicability of effects. Robustness tests (Clemens, 2017) and conceptual replications (Schmidt, 2009) are examples where the goal is to assess whether the causal effect varies when at least one or more replication features are changed such that that the replication assumption might be systematically or naturally violated (e.g., by varying the target population, A2.3, the setting A2.4, or research design, A3). However, an advantage of the causal replication framework is that it provides a unified perspective for understanding all different types of replication using a common set of assumptions.

The assumptions of the causal replication framework also relate directly to Simons et al.'s discussion of Constraints on Generality (COG; Simons, Shoda, & Lindsay, 2017). That is, given the specifics of a single study, they suggest that

researchers should be clear about to which populations, treatment-control contrasts, settings, and procedures the findings can or cannot be generalized. Studies with clearly stated COGs facilitate the design and implementation of meaningful direct and conceptual replication efforts. More generally, replication efforts can be framed from a *causal generalization* or *transportability* point of view (Bareinboim & Pearl, 2012; Stuart, Bradshaw, & Leaf, 2015; Tipton, 2013). Given that two studies are rarely identical in all population and setting characteristics, the question is whether the effect of one study can be transported to the population and setting of the other study. If this is possible, direct replication will be successful. Not surprisingly, many assumptions underlying the causal replication framework and causal transportability are identical. As for the causal replication framework, knowing the effect moderating population and setting characteristics is key for a valid generalization of causal effects.

Can Replication Efforts Be Successful in Practice?

Given the stringent assumptions required for direct replication of results, is it possible to conduct successful replication studies in actual research practice? This is a crucial question because direct replication is challenging. Our general sense is that causal inference with a single study, even with a randomized experiment, is challenging, so replicating the same causal estimand should be even more demanding.

The Shadish et al. (2008) study showed that it is possible to address assumptions through the thoughtful and creative use of research design elements in controlled settings. The question then becomes: Are such direct replication efforts, where the researcher successfully controls all possible confounds, informative for science? For example, imagine if the authors of the Shadish et al. (2008) were not interested in evaluating the performance of an observational method in their design-replication. Instead, their goal was to conduct a direct replication and implemented a second RCT instead of an observational study with self-selection. If replication failure was observed, then it would be explained by sampling error or randomization uncertainty, a factor that is rarely of interest. Would replication success or failure even be interesting in this context given how tightly controlled the replication study is?

Our view is that prospective replication designs are useful to the extent they are able to uncover systematic sources of effect heterogeneity (see also Stroebe & Strack, 2014). Shadish et al. held everything constant except for the two studies' (quasi)-experimental design. To investigate effect homogeneity or heterogeneity across settings and time, a replication study could hold everything constant except for variations in setting or time, respectively. This could be implemented through a switching replication design that we described earlier, or by randomly assigning participants into studies that are implemented in different settings. Carefully planned, prospective replication studies allow the researcher to learn from replication failure. This is because controlled variation of a single factor allows the researcher to infer the reason for the effect heterogeneity. Such replication efforts are considered conceptual rather than direct (National Science Foundation and Institute of Education Sciences, 2018).

More generally, we believe the causal replication framework is most useful for encouraging researchers to think systematically through assumptions when designing and implementing conceptual replications for identifying effect heterogeneity. In post hoc replication studies, the Framework highlights the challenges for interpreting results that may come from unplanned or simultaneous violations of assumptions. Viewed from this perspective, post hoc replications have an exploratory rather than confirmatory causal character. Results from post hoc replications can and should be used to prospectively design stronger replications that allow for a systematic investigation of potential effect heterogeneities.

For both prospective and post hoc designs, subject-matter theory about the presumed data-generating process – in particular knowledge about effect-moderating population and setting characteristics – play important roles in replication efforts. If two studies have different effect-moderating factors, direct replication success will be unlikely. Simons et al.'s (2017) COGs represent an important step towards building substantively-based theories about the scope of an effect's generality. It may also be used to design better replication studies by providing researchers with substantive and theory-based guidance on the types of measures needed for assessing violations to replication assumptions.

Conclusion and Future Directions

In this paper, we show that the causal replication framework may be used to identify and examine potential sources of effect heterogeneity or replication bias. In general, we recommend that researchers (1) identify specific and plausible threats to validity based on their substantive knowledge about the data and effect-generating processes, (2) hypothesize data patterns that should emerge if these threats are realized, and (3) construct empirical tests and diagnostic probes for ruling out such threats. In most cases, replication assumptions will be achieved by implementing high quality research designs (e.g., randomization of participants into the original and replication study arms) or by using statistical adjustment procedures with rich covariate information (i.e., reweighting of units in the replication

study such that they reflect the same distribution of characteristics in the original study).

Beyond the applications discussed here, we believe the causal replication framework provides important insights about other types of replication designs that are currently underutilized. For example, in reproducibility studies (Chang & Li, 2015), independent investigators examine whether results are correctly reported (A5) by examining whether results replicate from the same data and analysis code. In stepped-wedge designs, participants are randomized to receive treatments in successive waves over time (Hussey & Hughes, 2007), allowing the researcher to examine whether results replicate over time (A2). Finally, the causal replication framework may help avoid questionable research practices (e.g., *p*-hacking or HARKing) and publication bias by shifting the focus away from the repetition of methods and procedures to addressing replication design assumptions. Studies that fail to clearly explicate the causal estimand, to carefully defend the causal identification assumptions, and to discuss the estimation procedures may be less credible and thus less suited for replication. Further work is needed on establishing the methodological foundations for a replication science, including developing new measures for testing replication success across studies (Steiner & Wong, 2018), but we believe the causal replication framework provides a coherent perspective for continuing this work.

References

Bareinboim, E., & Pearl, J. (2012). Transportability of causal effects: Completeness results. In J Hoffmann & B. Selman (Eds.), *Proceedings of the Twenty-Sixth Conference on Artificial Intelligence* (pp. 698–704). Menlo Park, CA: AAAI Press.

Bollen, K., Cacioppo, J., Kaplan, R., Krosnick, J. A., & Olds, J. L. (2015). *Social, behavioral, and economic sciences perspectives on robust and reliable science*. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford, UK: Oxford University Press.

Chang, A. C., & Li, P. (2015). *Is economics research replicable? Sixty published papers from thirteen journals say "usually not" [Finance and Economics Discussion Series 2015-083]*. Washington, DC: Board of Governors of the Federal Reserve System. https://dx.doi.org/10.17016/FEDS.2015.083

Clemens, M. A. (2017). The meaning of failed replications: A review and proposal. *Journal of Economic Surveys, 31*, 326–342. https://doi.org/10.1111/joes.12139

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in empirical economics: The Journal of Money, Credit and Banking project. *The American Economic Review, 76*, 587–603. Retrieved from http://www.jstor.org.proxy01.its.virginia.edu/stable/1806061

Duvendack, M., Palmer-Jones, R., & Reed, W. R. (2017). What is meant by "replication" and why does it encounter resistance in economics? *American Economic Review, 107*, 46–51. https://doi.org/10.1257/aer.p20171031

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science, 351*, 1037–1037. https://doi.org/10.1126/science.aad7243

Hansen, W. B. (2011). Was Herodotus correct? *Prevention Science, 12*, 118–120. https://doi.org/10.1007/s11121-011-0218-5

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Boca Raton, FL: Chapman & Hall/CRC.

Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials, 28*, 182–191. https://doi.org/10.1016/j.cct.2006.05.007

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge, UK: Cambridge University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*, 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Kahneman, D. (2014). A new etiquette for replication. *Social Psychology, 45*, 310–311. https://doi.org/10.1027/1864-9335/a000202

Kim, Y., & Steiner, P. M. (2019). Gain scores revisited: A graphical models perspective. *Sociological Methods & Research, 1*, 1–11. Advance online publication. https://doi.org/10.1177/0049124119826155

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*, 142–152. https://doi.org/10.1027/1864-9335/a000178

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 151–159. https://doi.org/10.1037/h0026141

Madden, C. S., Easley, R. W., & Dunn, M. G. (1995). How journal editors view replication research. *Journal of Advertising, 24*, 77–87. https://doi.org/10.1080/00913367.1995.10673490

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*, 304–316. https://doi.org/10.3102/0013189X14545513

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge, UK: Cambridge University Press.

National Science Foundation and Institute of Education Sciences. (2018). *Companion guidelines on replication & reproducibility in education research*. Retrieved from https://www.nsf.gov/pubs/2019/nsf19022/nsf19022.pdf

Nosek, B. A., & Errington, T. M. (2017). Making sense of replications. *ELife, 6*, e23383. https://doi.org/10.7554/eLife.23383

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. https://doi.org/10.1126/science.aac4716

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis, 31*, 463–479. https://doi.org/10.3102/0162373709343964

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of*

Educational Psychology, 66, 688–701. https://doi.org/10.1037/h0037350

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100. https://doi.org/10.1037/a0015108

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*, 1334–1344. https://doi.org/10.1198/016214508000000733

Shadish, W. R., Cook, T. D., & Cambell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Cengage Learning.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*, 76–80. https://doi.org/10.1177/1745691613514755

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*, 1123–1128. https://doi.org/10.1177/1745691617708630

Steiner, P. M., Kim, Y., Hall, C., & Su, D. (2017). Graphical models for quasi-experimental designs. *Sociological Methods & Research, 46*, 155–188. https://doi.org/10.1177/0049124115582272

Steiner, P. M., & Wong, V. C. (2018). Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons. *Evaluation Review, 42*, 214–247. https://doi.org/10.1177/0193841X18773807

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*, 59–71. https://doi.org/10.1177/1745691613514450

Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science, 16*, 475–485. https://doi.org/10.1007/s11121-014-0513-z

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics, 38*, 239–266. https://doi.org/10.3102/1076998612441947

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., … Schinke, S. P. (2011). Replication in prevention science. *Prevention science, 12*, 103–117. https://doi.org/10.1007/s11121-011-0217-6

Wong, V. C., & Steiner, P. M. (2018a). Designs of empirical evaluations of nonexperimental methods in field settings. *Evaluation Review, 42*, 176–213. https://doi.org/10.1177/0193841X18778918

Wong, V. C., & Steiner, P. M. (2018b). *Replication designs for causal inference*. EdPolicyWorks Working Paper Series. Retrieved from http://curry.virginia.edu/uploads/epw/62_Replication_Designs.pdf

Wong, V. C., Wing, C., Steiner, P. M., Wong, M., & Cook, T. D. (2012). Research designs for program evaluation. In W. Velicer & J. Schinka (Eds.), *Handbook of psychology: Research methods in psychology* (2nd ed., Vol. 2) (pp. 316–341). Hoboken, NJ: Wiley.

Zwaan, R., Etz, A., Lucas, R. E., & Donnellan, B. (2017). Making replication mainstream. *Behavioral and Brain Sciences, 41*, E120. https://doi.org/10.1017/S0140525X17001972

History
Received November 16, 2018
Revision received May 16, 2019
Accepted June 3, 2019
Published online December 20, 2019

Funding
The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grant #R305B140026 and a collaborative research grant from the National Science Foundation, Grant #2015-0285-00.

Peter M. Steiner
Department of Human Development and Quantitative Methodology
University of Maryland
3942 Campus Drive
College Park, MD 20742
USA
psteiner@umd.edu