University of Virginia | Curry School and Batten School
EdPolicyWorks

Working Paper:

# Replication Designs for Causal Inference

*Vivian C. Wong[1] & Peter M. Steiner[2]*

Recent work has raised questions about the extent to which scientific results are replicable. Prior efforts to replicate results have led to disappointing rates of replicability (Aarts et al., 2015). Critics, however, have challenged the interpretation of these replication results (Gilbert, King, Pettigrew & Wilson, 2016). Recent debates about the "replication crisis" have raised questions about the essential design features for a replication study, and whether a true replication study is feasible (Aos et al., 2011). This paper addresses these challenges by formalizing replication as a research design. Instead of emphasizing procedural requirements for replication that may not generalize well across different fields of study, our approach introduces a general definition of replication by identifying research design assumptions needed for multiple studies to replicate the same causal effect. Our conceptualization provides a framework for how researchers may employ replication design variants to assess the robustness of effects, and to identify sources of treatment effect heterogeneity. The paper shows how replication design variants may be integrated throughout the research cycle to improve the veracity, robustness, and generalization of scientific claims.

[1]University of Virginia
[2]University of Wisconsin-Madison

*Updated April 2018*

EdPolicyWorks
University of Virginia
PO Box 400879
Charlottesville, VA 22904



*Acknowledgements*: This research was supported by a collaborative NSF grant #2015-0285-00.

## 1. Introduction

Efforts to promote evidence-based practices in decision-making assume that scientific findings are of sufficient validity to warrant its use. 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, it is established through systematic and transparent methods, results that can be independently replicated, and findings that are generalizable to at least some target population of interest (Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015).

Given the central role of replication in the accumulation of scientific knowledge, researchers have evaluated 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. The replication studies were conducted by independent researchers who collected data from new samples, using study materials from the original research protocol ("Estimating the reproducibility of psychological science," 2015). Overall, the OSC found that only 36% of these efforts produced results with the same statistical significance pattern as the original study, and 47% of the original effects fell within the 95% confidence interval of the replicated results. 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. 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 contribute to a growing sense of a "replication crisis" occurring in multiple domains of science,

including marketing (Madden, Easley, & Dunn, 1995), economics (Dewald & 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 replication efforts, there remains considerable disagreement about what constitutes as replication, how a replication study should be implemented, and how results from these studies should be interpreted. Gilbert, King, Pettigrew, and Wilson (2016) argue that OSC's conclusions about replication 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 be biased. For example, although the OSC attempted to replicate the same research procedures used in the original study, only 69% of their study protocols were endorsed by the original authors – suggesting substantial deviations in study factors across the original and replication efforts. In a reanalysis of the OSC data, Van Bavel, Mende-Siedlecki, Brady, and Reinero (2016) write that even small differences in contextual factors across studies can produce differences in the original and replication results.

But what conditions are needed for an original and replication study to produce identical treatment effects (within the limits of sampling error)? Currently, the social and health sciences lack consensus on what replication is, and what it is meant to demonstrate (Aos et al., 2011). We address these challenges by presenting replication as a formal research design using a nonparametric structural model (Pearl, 2009) and potential outcomes (Rubin, 1974). We define replication as a research design that tests whether two or more studies produce the same causal effect (within the limits of sampling error). Our approach focuses on the replication of *causal* treatment effects because researchers and policymakers are often interested in robust, scientific

results for programmatic and policy decision-making. However, the method extends easily to the replication of correlational and descriptive results as well, albeit with weaker assumptions.

This paper demonstrates the multiple benefits of conceptualizing replication as a research design. First, our approach draws upon an already well-established model for understanding research designs and their assumptions: the potential outcomes model. We will show that many research design features and empirical diagnostics used to improve causal inferences can be extended to the replication design context. Second, our definition of replication is applicable across diverse fields of study because it focuses on causal estimands of interest and assumptions, not on study procedures and operations that may vary with different outcome measures, units, treatments, and settings. Third, knowledge of research design assumptions can provide investigators with better guidance on the *planning* of replication studies. That is, researchers may incorporate prospective research design features and empirical diagnostic measures for addressing and/or probing replication assumptions. Replication design assumptions will also help readers evaluate when potential sources of biases may produce results that do not replicate. Finally, results from replication studies with well-defined treatment conditions and outcomes, clear causal quantities for well-specified target populations, and rigorous research designs, estimation, and reporting practices will improve the quality of meta-analytic results when they are synthesized across multiple studies.

## 2. Background

Given the importance of results replication in the accumulation of scientific knowledge, it is surprising that replication efforts are so rare. Makel and colleagues reviewed a history of the top 100 journals and found that only 0.13% of studies in education (Makel & Plucker, 2014) and 1.07% of studies in psychology (Makel, Plucker, & Hegarty, 2012) were replication efforts.

Technical advisory panels for the Institute of Education Sciences (IES) and the National Science Foundation (NSF) examined reasons for why there are so few replication efforts. One issue is that replication remains undervalued in science. Researchers find it more difficult to raise funds to support replication studies (Asendorpf et al., 2013), to publish results from replication efforts (Nosek, Spies, & Motyl, 2012; Valentine et al., 2011), and to receive promotion and tenure for their replication studies (Asendorpf et al., 2013). A recent review of 1,151 psychology journals found that only 3% indicated replication as an interest area, but 33% of journals emphasized the need for originality in its criteria for publication (Martin & Clarke, 2017). In acknowledging cultural barriers in the scientific community toward replication efforts, NIH Director Francis Collins and Deputy Director Lawrence Tabak wrote in *Nature* that "science has long been regarded as 'self-correcting,' … Over the long term, that principle remains true. In the shorter term, however, the checks and balances that once ensured scientific fidelity has been hobbled. This has compromised the ability of today's researchers to reproduce others' findings" (2014, pg. 612).

Compounding the cultural stigma related to "replication" is the issue that replication is not well established nor understood as a research methodology. An IES Technical Working Group (TWG) on "What Comes After an Efficacy Study?" observed that "The topic of replication in the education sciences is complex, in part because there are different definitions of replication studies" (2016, pgs. 9-10). This is despite efforts already made by the NSF Subcommittee on Replicability in Sciences in 2015 to institute a shared understanding of common terminology in the field. For example, the subcommittee defined *replicability* as, "the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected" (Bollen et al., 2015, pg. 4) and *reproducibility* as "the ability ... to

duplicate the results ... using the same materials and procedures" (Bollen et al., 2015, pg. 3). The difference here is that replication requires the collection and analysis of new data, whereas reproducibility involves reanalysis of original data and code files. However, these definitions are not yet widely adopted, even among federal agencies. The Director and Deputy Director of National Institutes of Health describe reproducibility more broadly to include "the design, data collection, and analysis of new data to replicate results from an original study" (Collins & Tabak, 2014). The OSC calls its own independent replications an effort to estimate the reproducibility rate in psychology.

Over the years, researchers have sought to clarify what is meant by replication by proposing new topographies and procedures for the method. The most common distinctions focus on the *purpose* of the replication effort. Schmidt (2009) differentiates between direct and conceptual replications. Generally, direct or statistical replications (Valentine et al., 2011) assess whether corresponding results are replicated within the bounds of sampling error. Here, the researcher attempts to repeat the exact same research and treatment protocols as the original study, but draws a new sample of participants. The goal of conceptual replications, however, is to assess whether the same result is obtained despite heterogeneity across units, time, outcomes, setting, treatment variants, and methods. Conceptual replications are informal, with multiple independent researchers asking broadly defined research questions such as, "What is the effect of grade retention on students' long-term achievement and behavioral outcomes?" or "What is the effect of pre-kindergarten on students' achievement scores?" These questions may be evaluated across multiple states and time periods, and often involve somewhat different target populations, treatment implementations, and outcomes.

Others have defined replication based on procedural characteristics that are varied across original and replication studies. Lykken (1968) highlights differences between replications that are conducted by (or in collaboration with) the original investigators, and those that are conducted by independent researchers. The latter has the benefit of reducing the threat of experimenter bias and error, but may be limited due to inadvertent deviations in the study protocol. Duncan, Engel, Classens, and Dowsett (2014) suggest procedures for conducting "within-study replications." Here, the original authors present replication results within the same paper, report, or data – where, the replicated results are obtained by systematically varying the subgroups under investigation, estimation methods for producing results, and data sources for conducting the analysis. In the social science literature, these procedures are often described as efforts to assess the "robustness" of results across variations in study characteristics and samples.

Despite repeated efforts to catalog different types of replication approaches, the most basic issue of what a replication is and what it is meant to demonstrate remains not well understood. This is because prior definitions of replication focus on *procedures* for conducting this type of study. A limitation of procedure-based definitions of replication is that they do not generalize well to different fields of study, or even within the same field but with very different types of treatments, units, contexts, and outcomes. In fact, the IES TWG report noted the methodological confusion is so great that "even co-investigators sometimes disagree about whether their study is new or a replication" (*Building Evidence: What Comes After an Efficacy Study?*, 2016, pg. 10).

Our approach differs from prior conceptualizations of replication in that it focuses on the stringent assumptions required for the replication of the same causal estimand/effect. Although our approach does not identify specific procedures needed for replication, addressing

replication assumptions does have practical implications for how the approach should be conducted in field settings. As we will see, replication assumptions provide a framework for deriving replication research design variants that allow researchers to evaluate the robustness of results, and to systematically identify sources of treatment effect heterogeneity.

## 3. Replication as a Research Design

In this section we describe components for a replication design with two study arms, an "original" and a "replication" study, and two treatment conditions – the treatment of interest and a control condition. Although our discussion focuses on replication designs with two studies and two treatment conditions, our conceptualization extends to study designs with more than one replication effort and treatment condition. Figure 1 provides a graphical depiction of the replication design with two study arms. Within each study arm, participants are assigned – or select into – a treatment or control condition. Treatment effects are compared across both study arms to determine whether results are sufficiently "close" (e.g. in terms of direction, size, statistical significance patterns). The goal of the study design is to evaluate whether causal effects replicate across both study arms. Below, we describe the assumptions required for two studies to produce identical treatment effects (within the limits of sampling error). Then we discuss a proposed framework for using replication design assumptions to improve the design of replication studies, and to identify systematic sources of variation.

### Conceptual Framework

We begin by describing a general model of the outcome-generating process in each study arm. Let $Y_{i0}$ and $Y_{j1}$ be the outcomes of interest for the original and replication study, respectively, where the subscript *i* denotes a unit from the target population in the original study (indexed by 0), and subscript *j* denotes a unit from the target population in the replication study

(indexed by 1). The outcomes can be written as nonparametric structural functions that depend on a unit's treatment status $Z \in \{0,1\}$, individual characteristics $X$, and the overall setting $S$ under which the original and replication study took place:

$$Y_{i0}(z_{i0}, x_{i0}, s_{i0}) = f_0(Z_{i0} = z_{i0}, X_{i0} = x_{i0}, S_{i0} = s_{i0}) \text{ for the original study and} \quad (1)$$

$$Y_{j1}(z_{j1}, x_{j1}, s_{j1}) = f_1(Z_{j1} = z_{j1}, X_{j1} = x_{j1}, S_{j1} = s_{j1}) \text{ for the replication study.}$$

$X$ and $S$ may be vectors of variables that also include exogeneous error terms. The setting $S$ includes all outcome-determining factors other than the treatment condition and individual characteristics of the unit. For example, study site characteristics like school urbanicity or percentage of free or reduced-price lunch and other factors such as cultural attitudes that may support or suppress the treatment effect are included in $S$. The setting $S$ can vary across studies, but it also can vary across units within a study's target population (as in a multisite RCT design). Nonparametric functions $f_0$ and $f_1$ also may differ across sites or time.[1] Subscripts 0 and 1 refer to the original and replication study, respectively, indicating that the two studies may occur at different sites, different time points, or both. The use of subscripts $i$ and $j$ suggests that the populations of the original and replication study do not overlap or be the same. However, in cases where participants in both study arms are sampled or shared from the same population (simultaneously or at different time points), then some or even all participants across both study arms will belong to the same target population.

The two structural equations (1) allow us to derive potential treatment and control outcomes for each study's target population. For example, the potential treatment outcome for

[1] We could also allow the outcome function to vary across subjects, but for notational simplicity we assume that all outcome variations across individuals can be expressed as variations in the functions' arguments.

the original study is given by $Y_{i0}(1) = Y_{i0}(1, x_{i0}, s_{i0})$. This is the outcome we would observe if unit *i* were exposed to the treatment condition in the original study. Similarly, $Y_{i0}(0) = Y_{i0}(0, x_{i0}, s_{i0})$ denotes the potential control outcome we would observe if unit *i* were exposed to the control condition in the original study. $Y_{j1}(1) = Y_{j1}(1, x_{j1}, s_{j1})$and $Y_{j1}(0) = Y_{j1}(0, x_{j1}, s_{j1})$ are the corresponding potential treatment and control outcomes for the replication study.

Each study arm in the replication design can have its own causal quantity of interest. The original study, for example, may use an RCT and estimate the average treatment effect, ATE = $E[Y_{i0}(1) - Y_{i0}(0)]$. The replication study, however, may use observational data from a different site and estimate the average treatment effect for the treated, ATT = $E[Y_{j1}(1) - Y_{j1}(0) \mid Z_{j1} = 1]$. However, even if both studies yield unbiased causal effects, the original result will not be replicated when the causal quantities differ across the two study designs. Replication of causal effects, therefore, requires much more stringent assumptions than what is needed for causal inference in single studies alone. Below, we briefly summarize assumptions required for identification and estimation of unbiased treatment effects for a single study. Then, we describe the stringent assumptions needed for causal effects to be replicated across multiple studies.

**Causal Identification and Estimation Assumptions for Treatment Effects in a Single Study**

The assumptions required to identify and estimate a causal effect depends on the study's design, the data collected, and the statistical methods used to estimate effects. RCTs, non-equivalent control group designs (NECGD), time series designs with control groups (including difference-in-differences), and instrumental variable approaches all rely on different assumptions to identify a causal effect. In an RCT, the identification of the ATE requires the assumption that potential outcomes are independent of treatment status, $(Y(1), Y(0)) \perp Z$, which may be achieved if random assignment is correctly implemented. In an NECGD, causal identification of the ATE

or ATT demands that the conditional independence assumption is met, such that $(Y(1), Y(0)) \perp Z \mid X^*, S^*$. This means that potential outcomes must be independent of treatment selection given a set of covariates $X^*$ and $S^*$ (which may be subsets of the outcome-generating sets *X* and *S*). The NECGD also requires the positivity assumption, $P(Z = 1 \mid X^*, S^*) > 0$, such that each unit's treatment probability, given $X^*$ and $S^*$, is greater than zero. In a comparative interrupted time series design, the common trend assumption must hold to identify the ATT, and for an instrumental variable approach, the exclusion restriction and monotonicity assumption must be met to identify the ATE for the latent subpopulation of compliers (Angrist, Imbens, & Rubin, 1996). Similar assumptions must be met for other research designs to identify a causal effect.

The causal interpretation of the effects identified by the aforementioned designs requires one additional assumption—the stable-unit-treatment-value assumption (SUTVA; Imbens & Rubin, 2015). SUTVA implies (a) uniquely defined and implemented treatment and control conditions, (b) the absence of peer or spillover effects, and (c) the absence of any effects due to the mode of treatment assignment or selection. Combined, these conditions are implicitly encoded in the structural outcome equations above. The treatment indicator *Z* can only take on values of 0 or 1, indicating a uniquely defined treatment and control condition. The outcome $Y_{i0}$ (or $Y_{j1}$) depends only on the unit's own treatment exposure but not on any other unit's study participation or treatment exposure (i.e., no peer and spillover effects). Because the mode of treatment selection or assignment (e.g., random selection or self-selection) is not included as an argument in the outcome-generating functions, it does not affect the potential outcomes. This implies that there are no preference effects from being allowed to select one's own treatment status.

Finally, identification of causal effects does not yet imply that it can be estimated without bias. A valid estimate of the causal quantity also requires an unbiased estimator – or at least a consistent estimator, provided that sample sizes are sufficiently large. Other technical assumptions such as the full rank of the design matrix (e.g., more observations than variables, no collinearity) are also needed. Moreover, Null Hypothesis Significance Testing (NHST) requires assumptions such as homoscedasticity, normality, and independence of observations. However, because this paper is about the identification and estimation of point estimates, we will limit our discussion to assumptions needed for point estimation, and do not discuss issues related to statistical inferences.[2]

**Causal Identification and Estimation Assumptions for the Replication of Treatment Effects**

In the above section, we showed that stringent assumptions are needed for a study to identify and estimate an unbiased causal effect. However, even in cases where multiple studies individually meet assumptions for producing causal results, it is possible that these results may not replicate, even within the limits of sampling error. A successful replication requires four additional *replication assumptions* (see Table 1 for a summary of assumptions and implications for practice).

Assumption A1. Treatment and Outcome Stability

The first replication assumption is that treatment and control conditions must be well defined, and that the outcome measure must be the same across both study arms. This assumption implies that there are no peer effects across studies, that participants do not react to how they were assigned to studies, and that there are no peer or spillover effects across studies.

[2] Valentine et al. (2011) addresses the issue of statistical tests in assessing the replication of results. Steiner and Wong (in press) discusses methods for assessing correspondence in results in design replication studies.

The assumption corresponds to SUTVA in individual studies, but in replication designs, it requires treatment and outcome stability *across multiple study arms*. We explicitly state the implications of treatment and outcome stability in more detail:

*A1.1 No Variation in Treatment and Control Conditions.* The treatment and control conditions must be identical across the original and replication study. That is, the treatment indicator $Z$ in the two structural equations in (1) refers to the same uniquely defined treatment and control conditions. This implies that all components of the treatment and control conditions are known and implemented in exactly the same way. If, for example, the replication study includes a treatment component that differs from the original study, or the replication study has a weaker treatment dosage, then this assumption is violated. The assumption is also violated if participants in the control condition of the replication study have and use alternative treatment options that were not available to controls in the original study.

*A1.2 No Variation in Outcome Measures*. The treatment and control conditions of both studies must be evaluated with respect to the same outcome $Y$. This means that the outcome in both studies must be measured with the same instruments, in the same settings, and at the same time points after the treatment was introduced. Any variation in the instrument, setting, or timing across studies may produce differences in treatment effects.

*A1.3 No Mode-of-Study-Selection Effects*. The potential outcomes are not affected by participants' selection into the original and replication study. That is, whether the participants are included in the study through self-selection or random selection does not affect the potential outcomes. This assumption may be violated if participants are randomly sampled from the target population for one study, and self-select into the second study arm. Then, if volunteers for the

study are especially motivated to respond to treatment conditions, differences in treatment effects may be introduced.

*A1.4 No Peer, Spillover, or Carryover Effects*. The potential outcomes depend only on each participant's exposure to treatment, and not on the treatment exposure of others in a different study arm. This assumption is violated if knowledge of treatments or peers from the original study affects the potential outcomes in the replication study. For example, a participant's motivation may increase after learning that her peers were assigned to the treatment (or control) condition in the original study. Another example occurs if treatment participants in the replication received advice from treatment participants in the original study. In cases where the same units participate in both the original and replication study at different times (as in a switching replication design), the assumption would be violated if effects of the treatment and control conditions persist from one study to the next.

*Implications for Practice.* To ensure stability in treatment conditions and outcomes across replication arms, investigators should consider in advance plausible validity threats. Will participants have knowledge of their treatment and study status, and will it affect their potential outcomes? Do participants in the original and replication study have opportunities to interact and share their experiences and knowledge? Can the treatment and outcome measures be implemented by multiple researchers in a consistent way, under the same time frame, and with high fidelity to the original protocol? If any of the assumptions are violated, we cannot expect to replicate a causal effect estimate.

Assumption A2. Equivalence of Causal Estimands

Successful replication of effect estimates (within the limits of sampling error) requires that the causal estimand of interest is the same across both study arms. That is, the original and

replication studies must have equivalent causal quantities for the same well-defined target population. In this section, we focus on the average treatment effect (ATE) but similar assumptions are needed for other potential quantities of interest across study arms, such as the average treatment effect for the treated (ATT).

To formalize equivalence in causal estimands, we begin by assuming additive treatment effects so that the potential treatment outcome in both studies may be defined as the sum of the potential control outcome and the treatment effect:

$$Y_{i0}(1, x_{i0}, s_{i0}) = Y_{i0}(0, x_{i0}, s_{i0}) + \tau_0(x'_{i0}, s'_{i0}), \text{ and} \tag{2}$$

$$Y_{j1}(1, x_{j1}, s_{j1}) = Y_{j1}(0, x_{j1}, s_{j1}) + \tau_1(x'_{j1}, s'_{j1}).$$

Here, the treatment functions $\tau_0$ and $\tau_1$ depend on unit characteristics $x'_{i0}$ and $x'_{j1}$ (respectively), and study setting factors $s'_{i0}$ and $s'_{j1}$ (respectively). These are characteristics that magnify or weaken the treatment effect, and they are subsets of all unit and study setting factors that affect the outcome ( $x'_{i0} \subseteq x_{i0}$, $x'_{j1} \subseteq x_{j1}$, $s'_{i0} \subseteq s_{i0}$, $s'_{j1} \subseteq s_{j1}$). Thus, the effect-generating functions include all individual and study setting characteristics that explain treatment effect variation.

The valid replication of causal effect requires equivalence of ATEs for the original and replication study:

$$\begin{aligned} E_P[Y_{i0}(1, x_{i0}, s_{i0}) - Y_{i0}(0, x_{i0}, s_{i0})] &= E_Q[Y_{j1}(1, x_{j1}, s_{j1}) - Y_{j1}(0, x_{j1}, s_{j1})] \\ E[\tau_0(x'_{i0}, s'_{i0})] &= E[\tau_1(x'_{j1}, s'_{j1})] \\ \text{ATE}_P &= \text{ATE}_Q \end{aligned} \tag{3}$$

where, the expectations are taken with respect to the target populations of the original (*P*) and replication (*Q*) study.

Equality (3) does not demand that the expected potential outcomes are equivalent across the original and replication study, but that the additive treatment effect must be identical across

both study arms. This means that participants in the replication study may be – for example – more advantaged and have higher average potential outcomes than those in the original study, but the treatment must have the same additive (and not multiplicative) effect in both study arms. The equivalence of causal estimands is achieved if four requirements hold:[3]

*A2.1 Same Causal Quantity of Interest*. The original and replication study must have the same causal quantity of interest. Here we assume that the parameter of interest is the ATE in both study arms. If, for example, the original study identifies the ATE and the replication identifies the ATT or the intent-to-treat effect, then equivalence in causal estimands is unlikely (unless additional assumptions such as constant treatment effects hold).

*A2.2 Identical Effect-Generating Processes*. The process that generates the treatment effect must be identical for both studies, $\tau_0 = \tau_1 = \tau$. This implies that the variable sets of individual characteristics and study setting that determine effect heterogeneity must be the same across studies ( $X'_0 = X'_1 = X'$and $S'_0 = S'_1 = S'$) and exert the same effect on the outcome, such that $\tau_0(x'_{i0}, s'_{i0}) = \tau_1(x'_{j1}, s'_{j1})$whenever $x'_{i0} = x'_{j1}$ and $s'_{i0} = s'_{j1}$.

*A2.3 Identical Distribution of Population Characteristics*. The target populations of the two studies must be identical with respect to the joint distribution of all individual characteristics $X'$ that modify the treatment effect. That is, target populations *P* and *Q* must have the same distribution of $X'$, but may differ with respect to other unit characteristics that do not moderate the magnitude of the treatment effect. If the distributions differ, then it must at least be possible

[3] Equivalence can be achieved even if the four requirements do not hold. But in this case the effects of violating the requirements must offset each other. Since this is unlikely or at least hard to assess in practice, we do not further discuss such situations.

to reweight or match the replication population $Q$ such that it then has the same distribution of $X'$ as $P$ (for further discussion see below).

*A2.4 Identical Distribution of Setting Variables*. Both studies must have the same joint distribution of setting variables $S'$ that moderate the treatment effect. If setting characteristics $S'$ do not vary across participants within a study, then these factors must be identical across the original and replication studies to achieve overlap in setting characteristics. When all setting factors are identical across study arms, then the above assumptions are sufficient for establishing equivalence in $ATE_P$ and $ATE_Q$. However, in cases where setting characteristics vary across participants within studies (i.e. multisite RCT designs), then the *joint* distribution of unit characteristics and settings ( $X'$, $S'$) must be identical across study arms.

*Implications for practice.* Although these assumptions are stringent, there are circumstances under which equivalence in causal estimands may be achieved. For example, in reproducibility efforts where the replication study shares the same observed data as the original study, such that $Y_{j1} = Y_{i1} = Y_{i0}$ (because $i = j$), the causal estimands are equivalent because the two study arms have the same causal quantity for the same target population. Equivalent causal estimands may also be achieved when units are randomly sampled from the same target population (at the same site and at the same time) into the original and replication studies. If treatment and outcome conditions are implemented in the same way, then random sampling of units into original and replication studies ensures that the effect-generating process, the target population, and the setting do not vary across study arms (provided assumption A1 holds). There also may be cases where it is reasonable to assume that the treatment effect is constant, or at least does not systematically vary with unit characteristics or setting variables. Then, even a lack of overlap in population and setting characteristics will not yield differences in causal estimands.

Finally, there may be cases where the original and replication arms do not share the same eligibility criteria for participating in the study, producing different underlying target populations whose distributions of $X'$ might not or only partially overlap. For example, the replication study may target a less advantaged population than those who were eligible for the original study. Thus, it may not be possible to achieve identical distributions of $X'$ through reweighting or matching of units, even if $X'$ would be fully observed. The causal estimands will differ across the study arms.

In the replication example with two studies with different eligibility requirements, it may be possible to achieve common support with respect to $X'$ (or *X*) for *subpopulations* of *P* and *Q* by trimming units that do not overlap on $X'$. For example, the researcher may replicate the causal effect for the subpopulation of units that are less advantaged. The researcher may define a "trimmed" replication population, *R*, for which common support on $X'$ (or *X*) can be assumed. This requires a re-analysis of the original study where the advantaged units are deleted based on clearly defined eligibility criteria. The replication population may refer to the characteristic distribution of the subset in *P*, in *Q*, or any other distribution to which we wish to extrapolate. Depending on the choice of the replication population *R*, the study populations *P* and *Q* may need to be reweighted or matched to reflect the distribution of $X'$ in *R*.

Thus far, we have only discussed the assumptions with respect to an additive treatment effect. In cases where the treatment effect is multiplicative, stronger assumptions are needed. When treatment effects are multiplicative, A2.2 requires identical *outcome*-generating functions, $f_0 = f_1$ (in equation 1), rather than identical effect-generating functions, and A2.3 and A2.4 require *identical distributions with respect to X and S* rather than $X'$ and $S'$. Finally, one might consider effect ratios instead of ATEs for evaluating the causal effect of a treatment.

A3. The Causal Estimand is Identified in Both Study Arms

Assumptions A1 and A2 restrict the data-generating process and the choice of a causal estimands to ensure direct replicability, at least in theory. In practice, we also need to make sure that the causal estimand, $ATE_R$, is identified in both studies. We addressed identification assumptions above (i.e. see section on the "causal identification for a single study"), but replication designs pose additional challenges for identification. One issue occurs when target populations in the original and replication studies are not equivalent in terms of the distribution of $X'$ and $S'$. Then, identification requires that one or both study populations need to be reweighted or matched with respect to a potentially trimmed replication population *R*. However, when $X'$ and $S'$ are not fully observed or reliably measured, then the $ATE_R$ will not be identified in at least one study arm. Thus, even if $ATE_P$ and $ATE_Q$ are identified, it does not imply that $ATE_R$ is identified because it requires reliable measures of $X'$ and $S'$ for reweighting or matching with respect to *R*.

*Implications for practice.* The identification of the same unique causal estimand, $ATE_R$, in both studies is facilitated if both studies rely on identical or very similar study designs that (a) require the least assumptions about inferring a causal effect, and (b) draw units from the same target population and site with identical eligibility criteria. For example, if both studies use a well implemented RCT, then each study's causal effect is identified with a minimal set of assumptions (independence and SUTVA). Moreover, if both RCTs draw their units from the same eligible target population, then the two study populations *P* and *Q* will be very similar with respect to *X* and *S*, such that no reweighting or matching with regard to a shared replication population *R* might be required.

When research designs across study arms differ and have different underlying target populations, it will be more challenging for the researcher to address A3. The researcher may not have complete knowledge and reliable measure of unit and setting characteristics. In addition, for study designs that rest on strong causal identification assumptions (e.g., NECGD or instrumental variable designs), replication success will also depend on the credibility of the research design in addressing these assumptions. For example, failure to replicate may be due to violations of the conditional independence or exogeneity assumption.

A4. The Causal Estimand is Estimable Without Bias in Both Studies.

Once $ATE_R$ is causally identified in each study arm, the effect must be estimated without bias in the original and replication study. Ideally, both studies use an unbiased estimator or at least a consistent estimator, provided sample sizes are large such that the bias is negligibly small. This assumption may be violated if, for example, the original study uses the mean difference between the outcomes of the randomized treatment and control group to estimate the causal effect, while the replication study uses observational data and a linear regression estimator with an additional covariance adjustment (for a set of observed covariates). Though the mean difference is unbiased for ATE, the regression estimator may be biased due to a violation of the linearity assumption and the implicit variance-of-treatment weighting (i.e., instead of ATE, a variance-of-treatment weighted ATE is estimated; Angrist & Pischke, 2008).

*Implications for practice.* Ensuring unbiased estimation of effects will depend on the estimation procedure and the corresponding assumptions. To avoid functional form and distributional assumptions, researchers should consider non- and semiparametric estimation procedures (despite the slight loss in efficiency). Replication studies also benefit from large

sample sizes: consistent estimators will be less biased and the replication of effect estimates may be tested with more power.

A5. Estimands, Estimators, and Estimates are Correctly Reported in Both Studies.

The last assumption focuses on correct reporting of the causal estimands, their estimators, and the corresponding estimates. Even after $ATE_R$ has been correctly estimated in both studies, incorrect reporting may lead to the conclusion that the replication study failed to replicate the causal effect estimate of the original study. A replication failure occurs if the two estimates diverge due to incorrect reporting. That is, in at least one of the studies, the published effect differs from the estimated effect. This may result if there is a typographical error in transcribing effects, or an inadvertent use of the wrong table of effects. It is also possible the causal estimand or estimator has been incorrectly reported in one study such that the original and replication study no longer seem comparable. Then, the results of a replication study will very likely be dismissed even if the effect estimates are almost identical.

*Implications for practice.* Reporting errors can never be entirely ruled out. However, if study protocols, code files of the analysis, and data are published in addition to the final report, then independent investigators may at least examine the reproducibility of results by (a) reading what actually has been done according to protocols and code files and (b) by reanalyzing the data provided. Thus, funding agencies and journal editors should increase requirements for data transparency to improve accessibility of files for reproducibility efforts.

**Evaluating Replication of Effects**

Researchers use multiple methods for evaluating the replication of effects. The most common set of measures look at the direction, magnitude, and statistical significance patterns of effects in the original and replication studies. These approaches are similar to vote counting

procedures in research synthesis, where correspondence in results is determined by whether treatment effects in the original and replication studies exceed a substantively important threshold or are statistically significant.

An alternative measure of correspondence in results includes estimating the difference in original and replication study results, which we will call "replication bias" ($\Delta_R$). Here, replication bias is defined as the expected difference in original and replication effect estimates: $\Delta_R = E(\hat{\tau}_0) - E(\hat{\tau}_1)$, where $\hat{\tau}_0$ is the effect estimate for the original study and $\hat{\tau}_1$ is the effect estimate for the replication study. Whether results replicate is determined through substantive and statistical criteria. The researcher may evaluate whether the effect difference is below some minimum threshold for which the results are assumed to be equivalent. To account for sampling error, the researcher may conduct formal statistical tests of difference, equivalence, or compare the confidence intervals of original and replication study results (see Steiner and Wong (2018) for assessing correspondence of results in design replication studies). When multiple replication results exist (i.e. the "Many Labs" Project), the researcher may use meta-analytic approaches for synthesizing effect sizes across multiple studies (Valentine et al., 2011).[4]

## 3. Replication Design Variants

We have demonstrated the stringent assumptions needed for multiple studies to produce the same effect. These assumptions are required for any direct or statistical replication study, where the goal is to implement the exact same study and treatment procedures on new random samples of participants. However, replication design assumptions may also be used as a

[4] Assessing the correspondence of results in an original and replication study is a central issue for a replication as a research methodology. However, because this paper is about replication as a research design, we address these methods briefly and refer readers to more extended discussions of these methods (Valentine et al., 2011).

framework for assessing the robustness of effects, and for identifying sources of treatment effect heterogeneity. For example, researchers may implement different replication design variants to probe whether there were violations in identification (A3) or estimation (A4) assumptions, or to evaluate whether results replicate over different sub-populations of participants (A2). In these cases, violations of any replication assumptions will produce differences in effect estimates. Therefore, the most interpretable replication designs will be ones that evaulate potential violations to specific design assumptions systematically, ensuring that all other assumptions are met.

In this section, we discuss how knowledge of replication assumptions can help researchers understand the purposes of different replication design variants. We highlight three common replication designs – *prospective*, *within-study*, and *matched* approaches – and discuss their basic design structures, their relative strengths and weaknesses, and examples of each method. Table 2 summarizes examples for each replication design variant, and how well replication design assumptions were addressed in each case.

**Prospective Replication Designs**

In prospective replication designs, both study arms – the original and replication studies – are planned in advance and conducted simultaneously. Although prospective replication designs are rare, they provide an important example for how researchers may incorporate research design features for addressing replication assumptions.

The basic structure of the prospective replication design is depicted in Figure 1. Here, participants are sampled from an overall target population and are randomly assigned into one of two study arms. This ensures that the target population will be the same across both study arms. Within each study arm, participants are randomly assigned again into treatment and control

conditions. The same treatments, measures, and research protocols are administered in both study arms, at the same time. Identical analytic procedures are used to estimate the causal estimand of interest. The same reporting protocol is used to describe the data, treatments, analysis code, and sensitivity tests. The two study arms may be implemented by independent – but coordinated – research teams.

*Example 1.* A variant of the prospective approach was introduced by Shadish, Clark and Steiner (2008) and implemented again by Shadish, Galindo, Wong, and Steiner (2011). We describe Shadish et al. (2008) to demonstrate how prospective replication designs may be implemented in real world settings, and what can be learned from this approach. In Shadish et al. (2008), university students within the same site were randomly assigned into two study arms. Within each study arm, they were assigned again into a short vocabulary or math workshop. The interventions were short, highly scripted, and implemented with high fidelity. Outcome measures of students' achievement in vocabulary and math were administered immediately after the intervention was completed.

Instead of two RCTs, Shadish et al. used an RCT only for the original study but a NECGD with self-selction for the replication study. They did so because their goal was to evaluate whether observational methods can identify and estimate causally valid treatment effects in practice. That is, Shadish et al. investigated whether the observational study can replicate the causal benchmark estimate from the RCT. For the NECGD, the research team allowed participants to select the mathematics or vocabulary training of their preference. They then used propensity score matching to estimate the average treatment effect from the NECGD and compared it to the RCT estimate. Any difference in results from the original and replication study was interpreted as "bias" from the observational method. This approach is sometimes

referred to as a "design replication" study. This is because the researcher interprets failure to replicate results as the result of poor performance in the observational design for identifying valid treatment effects (violation of A3).

The Shadish et al. (2008) study demonstrates the multiple advantages of prospective replication designs (Table 2). The treatment and control conditions were well defined across both study arms, and because the intervention was short and relatively low-stakes, there was not much opportunity for spillover or peer effects (A1). Outcomes were measured in the same way across treatment conditions and study arms, and implemented at the same time (A1). In both study arms, they aimed at the same causal estimand (ATE) and randomization into study arms ensured that target populations in the original and replication study were equivalent (A2). Treatment effects were estimated the same way for the matched and RCT samples, ensuring that there were no differences due to estimation procedures (A3). Subsequent reanalysis of the original data by independent investigators found no reporting errors of results (A4). Thus, assuming STUVA (e.g., absence of preference or peer effects), any difference in effect estimates was credibly interpreted as failure in the observational method to identify valid treatment effects (A3).

The prospective design allows researchers to evaluate violations to replication design assumptions, and to identify potential sources of treatment effect variation. In the Shadish et al. (2008) example, the researchers concluded that despite using different methods for identifying effects (RCT vs. observational methods), the study arms were able to replicate results. This finding implied that in this specific context at least, A3 was met (identification of equivalent causal estimands). It is also possible for prospective designs to examine other potential violations to replication assumptions, including whether results replicate when the intensity of treatment dosage varies across study arms; when participants with higher pretest scores are assigned to the

replication arm; or when different estimators are used to estimate treatment effects. The strength of the replication design rests on whether other replication design assumptions (that are not being evaluated) are met.

There are limitations of the prospective approach as well. The design must be planned in advance, so it may be time consuming and expensive for researchers to implement. If units are randomized into the original and replication studies, it requires sufficient sample sizes – often within the same site – to support multiple study arms. To ensure well defined treatment conditions that can be implemented with fidelity, it helps if the intervention is short and easily standardized with quick follow-up measures. For these reasons, this design may have limited applications in intervention evaluation contexts. Prospective approaches may be most appropriate for researchers who are interested in replication of results in highly controlled, laboratory-like settings with short interventions and follow-up measures.

**Within-Study Replication Designs**

In within-study approaches, the researcher evaluates whether results replicate using the same or similar data, analysis procedures, and samples across study arms, but tests for potential violations of replication assumptions by introducing systematic differences in study procedures across the two arms. For example, an independent investigator may use data and code files provided by the original investigator and attempt to reproduce a result that appeared in the original paper. Here, the study protocol to produce the original result is depicted by the first study arm in Figure 1; the independent investigator's attempt to reproduce the result is depicted by the second study arm. Both study arms share the same sample of participants, the same experimental conditions and outcome measures, the same research design and causal estimand of interest, the same analytic procedures, and the same setting and time. There should be no

differences in results. When lack of correspondence in results occur, it is because of independent investigators reported the results (A5). Within-study replication designs may be implemented by independent investigators, or by the same researcher across both study arms. Below we describe two examples of how within-study replications may be implemented.

*Example 2.* Chang and Li (2015) used a within-study replication design in their effort to reproduce results from 67 papers in 13 economics journals. Using data and code-replication files provided by the original authors, Chang and Li found that only 49% of the publication results could be reproduced (sometimes requiring additional assistance from the original investigators). For studies with results that could not be reproduced, the majority – 79% – could not be replicated because of missing or incorrect data or code to report the correct results.

This example shows two advantages of the within-study design (Table 2). First, it yields replication results with clear interpretations about why these effects did not replicate. The researcher has high confidence that the first four replication design assumptions were met (A1-A4) – any observed discrepancy was due to a reporting error (A5). Second, the design is straight-forward to implement and does not require additional resources in terms of data collection. It needs only data and coding files from the original study, and an independent investigator to assess the reproducibility of results. Because of the feasibility of this approach, journals in economics, medicine, and education are adopting data transparency policies that promote within-study replication efforts. Its limitation, however, is that it identifies replication bias only due to errors in reporting. If, for example, treatment effects were not causally identified in the original study (a violation of A3), the method would reproduce bias from the original study.

*Example 3.* Duncan, Engel, Claessens, and Dowsett (2014) examined a variant of the within-study approach (Table 2). The authors coded 50 articles from high ranking, peer-reviewed

journals in two disciplines: economics and developmental psychology. Their purpose was to compare across disciplines the percentage of published studies that examined the reproducibility of effects using different estimation procedures, data sources, and sub-populations. In this example, the main study result formulated the "original study arm" in Figure 1; results from additional analyses created the "replication arms." Here, replication design assumptions are systematically examined by comparing effects from different sub-populations, data sources, and methods. However, because the comparisons of effects occurred within the same study setting by the same researcher, most other replication design assumptions were met. This allowed the researchers to examine the sensitivity of effects to potential violations of specific replication design assumptions.

Duncan et al. (2014) looked at the percentage of studies that examined results sensitivity using different identification or estimation procedures (violations of A3 or A4), while maintaining same target populations, outcome measures, and treatment conditions (A1 and A2). They also looked at studies that assessed the replication of results using different datasets – which introduced variations in study conditions and outcome measures (violation of A1), as well as in individual and setting characteristics (violation of A2) – but employed the same identification and estimation methods as used in the original studies (A3 and A4). Finally, the subgroup comparisons allowed researchers to assess the robustness of results with different individual characteristics (violation of A2), while ensuring that treatment conditions and research methods remained the same across study arms (A1, A3, and A4). The advantage of this approach is that it is also straight-forward to implement, and results are easily interpretable. Indeed, Duncan et al. found evidence that in economics, the practice is already widely adopted. Between 68% and 90% of recently published papers examined the robustness of results. The

practice, however, is observed in only 18% and 32% of recent papers in developmental psychology. The limitation of this replication design is that because it is usually conducted by the original researcher, it may have little to say about replication bias due to reporting error.

**Matched Replication Designs**

In matched designs, the original and replication arms are conducted at different times, at different sites, and usually, by different investigators. There also may be differences in how treatment conditions are implemented, and in how treatment effects are identified, estimated, and reported across study arms. Matched designs differ from within-study or prospective designs because in the latter two approaches, the researcher introduces systematic variation across study arms through differences in research design features or statistical analysis. In matched designs, study arm differences occur naturally and are not researcher controlled. To address replication assumptions, the researcher may attempt to match on characteristics related to treatment conditions (A1), units and settings (A2), and methodology (A3 and A4). However, a successful replication of results can only be expected if all assumptions A1 through A5 are met. But often, researchers will lack sufficient knowledge about whether their matching procedures will address all replication assumptions. For example, it may not be clear which study factors moderate treatment effects (A2).

In some matched designs, it may be possible for the researcher to compare results for a sub-population of units where replication design assumptions are plausibly met, or for which constant treatment effects may be assumed. However, in many cases, it is impossible for the researcher to match on all possible study and treatment related characteristics that moderate the magnitude of effects. This may be because the factors are either unknown or unobserved by the researcher, or because there is no common support on these covariates in the original and

replication study. For example, the original study arm may include only boys while the replication study includes only girls; the original study may use an RCT for identifying the ATE while the replication uses an RDD to identify a local ATE at the cutoff score. There also may be cultural and setting differences across study arms that cannot be matched. In these cases, the researcher may be interested in estimating treatment effect variation due to potential violations in replication design assumptions. However, a meaningful interpretation of effect variations is only possible if the researcher knows which assumptions has been violated.

The extent to which the researcher will match treatment and study characteristics to address replication assumptions depends on the purpose of the replication effort. In studies that attempt to replicate the original procedures as closely as possible, the researcher may attempt to address all design assumptions by matching on all relevant study (A2 – A5) and treatment (A1) characteristics. In conceptual replications, the researcher may match on broadly defined characteristics related to the "treatment" (e.g. grade retention, early childhood education) (A1), but may be less concerned about matching on study, sample, setting, and method characteristics across the study arms.

One challenge with matched designs is that because of the post hoc nature of the approach, multiple replication design violations will often occur simultaneously. It may be impossible for the researcher to interpret results from matched designs, especially when effects do not replicate. Did the effect not replicate because of variations in contexts and settings, in sample participants, or some other source of hidden variation? It may be hard for the researcher to tell. Currently, replication and original studies are "matched" informally and qualitatively, if at all. There are rarely empirical diagnostics demonstrating how well replication design assumptions are met.

Below, we describe two examples of matched replication designs. The first is a "close" replication design, where the replicators attempted to match as many study features as possible; the second is an example of conceptual replications, where early childhood education researchers attempt to summarize findings from evaluations of pre-kindergarten programs.

*Example 4.* A collaborative of 36 independent research teams from 12 countries examined the reproducibility of 13 well-known effects in psychology, and the robustness of these effects across samples, settings, and cultures (Klein et al., 2014). The research team selected effects to be replicated based on several key criteria. First, effects 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 in duration and straight-forward for independent investigators to implement. This was to allow for multiple effects to be evaluated in a single testing session. Third, with the exception of a single correlational study, effects were evaluated using simple, two group designs. Fourth, the 13 effects were selected to represent variations in topics, time frames since the original study was conducted, and certainty of their reproducibility.[5]

In total, 6,355 individuals participated in the replication effort. 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. Overall, the researchers found that 10 of the 13 results were replicated based on multiple measures of correspondence; three had weak to no evidence of reproducibility. Statistical tests

[5] Some effects already had been shown to be reproducible by independent researchers, while other effects had not yet been investigated.

suggested that observed differences in sample characteristics and local contexts failed to explain fully the discrepancies in results for the three effects that did not replicate. This led the researchers to conclude that at least with this sample of studies, much of the variation was due to the effects themselves, rather than systematic differences in sites, samples, and contexts.

These findings highlight the challenge of matched designs of even close replication studies (Table 2). Replication bias may occur if any one of the design assumptions is violated. There may be unmeasured differences in treatment conditions across study arms, or from multiple study factors that interact to introduce hidden biases. There could have also been differences based on how long it had been since the original study was conducted, and how well these effects could be translated to online platforms. Finally, effects from one study could have influenced results from another because multiple effects were under investigations simultaneously. The challenge here is that without research design elements to control these factors systematically, it is hard to interpret the source of the effect variations in the replication designs. However, matched 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, samples, settings, and outcomes. In these cases, investigators should conduct empirical diagnostics to probe and describe each design assumption systematically.

*Example 5.* The Brookings Institute and the Duke Center for Child and Family Policy convened a panel of early childhood education experts to review research on conceptual replications of the impacts of state pre-kindergarten (pre-k) programs (Duke University & Brookings Institute, 2017). The panel examined results from the earliest, small scale RCT evaluations of the Perry Preschool (Schweinhart, Montie, Xiang et al., 2005) and the Carolina Abecedarian (Campbell & Ramey, 2010) Projects, to findings from the national RCT evaluation

of Head Start (Puma, Bell, Cook, et al., 2010), to more recent experimental and quasi-experimental evidence of state-specific pre-k programs (Gormley, 2007; Lipsey et al., 2013; Magnuson, Ruhm, & Waldfogel, 2007; Wong, Cook, Barnett, & Jung, 2008). The panel noted that although many studies found "greater improvements in learning at the end of the pre-k year for economically disadvantaged children," (Duke University & Brookings Institute, 2017, pg. 12) the magnitude of these effects varied. In addition, evaluations of early childhood education programs failed to produce consistent results about the longer-term impacts on children's academic and life outcomes.

The panel noted the multiple challenges with identifying reasons for why these studies arrived at different conclusions about the longer-term impacts of state pre-k (Table 2). The studies include tremendous variation in treatments, sample and setting characteristics, and research designs under which the preschool programs were evaluated. For example, the panel observed that the duration, intensity, and focus of the early childhood education has changed across time and program types (from Head Start to private childcare centers to state pre-k programs), with some interventions focused on "whole child" development that includes social-emotional, physical, and cognitive development, while others targeted children's early achievement outcomes (violation of A1). Alternative childcare options for children who are not enrolled in public preschool programs also varied across time, and even across sites within the same study (Feller, Grindal, Miratrix, & Page, 2016). This yielded important differences in the control conditions under which treatments were being evaluated against (violation of A1). Pre-k programs focused on different populations of students – with some targeting children who are at risk for school failure, and others enrolling all age-eligible children, regardless of income or need (violation of A2). In addition, educators may have become more aware of how to promote

children's early socio-emotional, cognitive, and physical development. This may have changed the fundamental process for how children learn and develop in and out of school, resulting in differences in the pre-k effect generating process over time (violation of A2).

Finally, the evaluations studies varied in how well treatment effects were identified and estimated (A3 and A4). Some employed rigorous RCT approaches while others used quasi-experimental methods, such as matching or interrupted series designs. An issue that plagues nearly all longer-term follow-up studies of state pre-k is that students attrite as they move from early preschool to elementary school, and from elementary school, to middle school and beyond. Despite the large body of research that exists on the effectiveness of early childhood education programs, the Brookings panel stated that "understanding the impact of is an extremely complicated endeavor" (Duke University & Brookings Institute, 2017, pg. 3). The Brookings report highlights the challenges with interpreting results from conceptual replications when none of the replication design assumptions are met. It may be impossible for the researcher to understand inside the "black box" of why treatment effects vary across studies.

## 4. Discussion

This paper highlights the multiple, stringent assumptions required for replication designs to produce the same effect (within the limits of sampling error). First, there must be stability in the outcomes and treatment conditions across study arms. This implies that treatment conditions are well defined, and that there are no peer or contamination effects across study arms. Second, the causal estimand of interest that is being compared must be same. That is, the causal quantity of interest, the effect data generating process, and the distributions of the population and setting characteristics that moderate treatment effects must all be the same. Third, the causal estimand of interest for the replication population must be well identified across both studies. This may be

achieved through an RCT or a well implemented quasi-experiment. Fourth, an unbiased – or consistent – estimator with sufficiently large samples must be used for estimating effects. Fifth, there should be no reporting error of results and methods. A fair and interpretable replication test requires careful consideration of all replication assumptions.

**Addressing Replication Assumptions**

Given the multiple ways in which replication assumptions may be violated, it is perhaps not surprising that replication rates in the social and health sciences are low (Ioannidis, 2005; Makel & Plucker, 2014; Mackel, Plucker, & Hegarty, 2012). However, the last fifty years of the causal inference literature has provided useful guidance about how researchers may consider and address replication design assumptions.

As a matter of practice, we recommend that researchers are proactive about identifying potential sources of replication bias and addressing replication assumptions. That is, they should identify specific and plausible threats based on substantive knowledge about the data and effect generating processes, hypothesize data patterns that should emerge if these threats are realized, and 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).

This paper provides real world examples for how researchers may address replication assumptions in field settings. To ensure stability in treatment conditions and outcomes across replication arms (A1), researchers should register and document their research procedures and outcome measures, incorporate fidelity measures of study conditions (treatment and control), and

assess whether participants have knowledge of study and treatment status. To ensure equivalence in causal estimands (A2), researchers may implement a prospective replication design and randomly assign units into study arms. They may also match or reweight participants so that the distribution of unit characteristics is similar across both study arms. Treatment effects may be identified through an RCT or quasi-experimental design (A3). Covariate balance tests of treatment and control units are helpful diagnostics for addressing this assumption. Ensuring unbiased estimation of effects will depend on the estimation procedure itself (A4), but a pre-registration plan for the analysis protocol will help document sensitivity tests that are conducted by the reseasrchers. It may also provide an opportunity for researchers to receive feedback on their sensitivity tests. To reduce reporting error of results, pre-registration of analysis plans that specify treatment effects of interest will help researchers decide in advance which results to report. Researchers may also improve the reproducibility of their own results by sharing original data and code files of published work. Finally, in cases where replication design assumptions are not met, baseline measures describing differences in treatment implementations, causal quantities of interest, unit and setting characteristics, and research methods are helpful for describing the extent to which assumptions have been violated.

**Selecting a Replication Design Approach**

As we have shown, replication design approaches vary in their comparative strengths and weaknesses. They also serve different purposes. A key benefit of prospective and within-study approaches is that these designs provide greater confidence that replication assumptions are met. They also help researchers understand why treatment effects vary when results do not replicate. Our sense is that these approaches may be most useful early in the research cycle, when the investigator is evaluating and assessing the robustness of results to potential violations in the

replication design assumption related to minor differences in settings and units, as well as in research methodology and reporting. However, these designs may be limited in that they fail to capture all potential sources of replication bias that may occur. For example, within-study replications may reproduce the same estimation bias across studies if the goal is to evaluate only whether there was reporting error.

Matched replication designs are also critical for the accumulation of scientific knowledge. There is a strong need to know whether treatment effects are replicable over natural sources of variation, and the extent of the variation in treatment effects. However, results from matched replication studies may be challenging for researchers and policy-makers to interpret given that diagnostics for replication assumptions are rarely reported.

To improve the replicability of results in science more generally, our recommendation is to encourage replication design variants at every phase of the research cycle. For example, as the intervention is being developed and piloted, prospective replication studies conducted by dependent and independent researchers may be appropriate for assessing the robustness of results in controlled lab settings. Once the intervention has been evaluated, data have been collected, and results are to be published, within-study replications may be warranted to assess the robustness of results to different methodological assumptions, and to detect reporting errors from the researcher. Matched replication designs are useful for understanding treatment effect heterogeneity under natural sources of variation in study and treatment conditions. The replicability of results from matched designs is especially important as policy-makers and program officials consider adopting an intervention or treatment for scale-up.

Finally, it is worth noting that replication assumptions are closely related to assumptions required for causal generalization (Cole & Stuart, 2010; Stuart, Cole, Bradshaw & Leaf, 2011;

Tipton, 2012), or transportability of effects (Pearl & Bareinboim, 2014). The replication design can be seen as an effort to empirically evaluate whether the causal effect found in the original study can be transported to the population, site, and time of the replication study (or vice versa). The only difference is that the replication design requires the same target population across both study arms (A2). This is not required in causal generalization methods because it's very purpose is to extrapolate treatment effects to target populations not included in the original study.

**Limitations**

This paper focuses on presenting replication as a research design for evaluating whether two study effects replicate (within the limits of sampling error). To this end, we have not addressed other important methodological issues related to the *implementation* and *analysis* of replication designs. Goodman, Fanelli, and Ioannidis (2016), for example, distinguishes between "methods" and "results" replications. Methods replication seeks to evaluate whether there is sufficient information for independent researchers to implement the methodological procedure used in the original study, while results replication evaluates whether the same conclusion is obtained by an independent investigator using similar methodological procedures. This paper examines conditions required for results replication. Although knowledge of replication design assumptions may improve "methods replication," we do not address the many implementation issues that arise in ensuring that methodological procedures are replicable across study arms. Future work is needed to address these questions.

Recent methodological work on replication has focused on statistical and substantive criteria for assessing correspondence in study results (Goodman, 1999; Simonsohn, 2015). In addition to replication bias, two studies may fail to produce the same treatment effect because of sampling error, or because either or both study arms are underpowered for comparing effects

(Simonsohn, 2015). There is no current standard for making statistical inferences about whether results from two studies replicate. Steiner and Wong (in press) discuss the benefits and limitations of conclusion-based (e.g. comparing the direction, size, and statistical significance patterns) and distance-based measures (e.g. estimates of replication bias) for assessing correspondence in results in design replication approaches with two study arms. They recommend an equivalence framework for assessing correspondence in results (Steiner & Wong, in press). Others have recommended Baysian methods for assessing correspondence in results (Goodman et al., 2016; Rindskopf, Shadish, & Clark, in press; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011), or changing the p-value for determining statistical significance under NHST (Benjamin et al., 2017). When results from multiple replications are available, meta-analytic methods are essential for estimating systematic and random sources of variation in treatment effects.

**Conclusion**

One of the great methodological innovations in the twentieth century has been the introduction of the potential outcomes framework for understanding causal inferences. The approach has the clear advantage of identifying causal estimands of interest, as well as helping researchers understand the formal assumptions needed for a research design to produce valid effects. We have extended this framework to replication designs, and the stringent assumption needed for two studies to produce the same effect. Although replication has yet to be established as a formal methology in its own right, we believe that conceptualizing the approach as research design will provide researchers with tools for understanding and improving replication studies in the future.

References


Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., … Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Aos, S., Cook, T. D., Elliott, D. S., Gottfredson, D. C., Hawkins, D., Lipsey, M. W., & Tolan, P. (2011). Commentary on Valentine, Jeffrey, et al. Replication in Prevention Science. *Prevention Science*, *12*(2), 121–122. https://doi.org/10.1007/s11121-011-0219-4

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Barnett, W. S., Lamy, C., Jung, K., Wong, V. C., & Cook, T. D. (2007). *Effects of five state pre-kindergarten programs on early learning*. National Institute for Early Education Research.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-017-0189-z

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*, 1–29. Retrieved from https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

*Building Evidence: What Comes After an Efficacy Study?* (2016). Washington, DC. Retrieved from https://ies.ed.gov/ncer/whatsnew/techworkinggroup/pdf/BuildingEvidenceTWG.pdf

Campbell, F. A., & Ramey, C. T. (2010). Carolina abecedarian project. In *Childhood Programs and Practices in the First Decade of Life: A Human Capital Integration* (pp. 76–98). https://doi.org/10.1017/CBO9780511762666.005

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*(83), 1–26. https://doi.org/10.17016/FEDS.2015.083

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, *172*(1), 107–115. https://doi.org/10.1093/aje/kwq084

Collins, F. S., & Tabak, L. A. (2014). NIH plans to enhance reproducibility. *Nature*. https://doi.org/10.1038/505612a

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*(4), 587–603. https://doi.org/10.2307/1806061

Duke University, C. for C. and F. P., & Institution, B. (2017). *The Current State of Scientific Knowledge on Pre-Kindergarten Effects*. *Center for Child and Family Policy, Duke University*.

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology*, *50*(11), 2417–2425. https://doi.org/10.1037/a0037996

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

Estimating the reproducibility of psychological science. (2015). *Science*, *349*(6251). Retrieved from http://science.sciencemag.org/content/349/6251/aac4716.abstract

Feller, A., Grindal, T., Miratrix, L., & Page, L. C. (2016). Compared to what? Variation in the impacts of early childhood education by alternative care type. *Annals of Applied Statistics*, *10*(3), 1245–1285. https://doi.org/10.1214/16-AOAS910

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*(6277), 1037 LP-1037. Retrieved from http://science.sciencemag.org/content/351/6277/1037.2.abstract

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341). https://doi.org/10.1126/scitranslmed.aaf5027

Gormley, W. T. (2007). Early childhood care and education: Lessons and puzzles. *Journal of Policy Analysis and Management*, *26*(3), 633–671.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference: For statistics, social, and biomedical sciences an introduction*. *Causal Inference: For Statistics, Social, and Biomedical Sciences an Introduction*. https://doi.org/10.1017/CBO9781139025751

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, *294*(2), 218–228. https://doi.org/10.1001/jama.294.2.218

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

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*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., Bilbrey, C., & Vanderbilt University, P. R. I. (PRI). (2013). Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and First Grade Follow-Up Results from the Randomized Control Design. Research Report. *Peabody Research Institute*, 0–105. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED566667&site=ehost-live

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3), 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*(4), 77–87. https://doi.org/10.1080/00913367.1995.10673490

Magnuson, K. A., Ruhm, C. J., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review*, *26*, 33–51.

Makel, M. C., & Plucker, J. A. (2014). Facts Are More Important Than Novelty: Replication in the Education Sciences. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10.3102/0013189X14545513

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, *8*(APR). https://doi.org/10.3389/fpsyg.2017.00523

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* . Cambridge: Cambridge University Press.

Pearl, J., & Bareinboim, E. (2011). External Validity and Transportability: A Formal Approach. In *JSM Proceedings* (pp. 157–171).

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start Impact Study. Final Report. Journal for Children & Families*. Retrieved from http://files.eric.ed.gov/fulltext/ED507845.pdf

Rindskopf, D., Shadish, W. R., & Clark, M. H. (n.d.). Using Bayesian correspondence criteria to compare results from a randomized experiment and a quasi-experiment allowing self-selection. *Evaluation Review*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

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*(2), 90–100. https://doi.org/10.1037/a0015108

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). The High / Scope Perry Preschool Study Through Age 40 Summary, Conclusions, and Frequently Asked Questions. *Lifetime Effects: The High/Scope Perrry Study through Age 40*, *40*, 194–215.

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*(484), 1334–1344. https://doi.org/10.1198/016214508000000733

Shadish, W. S., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A Randomized Experiment Comparing Random to Cutoff-Based Assignment. *Psychological Methods*, *16*(2), 179–191.

Simonsohn, U. (2015). Small Telescopes. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Steiner, P. M., & Wong, V. C. (n.d.). Assessing Correspondence between Experimental and Non-experimental Estimates in Within-study Comparisons. *Evaluation Review*.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *174*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Tipton, E. (2012). Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts. *Journal of Educational and Behavioral Statistics*. 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*(2), 103–117. https://doi.org/10.1007/s11121-011-0217-6

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual

sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. https://doi.org/10.1037/a0022790

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, *27*(1), 122–154. https://doi.org/10.1002/pam.20310

Figure 1. Replication Design and its Assumptions with Two Study Arms

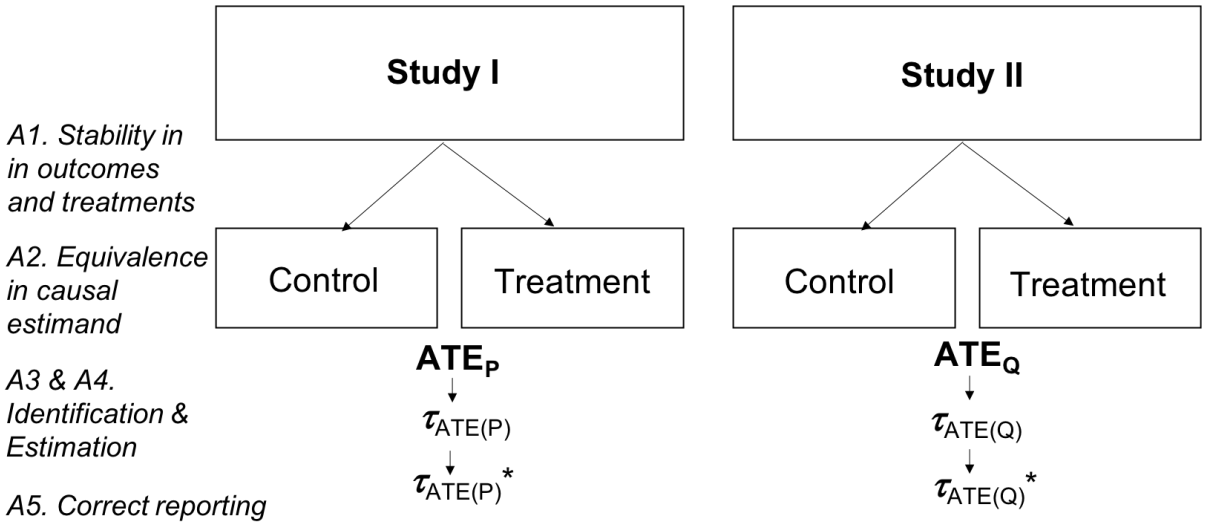

Table 1. Summary of Replication Design Assumptions and Their Implications for Practice

| | Assumption | Requirements | Implications for Practice |
|---|---|---|---|
| **A1** | **Outcome and Treatment Condition Stability** | 1. Well defined and identical treatment conditions across study arms.<br>2. Treatment effects are evaluated on the same outcome measure *Y*.<br>3. Potential outcomes are affected only by exposure to study and treatment conditions, and not by how units were assigned into conditions.<br>4. Non-interference between participants within each study arm, and across study arms. | 1. Evaluate theory and substantive knowledge of study setting to determine whether treatment conditions vary across study arms, whether participants have knowledge of their study and treatment status, whether participants are likely to react to their status, and share experience/knowledge with others.<br>2. Determine in advance outcome measures of interest, and plan for how outcome measures are administered.<br>3. Incorporate fidelity measures for documenting treatment and control conditions.<br>4. Document deviations from treatment and study protocol in pre-registration plan. |
| **A2** | **Equivalence in Causal Estimand** | 1. Both study arms have the same causal quantity of interest (e.g. ATE, ATT).<br>2. The effect-generating processes are identical across both study arms.<br>3. The target populations of both study arms must be equivalent with respect to the joint distribution of effect determining population characteristics *X*'.<br>4. Both study arms must have the same joint distribution of setting characteristics S' that moderate treatment effects.<br><br>If treatment effects are *multiplicative*, assume identical data generating process on the outcome and | *Five scenarios for ensuring equivalence in causal estimands:*<br>**Scenario 1:** The effect is constant across variations in *X* and *S* (no heterogeneous treatment effects).<br>**Scenario 2:** The same observed data of outcomes are used across study arms (reproducibility designs).<br>**Scenario 3:** Participants from an overall target population are randomly sampled into both study arms.<br>**Scenario 4:** Treatment effects are additive. Match, reweight, and trim units so that study arms share the same distribution of unit characteristics *X*'.<br>* Ensure that study arms have the same study and setting characteristics *S*'.<br>* Effect generating process must be the same across arms.<br>* Treatment effects are constant across any remaining and unobserved variations in *X* and *S*.<br>**Scenario 5:** Treatment effects are multiplicative. Match, reweight, and trim units on all values of *X* and *S*.<br>* Data generating process for outcomes must be the same across arms. |

| | | | |
|---|---|---|---|
| | | distributions with respect to all unit $X$ and setting $S$ characteristics. | |
| **A3** | **The Causal Estimand is Identified in Both Study Arms** | 1. The causal estimand of interest must be identified for the target population of interest in the both studies.<br>2. The identification assumption depends on the research design used in each study arm. | 1. Identification assumptions may be met through an RCT or a well-implemented quasi-experiment within each study arm (e.g. independence, conditional independence, exclusion restriction).<br>2. The research design does not need to be the same in both research arms; only the identification of treatment effects is required.<br>* In practice, this assumption is more credible in cases where the same valid research design is used in both study arms (e.g. RCTs in both study arms). |
| **A4** | **Estimator is Unbiased in Both Study Designs.** | 1. Use of unbiased or consistent estimator with sufficiently large sample sizes. | 1. Assumptions depends on estimator used for estimating effects (e.g. linearity assumptions in regression).<br>2. Each study arm includes large sample sizes for consistent estimators. |
| **A5** | **Correct Reporting of Estimands, Estimators, and Estimates** | 1. Error in transcribing results into a table.<br>2. Incorrect reporting of causal estimand or estimator. | 1. Pre-registration of results of interest can help prevent reporting bias.<br>2. Pre-registration of causal estimand of interest<br>3. Include multiple investigators to examine the reproducibility of results at the analysis and publication phase.<br>4. Adopt data transparency policies at the funding and publication phases. |

Table 2. Examples of Replication Design Variants

| | **Prospective Designs** | **Within-Study Designs** | | | | **Matched Designs** | |
|---|---|---|---|---|---|---|---|
| | Shadish et al. (2008) | Chang & Li (2015) | Duncan, Engel, Claessens, & Dowsett (2014) | | | Many Labs Project (2014) | Pre-k Consensus Report (2017) |
| | | | Different Datasets | Different Subgroups | Different Identification and Estimators | | |
| A1. Outcome and Treatment Condition Stability | ✓ | ✓ | X | ✓ | ✓ | X | X |
| A2. Equivalence in Causal Estimand | ✓ | ✓ | X | X | ✓ | X | X |
| A3. The Causal Estimand is Identified in Both Study Arms | X | ✓* | ✓* | ✓* | X | ✓ | X |
| A4. Estimator is unbiased in both study designs. | ✓ | ✓* | ✓* | ✓* | X | ✓ | X |
| A5. Correct reporting of results. | ✓ | X | ✓* | ✓* | ✓* | ✓ | X |

✓ indicates that the researchers had high confidence that this assumption was met; **X** Indicate that the replication design was meant to test sensitivity of results to this assumption; * indicates that it was unclear whether this assumption was met.