Design-Based Approaches to Causal Replication Studies

Authors
Vivian C. Wong (University of Virginia), Kylie Anglin (University of Virginia), and
Peter M. Steiner (University of Maryland)

Corresponding Author
Vivian C Wong, vcw2n@virginia.edu

January 2021

Abstract

Recent interest in promoting replication efforts assume that there is well-established methodological guidance for designing and implementing these studies. However, no such consensus exists in the methodology literature. This article addresses these challenges by describing design-based approaches for planning systematic replication studies. Our general approach is derived from the Causal Replication Framework (CRF), which formalizes the assumptions under which replication success can be expected. The assumptions may be understood broadly as replication design requirements and individual study design requirements. Replication failure occurs when one or more CRF assumptions are violated. In design-based approaches to replication, CRF assumptions are systematically tested to evaluate the replicability of effects, as well as to identify sources of effect variation when replication failure is observed. The paper describes research designs for replication and demonstrates how multiple designs may be combined in systematic replication studies, as well as how diagnostic measures may be used to assess the extent to which CRF assumptions are met in field settings.

Keywords: Replication, causal inference, open science

Introduction

Despite interest by national funding agencies to promote and fund systematic replication studies for validating and generalizing results (Department of Health and Human Services, 2014; Institute of Education Sciences, 2020; National Science Foundation, 2020), there is not yet consensus on what systematic replication is, how replication studies should be conducted, nor on appropriate metrics for assessing replication success (Institute of Education Sciences, 2016). The lack of methodological guidance on these issues is challenging for evaluators designing replications studies and for sponsors making decisions about whether research plans are of sufficient quality for funding.

This article addresses these concerns by describing design-based methods for planning systematic replication studies (that is, a series of prospectively planned individual studies). The approach extends methodological insights from experimental designs and the causal inference literature with individual studies (Rubin, 1974) to replication efforts with multiple studies. In individual evaluation studies, the researcher first chooses a causal estimand of interest, or the causal effect of a well-defined treatment-control contrast for a clearly determined target population and setting, and then selects an appropriate research design, such as a randomized- or quasi-experiment, to identify and estimate the causal estimand of interest (Imbens & Ruben, 2015; Morgan & Winship, 2014; Shadish et al., 2002). For a research design to yield valid results, stringent assumptions must be met (Angrist & Pischke, 2009). To assess these assumptions empirically, the researcher may report results from diagnostic probes, such as balance tests that demonstrate group equivalence at baseline in a randomized experiment. Results from diagnostic probes are critical for helping both the researcher and the reader evaluate the credibility of causal inferences from a study (Angrist & Pischke, 2009; Rosenbaum, 2017).

Design-based approaches to replication adapt and apply methodological principles from causal inference to planning and evaluating high-quality replication studies. This perspective addresses a critical challenge in replication – making appropriate inferences about *why* replication failure occurred.[1] Currently, when studies produce different results, it is often difficult for the researcher to discern whether this occurred because of bias and error in individual studies, or because of differences in target populations, treatments, outcomes, and settings that amplify or dampen the effect across studies. In the latter case, the source of effect heterogeneity cannot be determined because too many study factors are varied simultaneously across replication studies. To understand the sources of replication failure, we argue that researchers should: 1) define a causal estimand of interest across studies; 2) select an appropriate research design for replication that systematically tests hypotheses about potential sources of effect heterogeneities; and 3) analyze and report results from diagnostic tests that assess the assumptions of the replication design required for making causal inferences about variations in effect estimates.

To formalize the assumptions under which causal effect estimates can be expected to replicate and determine the systematic sources of effect variation in results across studies, we rely on the Causal Replication Framework (CRF; Steiner et al., 2019; Wong & Steiner, 2018). For direct replication efforts, CRF assumptions ensure that the same causal estimand is compared across studies, and that the effect is identified and estimable without bias, as well as correctly reported in each study. Since *direct replications* examine whether two or more studies with the same underlying causal estimand (that has the same treatment-control contrast, outcome variable,

[1] Currently, there is no standard approach for determining replication failure. Researchers often compare the direction, size, and statistical significance patterns of study effects; they have also examined statistical tests of difference and/or equivalence of study results. In this article, we will define replication failure as statistical differences in two or more study effect estimates.

target population, and setting) yield the same effect estimates within the margin of sampling uncertainty, they are useful for ruling out chance findings and biases or for reporting errors in individual studies (Simons, 2014).

However, if the goal is to evaluate whether variations in treatments, units, settings, and/or outcomes across studies yield different effect estimates, the researcher should select a *conceptual replication* design and evaluate whether the systematic variations in the causal estimand result in different estimates across studies. In introducing systematic variations, the researcher intentionally "violates" one or more CRF assumptions that would be required in a direct replication effort. If the effect estimates do not replicate (that is, differ significantly), the researcher concludes that the deliberately induced variations caused the differences in results. Results from replication efforts are most interpretable when the variations across studies are implemented in controlled settings and restricted to one or two factors only. Importantly, in design-based approaches to replication, "replication failure" is not a scientific failure – it is actually a success – so long as the replication design is able to identify which of the systematically varied study factors caused the heterogeneity in effect estimates.

In this article, we focus on research design variants for conceptual replications because identifying causal sources of effect variation is essential for theory development and generalizing of effects (Cartwright & Hardie, 2012; Cole & Stuart, 2010; Stroebe & Strack, 2014; Stuart et al., 2011; Tipton, 2012; Tipton & Olsen, 2018). We will also show that in cases where multiple sources of effect variation are hypothesized, the researcher may plan a series of different research designs for replication to test each potential source (or set of sources) of effect variation systematically.

Finally, because deviations from study protocols are common in field settings, design-based approaches to replication emphasize diagnostic probes for assessing the extent to which CRF assumptions are met in field settings. We argue that the researcher cannot make inferences about the source of potential effect heterogeneities without evidence from diagnostic probes that demonstrate how well the planned replication design was implemented. We also show that reporting of results from diagnostic probes has benefits for both individual and replication study efforts. For each individual study, reporting diagnostics of CRF assumptions can help researchers interpret the validity of individual study findings, as well as provide critical information for guiding future replication studies or research synthesis efforts. For replication studies, reporting diagnostics allow researchers to evaluate the extent to which CRF assumptions are met across studies for making inferences about the causal sources of effect heterogeneity.

The Causal Replication Framework

One challenge underlying the planning of many replication studies is that replication as a method has yet to be established. There is not yet agreement on the definition of replication nor on appropriate standards for determining "high quality" replication studies. The CRF defines replication as a research design that tests whether two or more studies produce the same causal effect within the limits of sampling error (Wong & Steiner, 2018). The core of the framework is based on potential outcomes notation (Rubin, 1974), which has the advantage of clearly defined causal estimands of interest and assumptions for the direct replication of results. Table 1 summarizes the five sets of assumptions under which replication success can be expected. These assumptions can be categorized into *replication assumptions* (R1-R2) and *individual study assumptions* (S1-S3). A full discussion of each assumption can be found in Steiner and Wong

(2018). Here, we briefly describe the assumptions and their implications for design-based replication.

Replication Assumptions

Replication assumptions (R1-R2) ensure that the same causal estimand is compared across all studies in the replication effort. This requires treatment and outcome stability (R1) and equivalence in causal estimands (R2). Treatment and outcome stability means that treatment conditions must be well-specified and implemented in identical ways across all studies (that is, there must be no hidden variations in both intervention and control conditions). The assumption is violated if there are variations in treatment and control conditions across studies or if outcome measures differ across studies, such as when different instruments are used, or when the same instrument is used, but administered at different times and settings. The second replication assumption (R2) requires an equivalence of causal estimands across studies. This implies that there must be identical joint probability distributions of all population and setting characteristics that may moderate the effect. This may be achieved by either ensuring that the studies sample from the same target population of interest, or by matching participants across studies to achieve an equivalent joint distribution of participant characteristics. Finally, equivalence in the causal estimand requires that all studies should focus on the same causal quantity. For instance, all studies should aim at the average treatment effect (ATE), the intent-to-treat effect (ITT), or the average treatment on treated effect (ATT). The ATE from one study should not be compared to the ITT or ATT from a different replication study. In cases where there is effect heterogeneity, comparing impacts for different subpopulations will likely result in replication failure.

Individual Study Assumptions

Individual study assumptions ensure that the causal estimand of interest is identified and estimated without bias and correctly reported *in each individual study*. This requires the identification of a causal estimand (S1), unbiased estimation of the causal estimand (S2), and correct reporting of the estimand, estimator, and estimate (S3). These assumptions are met if each individual study in the replication effort has a valid research design for identifying effects, appropriate estimators for estimating effects, and correct reporting of results. Assumptions S1 and S2 may be violated if, for example, a study fails to successfully address attrition, nonresponse or selection bias, or if a result is estimated by a misspecified regression model. Assumption S3 is violated when the same data and syntax file fail to yield the same reported study findings, as in reproducibility efforts (Chang & Li, 2015; LeBel et al., 2018). These are standard assumptions for any individual study to yield a valid causal effect. The *Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science* provide comprehensive recommendations for identifying and estimating a causal estimand with limited bias (S1-S2), as well as for reporting estimates (S3; Gottfredson et al., 2015). For the purposes of this paper, we focus our attention on replication assumptions (R1-R3), as they are less familiar to readers, but return to single study assumptions (S1-S3), in our discussion of diagnostic probes.

Implications for Design-Based Replication

Under the CRF, the goal of replication is to evaluate whether the same result is produced while addressing and testing replication (R1-R2) and individual study (S1-S3) assumptions. Here, the quality of the replication effort is based on the extent to which CRF assumptions are systematically met (or not met). Replication failure occurs when one or more assumptions are violated. In design-based approaches to replication, replication failure provides empirical evidence for sources of effect heterogeneity when constant treatment effects across units, contexts, and settings cannot be assumed. In this way, replication failure is essential for scientific discovery, but only when the researcher is able to determine why it occurred.

Causal Replication Designs

Design-based approaches to replication depend on well-specified research questions regarding the causal estimand of interest. This means that researchers need to specify the units, treatments, outcomes, and settings of interest, and which of these factors, if any, the researcher wishes to vary. Given the "innumerable" potential variations in causal estimands, Nosek and Errington advise researchers to ask, "which ones matter (Nosek & Errington, 2020, p. 4)?" The selection of factors for systematic testing, and the extent to which these factors should be varied, necessarily depends on the researcher's deep subject matter theory of the intervention, as well as their expert knowledge about the most important factors that are hypothesized to result in effect variation (Simons et al., 2017). These are conditions that the researcher believes are both necessary and sufficient for replicating a causal claim. Explicating these factors is needed for understanding how an intervention's effect is meant to generalize, as well as the limits of the intervention's effect under investigation (Nosek & Errington, 2020).

In design-based approaches to replication, applying subject matter theory for selecting a replication design is operationalized through the researcher's question regarding the causal estimand of interest. For example, direct replications seek to evaluate whether two or more studies with the same well-defined causal estimand yield the same effect. Although the most stringent forms of direct replication seek to meet all replication and individual study assumptions, the most informative direct replication approaches seek to test one or more individual study assumptions (S1-S3) for producing replication failure. High quality direct replications require that CRF assumptions R1 and R2 are met because these assumptions ensure that studies compare the same causal estimand, while introducing systematic sources of variation that test individual study assumptions (S1-S3). Examples include within-study comparison designs (Fraker & Maynard, 1987; Lalonde, 1986), which compare effect estimates from an observational study with those from an RCT benchmark with the same target population (S1); robustness checks (Duncan et al., 2014), which compare effect estimates for the same target population using different estimation procedures (S2); and reproducibility analyses (Chang & Li, 2015), which compare study results produced by independent investigators using the same data and syntax code. In all of these approaches, the researcher concludes that an individual study effect is biased or incorrectly reported (that is, a violation of individual study assumptions S1-S3) if replication failure is observed.

Conceptual replications, however, seek to examine whether two or more studies with potentially different causal estimands produce the same effect. To implement this approach, the researcher selects and introduces variations in units, treatments, outcomes, and settings (R1-R2) while attempting to ensure that all individual study assumptions (S1-S3) are met. The goal is to test and identify potential sources of effect variation based on subject matter theory, often for the purpose of generalizing effects for broader target populations (Clemens, 2017; Schmidt, 2009; Simons et al., 2017).

Definitions of conceptual and direct replications under the CRF complement existing, more heuristic approaches to replication (Brandt et al., 2014; LeBel et al., 2018). An advantage of the CRF, however, is that it provides a formal method for deriving replication designs that systematically test sources of effect heterogeneity, as well as for evaluating the quality of the replication design for making inferences. The remainder of this section focuses on research designs for conceptual replication. Although the designs we discuss are widely implemented in field settings, they are not currently recognized as replication designs. Understanding these approaches as replication designs demonstrate that it is both feasible and desirable to conduct high quality replication studies in field settings, as well as to make inferences about why replication failure occurred.

Multi-Arm RCT Designs

Multi-arm RCTs are designed to evaluate the impact of two or more intervention components in a single study. Participants are randomly assigned to one of multiple intervention arms with differing treatment components, or to a control group. This allows researchers to compare a series of pairwise treatment contrasts. For example, in a study evaluating the effectiveness of personalized feedback interventions for reducing alcohol-related risky sexual behavior, researchers randomly assigned participants to one of three arms: one arm received personalized information on alcohol use and personalized information on sexual behavior ("additive approach"); a second arm received personalized information on the relationship between alcohol and risky sexual behavior ("integrated approach"); and a third control arm received an unrelated information on nutrition and exercise (Lewis et al., 2019).

This multi-arm RCT may be understood as a replication design that purposefully relaxes the assumption of treatment stability (R1) to test whether the effect of a personalized feedback intervention replicates across variations in feedback content relative to the same control condition (an additive versus integrated approach). Because systematic variation is introduced within a single study, all other CRF assumptions other than treatment stability (R1) may be plausibly met: the same instruments are used for assessing outcomes at the same time and settings for all comparisons (R1); the control condition for evaluating intervention effects is the same for each comparison (R1); and, random assignment of participants into different intervention conditions ensures identical distributions of participant characteristics on expectation across groups (R2), and that the causal estimand is identified (S1). The researchers may also examine whether each pairwise contrast is robust to different model specifications, providing assurance of unbiased estimation of effects (S2). If all other CRF assumptions are met, and pairwise contrasts yield meaningful and significant differences in effect estimates, then the researcher may conclude with confidence that variation in intervention conditions caused the observed effect heterogeneity.

RCTs with Multiple Cohorts

RCTs with multiple cohorts allow researchers to test the stability of their findings over time. In this design, successive cohorts of participants are recruited within a single institution or a set of institutions, and participants within each cohort are randomly assigned to intervention or control conditions. As a concrete example, in an evaluation of a comprehensive teen dating violence prevention program, 46 schools were randomly assigned to participate in Dating Matters over two successive cohorts of 6th graders or to a business-as-usual control condition (Degue et al., 2020). Experimental intervention effects for each cohort were compared to evaluate whether the same result replicated over time. This multiple cohort design also facilitates recruitment efforts by allowing researchers to deliver intervention services and collect data over multiple waves of participants, which may be useful in cases where resources are limited.

RCTs with multiple cohorts may be considered a conceptual replication designed to test for effect heterogeneities across cohorts at different time points. To address CRF assumptions, the researcher would implement a series of diagnostic checks to ensure replication and individual study assumptions are met. For example, the researcher may check to ensure that the same instruments are used to measure outcomes, and that they are administered in similar settings with similar timeframes across cohorts (R1). The researcher may also implement fidelity measures to evaluate whether intervention and control conditions are carried out in the same way over time (R2) and whether there are no spill-over effects across cohorts (R2), and they may assess whether the distribution of participant characteristics also remain the same (R2). Finally, to address individual study assumptions (S1-S3), the researcher should ensure that a valid research design and estimation approach are used to produce results for each cohort, and that the results are verified by an independent analyst.

Because RCTs with multiple cohorts are often implemented in the same institutions with similar conditions, many characteristics related to the intervention, setting, participants, and measurement of outcomes will remain (almost) constant over time. However, some replication assumptions (R1, R2) may be at risk of violation. For instance, intervention conditions often change as interventionalists become more comfortable delivering protocols and/or as researchers seek to make improvements in the intervention components or in their data collection efforts. Moreover, intervention results may change if there are maturational effects among participants that interact with the treatment, or if there are changes in settings that may moderate the effect. The validity of the multiple-cohort designs may also degrade over time, as participants in entering cohorts become aware of the study from prior years. When participants have strong preferences for one condition over another, they may respond differently to their intervention assignments, which may challenge the interpretation of the RCT. Replication designs with multiple cohorts provide useful tests for examining treatment effect variation over time. However, the design is most informative when the researcher is able to document the extent to which replication assumptions are violated over time that may produce replication failure.

Switching Replication Designs

Switching replications allow researchers to test the stability of a causal effect over changes in a setting or context. In this approach, two or more groups are randomly assigned to receive an intervention at different time intervals, in an alternating sequence such that when one group receives treatment, the other group serves as control, and when the control later receives treatment, the original treatment group serves as the control (Shadish et al., 2002). Replication success is examined by comparing the treatment effect from the first interval with the treatment effect from the second interval. Helpfully, the design provides an opportunity for every participant to engage with the intervention, which is useful in cases where the intervention is highly desired by participants or when it is unethical or infeasible to withhold the intervention.

Though switching replications are relatively rare in prevention science, opportunities for their use are commonplace; many evaluations incorporate a waitlist control group design where the control group receives the intervention after the treatment group. Waitlist control groups have recently been used to evaluate the impact of parenting interventions (Keown et al., 2018; Roddy et al., 2020), mental health interventions (Maalouf et al., 2020; Terry et al., 2020), and healthy lifestyle interventions (Wennehorst et al., 2016). As a concrete example, in an evaluation of the Champion in Prevention (CHIP) Germany program, treatment participants met twice a week for 8-weeks receiving lessons aimed at preventing Type 2 diabetes and cardiovascular diseases. The control group was provided access to the same program after the 12-month follow-up period (Wennehorst et al., 2016). If the study researchers were additionally interested in the relative effectiveness of an online version of the program, this study could be easily adapted to become a switching replication. In this design, the waitlist control group would serve as a control for the first treatment group, but after the year follow-up, the waitlist group would participate in virtual

CHIP meetings while the first group served as the control. Health outcomes would be measured at the beginning of the study, after the first group receives CHIP, and after the second group receives the online version of CHIP.

In the switching replication design, the RCT in the second interval serves as a conceptual replication of the RCT conducted in the first interval. The primary difference across the two studies is the setting for how the healthy lifestyle intervention was delivered (in-person class versus online class). This allows the researcher to address multiple assumptions under the CRF. Because participants are shared across both studies, the same causal estimand is compared (R2); because participants are randomly assigned into conditions, treatment effects are identified for each study (S1). Reports of results from multiple estimation approaches and independent analysts can provide assurances that assumptions S2 and S3 were met. If replication failure is observed, the researcher may conclude that changes in how the intervention protocol was delivered was the cause of the effect variation.

However, results from the switching replication design are most interpretable when the intervention effect is assumed to be a causally transcient process – that is, once the intervention is removed, there should be no residual impact on participants' health (R1). The assumption may be checked by extending the length of time between the first and second intervals, and by taking measures of health immediately before the intervention is introduced to the second group. The design also requires that the same outcome measure is used for assessing impacts and for comparing results across study intervals (R1), that there are no history or maturation effects that violate CRF assumptions (R2), and no compositional differences in groups across the two study intervals (R2).

Combining Replication Designs for Multiple Causal Systematic Replications

On its own, a well-implemented research design for replication is often limited to testing a single source of effect heterogeneity. However, it is often desirable for the researcher to investigate and identify multiple sources of effect variation. To achieve this goal, a series of planned systematic replications may be combined in a single study effort. Each replication may be a different research design (as described above) to test a specific source of effect variation or to address a different validity threat. The researcher then examines the pattern of results over multiple replication designs to evaluate the replicability and robustness of effects.

As an example, Cohen, Wong, Krishnamachari, and Berlin (2020) developed a coaching protocol to improve teacher candidates' pedagogical practice in simulation settings. The simulation provides opportunities for teacher candidates to practice discrete pedagogical tasks such as "setting classroom norms" or "offering students feedback on text-based discussions." To improve teacher candidates' learning in the simulation setting, the research team developed a coaching protocol in which a master educator observes a candidate practice in the simulation session and then provides feedback on the candidate's performance based on a standardized coaching protocol. The teacher candidate then practices the pedagogical task again in the simulation setting. To assess the overall efficacy of the coaching protocol (the treatment condition), the research team randomly assigned teacher candidates to participate in a standardized coaching session or a "self-reflect" control condition, and compared candidates' pedagogical performance in the simulation session afterwards. Outcomes of candidates' pedagogical practice were assessed based on standardized observational rubrics of candidates' quality instructional practices in the simulation setting (Cohen et al., 2020).

To examine the robustness of effects across systematically controlled sources of variation, the research team began by hypothesizing three important sources of effect variation that included differences (a) in the *timing* of when the study was conducted, (b) in *pedagogical tasks* practiced in the simulator, and (c) in *target populations and study setting*. To test these sources of variation, the research team then implemented three replication designs that included a multiple-cohort design, a switching replication design, and a conceptual replication that varied the target population and setting under which the coaching intervention was introduced.[2] These set of replication designs were constructed from four individual RCTs that were conducted from Spring 2018 to Spring 2020. RCTs took place within the same teacher training program but were conducted over two cohorts of teacher candidates (2017-2018, 2018-2019) and an undergraduate sample of participants (Fall 2019).

Table 2a provides an overview of the schedule of the four RCTs. Here, each individual RCT is indexed by S_{ij}, where i denotes the sample (teacher candidate cohorts 1 or 2 or undergraduate sample 3), and j denotes the pedagogical task for which the coaching or self-reflection protocol was delivered (1 if the pedagogical task involved a text-based discussion; and 2 if the pedagogical task involved a conversation about setting classroom norms). Table 2b demonstrates how each replication design was constructed using the four RCT studies. Here, the research team designated S_{22} as the benchmark study for comparing results from the three other RCTs. For example, to assess the replicability of coaching effects over *time*, the research team looked at whether coaching effects were similar across two cohorts of teacher candidates (S_{22} versus S_{12}). To examine the replicability of effects across *different pedagogical tasks*, the research team implemented a modified switching replication design (S_{22} versus S_{21}). Here,

[2] The replication effort actually consisted of six individual RCTs and five replication study designs. We limit our discussion to include only the first three RCTs and replications studies because of space considerations. Results of the systematic conceptual replication study is available at Krishnamachari, Wong, and Cohen (in progress).

candidates were randomly assigned in Fall 2018 to receive the coaching or the self-reflection protocol in the "text-based discussion" simulation scenario; their intervention conditions were switched in Spring 2019 while they practiced the "text-based discussion" simulation scenario). Coaching effects for the fall and spring intervals were compared to assess the replicability of effects across the two different pedagogical tasks. Finally, to examine replicability of effects over a *different target population and setting*, the research team compared the impact of coaching in the benchmark study to RCT results from a sample of participants who had interest in entering the teaching profession but had yet to enroll in a teacher preparation program (S_{22} versus S_{32}). The sample included undergraduate students in the same institution enrolled in a "teaching as a profession" class but had not received any formal methods training in pedagogical instruction. Participants were invited to engage in pedagogical tasks for "setting classroom norms" and were randomly assigned to receive coaching from a master educator, or to engage in the self-reflection protocol. Table 3 summarizes the sources of planned variation under investigation for each replication design. Anticipated sources of variation are indicated by ╳; assumptions that are expected to be held constant across studies are indicated by ✓.

Combined, the causal systematic replication approach allowed the research team to formulate a theory about the replicability of coaching effects in the context of the simulation setting. The research team found large, positive, and statistically significant impacts of coaching on participants' pedagogical practice in the simulation setting. Moreover, coaching effects were robust across multiple cohorts of teacher candidates and for different pedagogical tasks. The magnitude of effects, however, were smaller for participants who were exploring teaching as a profession but had yet to enroll in the training program. These results suggest that differences in participant characteristics and background experiences in teaching resulted in participants benefiting less from coaching in the simulation setting (Krishnamachari, Wong, & Cohen, in progress).

Assessing and Reporting Assumptions for Replication Designs

Under the CRF, the quality of replication studies is determined by the extent to which replication and individual study assumptions are met. For most assumptions, there are no direct empirical tests for evaluating whether they are met in field settings, but it is often possible to use information from diagnostic measures to probe whether an assumption is violated. This can be done by using design elements and empirical diagnostics to rule out the most plausible threats to validity (Shadish, Cook, & Campbell, 2002).

Though replication designs such as the switching replication or the multiple cohort design can be used to address many CRF assumptions, replication designs are rarely able to protect against violations of all the necessary assumptions under the CRF. Moreover, replication designs are often implemented with deviations from their protocols in field settings. Therefore, diagnostic probes provide empirical information about the extent to which assumptions were actually met in replication settings.

Fortunately, the last thirty years of the program evaluation literature has recommended methods for assessing assumptions that can (a) be used to evaluate the plausibility of individual study (S1-S3) assumptions, and (b) be easily extended to evaluate replication (R1-R2) assumptions. As we will see, subject-specific knowledge about study characteristics that are most likely to moderate intervention effects across studies is essential for selecting appropriate diagnostic measures (Simons et al., 2017). Here too, the CRF provides a structured approach for helping researchers anticipate, plan, and conduct diagnostic measures to assess assumptions empirically. In this section, we discuss and describe examples of how researchers can probe and assess all replication and individual study assumptions in the context of a systematic replication study.

Assessing and Reporting Individual Study Assumptions

The individual study assumptions (S1-S3) require identification of a clearly defined causal effect, unbiased estimation, and the correct reporting of results. To facilitate the identification and unbiased estimation of causal effects, strong research designs such as RCTs or regression discontinuity designs are preferred, but well-designed non-equivalent comparison group designs, difference-in-differences or interrupted time series designs can produce credible impact estimates as well. While each research design requires a different set of assumptions for the causal identification of effects (S1), empirically-based methods for probing the respective assumptions exist. For example, to evaluate whether randomization results in comparable treatment and control groups, it is common practice to assess the balance of groups by comparing the distribution of baseline covariates (e.g., their mean and standard deviation). Such balance checks are even more important when attrition or nonresponse is an issue. If the balance checks indicate group differences due to attrition, a causal interpretation of the effect estimate might not be warranted. However, if subject matter theory suggests that the observed baseline covariates are able to remove attrition bias, then statistical adjustments can still enable the causal identification of the effect. Balance tests can provide reassurance for the researcher and the reader that the randomization procedure in an RCT or attrition and nonresponse did not result in meaningful differences in groups, such that causal inferences become credible. For other research designs, the same or similar techniques for probing the identification assumptions are possible (see Wong et al. (2012) for a review of methods). To address S1, systematic replication studies with RCTs should report – at a minimum – for each replication study balance statistics for a broad set of baseline covariates to demonstrate that the causal assumptions are likely met.

Individual study assumptions also require unbiased estimation (S2) and correct reporting of results (S3) for each study in the systematic replication effort. If, for instance, a regression estimator is used to estimate the effect, then residual diagnostics should be used to assess whether the functional form has been correctly specified. Residual diagnostics also help in assessing whether standard errors, confidence intervals and significance tests are unbiased (homoscedasticity, independence, normality). To probe potential model misspecifications, non-parametric analyses may be used to check the results' robustness. The unbiased estimation also requires that the researchers choose an unbiased or at least consistent estimator for the effects and their standard errors, and that they abstain from questionable research practices like fishing for significant results or HARKing (Hypothesizing After Results are Known; Kerr, 1998). Pre-specified analysis protocols and the pre-registration of studies help ensure that the assumptions are more likely met and easier to assess by independent researchers.

New conventions in reporting and transparency practices also help in improving and assessing the correct establishing sufficient and correct reporting of results. For example, recent Transparency and Openness (TOP) guidelines from the Center for Open Science suggest journal standards for pre-registration of analyses as well as standards for sharing and archiving data and code (Nosek et al., 2015). The guidelines include standards related to data transparency for the sharing and archiving of data, as well as code sharing, which include all data management and analysis files for producing study effects. TOP also includes standards for pre-registration, which encourage researchers to specify their analysis plan for addressing research questions in advance. Combined, these standards facilitate efforts from independent researchers to verify that published results are obtained by appropriate analyses and are correctly reported by making the intended analysis plan transparent, as well as making data and syntax files accessible for reproducing results.

Assessing and Reporting Replication Design Assumptions

While empirical diagnostics for probing study-specific threats have become more widely adopted in recent years (Angrist & Pischke, 2009), less obvious is how researchers should address *replication* assumptions. Here, it is possible to extend diagnostic approaches for checking study-specific assumptions to examine replication assumptions about treatment and outcome stability (R1) and the equivalence of causal estimands (R2).

To establish the equivalence of causal estimands across studies, researchers should ensure that they estimate the same causal quantity (e.g., the average treatment effect, ATE) for the same population in an equivalent setting. Probing these assumptions do not require that populations and settings have to be identical in every respect—which is impossible—but they have to be (almost) identical with regard to the effect-moderating variables. Thus, a thoughtful replication design uses subject matter theory about the presumed data-generating process to determine potential effect moderators and to measure them in both studies. Then, balance tests as described above should be used to assess the equivalence of study populations with regard to the effect moderators and other baseline covariates. The equivalence of the study setting is harder to assess because a single study is typically implemented in a single or only a few settings (e.g., sites). However, the successful implementation of a systematic replication effort demands that effect-moderating setting characteristics are determined based on subject matter theory, and then held constant across settings (provided they are not a planned variation in the replication design).

Careful reporting of the study settings, particularly of potential effect-moderating aspects, helps in assessing the extent to which this assumption is met.

The assessment of treatment and outcome stability (R1) requires researchers to demonstrate that the treatment-control contrasts and the outcome measures are identical across studies (unless deliberately varied as part of the design). A major step towards addressing the outcome stability assumption is using the same instrument and measurement setting across studies. This includes ensuring the same timing of the single or repeated measurements of outcomes after treatment implementation, and the same order of measurements in case of multiple outcome measures. Careful descriptions of the outcome measures and their implementation in measurement protocols facilitate the assessment of whether the same outcomes are studied. Following TOP guidelines, the instruments and protocols should be made available to other researchers. However, even in cases where the same instrument is used in all replication studies, researchers should ensure that the same construct (e.g., anxiety, depression, math achievement) is measured across different populations and settings. This assumption is referred to as measurement invariance. In systematic replication studies, researchers should assess whether measurement invariance holds across populations and settings involved in the evaluation (Widaman et al., 2010; Wu et al., 2007). If well-established outcome measures are used, published reports on measurement invariance can be used to assess whether the assumption might be met. With newly developed measures, their measurement variance may need to be established and tested.

The replication assumptions also require that treatment-control contrasts are equivalent across studies. To this end, researchers should clearly define a treatment protocol and measure the extent to which the intervention is delivered consistently across participants, sites, settings, and studies. Traditionally, researchers hire trained observes to rate each intervention session according to an adherence checklist (Nelson et al., 2012). However, monitoring intervention delivery is time consuming and expensive, particularly in systematic replication studies where interventions are delivered at multiple times, in multiple settings, and with multiple research teams. Anglin and Wong (2020) offer an alternative automated approach to measuring treatment adherence using a set of natural language processing techniques termed semantic similarity that quantify the similarity between texts. These methods can be used to assess treatment stability in highly-standardized interventions that are delivered through verbal interactions with participants by quantifying the similarity of a treatment transcript to a scripted treatment protocol.[3] The approach also has the benefit of being open, transparent and reproducible as long as the original transcripts and syntax files are made available.

Example

We now discuss examples of how researchers can probe the replication assumptions R1 and R2. Tables 4 and 5 provide examples of balance tests using the causal systematic replication study described above (Krishnamachari, Wong, & Cohen, in progress). Table 4 summarizes descriptive statistics on study factors that were intended to be systematically varied across the four RCTs (timing, pedagogical task, and target population and setting); Table 5 summarizes study factors that were intended to remain fixed across studies. For ease of discussion, study S_{22} is designated as the "benchmark study" for comparing results with to create the multiple

[3] A full review of how researchers may apply semantic similarity methods is beyond the scope of this paper, but we provide readers with an intuition for the approach here. To quantify the similarity between texts, researchers represent texts numerically by their relative word frequencies or by the extent to which they include a set of abstract topics. After each transcript is represented as a numerical vector, researchers calculate the similarity of vectors by measuring the cosine of the angle between them. Two texts that share the same relative word frequencies will have a cosine similarity of one and two texts that share no common terms (or concepts) will be perpendicular to one-another and have a cosine similarity of 0. Importantly, semantic similarity methods create continuous measures which can be used to identify studies where treatments were delivered more or less consistently, or with more or less adherence. Anglin and Wong (2020) describe the method and provide an example of how it may be used in replication contexts.

cohort design (S_{22} versus S_{12}), the switching replication design (S_{22} versus S_{21}), and the conceptual replication design (S_{22} versus S_{32}) with a different target population and setting.

In looking at Table 4, the goal of the conceptual replication effort (S_{22} versus S_{32}) was to evaluate the replicability of effects across a different target population and study setting. The descriptive table summarizes characteristics related to replication assumption *R2* (equivalence in the causal estimand). For the conceptual replication, the undergraduate sample in study S_{32} differed in multiple ways from the teacher candidate sample (S_{22}). The undergraduate sample included more males, was younger, was more likely to be from an urban area, and reported attending high schools with higher proportions of individuals from high SES and high achieving backgrounds. As discussed above, the undergraduate sample also had different training experiences before entering the simulation setting.

The descriptive tables also summarize shared characteristics across multiple studies. For example, the undergraduate sample in the conceptual replication study participated in the same pedagogical task (“setting classroom norms”) that teacher candidates experienced in the benchmark study (Table 3). Table 5 reports means and standard deviations of the outcome scores on observed “quality” of participants’ pedagogical practice in the simulation session (outcome stability assumption). These scores were scaled from 1 through 10, where 10 indicated high quality pedagogical practice on the observational rubric and 1 indicated lower quality practice. Across all four studies, the reliability of the quality score was generally consistent, ranging with an alpha level of 0.74 in study S_{21} to 0.88 in study S_{12}.

Table 5 also reports summary scores of treatment adherence to the standardized coaching protocol. Although the systematic replication studies included planned variations in target populations, pedagogical tasks, and settings, the coaching protocol was intended to be delivered in a standardized way. To evaluate whether this assumption was met, the research team applied the semantic similarity method proposed by Anglin and Wong (2020) to evaluate how similar transcripts of coaching sessions were to a benchmark coaching script. The adherence scale ranges from 0 to 1, where transcripts of intervention sessions with higher adherence to the protocol have higher scores, and those that stray from the protocol have lower scores. Adherence scores in Table 5 indicate that fidelity to the coaching protocol was generally similar across studies, though coaching fidelity was higher in the benchmark study S_{22} and switching replication S_{21} than for the multiple cohort study S_{12} and the conceptual replication study S_{31}.

Finally, the RCT design and estimation strategy were similar across both studies (*S1*-*S2*). Balance tables of covariates for each study (available in a Methodological Appendix by request) demonstrate that intervention and control groups were equivalent at baseline, and that estimated effects were robust to multiple model specifications. In reproducibility analyses, effect estimates for four studies were analyzed and verified by independent researchers blinded to original results (S3).

Tables 4 and 5 also describe the extent to which replication design assumptions were met or varied for other research designs in the systematic replication effort. For example, relative to the benchmark study S_{22}, the sample characteristics of participants were generally similar for the multiple cohort design (S_{12}) and switching replication design (S_{21}). The multiple cohort design used the same pedagogical task, coaching intervention, research design and estimation approaches across studies. The primary difference was that S_{12} took place one year before the benchmark study S_{22}. The switching replication design also succeeded in holding most study factors constant, with the exception of introducing systematic variation in the pedagogical task under which the coaching intervention was applied (setting classroom norms versus providing feedback on text-based discussion). Five additional participants joined the benchmark study in Spring 2019 (N = 98 for S_{22}, N = 93 for S_{21}). However, these participants did not change the overall distribution of sample characteristics across the two studies and were randomized into intervention conditions in study S_{21}.

Importantly, Tables 4 and 5 also report the limitations of the conceptual replication studies. Variations in study factors not under investigation occur because of logistical challenges and/or because of deviations in the study protocol. In this study, because of sample size limitations, each RCT was conducted at different time intervals, potentially confounding variations in study characteristics with the timing of when the study was conducted. Moreover, the adherence scores indicate that while coaching was delivered with similar fidelity levels (according to the semantic similarity measure), the intervention was not delivered in exactly the same way across all the studies. Finally, the team observed multiple differences in both population and setting characteristics for the conceptual replication study (S_{22} versus S_{32}). As such, the team was limited in identifying the specific causal factors that resulted in the substantially smaller effects that was observed for the undergraduate sample. In the end, the team concluded that the systematic replication study provided strong evidence of the robustness of coaching effects for individuals enrolled in the teacher preparation program, but subsequent replication studies were designed to evaluate whether coaching effects are less effective for the sub-population of students in the undergraduate study or because these students lack the training experience in pedagogical techniques to realize the benefits of coaching.

Reporting Results from Diagnostic Tests

Given space limitations in peer-reviewed journals, a common issue that arises is whether researchers are able to report results from the diagnostic probes of their systematic replication studies. Our general recommendation is that systematic replication studies should include balance tables similar to Tables 4 and 5 that report descriptive statistics and summary study characteristics for addressing replication and individual study assumptions. These tables provide concise presentations of the extent to which replication assumptions were addressed or varied across studies, as well as describe sample and study characteristics that were included in the systematic replication. Online methodological appendices are useful for including results from additional diagnostic tests, including balance tests for individual studies, attrition analyses, as well as effect estimates from multiple specifications.

Discussion: Considering Design-Based Approaches in the Context of Planning Replication Studies

In this article, we introduce design-based approaches for conducting a series of systematically planned causal replications. These approaches are derived from the CRF, which describes replication and individual study assumptions for the direct replication of results. Causal conceptual replication designs test systematically planned violations of one or more replication assumptions while seeking to meet and diagnostically address all other assumptions under the CRF. If replication failure is observed, the researcher may conclude that effect variation is due to planned changes in the causal estimand. A key advantage of the CRF is that it provides a theoretical basis for understanding how existing research designs may be utilized for conceptual replication and for understanding the assumptions required for conducting high quality replication studies. Because researchers are often interested in identifying multiple reasons why effects vary across studies, they may plan a series of replications that systematically vary presumed effect-moderating factors across studies while meeting all other replication assumptions. Results from such systematic replication approaches are most interpretable when the researcher has control over multiple study characteristics and is able to introduce systematic variations in each study.

The recent methodological literature has identified multiple considerations for conducting high-quality replication studies. Selecting a design-based approach is one component of conducting a high-quality replication study. To this end, Table 6 provides a summary of what we believe are the crucial decision-points for each phase of a replication effort. During the *design and planning* phase of the replication study, the research team should use subject-matter theory to identify potential reasons for why replication failure may occur and choose a causal estimand of interest (well-defined treatment-control contrast for a clearly defined target population and setting). These are the study factors that identify the “boundary conditions” for which intervention effects may or may not replicate (Nosek & Errington, 2020). Second, the researcher should determine which CRF assumptions are most interesting for testing and select research designs that are capable of evaluating the hypothesized moderating characteristics while assessing the plausibility of the remaining assumptions. Potential designs include, but are not limited to, multi-arm RCT designs, RCTs with multiple cohorts, multi-site designs, and switching replication designs. Third, the researcher should consider in advance potential sources of bias and moderators of effects so they can plan for collecting the data necessary to conduct diagnostic tests of assumptions (Simons et al., 2014).

During the planning phase, the researcher should also ensure that the replication study has adequate statistical power for detecting replication success/failure and pre-register study procedures, methods, and criteria for assessing replication success. Both topics are beyond the scope of this article, but we note that these issues are centrally related to selecting an appropriate research design. For example, in selecting a replication design, the researcher should consider the need to balance controlled variation in study characteristics with adequate statistical power for detecting replication failure. That is, when studies differ in multiple, substantive ways, they will likely have greater statistical power for concluding differences in effect estimates. However, when all factors vary simultaneously across studies (such that multiple CRF assumptions are violated), it is impossible for the researcher to conclude why replication failure occurred. Steiner and Wong (2018) suggest design-based approaches to replication that may address both concerns. They note, for example, that replication designs with the same units across multiple studies (e.g. switching replication designs, dependent arm within-study comparison designs) have greater statistical power for detecting replication success than replication approaches with independent units across studies. This result implies that while the current methodological literature has noted the limited power of most replication studies (Anderson et al., 2017; Anderson & Maxwell, 2017; Hedges & Schauer, 2019; Schauer & Hedges, 2020; Simonsohn, 2015), there are likely design-based approaches to replication that may be effective for addressing these challenges. However, these approaches require strong subject matter theory for guiding the selection of which factors should be systematically tested in the replication design.

During the *implementation and analysis phase* of the replication study, the research team is responsible for collecting measures that will allow them to assess the extent to which CRF assumptions were met or violated. In most field settings, whether due to chance or to systematic error, deviations from the study protocol are likely to occur. Results from diagnostic probes provide researchers (and readers) with empirical evidence for ruling out or acknowledging threats to validity in the replication design. Analysis approaches for determining replication success are critical as well. Because different analysis methods may often yield different conclusions about replication success, we recommend that researchers determine in advance criteria for determining replication success. Hedges and Schauer (2019), Rindskopf, Shadish, and Clark (2018), Schauer and Hedges (2020), and Steiner and Wong (2018) have written about common approaches, but more research on this topic is needed.

This article has highlighted the need for *standardized reporting* of diagnostic measures from CRF assumptions. The goal here is to report the extent to which study factors that are hypothesized to “matter” (based on the CRF assumptions) were varied or were held constant across studies. Without the reporting of this diagnostic information, neither the researcher nor the reader can have confidence in understanding why replication failure occurred when study results differ. Currently, there is not yet consensus on the best ways to report results from replication studies, though Lebel et al. (2018) provides useful examples and Spybrook et al. (2019) have discussed the importance to presenting analysis results according to the pre-registration plan. Finally, the Center for Open Science, ICPSR, and scholars such as Klein et al. (2018) and Nosek et al. (2018) have provided recommendations for ensuring that study materials, procedures, and data are open and transparent.

A Design-Based Perspective for Individual and Replication Study Efforts

Individual studies rarely provide sufficient evidence for identifying all conditions that are necessary and sufficient for replicating effects. Establishing credibility of scientific findings is a team enterprise that requires cooperation from both independent and inter-related investigators. The CRF provides a common framework for investigators to plan, implement, analyze, and report their findings.

For individual studies, once the intervention has been evaluated, data have been collected, and results are to be published, “within-study” replications such as robustness checks with multiple model specifications (Duncan et al., 2014) and reanalysis or reproducibility approaches with independent reporter (Chang & Li, 2015) may be used to assess individual study assumptions. Reporting within-study replication results provides the researcher and reader with confidence that study findings are not biased or incorrectly reported. These replication practices have already been adopted in economics (Duncan et al., 2014), but they could be incorporated in other fields of study including prevention science. Individual studies should also consider standardized reporting practices of diagnostic results related to replication assumptions, or key factors that are hypothesized to "matter" for amplifying or moderating the effect (Simons et al., 2014). At a minimum, diagnostic information in individual studies about treatment and outcome stability (R1), as well as characteristics that describe the causal estimand of interest (R2), would improve the interpretability of future studies designed to replicate the original study's results, as well as assist in any research synthesis efforts that require common coding schemes for pooling results across different studies.

Design-based approaches to replication also improve inferences from multiple integrated studies. Already, carefully planned series of causal replications are becoming more popular for assessing the replicability of effects. For example, a recently funded study from the Institute of Education Sciences plans a causal replication evaluation of a reading intervention that includes three research designs: an RCT with multiple cohorts for assessing the replicability of effects over time, a multi-site RCT for assessing the replicability of effects over variations in students and settings, and a multi-arm RCT for examining the replicability of effects over different treatment dosages (Solari et al., 2020). The study's purpose is to assess the replicability of effects for the reading intervention, as well as to identify causal sources of effect variation if replication failure is observed. In another recently funded example, the Special Education Research Accelerator (SERA) is an effort to build a platform for conducting crowdsourced replication studies in the area of special education (Cook et al., 2020). The goal here is to provide researchers with infrastructure supports for conducting descriptive systematic replication studies in special education, including diagnostic information for assessing all replication and individual study assumptions under the CRF.

Over the last several decades, the overarching mission of many prevention scientists has been to understand “what works” for improving healthy life outcomes. Mounting evidence across multiple disciplines in the social sciences suggest that results from many studies are fragile and hard to replicate. This paper has argued that while ad hoc replications are often difficult to interpret, design-based approaches can help researchers systematically test sources of effect variation to uncover conditions under which study findings replicate in real world settings.

Compliance with Ethical Standards

Funding
The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B140026 and Grant #R305D190043 to the Rectors and Visitors of the University of Virginia.

Ethics Approval

Approval was obtained from the ethics committee of University of Virginia. The procedures used in this study adhere to the tenets of the Declaration of Helsinki. (Ethics approval numbers: 2170, 2727, 2875, 2918).

Conflict of Interest/Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Consent to Participate

Informed consent was obtained from all individual participants included in the study.

References

Anglin, K. L., & Wong, V. C. (2020). *Using Semantic Similarity to Assess Adherence and Replicability of Intervention Delivery* (No. 73; EdPolicyWorks Working Paper Series, pp. 1–33). EdPolicyWorks. https://curry.virginia.edu/sites/default/files/uploads/epw/73_Semantic_Similarity_to_Assess_Adherence_and_Replicability_0.pdf

Angrist, J., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bareinboim, E., & Pearl, J. (2012). Transportability of causal effects: Completeness results. *Proceedings of the AAAI Conference on Artificial Intelligence*, *26*(1).

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*(1), 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Cartwright, N., & Hardie, J. (2012). *Evidence-Based Policy: A Practical Guide to Doing it Better* (p. 208). Oxford University Press.

Chang, A., & Li, P. (2015). *Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not"* (Finance and Economics Discussion Series) [2015-083]. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/econresdata/feds/2015/files/2015083pap.pdf

Clemens, M. A. (2017). The meaning of failed replications: A review and proposal. *Journal of Economic Surveys*, *31*(1), 326–342.

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher Coaching in a Simulated Environment. *Educational Evaluation and Policy Analysis*, *42*(2), 208–231. https://doi.org/10.3102/0162373720906217

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

Cook, B., Therrien, W., & Wong, V. (2020). *Developing Infrastructure and Procedures for the Special Education Research Accelerator*. NCSER. https://ies.ed.gov/funding/grantsearch/details.asp?ID=3356

Degue, S., Niolon, P. H., Estefan, L. F., Tracy, A. J., Le, V. D., Vivolo-Kantor, A. M., Little, T. D., Latzman, N. E., Tharp, A., Lang, K. M., & Taylor, B. (2020). Effects of Dating Matters® on Sexual Violence and Sexual Harassment Outcomes among Middle School Youth: A Cluster-Randomized Controlled Trial. *Prevention Science*. https://doi.org/10.1007/s11121-020-01152-0

Department of Health and Human Services. (2014). *PAR-13-383: Replication of Key Clinical Trials Initiative*. Grants and Funding. https://grants.nih.gov/grants/guide/pa-files/PAR-13-383.html

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

Fraker, T., & Maynard, R. (1987). The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs. *The Journal of Human Resources*, *22*(2). https://doi.org/10.2307/145902

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science: Next Generation. *Prevention Science*, *16*(7), 893–926. https://doi.org/10.1007/s11121-015-0555-x

Hedges, L. V., & Schauer, J. M. (2019). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*, *24*(5), 557.

Imbens, G. W., & Ruben, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences* (p. 644).

Institute of Education Sciences. (2016). *Building Evidence: What Comes After an Efficacy Study?* (pp. 1–17). https://ies.ed.gov/ncer/whatsnew/techworkinggroup/pdf/BuildingEvidenceTWG.pdf

Institute of Education Sciences. (2020). *Program Announcement: Research Grants Focused on Systematic Replication CFDA 84.305R*. Funding Opportunities; Institute of Education Sciences (IES). https://ies.ed.gov/funding/ncer_rfas/systematic_replications.asp

Keown, L. J., Sanders, M. R., Franke, N., & Shepherd, M. (2018). Te Whānau Pou Toru: A Randomized Controlled Trial (RCT) of a Culturally Adapted Low-Intensity Variant of the Triple P-Positive Parenting Program for Indigenous Māori Families in New Zealand. *Prevention Science*, *19*(7), 954–965. https://doi.org/10.1007/s11121-018-0886-5

Lalonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, *76*(4), 604–620.

Lebel, E. P., Mccarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A Unified Framework to Quantify the Credibility of Scientific Findings. *Advances in Methods and*

Practices in Psychological Science, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

LeBel, E. P., Mccarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A Unified Framework to Quantify the Credibility of Scientific Findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

Lewis, M. A., Rhew, I. C., Fairlie, A. M., Swanson, A., Anderson, J., & Kaysen, D. (2019). Evaluating Personalized Feedback Intervention Framing with a Randomized Controlled Trial to Reduce Young Adult Alcohol-Related Sexual Risk Taking. *Prevention Science*, *20*(3), 310–320. https://doi.org/10.1007/s11121-018-0879-4

Maalouf, F. T., Alrojolah, L., Ghandour, L., Afifi, R., Dirani, L. A., Barrett, P., Nakkash, R., Shamseddeen, W., Tabaja, F., Yuen, C. M., & Becker, A. E. (2020). Building Emotional Resilience in Youth in Lebanon: A School-Based Randomized Controlled Trial of the FRIENDS Intervention. *Prevention Science*, *21*(5), 650–660. https://doi.org/10.1007/s11121-020-01123-5

Morgan, S. L., & Winship, C. (2014). Counterfactuals and causal inference: Methods and principles for social research. In *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. https://doi.org/10.1017/CBO9781107587991

National Science Foundation. (2020). *Improving Undergraduate STEM Education: Education and Human Resources*. Funding. https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505082

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *Journal of Behavioral Health Services and Research*, *39*(4), 374–396. https://doi.org/10.1007/s11414-012-9295-x

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., … Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help promote transparency, openness, and reproducibility. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Rindskopf, D. M., Shadish, W. R., & Clark, M. H. (2018). Using Bayesian Correspondence Criteria to Compare Results From a Randomized Experiment and a Quasi-Experiment Allowing Self-Selection. *Evaluation Review*, 0193841X18789532-0193841X18789532. https://doi.org/10.1177/0193841X18789532

Roddy, M. K., Rhoades, G. K., & Doss, B. D. (2020). Effects of ePREP and OurRelationship on Low-Income Couples' Mental Health and Health Behaviors: A Randomized Controlled Trial. *Prevention Science*, *21*(6), 861–871. https://doi.org/10.1007/s11121-020-01100-y

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Schauer, J. M., & Hedges, L. V. (2020). Assessing heterogeneity and power in replications of psychological experiments. *Psychological Bulletin*.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128.

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

Solari, E., Wong, V., Baker, D. L., & Richards, T. (2020). *Iterative Replication of Read Well in First Grade*. NCSER. https://ies.ed.gov/funding/grantsearch/details.asp?ID=4404

Spybrook, J., Anderson, D., & Maynard, R. (2019). The Registry of Efficacy and Effectiveness Studies (REES): A Step Toward Increased Transparency in Education. *Journal of Research on Educational Effectiveness*, *12*(1), 5–9. https://doi.org/10.1080/19345747.2018.1529212

Steiner, P. M., Wong, V. C., & Anglin, K. L. (2019). A Causal Replication Framework for Designing and Assessing Replication Efforts. *Zeitschrift Für Psychologie / Journal of Psychology*, *227*(4), 280–292. https://doi.org/10.1027/2151-2604/a000385

Stroebe, W., & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science*, *9*(1), 59–71. https://doi.org/10.1177/1745691613514450

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

Terry, J. D., Weist, M. D., Strait, G. G., & Miller, M. (2020). Motivational Interviewing to Promote the Effectiveness of Selective Prevention: An Integrated School-Based Approach. *Prevention Science*. https://doi.org/10.1007/s11121-020-01124-4

Tipton, E. (2012). Improving Generalizations From Experiments Using Propensity Score Subclassification. *Journal of Educational and Behavioral Statistics*, *38*(3), 239–266. https://doi.org/10.3102/1076998612441947

Tipton, E., & Olsen, R. B. (2018). A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions. *Educational Researcher*, *47*(8), 516–524. https://doi.org/10.3102/0013189X18781522

Wennehorst, K., Mildenstein, K., Saliger, B., Tigges, C., Diehl, H., Keil, T., & Englert, H. (2016). A Comprehensive Lifestyle Intervention to Prevent Type 2 Diabetes and Cardiovascular Diseases: The German CHIP Trial. *Prevention Science*, *17*(3), 386–397. https://doi.org/10.1007/s11121-015-0623-2

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial Invariance Within Longitudinal Structural Equation Models: Measuring the Same Construct Across Time. *Child Development Perspectives*, *4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Wong, V. C., & Steiner, P. M. (2018). Replication designs for causal inference. In *EdPolicyWorks Working Paper Series* (No. 62; EdPolicyWorks Working Paper Series, Issue 62). EdPolicyWorks. https://curry.virginia.edu/sites/default/files/uploads/epw/62_Replication_Designs.pdf

Wong, V. C., Wing, C., Steiner, P. M., Wong, M., & Cook, T. D. (2012). Research designs for program evaluation. *Handbook of Psychology, Second Edition*, *2*.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data. *Practical Assessment Research & Evaluation*, *12*(3), 25. https://doi.org/10.7275/mhqa-cd89

Tables

Table 1. Assumptions of the Causal Replication Framework for the Direct Replication of Effects (Steiner, Wong, & Anglin, 2020; Wong & Steiner, 2018)

Assumptions	For Study 1 …	… Through Study *k*
Replication assumptions (*R1*-*R2*)	R1. Treatment and outcome stability R2. Equivalence of causal estimands	
Individual study assumptions (*S1*-*S3*)	S1. Causal estimand is identified S2. Unbiased estimation of effects S3. Correct reporting of estimators, estimands, and estimates	S1. Causal estimand is identified S2. Unbiased estimation of effects S3. Correct reporting of estimators, estimands, and estimates

Table 2a. Schedule of four RCT Studies for Constructing Systematic Replications

Spring 2018	Fall 2018	Spring 2019 Benchmark Study	Fall 2019
S_{12}	S_{21}	S_{22}	S_{32}

Table 2b. Combination of RCTs for creating systematic conceptual replication study

	S_{22} Benchmark Study
S_{12}	Multiple cohort design
S_{21}	Switching Replication Design
S_{32}	Conceptual Replication with Different Units and Settings

In Tables 2a and 2b, each individual RCT is indexed by S_{ij}, where *i* denotes the sample (teacher candidate cohorts 1 or 2 or "teaching as a profession" undergraduate sample 3), and *j* denotes the pedagogical task for which the coaching or self-reflection protocol was delivered (1 if the pedagogical task involved a text-based discussion; and 2 if the pedagogical task involved a conversation about setting classroom norms and managing disruptive student behaviors). Conceptual RCTs are described as the comparison of two RCTs (S_{22} versus S_{12} for the multiple cohort design). The research team selected S_{22} as the benchmark study for creating each of the conceptual replication designs.

Table 3: CRF Assumptions Tested in *Planned* Causal Replication Study (Krishnamachari, Wong, & Cohen, in progress)

	R1. Treatment / Outcome Stability	R2. Equivalent Causal Estimand	S1. Identification	S2. Estimation	S3. Reporting
Multiple Cohort (S_{22} vs S_{12})	Treatments ✓ Outcomes ✓	Participants ✓ Settings ✓ Causal quantity ✓ Time ✗	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓
Switching Replication (S_{22} vs S_{21})	Treatments ✓ Outcomes ✓	Participants ✓ Settings ✗ Causal quantity ✓ Time ✓	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓
Conceptual Replication with Different Units and Settings (S_{22} vs S_{32})	Treatments ✓ Outcomes ✓	Participants ✗ Settings ✗ Causal quantity ✓ Time ✓	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓

Table 4: Balance on Factors that are Systematically Varied

	S_{22}	S_{21}	S_{12}	S_{32}
	Benchmark Study	Switching Replication	Multiple Cohort	Conceptual Replication
Participant Characteristics				
GPA	3.46	3.51	3.42	3.54
Mothers' education				
% College or above	0.79	0.85	0.75	0.76
% Female	0.88	0.98	1.00	0.50
% Over the age of 21	0.16	0.19	0.18	0.08
% White	0.63	0.69	0.56	0.56
Location of high school attended				
% Rural	0.12	0.13	0.03	0.09
% Suburban	0.82	0.85	0.86	0.79
% Urban	0.06	0.02	0.11	0.13
Average SES of high school attended				
% Low SES	0.00	0.00	0.04	0.00
% Middle SES	0.61	0.68	0.59	0.57
% High SES	0.28	0.28	0.32	0.40
Majority race of high school attended				
% Primarily students of color	0.03	0.04	0.10	0.06
% Mixed	0.47	0.51	0.48	0.41
% Primarily white students	0.50	0.45	0.42	0.53
Average achievement level of high school attended				
% Primarily low achieving	0.00	0.00	0.06	0.03
% Primarily middle achieving	0.43	0.53	0.37	0.34
% Primarily high achieving	0.46	0.45	0.53	0.60

Setting Characteristics				
Pedagogical Task in Simulator	Setting Classroom Norms	Providing Text-based Discussion	Setting Classroom Norms	Setting Classroom Norms
Training Setting	Methods Course	Methods Course	Methods Course	Teaching as a Profession
Timing	Spring 2019	Fall 2018	Spring 2019	Fall 2019

Notes: Descriptive table adapted from Krishnamachari, Wong, & Cohen (in progress)

Table 5: Balance on Factors Intended to be Held Constant across Studies

	S_{22}	S_{21}	S_{12}	S_{32}
	Benchmark Study	Switching Replication	Multiple Cohort	Conceptual Replication
Outcome & Treatment Stability				
Outcome Stability (Pretest means & standard deviations)	3.46 (1.33)	3.90 (1.30)	3.64 (1.22)	2.89 (1.03)
Coaching Stability (Intervention adherence)	0.31	0.42	0.26	0.26
Individual Study Design Assumptions				
Research Design for Causal Identification	RCT Covariate balance ✓	RCT Covariate balance ✓	RCT Covariate balance ✓	RCT Covariate balance ✓
Estimation Strategy	Regression-adjustment Robustness checks ✓	Regression-adjustment Robustness checks ✓	Regression-adjustment Robustness checks ✓	Regression-adjustment Robustness checks ✓
Independent Reproducibility	Yes	Yes	Yes	Yes

Notes: To examine the validity of the RCT, the research team examined baseline equivalence on an array of baseline characteristics for each study. To assess the sensitivity of effect estimates to different model specifications, the research team reports the robustness of results with different control covariates included in the models. All effect estimates were reproduced by an independent analyst with access to the original data and syntax files but was blinded to original study results. Coaching stability was assessed using the semantic similarity approach described in Anglin and Wong (2020); a higher score indicates higher similarity to a benchmark scripted treatment protocol. Table adapted from Krishnamachari, Wong, & Cohen (in progress).

Table 6: Planning Systematic Replication Studies: Design, Implementation and Analysis, and Reporting Phases

Phase	Recommendation	Related Literature
Design and Planning Phase	1. Use subject-matter theory and the Causal Replication Framework to identify potential sources of effect variation and the causal estimand of interest.	Nosek & Errington (2020), Steiner, Wong, & Anglin (2020)
	2. Determine if planned study is a conceptual or direct replication study and select research design(s) for testing sources of variation.	Schmidt (2009), Wong, Anglin & Steiner (2021)
	3. Plan diagnostic measures for evaluating CRF assumptions.	Wong, Anglin & Steiner (2021)
	4. Identify appropriate statistical power for detecting replication failure/success.	Anderson & Maxwell (2017), Schauer & Hedges (2020), Simonsohn (2015)
	5. Pre-register replication study design, diagnostic measures, statistical power, and criteria for assessing replication success/failure.	Lei, Gelman, & Ghitza (2016), Nosek et al. (2018)
Implementation and Analysis Phase	6. Implement intervention, measures, and data collection procedures in ways that are open, transparent, and replicable.	Klein et al. (2018), Nosek et al. (2015)
	7. Use diagnostic measures to assess the extent to which CRF assumptions are met or varied in replication studies.	Shadish, Cook, & Campbell (2002), Wong, Anglin & Steiner (2021)
	8. Analyze study results for assessing replication success/failure.	Hedges & Schauer (2019), Rindskopf, Shadish, & Clark (2018), Steiner & Wong (2018)
Reporting Phase	9. Report findings on replication success/failure using criteria established in pre-registration.	Spybook, Anderson, & Maynard (2019)
	10. Report results of key diagnostic measures for assessing CRF assumptions.	Wong, Anglin & Steiner (2021)
	11. Report study materials, measures, procedures, data, and data analysis files such that results are open and transparent.	Nosek et al. (2015)