ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 20-311

Design-Based Approaches to Causal Replication Studies

Vivian C. Wong
University of Virginia

Peter M. Steiner
University of Maryland

Kylie L. Anglin
University of Virginia

Recent interest to promote and support 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. In direct replication designs, replication failure is evidence of bias or incorrect reporting in individual study estimates, while in conceptual replication designs, replication failure occurs because of effect variation due to differences in treatments, outcomes, settings, and participant characteristics. The paper demonstrates how multiple research 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.

VERSION: October 2020

Suggested citation: Wong, Vivian C., Peter M. Steiner, and Kylie L. Anglin. (2020). Design-Based Approaches to Causal Replication Studies. (EdWorkingPaper: 20-311). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/xsqw-c323

Design-Based Approaches to Causal Replication Studies

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

Corresponding Author
Vivian C Wong, vcw2n@virginia.edu

Abstract

Recent interest to promote and support 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. In direct replication designs, replication failure is evidence of bias or incorrect reporting in individual study estimates, while in conceptual replication designs, replication failure occurs because of effect variation due to differences in treatments, outcomes, settings, and participant characteristics. The paper demonstrates how multiple research 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 systemic 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 approaches for systematic replication studies. Our general approach is derived from the Causal Replication Framework (CRF), which formalizes the assumptions under which causal effect estimates can be expected to replicate and under what conditions the source(s) of effect variations across studies can be drawn (Steiner et al., 2019; Wong & Steiner, 2018). CRF assumptions ensure that the same causal estimand is compared across studies and that the effect is estimated without bias and correctly reported in each study. Here, a causal estimand is defined as the causal effect of a well-defined treatment-control contrast for a clearly defined target population and setting. Replication failure occurs when one or more CRF assumptions are not met. Importantly, under the CRF, "replication failure" is not a scientific failure – it is actually a success – so long as the replication study is well-designed to systematically test one or more CRF assumptions.

Under the CRF, it is straight-forward to derive design-based approaches for replication. In design-based approaches, the inferences that may be drawn from a study depend on the quality of the research design in addressing plausible threats to validity (Shadish, Cook, & Campbell, 2002). In replication, threats to a study's validity – and therefore the research design – depend on the researcher's questions regarding the causal estimand. For example, *direct replications* examine whether two or more studies with the same causal estimand yield the same effect. In this approach, the researcher designs their study so that treatment effects are compared for the same intervention and for the same target population of participants in comparable settings. When replication failure is observed, the researcher concludes that bias due to violations of CRF assumptions (e.g., attrition bias or incorrect reporting in at least one study) caused the differences in results. *Conceptual replications* examine whether studies with potentially different causal estimands yield the same effect. Here, the researcher designs their replication to introduce systematic variations in interventions, participants, settings, and outcome measures. When replication failure is observed, the researcher concludes that differences in the causal estimand – due to differences in treatments, units, settings, and/or outcomes – caused variation in effect estimates. Results from direct and conceptual replication studies are most interpretable when CRF assumptions are tested in controlled settings.

In this article, we focus on research designs for conceptual replication studies because identifying causal sources of effect variation is essential for generalizing effects (Cole & Stuart, 2010; Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2012; Tipton & Olson, 2019). We argue that replication studies should be judged in similar ways to how we assess research designs of empirical studies more generally – by looking at the extent to which assumptions are met and plausible threats to validity are ruled out for making inferences (Angrist & Pischke, 2009; Shadish et al., 2002). To this end, we demonstrate how diagnostic measures may be used to evaluate CRF assumptions and to appropriately interpret replication results. Finally, we highlight that a series of replication designs may be combined in a single evaluation effort for identifying multiple sources of effect variation, and for addressing different validity threats. We demonstrate the benefits and limitations of combining multiple systematic replication approaches using an applied example.

This article highlights an issue that has not yet been addressed in the methodology literature – design-based approaches for replication. There are additional methodological considerations related to planning replication studies, including ensuring adequate statistical power and analysis methods for determining replication success. These topics are beyond the scope of this paper, but we note that selecting an appropriate research design is central for all other planning decisions in replication studies.

The paper proceeds as follows. First, we introduce the CRF as a framework for planning and interpreting causal replications. Second, we describe research designs for conceptual replication under the CRF. Third, we demonstrate how multiple replication designs may be combined to assess effect variations. Fourth, we present diagnostics for assessing assumptions under the CRF, illustrating these diagnostics through a series of systematically planned replications. We conclude with recommendations for researchers seeking to establish causal validity in their approach to replication.

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. To help shed light on this issue, Wong and Steiner (2018) derived the CRF, which defines replication as a research design that tests whether two or more studies produce the same causal effect within the limits of sampling error. The CRF formalizes the conditions under which replication success can be expected. The core of the framework is based on potential outcomes notation (Rubin, 1974), which has the advantage of identifying clear causal estimands of interest and assumptions for the direct replication of results. A causal estimand is the effect parameter of interest for a well-defined treatment and control contrast for a clearly defined target population and setting.

Table 1 summarizes the five assumptions required for the direct replication of results. They include both *replication design assumptions* (*R1*-*R2*) and *individual study design assumptions* (*S1*-*S3*). Wong and Steiner (2018) describe implications of each assumption, but replication design assumptions may be understood broadly as the need for "treatment and outcome stability" (*R1*) and "equivalence in causal estimands" (*R2*). Treatment and outcome stability (*R1*) may be violated if there are variations in treatment and control conditions across studies. This means that treatment conditions must be well-specified and implemented in identical ways across all studies (i.e. there must be no hidden variations in intervention and control conditions). The assumption may also be violated 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 equivalent causal estimand 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 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 have the same causal quantity. For instance, both 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. Combined, replication assumptions *R1* and *R2* ensure that the same causal estimand for a clear treatment-control contrast and target population is compared across all studies. If either assumption is violated, then results cannot be expected to replicate.

Individual study design assumptions (*S1*-*S3*) require 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 ensure that for each individual study included in the replication effort, a valid research design is used for identifying effects, appropriate estimators are used for estimating effects, and effects are correctly reported. These are standard assumptions for any individual study design to yield a valid causal effect. Assumptions *S1* and *S2* may be violated if, for example, a study fails to successfully address attrition or nonresponse bias, or if a result is estimated by an incorrectly specified regression model. However, even when the effect is well-identified and estimated, replication failure may occur if the result is reported incorrectly or incompletely by the researcher. The investigator may report an incorrect result from the analyses or fail to report appropriate procedures and assumptions for an independent investigator to identify and estimate the same effect.

The emphasis on replication and individual study assumptions highlight a critical difference between traditional, procedure-based approaches to replication and replication under the CRF. In traditional approaches to replication, the goal is for replication studies to evaluate whether the same result is produced by implementing the same methods and procedures that were used to carry out the original study (Nosek & Errington, 2017). The quality of the replication study is determined by how closely studies are able to replicate methods and procedures from the original study (Brandt et al., 2014; Kahneman, 2014). Despite this seemingly straight-forward approach to replication, however, multiple challenges arise when implementing procedure-based approaches. For example, the authors of the original study may have failed to report all relevant methods and procedures for implementing the study, or the methods and procedures in the original study may be flawed or not perfectly implemented in a field setting such that no causal interpretations are warranted. In these cases, it is not obvious how the replicator should proceed.

Under the CRF, the goal is for replication studies to evaluate whether the same result is produced while addressing replication (*R1*-*R2*) and individual study (*S1*-*S3*) assumptions. Here, the quality of the replication study is based on the extent to which necessary CRF assumptions are met (or not met). Replication failure occurs when one or more assumption is violated. However, under the CRF, replication failure is not viewed as being inherently bad for science, as long as the researcher is able to identify why it occurred. This is because replication failure resulting from violations in replication assumptions (*R1* or *R2*) is evidence of effect variation, which is essential for understanding to generalize effects to broader target populations of interest (Cole & Stuart, 2010; Stuart et al., 2011; Tipton, 2012; Tipton & Olsen, 2018). Thus, replication studies may be considered as a core method for understanding and identifying effect variation when constant treatment effects across units, contexts, and settings cannot be assumed. In the following section, we describe research designs for identifying sources of effect variation in replication studies.

Research Designs for Replication

In design-based approaches, the researcher uses research designs for systematically testing and addressing assumptions under the CRF. If replication failure is observed—and all other assumptions are met—then the researcher may infer that the tested assumption was violated and resulted in treatment effect variation. As mentioned in the introduction, there are two well-known approaches to replication: direct and conceptual replications. The CRF provides a formal way to understand each of these approaches. *Direct replications* seek to examine whether two or more studies with the same well-defined causal estimand yield the same effect. The most stringent forms of direct replication seek to address *all* replication and individual study design assumptions. That is, these approaches attempt to hold all study characteristics fixed, while drawing new random samples for each replication study. When all assumptions are met, comparison of study results may be considered a test of statistical replication (Schmidt, 2009; Valentine et al., 2011). However, on their own, statistical replications are rarely of interest in the social sciences (Valentine et al., 2011). This is because statistical theory already provides strong guidance on the probability of replication failure through Null Hypothesis Significance Testing. Moreover, statistical replications are rarely feasible in field settings because it is often impossible to reproduce the same exact conditions over multiple studies, even for the simplest interventions (Hansen, 2011).

More informative are direct replications that seek to test one or more individual study assumption (*S1*-*S3*). 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 (i.e. a violation of individual study assumptions *S1*-*S3*) if replication failure is observed. Wong and Steiner (2018) describe examples of direct replication designs.

Conceptual replications, on the other hand, seek to examine whether two or more studies with potentially different causal estimands produce the same effect. To implement this approach, the researcher 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 identify potential sources of effect variation, often for the purpose of generalizing effects for broader target populations (Clemens, 2017; Schmidt, 2009). The remainder of this section focuses on research designs for conceptual replication. Although these designs 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. Below, we discuss examples of research designs for conceptual replication, and how they may be understood under the CRF.

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 make a series of pairwise contrasts for addressing questions – they may make contrasts for each intervention condition with the control group, or with other intervention conditions. The approach has been implemented to evaluate the relative efficacy of multiple reading interventions for struggling readers (Torgesen et al., 2007) and to identify approaches for promoting teacher-parent communications (Kraft & Rogers, 2015). Multi-armed RCTs have also been applied in international contexts. For example, Leventhal, DeMaria Gillham, Andrew, Peabody, and Leventhal (2015) tested in India whether adding different components of a social-emotional intervention to an adolescent health intervention would improve girls' emotional, social, and physical well-being.

Under the CRF, a multi-arm RCT may be understood as a replication design that purposefully relaxes the assumption of treatment stability (*R1*) to test whether results hold across variations in interventions. The design is considered a conceptual replication approach because the researcher evaluates whether two intervention contrasts with different causal estimands produce the same result. However, because systematic variation is introduced within a single study, all other CRF assumptions 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 unbiased identification of the causal estimand (*S1*). The researcher 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 difference in effect estimates, then the researcher may conclude with confidence that variation in intervention conditions resulted in the replication "failure" – that is, the different interventions produce different effects.

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, researchers may randomly assign ninth grade students in a school to a social-emotional intervention over three successive cohorts. Treatment effects for each cohort may be compared to evaluate whether the same result replicates over time. The 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.

Under the CRF, RCTs with multiple cohorts may be considered a conceptual replication designed to test natural violations in assumptions that occur when studies are repeated 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 different 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 effect 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 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. For example, in a social-emotional intervention that is implemented through text messages to students' cell phones, behavioral prompts that nudge participants to adopt more positive attitudes may become less effective over time if participants in each successive cohort become less engaged with text message prompts. The validity of the study design 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, Cook, & Campbell 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. In a hypothetical evaluation examining the efficacy of exercise on reducing anxiety, participants are randomly assigned into two groups at the beginning of the first interval – where one group is asked to attend a daily exercise class for a week and the second group is asked to engage in their regular activities. Here, the second group serves as a control for the first. At the end of the week, participants in the first group are asked to return to their regular routine while the second group is asked to participate in a weekly exercise protocol for the second week. However, the context for how the exercise protocol is delivered changes. In the second week, the same exercise protocol is delivered through a streaming app on participants' phones instead of through in-person exercise classes. Measurement of participants' anxiety levels are administered at the beginning of the study, after the first week of exercise, and following the weeklong exercise protocol for the second group.

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 exercise protocol was delivered (in-person class versus online app). 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 provide assurances that assumptions *S2* and *S3* were met. If replication failure is observed, the researcher may conclude that changes in how the exercise protocol was delivered was the cause of the effect variation.

For the switching replication design to be valid, however, requires that once the daily exercise protocol is removed, there should be no residual impact on participants' anxiety levels (*R1*). The assumption may be checked by extending the length of time between the first and second intervals, and by taking measures of anxiety immediately before the exercise protocol is introduced to the second group. The design also requires that the same outcome measure is used for assessing impacts and comparing results across study intervals (*R1*), and that there are no history or maturation effects that violate CRF assumptions (*R2*). For example, the second week of the intervention may coincide with a social media campaign to reduce anxiety through mindfulness practices. The social media campaign is a history threat for the causal interpretation of replication results only if intervention and control groups are affected differentially by the introduction of the campaign.

Combining Research Designs in Multiple Causal Systematic Replication Studies

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. 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 research designs to evaluate the replicability and robustness of effects.

The choice of research designs will depend on the researcher's subject matter knowledge about study characteristics that are hypothesized to produce different causal estimands or introduce bias in individual study results. Because causal replication designs often require controlled settings for manipulating study factors, the approach is most easily implemented by the same team of investigators, or through a team of investigators working collaboratively. Results from each replication study can be used to refine and improve intervention components and delivery mechanisms. Often, the goal is to develop and understand theory for describing the units, contexts, and conditions under which replicability of effects can be expected. As such, causal systematic replications may be especially helpful during the "development and innovation" phase of an intervention.

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 business-as-usual (BAU) control condition and compared candidates' pedagogical performance in the simulation session afterwards. The BAU condition consisted of the same practice opportunities in the simulation setting, but instead of receiving feedback from a coach, teacher candidates were asked to "self-reflect" on their performance through a series of structured question prompts. 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.[1] 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).

Figure 1a provides an overview of the schedule of the four RCTs. Here, each individual RCT is indexed by S_{ij}, where i denotes the sample (i.e. 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

[1] 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).

2 if the pedagogical task involved a conversation about setting classroom norms). Figure 1b 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 (e.g. S_{22} versus S_{12}). As we will discuss below, the team used empirical diagnostics of baseline characteristics to evaluate whether the distribution of participant characteristics actually remained fixed over time. To examine the replicability of effects across *different pedagogical tasks*, the research team implemented a modified switching replication design (e.g. S_{22} versus S_{21}). Here, candidates were randomly assigned in Fall 2018 to receive the coaching or the self-reflection protocol in the "text-based discussion" simulation scenario. However, instead of switching coaching conditions across groups, the research team re-randomized coaching conditions in Spring 2019 to explore evidence of effect variation for different levels of coaching over the two interval periods. Since there was no evidence of effect variation across the different coaching levels, the research team proceeded by analyzing the design as a "switching replication" approach. Coaching effects for the fall and spring intervals were compared to assess the replicability of effects across 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 2 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 Assumptions under the CRF

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 met. This can be done by using research design elements and empirical diagnostics to rule out the most plausible threats to validity (Shadish, Cook, & Campbell, 2002).

Though research designs such as the switching replication or the multiple cohort design can be used to address many CRF assumptions, research designs on their own are rarely able to address all assumptions under the CRF. Moreover, research designs are often implemented with deviations from their protocols in field settings. Unplanned variations in treatment conditions and in target populations may introduce violations of replication assumptions; individual studies may suffer from differential attrition or non-equivalence in treatment and control groups at baseline. Diagnostic probes are needed to assess the extent to which assumptions were actually met in replication settings.

Fortunately, the last thirty years of program evaluation literature has recommended methods for assessing assumptions that can (a) be directly used to assess the plausibility of individual study assumptions and (b) be easily extended to evaluate replication 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. 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.

Individual Study Design 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 (i.e., 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. 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 establishing sufficient and correct reporting of results. For example, recent Transparency and Openness (TOPS) guidelines from the Center for Open Science suggest standards for journals to support the replication and reproducibility of findings (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. TOPS 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 transparent the intended analysis plan, as well as making data and syntax files accessible for reproducing results.

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 TOPS 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.

A key issue in all replication studies is understanding which components constitute the treatment – as well as control – conditions and how they are delivered across studies. The challenge of assessing treatment stability is evident in debates regarding the extent to which failed replications should be considered direct replications given the adaptations made from one study to another (Gilbert et al., 2016a, 2016b, 2016c). Ideally, adherence to a clearly defined intervention protocol would be measured for every intervention session across all sites and studies. Traditionally, this requires hiring trained observers to rate each intervention session according to a rubric measuring fidelity. However, there are substantial logistical, methodological and budgetary constraints when attempting to monitor intervention delivery. For example, 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. Moreover, there is limited practical guidance about the best ways to assess intervention fidelity in field settings (Roberts, 2017), and even less guidance on how to quantify variation in delivery that may not be captured by the original study protocol. Perhaps because of these challenges, systematic replications rarely measure intervention fidelity and replicability, instead relying on researcher descriptions of changes made to the protocol (see, for example, Klein et al., 2014).

To address the challenge of assessing treatment stability in replication settings, Anglin and Wong (2020) introduce automated measures of treatment adherence and treatment replicability using natural language processing (NLP). These measures are calculated using a set of NLP techniques termed semantic similarity, which quantify the similarity between two or more texts. In evaluation contexts, semantic similarity methods can be used to assess treatment stability in highly-standardized interventions that are delivered through verbal interactions with participants. Here, treatment adherence is defined as the extent to which interventionalists stick to a scripted “benchmark” treatment protocol, setting aside inconsequential changes in language. Treatment replicability is defined as the extent to which the intervention language remains consistent across participants, sites, and studies. Then, semantic similarity techniques can be easily adapted to quantify the similarity between transcripts of intervention sessions as a measure of treatment replicability or between intervention transcripts and an ideal scripted protocol as a measure of treatment adherence.[2] Importantly, the semantic similarity approach is context agnostic, which means that it can be applied in any replication setting in which the intervention protocol is highly standardized, scripted, and delivered primarily through verbal interactions, and where transcripts of intervention sessions are available.

Example

[2] 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.

We now discuss examples of how researchers can probe the replication assumptions *R1* and *R2*. Tables 3 and 4 provide examples of balance tests using the causal systematic replication study described above (Krishnamachari, Wong, & Cohen, in progress). Table 3 summarizes descriptive statistics on study factors that were intended to be systematically varied across the four RCTs (e.g. timing, pedagogical task, and target population and setting); Table 4 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 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 3, 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. These participants were enrolled in a class exploring teaching as a profession, while samples from the other replication studies included teacher candidates who were already taking methods classes for pedagogical instruction.

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 4 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 4 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 4 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}.

In addition to comparing coaching transcripts to a gold-standard treatment benchmark to assess treatment adherence, the team also used semantic similarity methods to examine how similar intervention sessions were across studies. That is, instead of comparing session transcripts to a single gold standard benchmark transcript, the researchers assessed the similarity of transcripts with each other. These scores provide the researcher with direct summary measures of intervention "replicability." Table 5 provides a summary of replicability scores for coaching sessions across the four studies. The replicability score scale again ranged from 0 and 1, with higher scores indicating greater replicability of coaching sessions across the two studies. As a basis of comparison, we provide replicability scores of transcript sessions *within* each study (as indicated by the dark shaded boxes on the diagonal). Overall, coaching sessions were more different across two different studies (replicability scores reported in the off-diagonal) than coaching sessions within a study (replicability scores reported on the diagonal). We also see that despite having the same coaching protocol across all studies, coaching sessions were more similar in studies with the same pedagogical task (as indicated by the lightly shaded cells) compared to coaching sessions delivered in studies with different pedagogical tasks (as indicated by white cells). Table 5 also indicates that from a standpoint of treatment stability, the conceptual replications and multiple-cohort replications are better replications of the benchmark than the switching replication.

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 3 and 4 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 3 and 4 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.

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 3 and 4 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

In this paper, 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 required for the direct replication of results. Causal conceptual replication designs systematically test 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 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 are able to introduce systematic variations in each study.

To that end, we offer the following recommendations. First, the research team should choose a causal estimand (well-defined treatment-control contrast for a clearly defined target population and setting) with an effect that has been established, but perhaps not replicated. Second, the researchers should consider upfront which of the assumptions in the CRF they are most interested in testing and select research designs that are capable of testing 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, and switching replication designs. Fourth, researchers should outline upfront potential sources of bias and moderators of effects so that they can collect the data necessary for including diagnostic tests of assumptions. Finally, researchers should consider incorporating multiple research design elements at various phases of the research cycle. For example, as an intervention is being developed and piloted, switching replication and multiple cohort designs may be appropriate for assessing the replicability of results in highly controlled settings. 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 approaches with independent reporter (Chang & Li, 2015) may be used to assess individual study assumptions. As evaluations are scaled-up to reach more diverse populations, multi-site RCT designs may be used to examine the replicability of effects across sites with planned and unplanned sources of variation in populations, treatments, and settings.

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 goal here 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.

Finally, although it is beyond the scope of this paper to address methodological issues related to the statistical power and the analysis of replication approaches, we note that these issues are centrally related to the selection of an appropriate research design. For example, Steiner and Wong (2018) have noted that replication approaches with the same units across multiple studies (i.e. 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 (Hedges & Schauer, 2018; Simonsohn, 2015), there are likely design-based approaches to replication that may be effective for addressing these challenges. Future research should continue to expand the methodological foundations for the design, implementation, and analysis of replication approaches.

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.

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. https://doi.org/10.1111/joes.12139

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

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

Gilbert, D., King, G., Pettigrew, S., & Wilson, T. (2016a). *A Response to the Reply to our Technical Comment on "Estimating the Reproducibility of Psychological Science."*

Gilbert, D., King, G., Pettigrew, S., & Wilson, T. (2016b). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad7243

Gilbert, D., King, G., Pettigrew, S., & Wilson, T. (2016c). *More on "Estimating the Reproducibility of Psychological Science."* https://gking.harvard.edu/files/gking/files/gkpw_post_publication_response.pdf

Hansen, W. B. (2011). Was Herodotus Correct? *Prevention Science*, *12*(2), 118–120. https://doi.org/10.1007/s11121-011-0218-5

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

Kraft, M. A., & Rogers, T. (2015). The underutilized potential of teacher-to-parent communication: Evidence from a field experiment. *Economics of Education Review*, *47*, 49–63.

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

Leventhal, K. S., Demaria, L. M., Gillham, J., Andrew, G., Peabody, J. W., & Leventhal, S. (2015). *Fostering emotional, social, physical and educational wellbeing in rural India: The methods of a multi-arm randomized controlled trial of Girls First*. *16*(1). https://doi.org/10.1186/s13063-015-1008-3

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

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

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

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.

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

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. *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

Torgesen, J., Schirm, A., Castner, L., Vartivarian, S., Mansfield, W., Myers, D., Stancavage, F., Durno, D., Javorsky, R., & Haan, C. (2007). National Assessment of Title I. Final Report.

Volume II: Closing the Reading Gap--Findings from a Randomized Trial of Four Reading Interventions for Striving Readers. NCEE 2008-4013. In *National Center for Education Evaluation and Regional Assistance*. National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED499018

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 and Figures

Figure 1a. 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}

Figure 1b. 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 Figures 1a and 1b, each individual RCT is indexed by S_{ij}, where *i* denotes the sample (i.e. 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 (e.g. 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 1. Causal Replication Framework for the Direct Replication of Effects (Steiner, Wong, & Anglin, 2020; Wong & Steiner, 2018)

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

Table 2: 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 3: 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 4: 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 5: Replicability Matrix for Evaluating Treatment Stability

	S_{22}	S_{21}	S_{12}	S_{32}
	Benchmark Study	Multiple Cohort	Switching Replication	Conceptual Replication
S_{22} Benchmark	0.40	0.31	0.24	0.33
S_{12} Multiple Cohort	0.31	0.40	0.22	0.27
S_{21} Switching Replication	0.24	0.22	0.50	0.23
S_{31} Conceptual Replication	0.33	0.27	0.23	0.37

Notes: The replicability index is calculated by calculating the pairwise similarity of each transcript in the study indicated in the first row to each transcript in the study indicated by the first column. Cosine similarity was calculated using a document-term matrix with latent-semantic analysis, no stop words, and term-frequency-inverse-document-frequency weighting.

Reproduced and adapted from Anglin and Wong (2020)