EdPolicyWorks

Working Paper:

Designs of Empirical Evaluations of Non-Experimental Methods in Field Settings

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

Over the last three decades, a research design has emerged to evaluate the performance of non-experimental designs and design features in field settings. It is called the within-study comparison (WSC) approach, or design replication study. In the traditional WSC design, treatment effects from a randomized experiment are compared to those produced by a non-experimental approach that shares the same target population. The non-experiment may be a quasi-experimental design, such as a regression-discontinuity or an interrupted time series design, or an observational study approach that includes matching methods, standard regression adjustments, and difference-in-differences methods. The goals of the WSC are to determine whether the non-experiment can replicate results from a randomized experiment (which provides the causal benchmark estimate), and the contexts and conditions under which these methods work in practice. This paper presents a coherent theory of the design and implementation of WSCs for evaluating non-experimental methods. It introduces and identifies the multiple purposes of WSCs, required design components, common threats to validity, design variants, and causal estimands of interest in WSCs. It highlights two general approaches for empirical evaluations of methods in field settings, WSC designs with independent and dependent benchmark and non-experimental arms. The paper highlights advantages and disadvantages for each approach, and conditions and contexts under which each approach is optimal for addressing methodological questions.

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

Updated April 2016

EdPolicyWorks
University of Virginia
PO Box 400879
Charlottesville, VA 22904

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

EdPolicyWorks Working Paper Series No. 44. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

DESIGNS OF EMPIRICAL EVALUATIONS OF NON-EXPERIMENTAL METHODS IN FIELD SETTINGS

Vivian C. Wong & Peter M. Steiner

Introduction

Across the disciplines of economics, political science, public policy, and now, education, the randomized controlled trial (RCT) is the preferred methodology for establishing causal inference about program impacts. But randomized experiments are not always feasible because of ethical, political, and/or practical considerations, so non-experimental methods are also needed for identifying "what works." Given the widespread use of non-experimental approaches for assessing program, policy, and intervention impacts, there is a strong need to know whether non-experimental approaches are likely to yield unbiased treatment effects, and the contexts and conditions under which non-experimental methods perform well.

Over the last three decades, a research design has emerged to evaluate the performance of non-experimental designs in field settings. It is called the within-study comparison (WSC) design, or design replication study. In the WSC design (see Figures 1 and 2 for overview of designs), treatment effects from a benchmark design are compared to those produced by a non-experimental (NE) approach that shares the same target population. The non-experiment may be a quasi-experimental (QE) design, such as a regression-discontinuity (RD) or an interrupted time series (ITS) design, or an observational study (OS) that includes matching methods, standard regression adjustments, and difference-in-differences methods. The goals of the WSC are to determine (1) whether the non-experiment can replicate results from a randomized experiment (which provides the causal benchmark estimate) in field settings, and (2) the contexts and conditions under which these methods work in practice.

As statistical theory on non-experimental methods continues to develop, more WSCs will be needed to assess whether these methods are suitable for causal inference in field settings, that is, whether the underlying assumptions required for identification and estimation are likely to be met. Moreover, researchers have only recently begun to use WSCs for validating non-experimental designs common in social and behavioral science evaluations that include pretest-posttest, difference-in-differences, and interrupted time series approaches (McConeghy, Steiner, Wing, & Wong, 2013; St. Clair, Cook, & Hallberg, under review; Somers, Zhu, Jacob, & Bloom, 2012). Given that non-experimental methods continue to be crucial for uncovering "what works" in program evaluation, high-quality empirical tests of non-experimental methods are needed to understand contexts and conditions under which these approaches are likely to produce trustworthy results. WSCs may also address questions about the current practice of non-experimental methods. For substantive social and behavioral science researchers, WSCs provide opportunities to check hypotheses about non-experimental methods for particular contexts and outcomes when random assignment is not feasible (Angrist, Autor, Hudson, & Pallais, 2015). At stake is a methodology that allows program and policy evaluators to develop and refine empirically-based "best practices" for non-experimental approaches.

Despite the opportunities and reasons for conducting WSCs, the approach is underutilized for improving research practice in social and behavioral science settings. This is because there is no coherent framework for understanding the design of WSCs as a method for evaluating non-experimental approaches. The lack of guidance on the design, implementation, and analysis of WSCs is problematic for a number of reasons. First, for researchers who wish to use WSCs to investigate non-experimental methods, the only available resources are examples of WSCs scattered across the social and behavioral sciences that includes job training, criminology, political science, international development, health policy, and education. With the exception of a brief discussion by Cook, Shadish, and Wong (2008), who present six criteria for a causally valid WSC, there is no methodological paper devoted to the appropriate design and analysis of the WSC itself. As a result, the existing WSCs are of heterogeneous quality, with researchers using ad hoc designs and methods that may or may not be appropriate for addressing the research question of interest (Shadish, Steiner & Cook, 2012). Second, without a general framework for considering WSC designs, researchers may not understand different types of WSC approaches, their relative strengths and limitations, and other methodological considerations for implementing high quality empirical validation studies. Finally, thus far, the WSC design has been used by a relatively small cadre of research methodologists (and scholars interested in methods) who wish to understand the performance of non-experimental methods in field settings. For methodological researchers in program evaluation, a general WSC framework would provide an important resource on the design, implementation, and analysis of WSCs for evaluating non-experimental approaches in field settings. For researchers interested in substantive social and behavioral science issues, the WSC framework offers guidance for incorporating WSCs in current research programs to empirically investigate the performance of non-experimental methods with particular target populations, treatments, selection processes, and outcomes of interest.

The purpose of this paper then is to present a coherent framework for considering WSC research designs for evaluating non-experimental methods in field settings. In empirical evaluations of methods, there are two general approaches for evaluating non-experimental methods in field settings. They are: 1) research designs with *independent* WSC arms, where units are randomly assigned into the benchmark and non-experimental arms, and treatment effects from each arm are compared (Panel 1, Figure 1); and 2) research designs with *dependent* WSC arms, where the benchmark and non-experimental arms share some portion of the sample, and non-experimental effects are compared to benchmark results (Panel 2, Figure 1; Panels 1 & 2, Figure 2). For dependent WSC designs, researchers have introduced multiple design variants, which we will discuss in this paper. For the remainder of this paper, we will refer to the benchmark design as an RCT or an experiment. However, this needs not always be the case. The benchmark design may be a well-designed quasi-experiment, such as a regression discontinuity design, that the researcher believes to provide a credible benchmark result. Below, we highlight what is required of "credible" benchmark designs for evaluating non-experimental methods.

The paper proceeds by describing WSC approaches with independent and dependent arms. To this end, we discuss the overall set-up and logic of the designs, assumptions required for valid empirical tests of methods, and the causal estimands of interest. For each design, we highlight an application to highlight the approach's advantages and limitations. We conclude by offering researchers practical guidance for choosing and implementing appropriate WSC designs in field settings.

Background on Empirical Evaluation of Methods

Introduced by LaLonde (1986), the earliest WSC designs used data from job training evaluations to compare results from a non-experimental study with those from an experimental benchmark that shared the same treatment group. The early conclusion from these studies was that non-experimental methods failed to produce results that were comparable to their experimental benchmark estimates (Fraker & Maynard, 1987; Friedlander & Robins, 1995). In the early 2000s, Glazerman, Levy and Myers (2005) meta-analyzed 12 WSCs that used data from a series of job training experiments and found that overall, although non-experimental approaches sometimes replicated experimental benchmark results, they often produced effects that were "dramatically different from the experimental benchmark" (p. 86). Although Glazerman et al. wrote that results from the meta-analysis did not resolve "longstanding debates about non-experimental methods," for many readers, the take-home message was clear – non-experimental methods could not be trusted to produce credible causal estimates in field settings.

Results from early WSCs had profound influence on research practice and priorities, especially in the field of education. The Office of Management and Budget cited results from early WSCs in their 2004 recommendation that federal agencies should use randomized experiments for evaluating program impacts, cautioning against the use of "comparison group studies" that "often lead to erroneous conclusions" (OMB, 2004). The U.S. Department of Education also identified random assignment as the preferred method choice for "scientifically-based research" in a 2005 issue of the *Federal Register*. In responding to critiques that random assignment was "not the only method capable of generating causal effects," Secretary Paige cited WSC results, stating that "conclusions about causality based on other methods, including the quasi-experimental designs included in this priority, have been shown to be misleading compared with experimental evidence" (Federal Register, 2005, pg. 3588). Overall, these findings reified a clear preference in methodology choice for government funding policy – randomized experiments whenever possible, then RD, and finally if at all, observational approaches such as matching or regression.

However, the empirical evidence from early WSCs was suspect in a number of ways. Experimental and non-experimental groups in the early WSCs differed in many more ways than in the mode of assignment, possibly confounding interpretation of results. The field experiments that served as the benchmark for evaluating the non-experiments suffered from their own implementation problems in the field (e.g. differential attrition), and raised questions about whether they provided a valid causal benchmark. To date, only Cook, Shadish, and Wong (2008) address methodological theory for improving the quality of WSCs. In their qualitative review of WSCs from 2003 through 2007, the authors present their criteria for high quality WSCs (see also Cook & Wong, 2008; Shadish, Steiner & Cook, 2013). They developed these criteria for the critical assessment of the WSCs they reviewed, but they also aimed at helping readers and implementers of WSC to interpret findings from these studies, as well as to plan future WSCs. Cook et al. urged WSC analysts to ensure that: the RCT is well-executed to warrant its causal benchmark status; there is no third-variable confounder that challenges the causal interpretation of WSC results; the experiment and non-experiment have the same causal estimand (e.g. the experimental ATE is compared to the non-experimental ATE); there is a clear criteria for inferring correspondence in experimental and non-experimental results; among other criteria.

Although the criteria have been useful in producing more causally interpretable WSCs in recent years (Ferraro & Miranda, 2013; Fortson, Verbitsky-Savitz, Kopa, & Gleason, 2012; Gleason, Resch, & Berk, 2012), continued interest in WSCs – as well as emerging applications of the design (Peck, Bell, & Werner, 2013) – suggest the need for theory to describe the various design approaches for evaluating methods in field settings, formalize the assumptions required for each WSC approach to produce interpretable results, the causal quantities of interest, and the relative strengths and weaknesses of each approach.

Independent WSC Designs

Introduced by Shadish et al. (2008), the researcher begins by randomly assigning units into benchmark and non-experimental arms in WSC designs within independent benchmark and non-experimental arms (Panel 1, Figure 1). Participants in the benchmark arm are randomly assigned again into treatment conditions; participants in the non-experimental arm self-select or are selected by a third party into a preferred treatment option. Random assignment into WSC arms ensures that participants in the benchmark and non-experimental conditions are equivalent on expectation; random assignment in the benchmark design helps ensure that the non-experiment is evaluated against a valid estimate of the impact (given that the RCT has been perfectly implemented). Because the benchmark and non-experimental arms each have a treatment and control condition, the approach is sometimes referred to as the "four-arm" design (Shadish, Clark, Steiner, 2008; Shadish, Galindo, Wong, Steiner, & Cook, 2010).

Because researchers prospectively randomly assign units into WSC benchmark and non-experimental conditions, the researcher has control over many study attributes that often confound the interpretation of results in other WSC design approaches. For example, the researcher can ensure that participants in the benchmark and non-experimental arms have the same eligibility requirements, are measured on the same outcomes at the same time, and have undergone similar experiences in the WSC benchmark and non-experimental arms. The design also provides opportunities for the researcher to define treatment contrasts of substantive interest within each WSC arm. Units may be assigned to a reading or math intervention, to participate in a weight-loss program, or to receive electronic behavioral nudges for desired behaviors. Here, the WSC researcher has the opportunity to construct a prospective non-experimental design that mimics units' real world selection into treatment conditions. Moreover, they are able to test the performance of non-experimental methods in addressing these selection processes. Because of random assignment of units into WSC arms – as well as in the benchmark condition – the independent arm design is considered the "gold standard" approach for producing internally valid estimates of non-experimental bias. Despite these advantages, the method is often challenged in terms of implementation feasibility in field settings, and generalizability of results.

Conceptual framework. We begin by considering an independent design with two study arms $W_i \in \{0, 1\}$, where W_i is coded 0 if unit *i* belongs to the benchmark condition, and 1 if the unit is in the non-experimental condition. Within each arm, there are two treatment conditions $T_i \in \{0, 1\}$, such that $T_i = 0$ if the unit is in the control condition, and 1 if it is in the treatment. This means that in a WSC design, each individual *i* has four potential outcomes, $Y_i(T_i = t, W_i = w)$, which include control and treatment outcomes in the benchmark, $Y_i(0,0)$ and $Y_i(1,0)$, and control and treatment outcomes in the non-experiment, $Y_i(0,1)$ and $Y_i(1,1)$.

With this framework in mind, we can define the causal estimands of interest for the independent arm design. For clarity in notation, we continue to assume that the benchmark design is an RCT, and the non-experimental approach is an observational study in which individuals self-select into treatment conditions[1]. For the benchmark RCT ($W = 0$), the causal estimand of interest is the average treatment effect ATE($W = 0$) for the WSC population, such that: $ATE(0) = E(Y_i(1,0)) - E(Y_i(0,0)) = E(Y_i(1,0) - Y_i(0,0))$. Here, the expectation is taken over all individuals in both the RCT and observational study, that is, the ATE if all units were assigned to the RCT. Correspondingly, we define the ATE(1) for the observational study ($W = 1$) as: $ATE(0) = E(Y_i(1,1)) - E(Y_i(0,1)) = E(Y_i(1,1) - Y_i(0,1))$, where again, the expectation is taken over the entire WSC population.

The goal of the WSC design is to assess whether the non-experimental and the benchmark designs produce equivalent treatment effects[2]. However, the comparison is interpretable only if five key assumptions about the WSC study are met. Combined, these assumptions ensure equivalence in causal estimands of both WSC study conditions, validity of the benchmark result, and comparability of benchmark and non-experimental conditions. The first two assumptions for the WSC design ensure that the potential outcomes depend solely on WSC status and treatment condition that each unit itself receives. The two assumptions formulate an exclusion restriction and are frequently referred to as the stable-unit-treatment-value assumption (SUTVA). Here, the first part of SUTVA formulates a no-interference assumption; the second part excludes hidden variations in WSC and

[1] In practice, the benchmark may be a credible non-experimental design, such as a regression-discontinuity design (Somers, Zhu, Jacob, & Bloom, 2013), and the non-experiment may be a method or research design that the WSC analyst wishes to test in a field setting.

[2] Steiner & Wong (under review) discuss methods for assessing correspondence in WSC designs.

treatment conditions. Both are standard assumptions in most RCT and non-experimental studies, but they have special implications in the WSC context.

Assumption A1 (SUTVA: No Interference): A unit's potential outcomes depend only on its own WSC and treatment status W_i and T_i, but do not depend on other units' statuses. We formalize this by stating, $Y_i(T_i,\mathbf{T}_{-i},W_i,\mathbf{W}_{-i}) = Y_i(T_i,W_i)$ where $\mathbf{T}_{-i}$ and $\mathbf{W}_{-i}$ are the assignment vectors of all other units except for unit *i*.

In the WSC context[3], assumption A1 implies the following: A1.1: Potential outcomes are unaffected by others' participation in one of the WSC conditions and treatment conditions. This condition is met only if units do not react to others' participation in WSC conditions, nor do they react to the treatment status of others within each WSC arm. The no-interference assumption is violated if, for example, units' feel demoralized because friends and peers are assigned into a different WSC or treatment condition (because demoralization alters their potential outcomes).

Assumption A2 (SUTVA: No Hidden Variation of WSC conditions and Treatments): The WSC and treatment statuses, W and T, are uniquely defined, that is, there are no different versions of each WSC and treatment condition. Thus, the potential outcomes depend only on the uniquely defined and administered WSC and treatment conditions, and not on other third variables. This assumption is sometimes referred to as the exclusion restriction. We formalize this assumption by introducing two additional terms representing third variables. Let *G* be a third variable with respect to the WSC conditions, and *Z* a corresponding third variable with respect to the treatment conditions within each condition of the WSC. Then, this assumption implies that $Y_i(T_i,W_i,Z_i=z,\mathrm{G}_i=g) = Y_i(T_i,W_i,Z_i=z',\mathrm{G}_i=g') = Y_i(T_i,W_i)$ with $z \neq z'$ and $g \neq g'$.

Assumption A2 implies the following in the WSC context:

A2.1: Treatment and control conditions are uniquely defined in both WSC arms and do not vary across WSC conditions. This means that there are stable, clearly defined RCT and non-experimental conditions, as well as treatment and control conditions (Rubin, 1980; Rubin, 1986). Moreover, it requires units receive identical treatment and control conditions in the benchmark and non-experimental arms of the WSC. This assumption is violated if, for instance, RCT units receive a stronger treatment dosage than treated cases in the non-experimental arm, or, if comparison cases in

[3] The SUTVA assumption is stronger than what is actually required in this context. For a valid WSC design we only require that there is no differential violation of SUTVA between the two study arms. For instance, if peer effects are present to the same extent in the RCT and the observational study the causal quantities would be biased but still equivalent (which is sufficient). However, for clarity we maintain SUTVA throughout the paper.

the non-experimental study have additional alternative treatment options than what is available for control cases in the benchmark.

A2.2: Potential outcomes are unaffected by the mode of assignment into benchmark and non-experimental conditions, and into treatment and control conditions within each WSC arm. This requires that units do not respond differentially to randomization into benchmark or non-experimental conditions, nor do they react differentially to the treatment assignment mechanism in the benchmark and observational conditions. The assumption is violated if benchmark units react negatively to random assignment into treatment conditions, as opposed to being allowed to self-select into a desired treatment condition. Or, as another example, the benchmark and non-experimental units need to be measured on the same outcomes at the same time in the same setting and contexts.

When A1 and A2 are met, then the observed outcome is a function of each unit's potential outcomes, and its status T and W (i.e., whether the unit is assigned to the treatment or control conditions within the benchmark or non-experimental arms). Thus, for each unit i, the observed outcome may be written as:

$$Y_i = Y_i(0,0)(1-T_i)(1-W_i) + Y_i(0,1)(1-T_i)W_i + Y_i(1,0)T_i(1-W_i) + Y_i(1,1)T_iW_i \,. \qquad [1]$$

Moreover, SUTVA ensures that potential control and treatment outcomes are identical in the benchmark and non-experimental arm, $Y_i(0,1) = Y_i(0,0)$ and $Y_i(1,1) = Y_i(1,0)$, and the individual treatment effects are the same $Y_i(1,0) – Y_i(0,0) = Y_i(1,1) – Y_i(0,1)$. This implies equivalence in the causal estimands in the benchmark and non-experimental conditions, such that ATE(0) = ATE(1).[4] Violations of the above assumptions usually produce differences in potential control and treatment outcomes across WSC conditions, $Y_i(0,1) \neq Y_i(0,0)$ or $Y_i(1,1) \neq Y_i(1,0)$, creating non-equivalence in the causal estimands, ATE(0) ≠ ATE(1), rendering the WSC results causally uninterpretable for the purpose of probing the performance of non-experimental methods.

Causal Identification. Thus far, we have defined $ATE(0)$ and $ATE(1)$ in terms of their expected potential outcomes. In practice, however, these quantities cannot be computed directly because the researcher observes only one of the four potential outcomes for each unit. For control units in the benchmark RCT ($T_i = 0$, $W_i = 0$), the observed outcome is $Y_i(0, 0)$, and for treatment units ($T_i = 1$, $W_i = 0$), it is $Y_i(1, 0)$.

[4] We could slightly relax the SUTVA assumption and still have equivalence in potential outcomes. Instead of requiring the equivalence of potential control (treatment) outcomes across the two WSC conditions, it would be sufficient to assume that individual treatment effects are the same in both WSC conditions, $\mathrm{Y_i}(1,0) – \mathrm{Y_i}(0,0) = \mathrm{Y_i}(1,1) – \mathrm{Y_i}(0,1)$.

With additional (strong) assumptions, we can show that the expectations of the four observed outcomes, $E(Y_i \mid T_i = 0, W_i = 0)$, $E(Y_i \mid T_i = 1, W_i = 0)$, $E(Y_i \mid T_i = 0, W_i = 1)$, and $E(Y_i \mid T_i = 1, W_i = 1)$, may be used to identify ATEs for the benchmark and non-experimental conditions. To do this, we begin by showing that the difference in observable expectations for the RCT $\tau(0) = E(Y_i \mid T_i = 1, W_i = 0)$ - $E(Y_i \mid T_i = 0, W_i = 0)$ is equivalent to the WSC $ATE(0) = E(Y_i(1,1)) - E(Y_i(0,1))$ if two additional assumptions are met:

Assumption A3. The potential outcomes are independent of WSC conditions, such that: $(Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1)) \perp W_i$. This assumption is met if units are randomly assigned into benchmark and non-experimental conditions, as is the case in the independent arm design, and if units comply with the assigned WSC status (full compliance)

Assumption A4. For the benchmark ($W = 0$), potential outcomes are independent of treatment: $(Y_i(0,0), Y_i(1,0)) \perp T_i \mid W_i = 0$. In the benchmark, this assumption is met through random assignment[5] of units into treatment conditions and full compliance with the assigned treatment status.

When all four assumptions A1-A4 are met, the expectations of the observed outcomes may be used to identify the ATE in the benchmark (ATE(0)), such that:

$$\begin{aligned}\tau(0) &= E(Y_i \mid T_i = 1, W_i = 0) - E(Y_i \mid Y_i = 0, W_i = 0)\\ &= E(Y_i(1,0) \mid T_i = 1, W_i = 0) - E(Y_i(0,0) \mid T_i = 0, W_i = 0)\\ &= E(Y_i(1,0) \mid W_i = 0) - E(Y_i(0,0) \mid W_i = 0)\\ &= E(Y_i(1,0)) - E(Y_i(0,0) = ATE(0)\end{aligned}$$

Here, line 1 is the difference in the expectations of the observed treatment and control outcomes in the benchmark. Line 2 follows from SUTVA (A1 and A2, Equation [1]), line 3 from the independence between potential treatment and control outcomes (A4), and line 4 from the independence of WSC status and potential outcomes (A3).

We next show that the difference in expectations of the observed potential outcomes in the non-experimental arm, $\tau(1) = E_X\{E(Y_i \mid T_i = 1, W_i = 1, X = x)\} - E_X\{E(Y_i \mid T_i = 0, W_i = 1, X = x)\}$, may be used to identify ATE(1) if the following assumptions are met:

Assumption A3. The potential outcomes are independent of WSC conditions. This condition is the same as above (A3).

[5] Treatment may be conditionally independent of potential outcomes in the benchmark, such that $T_i \perp (Y(0,0), Y(1,0) \mid X)$. In these cases, the benchmark is a quasi-experiment such as a regression-discontinuity design in which the conditional independence assumption has been shown to be plausible and has been perfectly implemented.

*Assumption A5. For the non-experimental arm (W = 1), potential outcomes are independent of treatment conditional on a set of covariates **X** (strong ignorability)*: $Y_i(0,1), Y_i(1,1) \perp T_i \mid X_i = x, W_i = 1$ and $0 < P(T_i \mid X_i = x, W_i = 1) < 1$. This assumption is met only if the observed covariates of the non-experimental design (e.g. an observational study) remove all confounding bias due to differential selection.

Once A1, A2, A3, and A5 are met, we can show that the observed potential outcomes may be used to identify ATE in the non-experimental arm:

$$\begin{aligned}\tau(1) &= E_X\{E(Y_i \mid T_i = 1, W_i = 1, X = x)\} - E_X\{E(Y_i \mid T_i = 0, W_i = 1, X = x)\}\\ &= E_X\{E(Y_i(1,1) \mid T_i = 1, W_i = 1, X = x)\} - E_X\{E(Y_i(0,1) \mid T_i = 0, W_i = 1, X = x)\}\\ &= E_X\{E(Y_i(1,1) \mid W_i = 1, X = x)\} - E_X\{E(Y_i(0,1) \mid W_i = 1, X = x)\}\\ &= E(Y_i(1,1) \mid W_i = 1) - E(Y_i(0,1) \mid W_i = 1)\\ &= E(Y_i(1,1)) - E(Y_i(0,1)) = ATE(1)\end{aligned}$$

Here, line 1 is the difference in the expectations of the observed outcomes for treatment and comparison cases in the non-experimental arm. Line 2 is equivalent due to SUTVA (A1 and A2, Equation [1]), line 3 follows from A5 (strong ignorability in the non-experiment), line 4 takes the expectation over the distribution of **X**, and line 5 follows from A3 (independence of WSC conditions). Note that in both sets of equations, expectations that are not conditional on W are taken over units in the RCT and non-experiment, respectively, producing the ATE of the WSC population.

Evaluating non-experimental methods in field settings. The goal of the independent arm design is to examine whether the ATE estimator for the non-experiment is equivalent (in expectation)to from the ATE estimator for the benchmark design, that is, whether $E(\hat{\tau}(1)) = E(\hat{\tau}(0))$. If the estimators differ, $E(\hat{\tau}(1)) \neq E(\hat{\tau}(0))$, it implies that at least one of the five assumptions used to establish equivalence between benchmark and non-experimental treatment results is violated, such that $\tau(1) \neq \tau(0)$, or that at least one of the two ATE estimators is biased, $E(\hat{\tau}(1)) \neq \tau(1)$ or $E(\hat{\tau}(0)) \neq \tau(0)$. A violation of the identifying assumption occurs, for example, when the set of covariates **X** may not establish statistical independence between treatment status and the potential outcomes in the non-experimental arm (violation of A5); randomization may not have been perfectly implemented in the RCT benchmark (violation of A4); units' outcomes were measured at different times in the benchmark and non-experimental conditions such that the potential outcomes in the two arms are no longer equivalent due to maturation effects (violation of A2); or, peer effects were larger in the non-experimental condition than in the benchmark (violation of A1). Moreover, bias in the ATE estimators is possible due to model misspecification (incorrect functional form assumptions), or due to the use of a consistent estimator (where the bias vanishes as the sample size goes to infinity). If the benchmark estimate comes from an RCT, standard regression-based estimators are usually unbiased, but regression or matching estimators typically depend on a correctly specified functional form or the size of the treatment and comparison groups.

In the independent arm design, the difference in average treatment effects between the non-experimental and benchmark studies, $E(\hat{\tau}(1)) = E(\hat{\tau}(0))$ = β, may be interpreted as non-experimental bias only if the difference is due to a biased non-experimental ATE estimator ($E(\hat{\tau}(1)) \neq \tau(1)$) or a violation of the strong ignorability assumption A5 in the non-experiment (or both together), and not because of a violation of one or more of the other four WSC design assumptions: SUTVA (A1, A2), independence of potential outcomes and WSC conditions (A3), and, for the benchmark design, independence of potential outcomes and treatment status (A4). Without these conditions, interpretation of WSC results as a test of non-experimental methods is challenged due to ambiguity in knowing which assumption was violated in the field setting.

Example. In 2008, Shadish, Clark, & Steiner (2008) introduced the independent arm design in which 445 college students were randomly assigned into the RCT benchmark (N=235) or an observational design (N=210). Once students were randomly assigned into benchmark and observational conditions, those in the benchmark arm were randomly assigned again into vocabulary and math interventions, while those in the non-experiment were asked to select the intervention of their choice. The WSC included two possible treatments where students could learn either about 50 advanced vocabulary terms or five algebraic concepts. To ensure treatment conditions were the same across WSC arms, lecturers used the same scripts and overhead transparencies to deliver vocabulary and math trainings in both WSC conditions. Outcome measures included a 30-item vocabulary test and a 20-item mathematics test that were administered to all participants. For the benchmark, the researchers used ANCOVA to estimate the average treatment effect. For the non-experiment, the researchers used rich baseline covariates to conduct propensity score matching for creating equivalent groups, and ANCOVA to estimate the average treatment effect. Bias in the non-experimental methods was assessed by comparing the observed ATE in the non-experimental arm to the observed ATE obtained from the RCT benchmark.

This study exemplifies several characteristics common to independent arm designs for evaluating non-experimental methods. First, it is a prospective design, so that researchers had maximum control over the assignment of units into WSC conditions, their assignment into vocabulary and math trainings (e.g. random assignment versus self-selection), and the conditions under which baseline and post-test measures were administered. They also were able to collect a rich assortment of baseline information that plausibly described the multiple constructs related to treatment selection and the outcome. Finally, they were able to track students' responses to random assignment into WSC and treatment conditions in the benchmark, ensuring that randomization was correctly implemented. Overall, the researchers concluded that propensity score and ANCOVA methods performed well in replicating benchmark results, and that results were robust to the choice of propensity score techniques (e.g. stratification, covariate adjustment, weighting).

However, as the researchers acknowledge, the approach was implemented under "laboratory-like" conditions, limiting the generalization of results. And, because treatments involved low-intensity vocabulary and math trainings, the selection process in the observational arm may have been relatively straight-forward to model in a propensity score analysis. This is in contrast to selection processes that researchers typically encounter in field settings, such as students' decision to enroll in a charter school or an after-school program, welfare recipients' participation in a job training program, or schools' adoption of a formative assessment program. Moreover, the sample size obtained in the WSC was small – 235 students in the RCT and 210 students in the non-experiment. Here, interpreting correspondence in benchmark and non-experimental results may be challenged if the WSC was underpowered to detect statistical differences in results.

Thus far, the Shadish et al. study has been replicated with a sample of 205 students in Jena, Germany (Pohl, Steiner, Eisermann, Soellner, & Cook, 2009). Another variation of this design was implemented by Shadish, Galindo, Wong, Steiner, and Cook (2011), who randomly assigned units either into an RCT benchmark condition, or a regression-discontinuity design in which participants were assigned into a vocabulary or math intervention based on a vocabulary pretest score and a cutoff threshold. Here, the benchmark ATE at the cutoff was compared to an RD treatment effect at the same location.

The independent arm design has been used to address a variety of questions about non-experimental practice, including evaluating the performance of different propensity score techniques (e.g. matching, stratification, weighting) (Shadish et al., 2008) and choice of covariates in estimating the propensity score (Steiner et al., 2010). Moreover, Marcus, Stuart, Wang, Shadish, and Steiner (2012) re-analyzed the Shadish et al. data to assess the presence of preference effects. Preference effects are relevant in cases where researchers may be concerned about the generalization of treatment effects from RCTs because outcomes for individuals who agreed to be randomized may differ from those who have strong preferences for treatment. In Marcus et al., the researchers assumed that the covariate set (**X**) was sufficient for meeting the strong ignorability assumption in the observational arm (A5), but hypothesized that units respond to randomization into treatment and control conditions (a violation of STUVA, A2). As such, differences in the observed benchmark and non-experimental arms were interpreted as evidence of preference effects by units in the WSC study. This example illustrates both the rigor that is needed for a well-implemented WSC design, as well as its potential flexibility in testing methodological assumptions in field settings. However, the validity of conclusions drawn from a WSC depends on whether the required assumptions are met. If two assumptions are violated simultaneously, say A5 because not all confounding covariates are observed and A2 because of preference effects, then an observed difference in the experimental and non-experimental effect estimate cannot uniquely be attributed to a violation of the strong ignorability assumption (A5).

Dependent WSC Designs

In dependent WSC designs, the benchmark and non-experimental arms share some portion of the sample, creating dependency in the data structure between the two WSC conditions. Panel 2 of Figure 1 depicts a common variant of a dependent WSC, which we call the simultaneous approach. In this example, units from an overall population select into the benchmark condition – often an RCT – or do not participate in the RCT and, thus, may be included for the non-experimental condition. In the benchmark RCT, units are randomly assigned into conditions, and the obtained treatment effect again serves as the causal benchmark result for evaluating the non-experimental method. In the non-experimental arm, WSC analysts designate the RCT treatment units as the non-experimental treatment units, but draw comparisons from samples *not in the RCT* (i.e., from the population that did not participate in the RCT). Comparisons may be drawn from an observational study (LaLonde, 1986; Fraker & Maynard, 1987), from the control group of a different RCT evaluation (Dong & Lipsey, under review), or from administrative data (Hallberg, Wong, & Cook, under review). What is required is that non-equivalent comparisons were not exposure to the treatment in the benchmark design, and were measured on the same outcomes at the same time as treatment units in the RCT.

Non-equivalent groups in the non-experiment are created by units' selection into the benchmark and non-experimental arms of the WSC. Thus, in contrast to the independent WSC design, where differential *treatment selection* occurs in the non-experimental arm, the dependent WSC design is characterized by differential *selection into WSC conditions* (instead of treatments). The goal of the non-experimental method, then, is to find a valid comparison group for the RCT treatment units from a pool of non-equivalent comparison units. Because treatment units are held constant between benchmark and non-experimental conditions, the interpretability of WSC results hinges only on the comparability of the non-experimental comparison group and the benchmark control group – a somewhat weaker requirement than for the independent WSC design which also requires the comparability of treatment conditions.

There are three design variants of the dependent arm approach. In all of these cases, some portion of the treatment and comparison units are held constant between the benchmark and non-experimental conditions, which helps in ruling out irrelevant confounders (e.g., variations in the experimental and "non-experimental" treatment or control condition). However, the designs differ based on the initial selection process in the non-experimental design. For example, in the "simultaneous" design described above, the non-experimental process under investigation occurs when units select into WSC conditions (benchmark or non-experiment). In a "multi-site simultaneous" design, the non-experiment is based on units' (natural) selection into different sites in a benchmark RCT. In a "synthetic" design, the researcher begins with data from an RCT benchmark and creates a non-experiment by deleting some portion of the experimental sample to generate a non-equivalent comparison group. Below, we discuss how these designs are constructed and their advantages and limitations.

In general, dependent arm designs are the most popular method for evaluating non-experimental methods in field settings. It is an ad-hoc approach with minimum data requirements – it needs only a valid benchmark (often an RCT) and a non-experimental comparison sample with the same outcome.

Conceptual framework. We begin by formalizing the simultaneous design, where treatment units in the RCT benchmark are shared "simultaneously" with the non-experimental design. For now, we continue to assume that the benchmark is an RCT, and the non-experiment is an observational study in which RCT treatment units are matched to comparisons that did not select into the experimental sample. Using the same potential outcomes framework introduced above, we show that the average treatment effect for the treated in the RCT, $ATT(W = 0)$, is the causal estimand of interest in the dependent WSC design. This is in contrast to the independent WSC design, where the causal estimand of interest is the ATE of the overall WSC study population. This is because without random assignment of units into WSC conditions, the benchmark ATE cannot be recovered in the non-experimental arm, so the causal quantity of interest is the average treatment on treated in the RCT and non-experiment.

The goal of the WSC is to assess whether the non-experimental method replicates the ATT effect from the WSC benchmark, such that ATT(W = 0) = ATT(W = 1). We begin by defining the ATT(W = 0) in a simultaneous design. This is the expected difference of treated units' potential treatment and control outcomes in the RCT benchmark, where:

$$\begin{aligned} ATT(0) &= E(Y_i(1,0) - Y_i(0,0) \mid T_i = 1, W_i = 0) \\ &= E(Y_i(1,0) \mid T_i = 1, W_i = 0) - E(Y_i(0,0) \mid T_i = 1, W_i = 0) \end{aligned}$$

Here, the sub-population of interest is treated units in the benchmark arm, so the expectations are taken conditional on $T_i = 1$ and $W_i = 0$.

In the non-experimental arm, the causal estimand of interest is also the average treatment effect of treated (ATT) units. The ATT(W = 1) is defined by comparing potential treatment outcome, for treated units in the benchmark arm, $Y_i(1,0)$, with their potential control outcomes from the non-experimental arm, $Y_i(0,1)$, that is, the potential control outcomes for benchmark treated units if they would have been in the non-experimental arm of the WSC. Thus, the ATT for the observational arm is:

$$ATT(1) = E(Y_i(1,0) \mid T_i = 1, W_i = 0) - E(Y_i(0,1) \mid T_i = 1, W_i = 0)$$

where again, the expectations in the potential outcomes are conditional on $T_i = 1$ and $W_i = 0$, that is, the treated cases in the RCT.

Because ATT(0) and ATT(1) differ only with respect to the expectations of the potential control outcomes, but are identical on the expected potential treatment outcomes, we can show that the two ATT estimands are equivalent, ATT(0) = ATT(1), as long as the following assumption is met:

Assumption B1 (SUTVA: No Interference): A unit's potential outcomes depend only on its own WSC and treatment status W_i and T_i, but do not depend on other units' statuses.

Assumption B2 (STUVA: No Hidden Variations in WSC Conditions and Control Conditions): The WSC and Control statuses, W and T (for $T = 0$ only), are uniquely defined, that is, there are no different versions of each WSC and control condition. Thus, the potential outcomes only depend on the uniquely defined and administered WSC and control conditions, but no other third variables.

Note that SUTVA is essentially the same as for the independent WSC design but it is slightly weaker here because treatment units are held constant in the WSC arms, so the assumption pertains only to benchmark control and non-experimental comparison conditions but not to the treatment condition (which is the same for both WSC arms). When these conditions are met, then $Y_i(0,0) = Y_i(0,1)$.

Identification. To show that the expectations of the observed outcomes may be used to identify the ATT for the benchmark and non-experimental conditions, additional assumptions are required. First, to show that the benchmark yields a valid result for evaluating the non-experimental benchmark requires:

Assumption B3. In the benchmark, potential outcomes are independent of treatment (T). As in the independent WSC design, this assumption often is met through random assignment of units into treatment and control conditions. However, it may be met through other quasi-experimental approaches that produce credible treatment effects.

When Assumptions B1, B2 and B3 are met, we can show the difference in observable expectations identifies ATT of the benchmark condition, such that:

$$\begin{aligned}\tau_T(0) &= E(Y_i \mid T_i = 1, W_i = 0) - E(Y_i \mid T_i = 0, W_i = 0) \\ &= E(Y_i(1,0) \mid T_i = 1, W_i = 0) - E(Y_i(0,0) \mid T_i = 0, W_i = 0) = ATT(0)\end{aligned}.$$

Here, the logic is similar as to the independent design (see Appendix A). Note that in the case where randomization is perfectly implemented and there is no treatment non-compliance in the benchmark, then the ATE(W = 0) is equivalent to the ATT(W = 0), $\tau(0) = \tau_T(0)$.

For the non-experiment to produce an ATT that is equivalent to benchmark ATT, we require:

*Assumption B4: Units' potential outcomes are independent of their selection into WSC study conditions (W), conditional on a set of covariates **X** (strong ignorability)*: $(Y_i(0,0), Y_i(1,0), Y_i(0,1)) \perp W_i \mid X_i = x$, and $0 < P(W_i \mid X_i = x) < 1$. This assumption implies that all factors related to units' selection into WSC study conditions and the outcomes are observed and measured reliably by the researcher. Note that this is different to the independent WSC design, where we required ignorability with respect to treatment selection for the non-experimental arm. Here, ignorability with respect to selection into the experimental or non-experimental WSC arm is needed.

When Assumption 4 is met, then:

$$\begin{aligned}\tau_T(1) &= E_X\{E(Y_i \mid T_i = 1, W_i = 0, X_i = x)\} - E_X\{E(Y_i \mid T_i = 0, W_i = 1, X_i = x)\} \\ &= E(Y_i \mid T_i = 1, W_i = 0) - E(Y_i \mid T_i = 0, W_i = 1) = ATT(1)\end{aligned}.$$

By the same reasoning as above, $\tau_T(1) = ATT(1)$ when Assumptions B1, B2, B3, and B4 hold (see Appendix A). The difference here is that we take the expectation over the treated units' distribution of the covariates **X**.

Evaluating non-experimental methods in field settings. In a simultaneous design, non-experimental bias is given as the difference in the ATT estimators for the non-experiment and RCT benchmark, such that: $\beta_T = E(\hat{\tau}_T(1)) = E(\hat{\tau}_T(0))$. Because the treatment group is redundant in the simultaneous design, bias may also be calculated as the expected difference in outcomes for the non-experimental comparison and RCT control group, $\beta_T = E(\bar{Y}(1)) = E(\bar{Y}(0))$, where the mean $\bar{Y}(1)$ represents the adjusted mean of the non-experimental comparison group and $\bar{Y}(0)$ the mean of the experimental control group. Depending on the type of the non-experiment, adjustments for the non-experimental comparison group mean can be done via matching or weighting, for instance.

For the difference in ATTs in the non-experiment and benchmark to be interpreted as non-experimental bias requires that there are no alternative explanations for β_T except for violations to Assumption B4 or a biased estimator for the non-experimental effect or comparison group mean. Although in theory, SUTVA is a weaker assumption for the simultaneous design than it is for the independent design (because the assumption only applies to the benchmark control and non-experimental comparison, given that treatment units are held constant in both design); in practice, this requirement is often violated in field settings. For example, when benchmark controls and non-experimental comparisons are drawn from different study samples, there are often systematic differences in the method and timing of how outcomes were measured, in selection criteria for participation in study conditions, or in alternative treatment options available to benchmark controls and non-experimental comparisons. Moreover, if there are spillover effects in the RCT, it renders the interpretation of WSC results as ambiguous. As such, WSC analysts employing simultaneous designs should consider carefully the validity of the benchmark estimates and examine any irrelevant differences between benchmark control and non-experimental comparison conditions that may confound results.

Example. When LaLonde (1986) introduced the WSC design in 1986, he used RCT data from the National Supported Work Demonstration (NSW) evaluation to provide the experimental benchmark, and non-experimental data from the Panel Study of Income Dynamics (PSID) and the Current Population Surveys (CPS). In the NSW evaluation, job-training participants were randomly assigned to treatment or control conditions from April 1975 to August 1977 in each of 10 sites[6] from across the United States. Eligible applicants included AFDC women, ex-drug addicts and criminal offenders, as well as high school dropouts. The NSW intervention guaranteed treatment group members jobs for 9 to 18 months, depending on the target group and site. The outcome of interest for the WSC was participants' earnings four years after random assignment. For the non-experimental arm, LaLonde used data from large, stratified random samples in the PSID and CPS. He drew comparisons from the full survey samples, as well as subsets of samples that met similar – but not the same – eligibility requirements as those participating in the NSW evaluation. These subsamples included men and women who had received AFDC payments or were not employed prior to random assignment but had not participated in the NSW evaluation.

To construct the non-experimental arm, LaLonde began by deleting information from the RCT control group. He then used RCT treatment cases from the NSW and comparison cases from extant datasets, including the PSID and CPS. Overall, LaLonde tested non-experimental cross-sectional methods, panel data methods, and Heckman "two-step" selection approaches for addressing confounder effects. He found that models using longitudinal data performed better than models that relied on cross-sectional data, and that the two-step method did not perform worse than the one-step approaches, and in some cases did better. Still, specification tests using pre-training earnings data failed to flag better performing non-experimental methods and samples, leading LaLonde to conclude, "that many of the econometric procedures and comparison groups used to evaluate employment and training programs would not have yielded accurate or precise estimates of the impact of the National Supported Work Program…" (pg. 617). Since LaLonde's original study, the data have been reanalyzed at four times, including by Heckman and Hotz (1989), Dehijia and Wahba (2002), Smith and Todd (2005), and Diamond and Sekhon (2012), with each set of authors arriving at different conclusions based on the non-experimental method tested, and the comparisons employed.

The LaLonde's (1986) study demonstrates the multiple advantages of the simultaneous design. A key strength is that it is an ad hoc approach with minimal data-requirements. It only needs RCT data, and non-experimental cases that did not receive treatment but shared similar outcome measures. Here, LaLonde was able to take advantage of public, nationally representative datasets that are accessible to most researchers. Another benefit of the approach is that it is not based on a

[6] Sites included Atlanta, Chicago, Hartford, Jersey City, Newark, New York, Oakland, Philadelphia, San Francisco, and Wisconsin.

treatment contrast in a laboratory-like setting, but on units' real world selection process into the benchmark design. Assuming that units entered into the experimental sample with a desire to participate in the treatment, it's very well plausible that the selection mechanism captured in this WSC mirrors the process by which units choose to enter a job training program. Furthermore, although the benchmark RCT sample sizes for the LaLonde example were comparable to those obtained in Shadish et al. (2008), the PSID and CPS provided much larger pools for drawing comparison cases. In addition, it was unlikely that units reacted to their WSC status, though it is possible that comparison units may have reacted negatively if they were aware of the NSW evaluation, and resented that they could not participate in the RCT. Finally, the WSC analyst may use RCT data from multiple sites within an evaluation, helping improve the generalization of results that may be less feasible in a prospectively planned independent arm design.

Despite the multiple strengths of this approach, there are also challenges with the simultaneous design. The most common concern is differences in study conditions between benchmark controls and non-experimental comparisons. For example, as Smith and Todd (2007) point out, outcome measures were comprised of different survey instruments and administrative sources, and were taken at different times. And, because the comparison sample was drawn from a national survey, non-experimental comparisons were likely not in similar labor markets, experiencing the same conditions as RCT control units. These differences between RCT controls and comparisons potentially challenge the interpretation of WSC results.

Variants of the Dependent Arm Design

Multisite simultaneous design. A limitation of the simultaneous design for testing non-experimental methods in field settings is the multiple threats to validity that challenge the interpretation of results. Of paramount concerns are violations to B2, which states that the control potential outcomes in the benchmark and non-experimental arms must be the same and that there can be no other differences between control and comparison conditions besides random assignment to treatment conditions in the benchmark. Given that the simultaneous design is an ad hoc approach, it is rare that the WSC analyst can ensure that all study conditions were exactly the same in the benchmark control and non-experimental comparison conditions.

A variation of the simultaneous design is the multi-site design (Panel 1, Figure 2), which uses RCT data to construct the benchmark and non-experimental arms of the WSC. In this approach, the WSC analyst begins with a multi-site RCT trial in which randomization occurs within-sites, and constructs the non-experiment by comparing average outcomes from RCT treatment cases in target sites (*s_t*), and RCT control cases from other sites in the study (*s_{-t}*). Here, selection in the non-experimental arm occurs because units sort differentially into sites within the same experiment (i.e., the composition of units differs from site to site due to the selection of sites with fixed units or due to units selection into fixed sites). The WSC analyst selects the targeted sites of interest within the benchmark. These sites may be selected because the RCT was implemented well enough to serve as the causal benchmark, because there were large sample sizes within the site, or because the sites are of substantive interest to the researcher.

For the multi-site WSC, the causal quantity of interest is the average treatment on treated effect for target site s_t in the benchmark ($ATT_S(W_i = 0)$), such that:

$$\begin{aligned} ATT_S(0) &= E(Y_i(1,0) - Y_i(0,0) \mid T_i = 1, W_i = 0, S_i = s_t) \\ &= E(Y_i(1,0) \mid T_i = 1, W_i = 0, S_i = s_t) - E(Y_i(0,0) \mid T_i = 1, W_i = 0, S_i = s_t) \end{aligned}$$

For the non-experimental arm, the average treatment on treated effect for site s is ($ATT_S(W_i = 1)$):

$$ATT_S(1) = E(Y_i(1,0) \mid T_i = 1, W_i = 0, S_i = s_t) - E(Y_i(0,1) \mid T_i = 1, W_i = 0, S_i = s_t)\,.$$

Equivalent to the simultaneous design, the causal estimand of interest is the average treatment on the treated, but in the multi-site design, it is the ATT for treated units in targeted sites within the RCT. Thus, $ATT_S(0) = ATT_S(1)$ when there is no interference (B1) between units in the RCT control and non-experimental comparison conditions, and when there is no hidden variation in WSC and control conditions (B2). Moreover, we use the same logic as in the simultaneous design to identify ATT_S of the benchmark using the expectations of observed outcomes when assumptions B1, B2, and B3 are met, such that: $\tau_{TS}(0) = ATT_S(0)$. For the non-experiment, ATT_S is identified when B1, B2, B3, and B4 are met, such that: $\tau_{TS}(1) = ATT_S(1)$.

When one of the above assumptions is violated or one of the ATT estimators is biased, then the ATT estimators for the benchmark site and non-experiment are not equivalent: $\beta_{TS} = E(\hat{\tau}_{TS}(1)) - E(\hat{\tau}_{TS}(0)) \neq 0$. β_{TS} is interpreted as non-experimental bias only when the non-experimental estimator is biased or assumption B4 (potential outcomes are independent of their selection into sites, conditional on a set of covariates) is violated, but all other assumptions SUTVA (B1 and B2), and independence in the benchmark (B3) hold.

In cases where there are multiple target sites in the benchmark one can estimate a separate β_{TS} for each site, and then examine pooled results across sites. There are multiple ways to pool results – one approach is to simply take the average across all sites (Wilde & Hollister, 2007); another is to take a weighted average by the number of treatment cases in each target site (Black, Galdo, & Smith, 2007); a third approach is to meta-analyzed differences across sites in a meta-analytic framework, weighting observed bias estimates by their inverse variances (Gleason et al., 2012).

Example. Wilde and Hollister (2007) used RCT data from the Tennessee's Student Teachers Achievement Ratio Project (Project STAR) (Word et al., 1990), which examined the effects of class size on student reading and math achievement outcomes. Because random assignment of students within classrooms occurred within schools, the authors used RCT results from 11 of the 79 schools in the original Project Star experiment as the causal benchmark for the WSC. To construct the comparison, they matched 367 experimental treatment students in the 11 WSC benchmark schools to a potential pool of 4,589 control group students in the 68 *other* Project Star schools. To estimate observational treatment effects, the authors used nearest neighbor PSM with regression-adjustment. They compared ATT estimates for each of the 11 target schools, with bias estimates ranging from -2.58 sds to 1.85 sds, where in 6 of the 11 cases, the difference between experimental and observational results were statistically significant. They also reported average bias across the 11 sites, which was .17 sds. These results led Wilde and Hollister to conclude that the observational results did not succeed in replicating benchmark results.

The Wilde and Hollister (2008) example demonstrate a strong advantage of multi-site designs. Here, validity threats about irrelevant differences between the benchmark and non-experimental arms are reduced because benchmark control and non-experimental comparisons are involved in the same study design with the same measures, seligibility requirements, and conditions. Moreover, the original experiment may have documented deviations in treatment conditions for control groups across sites, allowing WSC analysts to assess the plausibility of validity threats that often are not possible when extant datasets are used.

However, there are concerns with the approach as well. One issue is that the selection process in the non-experimental study is units' selection into different sites within the RCT. But in practice, the selection process of interest to the WSC researcher involves units' preference for treatment conditions, so multi-site designs may not be as helpful for identifying whether non-experimental estimators overcome units' selection into treatments. Another issue is that the sample size for the comparison pool is limited to the number of available control cases within the RCT itself, limiting the power of the non-experiment for detecting effects, as well as for assessing correspondence in benchmark and non-experimental results (see Steiner & Wong (under review) for further discussion of power and correspondence measures in WSC designs). Taken together, although the multi-site design provides strong advantages for improving the internal validity of the

WSC design, there are challenges related to generalization of the selection process under investigating in the non-experiment, as well as threats to statistical conclusion validity due to limitations in the non-experimental sample size.

Synthetic Designs. In synthetic WSC designs, the researcher begins with data from an RCT, and constructs a non-experiment by simulating a selection process in which she deletes information from the RCT treatment and/or control group to create non-equivalent groups (Panel 2, Figure 2). An example of this method, which we discuss in more detail below, uses an RCT dataset to evaluate the performance of regression-discontinuity approaches by creating an RD design synthetically. Here, the researcher designates a variable in the RCT dataset to be the RD assignment variable –a pretest, for example – and establishes an artificial cutoff along the distribution of the pretest assignment variable. To create the RD, the researcher deletes RCT treatment cases above (or below) the arbitrary RD cutoff, and RCT control cases below (or above) the cutoff. The researcher uses parametric, semi-, or non-parametric approaches to estimate treatment effects at the cutoff for the RD sample. To assess the performance of the RD, the researcher compares the average treatment on treated effect at the RD threshold for experimental cases $ATT_c(W = 0)$, with the average treatment on treated effect at the cutoff for the RD sample $ATT_c(W = 1)$.

Synthetic designs are similar to simultaneous and multi-site designs in that some portion of the treatment and comparison cases are shared between the benchmark and non-experimental arms. Moreover, because data for the benchmark and non-experimental conditions in the WSC are from the same RCT, the design has many of the same advantages as multi-site designs. However, a key difference between multi-site and synthetic designs is that in multi-site designs, the non-experiment is based on units' selection into different RCT sites, but in synthetic designs, the non-experiment is based on a selection process defined by the researcher. As such, the synthetic design is akin to a simulation study, in which the researcher uses real world data for creating non-experimental conditions to test method performance. As we will see, this type of approach is well-suited for addressing some types of questions, such as the performance of analytic methods, but may be less helpful for examining whether research design assumptions often are met in real world settings.

For the synthetic design, the treatment effect of interest in the benchmark and non-experiment is the average treatment on treated effect for some target population of interest with $Z = z_t$, as determined by the WSC analyst. Thus, for the benchmark, the causal quantity of interest is:

$$\begin{aligned} ATT_Z(0) &= E(Y_i(1,0) - Y_i(0,0) \mid T_i = 1, W_i = 0, Z = z_t) \\ &= E(Y_i(1,0) \mid T_i = 1, W_i = 0, Z = z_t) - E(Y_i(0,0) \mid T_i = 1, W_i = 0, Z = z_t) \end{aligned}.$$

This is similar to the ATT for the multi-site design except that Z represents an indicator variable or a vector of variables that is used to define a synthetic subpopulation of the benchmark study. For the non-experimental arm, we are also interested in the average treatment on treated effect at z_t, such that:

$$ATT_Z(1) = E(Y_i(1,0) \mid T_i = 1, W_i = 0, Z_i = z_t) - E(Y_i(0,1) \mid T_i = 1, W_i = 0, Z_i = z_t).$$

As in the simultaneous and multi-site designs described above, $\text{ATT}_z(0) = \text{ATT}_z(1)$ when SUTVA (B1 and B2) is met. However, since the synthetic design does not have real-world selection into WSC arms (data of the benchmark arm are synthetically assigned to a non-experimental arm), assumptions B1 and B2 are automatically met with respect to the WSC status W. If a violation of SUTVA with respect to the treatment status T in the benchmark arm transfers to the same extent into the synthetic non-experiment, SUTVA for T is not required either. However, a systematic selection of non-experimental units from the benchmark arm very likely induces a differential violation of SUTVA. Thus, SUTVA should ideally be met in the benchmark arm (then it is automatically met for the synthetic non-experiment).

To identify $\text{ATT}_z(0)$ in the benchmark, B1, B2, and B3 are also required, then $\tau_{TZ}(0) = \text{ATT}_Z(\text{W=0})$. Here, because of random assignment into treatment conditions in the initial RCT, the distribution of the baseline assignment variable(s) (Z) is independent of units' treatment status, creating equivalent treatment and control groups. To identify $\text{ATT}_z(1)$ in the non-experiment, B1, B2, B3, and B4 are also needed. Given that the synthetic nonexperimental group is selected on the basis of observed covariates from the benchmark arm, B4 is automatically met unless the synthetic selection is very extreme such that units with specific covariate values have a treatment probability of zero or one. When these assumptions are met, then $\tau_{TZ}(1) = \text{ATT}_Z(\text{W=1})$. In cases where one of above assumptions is violated or one of the ATT estimators is biased, then $E(\hat{\tau}_{TZ}(1)) \neq E(\hat{\tau}_{TZ}(0))$. Bias is defined by comparing the difference in the ATT estimates in the benchmark and non-experimental arms, such that $\beta_{TZ} = E(\hat{\tau}_{TZ}(1)) - E(\hat{\tau}_{TZ}(0))$. Again, this result is interpretable as non-experimental bias only when B1 through B3 are met and the experimental estimator is unbiased. Since B4 is regularly met in a synthetic design, the non-experimental bias only assesses the bias in the non-experimental estimator $\hat{\tau}_{TZ}(1)$, that is, whether the functional form of an outcome or propensity score model has been correctly specified, or whether for a consistent estimator the sample was not large enough to obtain approximate unbiasedness.

Example. Gleason et al. (2013) conducted a synthetic WSC examining the performance of the regression-discontinuity design using two experimental datasets – one that used RCT data from an evaluation of education technology products (Ed Tech) in classrooms; another that looked at experimental evidence from an evaluation of Teach for America. We focus on the Ed Tech example here because the implementation of the WSC designs was similar in both cases. In the Ed Tech evaluation, teachers who were randomized to the treatment condition (*N* = 238) were asked to incorporate a suite of education technology products in their classrooms, with the goal of improving students' math achievement scores. Teachers who were randomized to the control condition (*N* =190) were allowed to use technology and computers in their classroom generally, but were not allowed to use specific study products under investigation. To construct the regression-discontinuity design, the WSC analysts designated students' pretest score as the RD assignment variable and the median as the RD threshold value. Here, students with pretest scores above the cutoff were designated as RD treatment students, and students with pretest scores below the cutoff were RD comparison students. Next, to create the RD sample, the researchers discarded experimental treatment students with pretest scores below the median, and experimental control students with pretest scores above the median. The WSC analysts then used parametric and non-parametric approaches for estimating treatment effects at the cutoff for the RD sample. To analyze the experiment, the WSC researchers used the complete RCT sample, as well as parametric and non-parametric approaches, to estimate the conditional mean differences in outcomes at the RD cutoff location. Bias was assessed by comparing treatment effects at the cutoff in the RCT to those obtained from the RD design at the RD cutoff. Overall, they found that the RD results were not statistically different from RCT results, with the magnitude ranging from -0.09 to +0.10 standard deviations. Because of the synthetic design approach, the authors were able to replicate the WSC study design using the discarded data (treatment students below the cutoff, and control students above the cutoff) from the original RD.

Synthetic designs have the benefit of being causally interpretable, cost-effective, and replicable. As the Gleason et al. example demonstrates, in synthetically constructed RE-RD WSCs, researchers can probe the robustness of their results by reanalyzing the RD using different threshold cutoffs, assignment variables, and outcomes (Gleason et al., 2012; Wing & Cook, in press). In fact, it is possible to construct WSCs using survey data without an intervention at all. In this case, there is no need to estimate an experimental benchmark (the true effect is zero), and the RD could be constructed synthetically with different outcomes, assignment variables and thresholds. In RE-RD synthetic designs, researchers are able to compare the same causal estimand in the experiment and non-experiment, which is often challenging in simultaneous RE-RD WSCs designs, which often compare RD treatment effects at a cutoff, with experimental evidence for a sub-population of students slightly above or below the cutoff (Buddelmeyer & Skouffias, 2005). Finally, because the non-experiment is constructed using the same RCT data, most potential confounders are held constant between the two study conditions. The only difference between the two conditions is the mode of treatment assignment.

However, synthetic designs cannot address many research questions that are of interest in WSC studies. Because the WSC researcher constructs the non-experiment, she is only able to examine selection processes that are created artificially and may simplify the complexities of individuals' real-life choices into treatments. The solution here may be to include a mechanism in the research protocol where an independent research team constructs the non-experiment using a selection model with rich covariate information and complex interaction and higher order terms that is not easily modeled. Still, given the artificial conditions of this approach, the synthetic design has more in common with simulation studies than other WSC approaches, such as the independent arm design. Moreover, the design cannot be used to investigate implementation challenges that often occur with research designs in field settings, such as potential sorting behaviors by units in an RD design. Finally, similar to multi-site designs, statistical power in synthetic WSC designs are often a concern in the non-experiment. This is because most non-experiments are constructed by deleting some portion of cases from the experimental data. Given limited resources, most RCTs are designed to detect minimum effects that are scientifically or policy relevant, suggesting that the non-experiment in a WSC will often be underpowered for detecting effects.

Discussion: Comparison of WSC Approaches

This paper describes WSC approaches for evaluating non-experimental methods, highlighting the stringent assumptions required for each approach to provide a fair test of methods. In considering the multiple WSC designs options, one might consider these designs on a continuum where prospectively planned, independent arm designs are on one side, synthetic designs are on the other side (Table 1). The continuum describes the advantages and tradeoffs of each design from the perspective of feasibility, internal validity, statistical conclusion validity, and relevance in terms of selection process under investigation.

As we have shown, the main benefit of independent designs is that because the approach is prospectively planned, the researcher has control over many aspects of the design that may otherwise confound estimates of non-experimental bias. The researcher can ensure independence between WSC conditions and units' potential outcomes through random assignment of units into the benchmark and non-experimental conditions. She may establish protocols to ensure comparability of conditions across WSC arms, and to check that randomization across WSC conditions and within the benchmark was perfectly implemented. In some cases, the researcher may be able to construct a non-experimental study that mimics a real world selection process of interest. For example, units may be asked to select whether or not to receive text message reminders to exercise, to participate in math or reading training, or to enroll in an afterschool tutoring program. Thus, as long as SUTVA (A1 and A2), and independence in the WSC and benchmark conditions are met (A3 and A4), then the approach provides a credible estimate of non-experimental bias.

However, as demonstrated in the Shadish et al. study, the prospective nature of the design raises implementation challenges in field settings. The approach has high data requirements given that the researcher must collect information prospectively on benchmark and non-experimental cases. And, when the design is implemented in laboratory-like conditions, the generalization of results also may be limited. Here, treatment contrasts are constrained to include only options that may be implemented in an RCT and non-experimental setting simultaneously. This may necessitate examining selection processes that are relatively straight-forward to observe and model, and do not generalize well to selection challenges that researchers encounter in real world settings. Finally, the independent arm design often has weak statistical power for detecting differences in effects between the benchmark and non-experiment. This is due to the variation from both treatment and control groups in the benchmark and non-experimental arms. Perhaps one way to address these concerns is to consider contexts in which doubly randomized designs may be implemented feasibly and provide useful information to researchers for cases when RCTs are not feasible. For example, the design may be incorporated into evaluations of low-cost behavioral interventions that assign electronic reminders, such as a text message or email, to participants.

Dependent arm designs provide strong comparative advantages over independent arm approaches. The method is ad hoc, sometimes requiring data only from a well-implemented RCT to serve as a credible benchmark. Simultaneous designs also require observational data that share the same outcome measure as the benchmark, and units not exposed to treatment. However, if the outcome is a standardized measure such as annual state assessment scores, household water consumption, or voting record, then it is often possible for researchers to obtain this data from administrative sources or survey information. Moreover, simultaneous designs often have most power for detecting effects in the experimental and non-experimental designs, as well as in the difference in estimates between the benchmark and non-experiment. This is because of the reduced variation in the outcome due to the shared treatment group between both arms of the design, as well as the often relatively larger sample sizes in the benchmark and non-experiment (compared to other WSC approaches). Finally, the chief advantage of the simultaneous design is that the non-experiment is based on a natural – and possibly complex – selection process that is plausibly related to units' decisions to uptake treatment – that is, their decision to participate in the RCT. These advantages – feasibility, improved statistical power, and relevance to real world selection processes – make simultaneous designs an appealing approach for many WSC analysts.

In theory, the dependent arm design has weaker assumptions than the independent design. This is because SUTVA (B1 and B2) applies only to units that are not shared between the benchmark and the non-experiment. As such, the WSC analyst needs only to ensure that all measurement and study conditions are the same between benchmark control and non-experimental comparisons. In practice, however, this assumption is often not met in field settings. Because simultaneous designs are ad hoc approaches that take advantage of data that were collected for different study purposes, the WSC researcher often lacks control over study characteristics that vary between the benchmark and non-experimental conditions. For example, if earnings outcomes in the benchmark and non-experiment were measured on different scales at different times, if treatment spillover occurred in the RCT control, or if the non-experimental comparisons were provided with alternatives to the treatment (e.g. job fairs at community colleges available to non-experimental comparisons), then the interpretation of the WSC result is challenged. Is the difference in non-experimental and benchmark results due to bias, or because of other WSC design violations? Is this difference due to non-experimental bias, or because another violation in the WSC design was violated?

To address these concerns, WSC researchers have adopted multi-site approaches. Because non-experimental comparisons are drawn from control units within the same RCT, many irrelevant study differences that plague the simultaneous design are ameliorated in the multi-site context. Moreover, the WSC researcher may have study information that documents deviations from the RCT protocol, as well as describes treatment and control conditions across sites. This allows the analyst to assess plausible validity threats in the WSC context. However, the advantages for internal validity of the WSC design may require strong tradeoffs in terms of statistical conclusion validity and relevance of the selection process under investigation. When these limitations are present, the WSC analyst should be cautious in her interpretation of WSC results as a test of non-experimental methods.

Finally, synthetic designs are easily implemented because it requires only data from a credible RCT benchmark. The approach also has strong advantages in terms of internal validity of WSC results – nearly all factors are held constant between the benchmark and non-experimental arms. However, because the researcher creates the selection process in the non-experiment, synthetic designs are more similar to Monte Carlo experiments. Our view is that synthetic designs are useful for studying methods in field settings *where the selection process is completely known to the researcher* – as in the case of the regression-discontinuity design. The design provides less compelling evidence about method performance when the underlying data generating mechanism is unknown.

Because of space considerations, this paper was not able to address all implementation challenges that arise in conducting WSC evaluations. A common issue in many WSC designs occurs treatment non-compliance occurs in the benchmark. In an RCT design, this occurs when units are randomly assigned to treatment and fail to take up the intervention, or when units are randomly assigned to control receive the treatment anyway. When treatment non-compliance occurs in the benchmark, it is a violation of the independence assumption in the benchmark (A3 and B3 for independent and dependent arm designs, respectively). In the WSC context, we may modify these assumptions to allow for one-sided treatment non-compliance, where there no units in the control group received treatment, but some treatment units failed to show up for the intervention. When treatment non-compliance occurs for dependent and independent arm designs, the intent to treat effect is the causal estimand of interest in both the benchmark and in the non-experimental sample. Here, the goal of the non-experiment is to find valid counterfactual cases for units that were randomly assigned to treatment in the benchmark.

Taken together, WSC designs provide important information for mapping out of contexts and conditions under which non-experimental methods perform well. However, each design has comparative advantages and limitations in terms of their feasibility, validity, and relevance of results. This paper provides WSC analysts with an overview of methods to help choose designs that are well aligned with their research questions, as well as to address common threats that often challenge the interpretation of these designs.

References

Angrist, Autor, Hudson, & Pallais (2015). Evaluating Econometric Evaluations of Post-Secondary Aid. American Economic Review: Papers & Proceedings. 105(5): 502-507.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies often produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750.

Cook, T. D. & Wong, V. C. (2008). Empirical tests of the validity of the regression discontinuity design. *Annales d'Economie et de Statistique.*

Federal Register (2005). Scientifically Based Evaluation Methods. Federal Register. 70(15), 3586-3589.

Ferraro, P. & Miranda, J. (under review). Can Panel Designs and Estimators Substitute for Randomized Control Trials in the Evaluation of Social Programs?

Fortson, Kenneth, Natalya Verbitsky-Savitz, Emma Kopa, and Philip Gleason. "Using an Experimental Evaluation of Charter Schools to Test Whether Nonexperimental Comparison Group Methods Can Replicate Experimental Impact Estimates." NCEE Technical Methods Report 2012-4019. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2012.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources, 22*(2), 194-227.

Friedlander, D., & Robins, P. (1995). Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *American Economic Review, 85*(4), 923-937.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy, 589*, 63-93.

Gleason, Philip M., Alexandra M. Resch, and Jillian A. Berk. "Replicating Experimental Impact Estimates Using a Regression Discontinuity Approach." NCEE Reference Report 2012-4025. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2012.

Heckman, J. J., Ichimura, H., Smith, J. C., & Todd, P. (1998). Characterizing selection bias. *Econometrica, 66*(5), 1017-1098.

LaLonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. *The American Economic Review, 76*(4), 604-620.

Marcus, S.M., Stuart, E.A., Wang, P., Shadish, W.R., and Steiner, P.M. (2012). Estimating the causal effect of randomization versus treatment preference in a Doubly-Randomized Preference Trial. *Psychological Methods* 17(2): 244-254.

McConeghy, Steiner, Wing, & Wong. (2013). Evaluating the Performance of Interrupted Time Series Approaches in Replicating Experimental Benchmark Results. Presentation at Association for Public Policy Analysis and Management. Washington, DC.

Office of Management and Budget, What Constitutes Strong Evidence of a Program's Effectiveness? (Washington, D.C.: September 2005) is at [hyperlink, http://www.whitehouse.gov/sites/default/files/omb/part/2004_program_eval.pdf]

Peck, Bell, & Werner. (2013). Learning 'What Works' from Multi-Site Experiments by Combining Natural Site-Level Variation with Randomized Individual-Level Variation in Program Features. Paper presented at the Association for Public Policy Analysis and Management.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased Causal Inference From an Observational Study: Results of a Within-Study Comparison. *Educational Evaluation and Policy Analysis, 31*(4), 463-479.

Shadish, W.R., Galindo, R., Wong, V.C., Steiner, P.M., & Cook, T.D. (2011). A randomized experiment comparing random to cutoff-based assignment. *Psychological Methods*, 16(2), 179-191.

Shadish, W., P. Steiner, and T. D. Cook. 2012. A Case Study about Why It Can Be Difficult to Test Whether Propensity Score Analysis Works in Field Experiments. *Journal of Methods and Measurement in the Social Sciences* 3(2): 1–12.

Smith, J. C., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*, 305-353.

Steiner, P.M., Cook, T.D., Shadish, W.R., & Clark, M.H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15(3), 250-267.

Somers, M., Zhu, P., Jacob, R., & Bloom, H., (2013) The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation. *MDRC working paper in research methodology*. New York, NY.

St. Clair, Cook, & Hallberg (in press). Examining the Internal Validity and Statistical Precision of the Comparative Interrupted Time Series Design by Comparison with a Randomized Experiment

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Testing nonexperimental estimates of impact against experimental estimates of impact with education test scores as outcomes. *Journal of Policy Analysis and Management, 26*(3), 455-477.

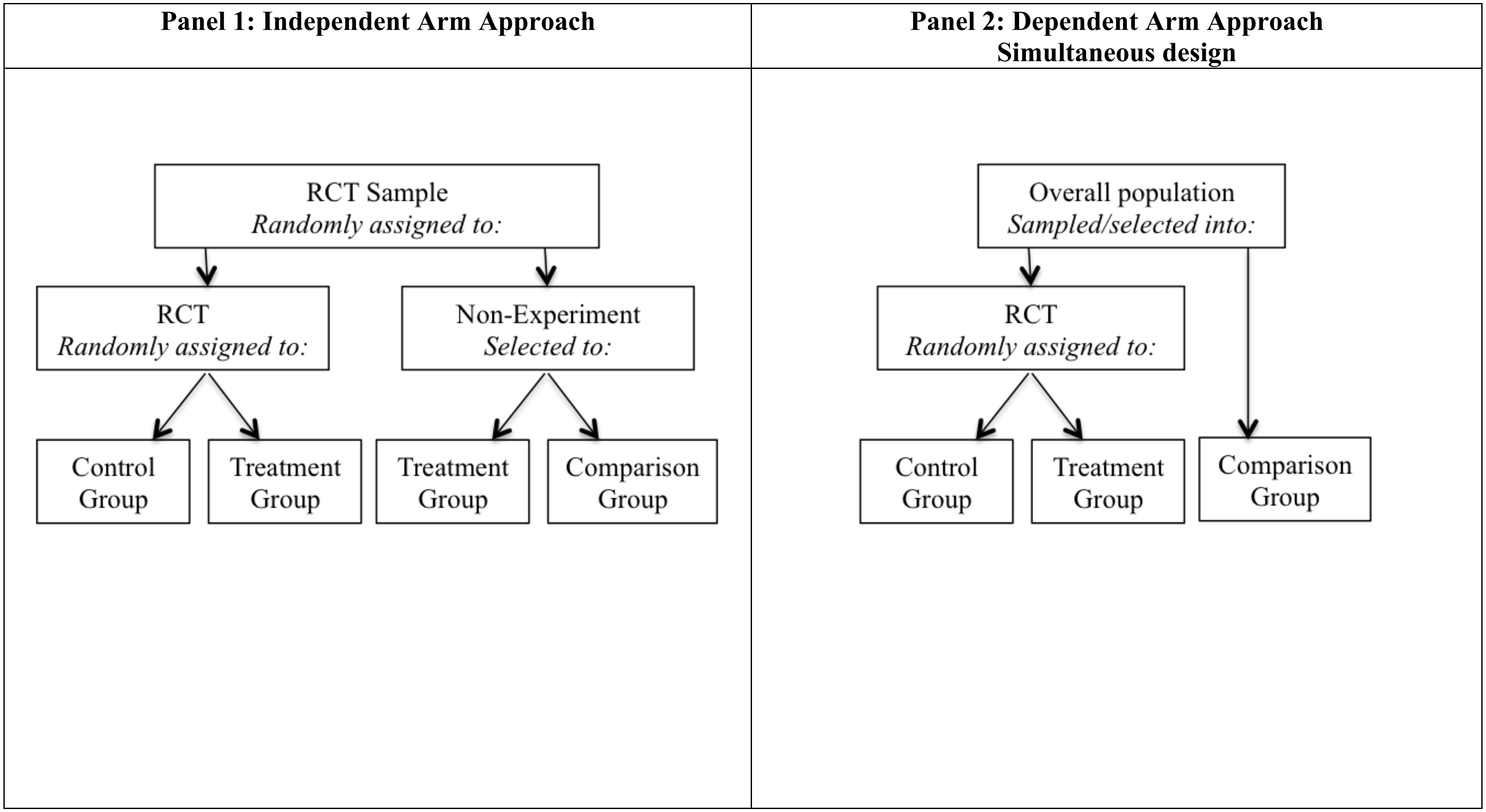
Wing, C. & Cook, T.D. (in press). Strengthening The Regression Discontinuity Design Using Additional Design Elements: A Within-Study Comparison. *Journal for Policy Analysis and Management.*

Table 1: Comparison of WSC Design Approaches

Prospective designs ***Ad hoc designs***

	Independent Arm Designs	Dependent Arm Designs		
		Simultaneous Designs	Multi-site Designs	Synthetic Designs
Selection process examined	Units select into treatment conditions	Units select into benchmark condition	Units select into benchmark sites, researcher selects sites	Researcher selects sub-populations from benchmark condition
Causal estimand of interest	ATE of WSC study population	ATT of benchmark sample	ATT of targeted sites	ATT of researcher targeted subpopulation
Advantages	- Mimics real world selection mechanisms - Third variable confounders less likely to be a problem in lab setting	- *May* mimic real world selection mechanisms - Less intensive data requirements - Larger sample sizes	- Units' real world selection into blocks - Third variable confounders less likely to be a problem - Less intensive data requirements	- Minimal data requirements - Third variable confounders less likely to be a problem - Replicable
Disadvantages	- Prospectively planned - Intensive data requirements - Results may be less generalizable due to laboratory-like conditions - Weak power	- Possible confounders between benchmark and NE arms -Restricted generalizability to benchmark rather than treatment selection	- Selection process may not be of interest to researchers - Reduced sample sizes for detecting effects - Restricted generalizability due to site rather than treatment selection	- Rarely mimics real world selection process which restrict the generalizability of results - Difficult to test implementation issues in the design

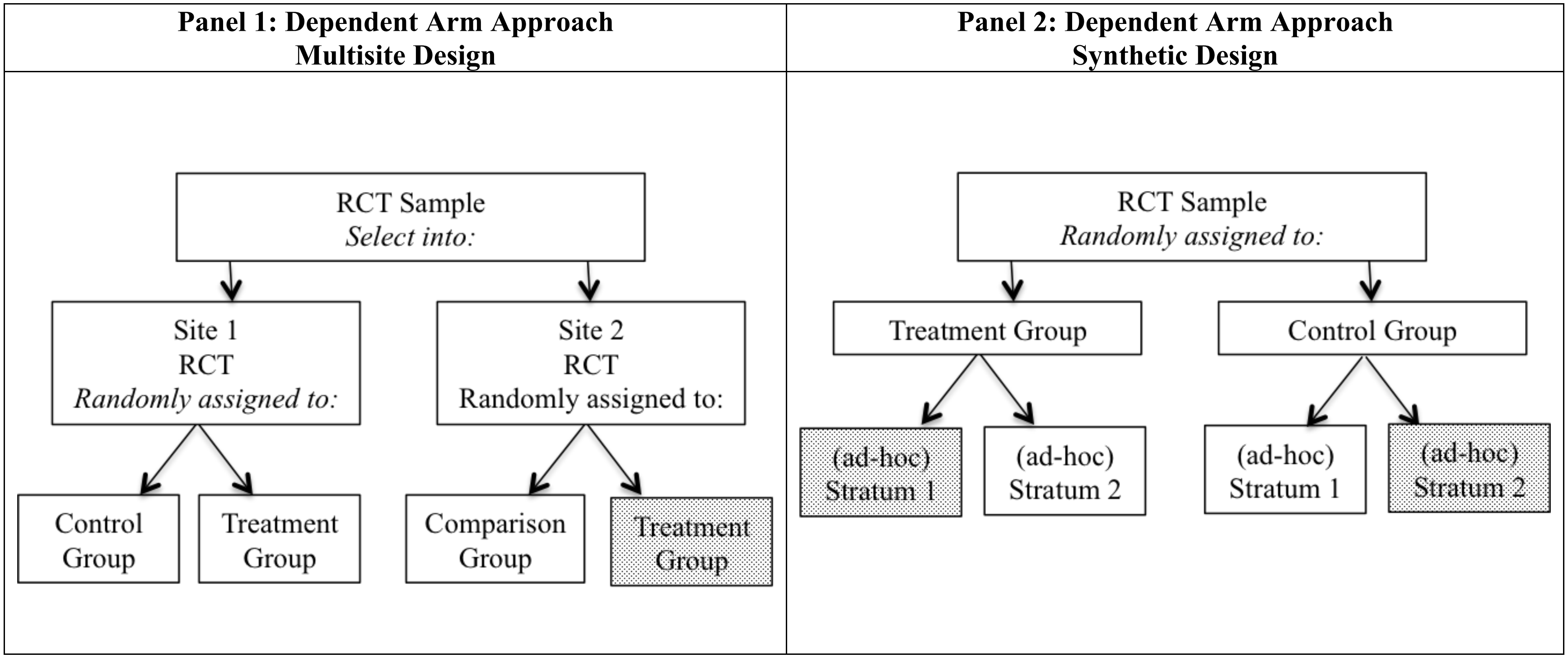
Figure 1: Independent versus Dependent Arm Within-Study Comparison Approaches



Panel 1: Independent Arm Approach

Panel 2: Dependent Arm Approach
Simultaneous design

Figure 2: Multisite and Synthetic Dependent Arm Within-Study Comparison Approaches



Panel 1: Dependent Arm Approach Multisite Design

Panel 2: Dependent Arm Approach Synthetic Design

Appendix A: Proof for Identification of ATT in Simultaneous Designs

Simultaneous Design

We show that the difference in observed expectations in the benchmark identifies ATT when the following assumptions are met: B1) SUTVA: no interference between units; B2) SUTVA: no hidden variation in treatment and WSC conditions; and B3) Independence: potential outcomes are independent of treatment status in the benchmark. That is, $\tau_T(0)$ = ATT(0) when:

$$\begin{aligned}\tau_T(0) &= E(Y_i \mid T_i = 1, W_i = 0) - E(Y_i \mid T_i = 0, W_i = 0)\\ &= E(Y_i(1,0) \mid T_i = 1, Wi = 0) - E(Y_i(0,0) \mid T_i = 0, W_i = 0)\\ &= E(Y_i(1,0) \mid T_i = 1, W_i = 0) - E(Y_i(0,0) \mid T_i = 1, W_i = 0) = ATT(0)\end{aligned}$$

Where, the equality of the second line follows from Assumptions B1 and B2 and Equation [1], and the third line's equality is met through independence in the benchmark assumption. We next show that $\tau_T(1)$ = ATT(1):

$$\begin{aligned}\tau_T(1) &= E_X\{E(Y_i \mid T_i = 1, W_i = 0, X_i = x)\} - E_X\{E(Y_i \mid T_i = 0, W_i = 1, X_i = x)\}\\ &= E_X\{E(Y_i(1,0) \mid T_i = 1, W_i = 0, X_i = x)\} - E_X\{E(Y_i(0,1) \mid T_i = 0, W_i = 1, X_i = x)\}\\ &= E_X\{E(Y_i(1,0) \mid T_i = 1, W_i = 0, X_i = x)\} - E_X\{E(Y_i(0,1) \mid T_i = 0, W_i = 0, X_i = x)\}\\ &= E_X\{E(Y_i(1,0) \mid T_i = 1, W_i = 0, X_i = x)\} - E_X\{E(Y_i(0,1) \mid T_i = 1, W_i = 0, X_i = x)\}\\ &= E(Y_i(1,0) \mid T_i = 1, W_i = 0) - E(Y_i(0,1) \mid T_i = 1, W_i = 0) = ATT(1)\end{aligned}$$

The first line formulates the ATT in terms of expectations of the observed outcome of the treated individuals in the RCT ($T_i = 1, W_i = 0$) and the observed outcome of the control individuals in the observational arm ($T_i = 1, W_i = 1$). The equality in the second line follows because of B1 and B2 (SUTVA), and Equation [1]. The third line, where the value of W switches from 1 to zero, follows from the independence assumption of units' selection into WSC arms (B4) – that is, conditional on **X**, the distribution of the potential control outcomes is the same for the control individuals in the observational arm and control individuals in the benchmark. The fourth line, where the value of T switches from 0 to 1, follows from independence in the benchmark arm (i.e., the distribution of the potential control outcomes is the same for the control and treatment individuals in the RCT (B3)).

Proofs for multisite and synthetic designs follow similar logic.