What Can Be Learned from Empirical Evaluations of Non-experimental Methods?

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

July, 2018

Citation:

Wong, V. C., Steiner, P. M., Anglin, K. L. (in press). What Can Be Learned from Empirical Evaluations of Non-experimental Methods? Evaluation Review.

[1] University of Virginia, Curry School of Education
[2] University of Wisconsin at Madison

This research was supported by a collaborative NSF grant #2015-0285-00. The opinions expressed are those of the authors and do not represent views of NSF.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B140026 to the Rectors and Visitors of the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Given the widespread use of non-experimental (NE) methods for assessing program impacts, there is a strong need to know whether NE approaches yield causally valid results in field settings. In within-study comparison (WSC) designs, the researcher compares treatment effects from an NE with those obtained from a randomized experiment that shares the same target population. The goal is to assess whether the stringent assumptions required for NE methods are likely to be met in practice. This essay provides an overview of recent efforts to empirically evaluate NE method performance in field settings. We discuss a brief history of the design, highlighting methodological innovations along the way. We also describe papers that are included in this two-volume special issue on WSC approaches, and suggest future areas for consideration in the design, implementation, and analysis of WSCs.

Introduction

Over the last fifty years, two advances have improved methodological rigor for making causal inferences. The first advance was acknowledging the primacy of research design, such as the randomized experiment or the regression-discontinuity design (RDD), over statistical adjustment procedures for establishing causal inference (Angrist & Pischke, 2008; Morgan & Winship, 2007; Shadish, Cook, & Campbell, 2002). The second advance was using potential outcomes to define causal quantities of interest and to formulate identification assumptions for various research designs (Rubin, 1974, 2005). Together, these developments have provided researchers with a formal understanding of the assumptions required for research designs to produce valid causal results. These two advances have also helped researchers develop empirical diagnostics to partially probe whether these assumptions are likely to be met.

However, it is rarely possible for a researcher to test whether the stringent assumptions needed to identify and estimate a causal quantity for a given research design are actually met in field settings. In an RDD, we never know whether parametric and non-parametric estimation methods correctly model the relationship between the assignment and outcome variables. In a non-equivalent comparison group design (NEGD), we rarely know whether all confounding covariates that are simultaneously related to treatment assignment and the outcome have been reliably measured. In comparative interrupted time series designs, we never know whether units in the treatment and comparison group share “common trends” over time in the absence of treatment.

The within-study comparison (WSC) design has emerged as a method for assessing whether the stringent assumptions needed to identify and estimate causal quantities are met in practice. In a traditional WSC design, treatment effects from a randomized control trial (RCT) are compared to those produced by a non-experiment (NE) that shares the same target population, outcomes, and intervention. The NE may be an RDD, a matching design, or a difference-in-differences or interrupted time series approach. The goals of a WSC are to determine whether and under which conditions the NE method succeeds in reproducing results from a high quality RCT with the same target population. Table 1 provides a summary of more than 70 WSCs from 1986 to 2017.

Results from early WSCs had a profound influence on research practice and priorities in program and policy evaluation (see WSC studies under "Job Training" in Table 1). These studies reified a clear preference in methodology choice for government funding agencies and evaluation policy: RCT whenever possible, RD when RCTs are not feasible, and finally if at all, observational approaches such as matching or regression adjustment (see What Works Clearinghouse Evidence Standards, 2008a, 2008b, 2011). The Office of Management and Budget cited results from early WSCs in their 2004 recommendation that federal agencies should use RCTs for evaluating program impacts, cautioning against the use of "comparison group studies" that "often lead to erroneous conclusions" (2005, p. 5). The U.S. Department of Education also identified random assignment as the preferred method for "scientifically-based research" in a 2005 issue of the *Federal Register* (2005). In responding to critiques that random assignment was "not the only method capable of generating causal effects," Rod Paige, the Education Secretary under the George W. Bush Administration, 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" (2005, p. 3588).

Despite the importance of WSCs in providing researchers, funders, and decision-makers with guidance about NE methods' performance in practice and designing valid program evaluations, a number of questions about the best ways to implement and analyze the WSC itself remain. For example, what are the requirements for a WSC design to yield interpretable results, and how can researchers design a valid and reliable WSC? What criteria should researchers use to determine if results from the NE replicate results from the RCT benchmark? And perhaps most importantly, how should we interpret results from one WSC to understand NE method performance in other contexts and settings?

In this essay, we provide a brief historical overview of WSC designs. To this end, we describe the special contributions of WSCs to the program evaluation literature, and common methodological challenges that arise in the design, implementation, analysis, and interpretation of the approach. We then highlight papers that appear in this two-volume special issue of *Evaluation Review*. These papers add to our knowledge of NE method performance; they also address important methodological considerations in the design and analysis of WSCs. The essay concludes by considering future directions for how WSCs may be used to improve NE theory and practice.

History of WSCs

Statistical theory formulates the assumptions needed for a causal method to work. That is, theory shows when a method *can* yield unbiased causal effects. Simulation studies help researchers understand the statistical properties of the method under specific, well-defined conditions. Simulation studies, however, rarely capture the full complexity of real world data, and have little to say about whether a research design's assumptions are actually met in field settings. Addressing these methodological questions requires *empirical* evaluations of NE methods in real world evaluations.

Introduced by LaLonde (1986) and Fraker and Maynard (1987), the earliest WSC designs used data from job training evaluations to compare results from an NE with those from an RCT benchmark. To construct the WSCs, LaLonde and Fraker and Maynard used RCT data from the National Supported Work Demonstration program (NSW) (MDRC, 1980). The NE was created by deleting RCT control cases from the NSW, and replacing them with no-treatment comparisons from the Current Population Survey (CPS) or the Panel Study of Income Dynamics (PSID). The interest was methodological – to see whether econometric techniques could be used with nationally representative datasets to reproduce RCT results. But the goal was policy driven – to discover whether there were more cost-efficient methods than RCTs for estimating program impacts.

The early WSCs examined the performance of regression, difference-in-differences, matching, and instrumental variable models. Researchers estimated NE bias by comparing NE results with those obtained from the RCT benchmark. Because the treatment group was shared across the RCT and NE arms, researchers also assessed bias by directly comparing conditional outcomes from NE comparisons and RCT controls (Fraker & Maynard, 1987; Bloom, Michaeloupoulos, & Hill, 2004). The general conclusion from these studies was that NE methods fail to reproduce RCT benchmark results (Fraker & Maynard, 1987; Friedlander & Robins, 1995). Fraker and Maynard summarized their findings by writing, "the results of our study indicate that NE design evaluations cannot be relied on to estimate the effectiveness of programs like Supported Work with sufficient precision (and in some cases unbiasedness) to provide policymakers with adequate information to guide decisions" (p. 196, 1987).

A decade later, Dehejia and Wahba (1999) claimed to overturn that conclusion. They reanalyzed the NSW data and concluded that propensity score matching methods did succeed in reproducing RCT benchmark results. However, Smith and Todd (2005) showed that these estimates were highly sensitive to the choice of covariates used for estimating the propensity score and the analysis sample used. Subsequent WSC results also demonstrate the importance of covariate selection in matching procedures (Steiner, Cook, Shadish, & Clark, 2010).

Heckman and colleagues (Heckman and Hotz, 1989; Heckman, Ichimura and Todd, 1997; Heckman, Ichimura, Smith and Todd, 1996; 1998) reanalyzed the NSW data, and conducted new WSCs with RCT data from the Job Training Partnership Act (JTPA) evaluation. For the JTPA data, they constructed the NE comparison group from observational data of individuals who qualified for JTPA but chose not to participate in the intervention. Using results from WSCs, Heckman and colleagues highlighted conditions under which NE bias can be successfully addressed – in job training settings at least. NE estimates were less biased when rich covariate information was available for matching units, when comparisons were drawn from the same local labor markets, and when dependent variables were measured in the same way for all participants. They also observed that difference-in-differences estimators address selection bias better than cross-sectional estimators, and that specification tests using pre-treatment outcomes often succeeded in eliminating the most biased estimators. However, Heckman et al. also concluded that while these approaches often succeeded in *reducing* bias, there was no assurance that they reliably *eliminated* bias.

Two studies provided further surveys of WSC results, with similar conclusions. Glazerman, Levy and Myers (2003) meta-analyzed 12 within-study comparisons that used data from a series of job training experiments; Bloom, Michaeloupoulos and Hill (2004) provide a qualitative summary of WSC results from early job training studies. Both reviews found that although NE approaches sometimes replicated RCT 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 – NE methods could not be trusted to produce credible causal estimates in field settings (2003, p. 86).

Methodological Challenges with WSCs

Results from early WSCs prioritized RCTs as the main research design for program evaluation. This was especially true in fields such as education which, prior to 2001, did not have a tradition of using experiments (Angrist, 2004; Cook, 2007). However, despite the sound theoretical reasons to prefer RCTs and some types of quasi-experimental designs, results from early WSCs were also suspect in a number of ways. Incorrect conclusions about the empirical performance of NE methods could have occurred due to invalid WSC designs, or the choice of an inappropriate metric for assessing NE performance. Below we highlight five common methodological challenges that arose in the design and analysis of early WSCs.

1. *Study differences between the RCT and NE.* In many early WSCs, the RCT and NE differed in ways beyond the mode of treatment assignment (i.e., random assignment versus self-selection). For example, comparison units in the CPS or PSID may have been drawn from remote locations (instead of within the same locale as treatment cases), measured at different time points, and in some cases, may not have shared the same outcome measures. Comparison units in the NE may also have had alternative job training options than what was available to control cases in the RCT. When the RCT and NE arms have extraneous study differences, it is difficult for the researcher to draw

conclusions about how well the NE actually performed. Lack of correspondence in NE and RCT results could have occurred because of bias in the NE estimate, or because the outcome measure was not assessed in the same way across the two study arms. It would be impossible for the researcher to tell.

2. *Differences in causal estimands.* WSC results were sometimes confounded by comparisons of different causal quantities from each study condition. For example, the experimental average treatment effect (ATE) may have been compared to an RD average treatment effect at the cutoff. If treatment effects are heterogeneous among sub-populations of units, then comparing two causal quantities may produce different effect estimates for reasons not related to bias in the NE.
3. *Weak causal benchmark for evaluating NE.* The RCT benchmark may have suffered from its own implementation problems in the field. Differential attrition, treatment non-compliance, or individuals trying to subvert the randomization process in the RCT may invalidate the RCT's benchmark status, that is, the RCT was not well enough implemented to serve as the standard for evaluating NE performance.
4. *Inappropriate metrics for assessing NE method performance.* Early WSCs lacked consensus on how close RCT and NE results needed to be for the researcher to judge that the NE method succeeded in reproducing the RCT effects. Some studies compared the direction and magnitude of effects (Aiken, West, Schwalm, Carroll, & Hsuing, 1998), while others examined patterns of statistical significance (Agodini & Dynarski, 2004; Diaz & Handa, 2006), and still others observed whether estimates differed by more than some policy-relevant threshold (Glazerman, Levy, & Myers, 2003). One challenge with these measures is that they may conclude that the NE fails to reproduce RCT results, even

when the effect estimates are identical or very similar. For example, if the RCT estimate is slightly greater than zero and the NE estimate is slightly less than zero, then comparing direction of effects may suggest lack of correspondence in results, even though the point estimates themselves may be considered as equivalent. In another example, the RCT and NE point estimates may be exactly identical, but the benchmark result is statistically insignificant while the NE result is significant. Although comparing significance patterns informs researchers about whether a policy-maker would arrive at the same decision from an RCT and NE design, these measures may be less useful for assessing the performance of the NE method itself.

5. *Limited generalization about NE method performance.* Although results from early WSCs provided information about NE performance in job training contexts, there were questions about the extent to which these findings could be generalized to NEs with different target populations, treatments, outcomes, selection mechanisms, baseline information, and research designs.

Glazerman and colleagues acknowledged the limitations of early WSCs by writing that their "summary of findings gives only part of the picture, and it does so for a specific area of program evaluation research: the impacts of job training and welfare programs on participant earnings" (2003, p. 87). Taken together, these concerns suggested that not only were more WSCs needed in different field settings, but WSCs of higher methodological quality for drawing valid conclusions about NE methods' ability to estimate causal effects in practice.

WSC Methodological Innovations

Since the Glazerman et al. (2003) review, researchers have introduced WSC design innovations to address the five methodological limitations in the earlier numbered list. To reduce study differences in the RCT and NE (issue #1 from above), researchers drew NE comparison units from the same target population as in the RCT. Bloom, Michalopoulos and Hill (2005) used RCT data from the multi-state, multi-site National Evaluation of Welfare-to-Work Strategies (NEWWS) to construct a WSC. In the RCT arm, welfare recipients were randomly assigned to job training services within sites; in the NE arm, RCT controls from other NEWWS sites (often within the same city) were used to form the comparison group. Because all participants were involved in the same study protocol, they met the same eligibility criteria, provided the same baseline and outcome information, and experienced the same macroeconomic and labor market conditions at the same time. The consistency in research protocols across both study arms reduced the threat of confounders that might otherwise explain differences in RCT and NE results.

Shadish, Clark, and Steiner (2008) introduced another WSC design variant that bolstered the interpretation of results. They ensured that the RCT and NE compared equivalent causal estimands (issue #2) for the same target population by randomly assigning study participants into the RCT or NE arm of the WSC. Once assigned into study arms, participants in the RCT were randomly assigned again into the reading or math intervention while those in the NE were allowed to select an intervention of their preference. NE bias was computed by comparing effect estimates of the ATE across both study arms. The researchers also were able to ensure that the RCT was well implemented by analyzing baseline and fidelity measures (issue #3). And, because the WSC was prospectively planned and took place within a controlled laboratory-like setting, the researchers were able to implement the same study procedures across the RCT and NE arms (issue #1). This meant delivering identical, scripted treatment and control interventions in the RCT and NE studies, and using the same outcome measures for assessing impacts of the interventions. Subsequent analyses found no evidence of differential attrition within the RCT, and across the RCT and NE arms.

Later WSCs introduced new approaches for assessing comparability between RCT and NE results (issue #4). These studies acknowledged that, because of sampling error, even close replications of the same RCT would not result in identical treatment effects. And although most studies assessed comparability by examining statistical significance patterns between the RCT and NE, some began using direct statistical tests of difference between RCT and NE results. Other new methods for assessing correspondence included looking at the percent of bias reduced from the initial naïve comparison (Shadish et al., 2008), percent difference in the RCT and NE estimate (Wilde & Hollister, 2007), the mean squared error (Wing & Cook, 2013), the effect size differences between RCT and NE results (Hallberg, Wong, & Cook, 2014), or the relative performance of different NE approaches across multiple bootstrap replications (Hallberg, Wong, & Cook, 2014). Bell and Orr used a Bayesian framework to compute the probability of an incorrect policy decision for different magnitudes of true effect sizes (Solari, Nisar, Bell, & Orr, 2017). All of these approaches have their advantages and limitations. However, the lack of consensus in the WSC literature on how correspondence should be assessed has led to ambiguity and challenges in synthesizing the literature.

Finally, a common critique of WSC evaluations concerns their generalizability. Researchers want to know how well results from one study setting apply to NE method performance in other contexts, with different outcomes and treatment selection mechanisms (issue #5). Although this issue is not unique to WSCs – the same concern arises in RCT evaluations – results from a single WSC study have little to say about general method performance. But results from *multiple* WSCs may provide insights as to how well these methods perform for similar outcomes and settings of particular interest.

Over the years, researchers have conducted qualitative and quantitative summaries of WSC results with the goal of providing advice for better NE practice. Some summaries have focused on observational method performance in particular disciplines or fields, with a narrowly defined set of outcomes. Glazerman, Levy, and Meyers (2003) and Bloom, Michaeloupoulos, and Hill (2004) reviewed WSC results in the job training literature, where the outcome of interest was participants' annual earnings. Both reviews confirmed Heckman et al.'s findings that NE methods produced less biased estimates when comparison groups were local, when covariate sets were rich and included pretest measures, and when researchers combined multiple design features (e.g. difference-in-differences with matching) for estimating effects.

Wong, Valentine, and Miller-Bains (2017) examined results from 12 WSCs in education settings with standardized reading or math outcomes. Their goal was to assess performance of common covariate-types used in observational studies in education. As in the job training literature, Wong et al. found that the pretest often reduced a major portion of the bias but it did not always eliminate it. However, matching units from similar geographic locales did not provide the same benefit within education contexts as it did in job training settings. This was likely because the selection process into education interventions varied across settings, as did the definition of "local" comparisons in these evaluations. Wong et al. also noted that when rich covariate sets were available, NE methods replicated RCT benchmark estimates more closely in educational contexts, but the authors noted that further replications are needed in this area.

Other summarizes have reviewed WSC results from multiple disciplines to assess method performance more generally. Cook, Shadish, and Wong (2008) looked at 12 WSCs from 2002 to

2007 that spanned the fields of education, international development, and public health. The authors observed three conditions under which the NE method appeared to remove all or at least a major part of the bias. The first condition was when treatment and comparison units were assigned to treatment conditions based on an assignment variable and a cutoff, as in the RDD. In a more recent review, Chaplin et al. (2018) meta-analyzed results from 15 WSCs looking at RD performance across various fields. They found that the average NE bias was small – less than 0.01 SDs, providing further evidence for Cook et al.'s hypothesis.

Cook, Shadish, and Wong's second and third conditions describe contexts under which NE methods appeared to remove most if not all the bias. Those contexts include when the selection process was known and observed by the researcher, as in students' selection into a math or vocabulary intervention in the Shadish et al. WSC described above, or when "intact groups" (e.g. schools, villages) were matched using rich covariate information, or within the same geographic area. However, these results have yet to be confirmed by more recent WSCs, so more research is needed in this area.

This Special Issue

This two-volume special issue of *Evaluation Review* contributes to the WSC literature in two distinct ways. First, the February issue presents four additional case-study evaluations of NE method performance in educational contexts. **Gleason, Resch, and Berk (2018)** examine parametric and non-parametric method performance in an RDD. The authors use RCT data from evaluations of Ed Tech and Teach for America to construct RD designs synthetically. They created the RD by selecting a hypothetical cutoff on a baseline covariate, and systematically deleting RCT treatment or comparison observations above and below the designated cutoff. A useful innovation of this paper is that the authors replicated their RCT results across multiple datasets, as well as multiple cutoffs within each dataset, and pooled their results through a systematic meta-analysis. **Dong and Lipsey (2018)** assess covariate performance in an observational study within the context of early childhood education (ECE). This is one of the few studies in the WSC literature that examines covariate performance in an ECE setting with outcomes of students' emerging literacy and math skills. They also looked at the performance of different matching estimators when comparisons were drawn from within and across states. **Kisbu-Sakarya, Cook, and Tang (2018)** also examined NE method performance in the context of ECE, but their WSC evaluates the performance of a comparative RD (CRD) design to an RCT benchmark from the Head Start Impact study. Finally, **Tang and Cook (2018)** show the benefits of the CRD design by comparing the statistical precision of CRD results with RD and RCT results from the Head Start Impact study.

The April issue includes a series of methodological papers that seek to improve the design and analysis of the WSC approach itself. To this end, **Wong and Steiner (2018)** formalize the WSC design using a potential outcomes framework. They explicate the required design components and assumptions needed for the approach to yield a valid interpretation of NE method performance. The paper also describes three different design variants for evaluating NE methods, and the benefits and limitations of each approach. **Steiner and Wong (2018)** next address the issue of how one should assess correspondence between RCT and NE results. That is, they address the question first posed by Wilde and Hollister (2008) of "how close is close enough" for the NE to have successfully replicated benchmark results? Through a series of simulation studies, the authors demonstrate the benefits and limitations of common criteria for assessing correspondence in RCT benchmark and NE results, and propose a new framework for assessing NE method performance: the correspondence test, which incorporates both frequentist tests of difference and equivalence in the same framework. **Rindskopf and Shadish (2018)** propose an alternative criterion for assessing correspondence between RCT and NE results using a Bayesian approach. Their method involves calculating the probability that the absolute value of the difference between the RCT and NE result is less than some threshold determined to be close enough to zero. They argue that the Bayesian criteria improve the power of WSCs by allowing for the incorporation of prior information into the analysis, and provide more varied, nuanced, and informative answers to questions of correspondence.

New Frontiers for WSC Approaches

Although the WSC literature has made strong advances since the early job training studies, our reading of the literature suggests four emerging areas for improving the design, analysis, and practice of NE evaluations:

Issue 1: Establish Research Protocols for the Design and Analysis of WSC Results. One issue with the implementation of WSCs is that knowledge of the benchmark result may inadvertently skew the many decisions researchers must make in the analysis of the NE. For example, in observational studies, the researcher has choices about covariate selection for estimating the propensity score (Smith & Todd, 2005), and about the type of estimator used to produce treatment effects (e.g., matching, stratification, or doubly robust estimators). Cook et al. (2008) recommend that two independent research teams should analyze the benchmark and NE separately, and that the analysts of the NE should be blinded of the benchmark results. This is generally good practice, but it may not be specific enough to be feasible. Research teams may wish to coordinate which causal estimands they will compare, and the analytic models they will use to estimate treatment effects (e.g., should the RCT and NE treatment effects be estimated using regression-adjusted (doubly robust) models or not?).

In future implementations of WSCs, research teams should establish and describe a protocol in advance of data collection or analysis. Developing a WSC research protocol is similar to preregistration of research plans for RCTs or meta-analyses (see Chuang, Wykstra, & Knowledge Management, 2015 for guidance). One benefit of a WSC protocol is that it would provide pre-specified guidance to researchers on questions that naturally arise in the design and analysis of WSCs. In cases where the NE and RCT are analyzed by independent teams of researchers, developing a research protocol can provide opportunities for investigators to come to a common understanding of the study plan. The research protocol could also allow for WSC researchers to obtain feedback and advice on their data collection and analysis plans, prior to revealing any results.

Generally, the WSC protocol should address the following topics: (1) confirmatory versus exploratory research questions in the WSC context; (2) diagnostics for assessing assumptions of the WSC design; (3) potential deviations from the intended research protocol; and (4) criteria for determining correspondence in results. The protocol should recommend that analysts of the RCT and NE document all analysis procedures; it should also provide a place for the researchers to document any problems or questions that arise, and how these questions were resolved. Finally, the protocol should provide guidance on when it is appropriate for RCT and NE analysts to consult with each other, and when their analysis should be conducted independently.

Issue 2: Consider Statistical Power for WSC Designs. Another critical issue in the planning of WSCs is ensuring that the design has sufficient statistical power for detecting comparability in treatment effects between the RCT and NE. In fact, WSCs usually have much greater power requirements than do the RCT or NE for detecting impacts. To understand why WSCs usually require larger samples, consider a scenario where the criterion for assessing correspondence in RCT and NE effects is to determine whether the two study conditions produce the same test result in a null hypothesis test of the treatment effect. In other words, do the RCT and NE result in the same conclusion about the presence of a treatment effect? In an independent WSC design (i.e., units were randomly assigned into RCT and NE conditions), the probability of rejecting the null in both study conditions depends on the statistical power in the RCT and NE. Here, a well-powered RCT and NE, with both having a statistical power of 0.80 to detect the true but unknown effect, produce the same pattern of statistical significance with a probability of 0.68 only (= 0.8×0.8 + 0.2×0.2, i.e., the probability of obtaining a significant effect estimate in both studies plus the probability of obtaining an insignificant result in both studies). But when—as is not uncommon—the RCT or NE is underpowered for detecting significant effects (e.g., both having a power of 0.2), the probability of obtaining corresponding significance patterns is again 0.68. But now correspondence is most likely due to obtaining insignificant (0.8×0.8) rather than significant (0.2×0.2) effect estimates in both studies. Thus, when there is no significant treatment effect for the NE and RCT, the researcher may incorrectly conclude that the NE lacks bias – but this may be because both study conditions are underpowered for detecting effects!

Future WSCs should consider statistical power for assessing comparability of results in the design phase of the evaluation. Three papers in the March issue provide guidance on statistical power. **Wong and Steiner** show that WSC design variants (e.g. WSCs with independent versus dependent data structures in the RCT and NE arms) have different statistical power for assessing correspondence in results; and **Steiner and Wong** suggest a method for assessing statistical power in the design phase through the correspondence framework. **Rindskopf and Shadish** suggest that Bayesian approaches for assessing correspondence of RCT and NE results have improved statistical power over frequentist approaches.

Issue 3: Continue to Explore the External Validity of WSC Results. The existing WSCs represent a heterogeneous mix of studies from different disciplines, research designs, and outcomes. Currently, the authors have identified more than 70 WSC studies (see Table 1). These studies include substantial variation in contexts, NE methods examined, as well as outcomes and treatment selection mechanisms. As more studies continue to be added to the literature, ongoing quantitative synthesis of results can provide important descriptive information about NE method performance in field settings, and the contexts and conditions under which these methods may perform well. Meta-analysis of WSC results may also address an important challenge that many standalone studies face – lack of statistical power for assessing correspondence in results.

However, we note that a rigorous synthesis of WSC results also requires more systematic reporting of study procedures and outcomes, as well as consistent criteria for assessing correspondence in results. For example, it would be useful for WSC analysts to report estimates of NE bias, and the standard error of their bias estimates. Moreover, in WSC designs where units are shared between the RCT and NE arm, the standard errors should account for dependencies in the data structure (see discussion by **Steiner and Wong**). In addition, because the direction and strength of the selection processes in the NE vary across WSC studies, analysts should always report the initial, unadjusted selection bias (i.e., the difference between the unadjusted NE estimate and the RCT estimate). This allows for an assessment of the sign and magnitude of the selection bias before making any statistical adjustments.

Meta-analysis of WSC results has tremendous promise in revealing new insights about good NE practice. However, given the heterogeneity of WSCs in terms of study designs, samples, outcomes, and selection processes, a rigorous meta-analysis should synthesize or pool results only when substantively or theoretically appropriate. To this end, WSC analysts should document and report study procedures and contextual factors that may be related to NE bias.

Issue 4: Using RCT Benchmark Results for Examining Treatment Effect Variation and Generalization. Recently, researchers have applied WSC designs to address research questions of programmatic and policy relevance. For example, an RCT benchmark may be used to validate an NE model that is then used to estimate treatment effects for a more general target population of interest. This method has been applied to generalize treatment effects across different units (Angrist & Rokkanen, 2012; Wing & Clark, 2016), treatments (Bell, Harvill, Moulton, & Peck, 2017; J. V Hotz, Imbens, & Mortimer, 2005) and settings (Abdulkadiroğlu, Angrist, Dynarski, Kane, & Pathak, 2011).

For example, Abdulkadiroğlu et al. (2011) used a WSC design to assess the external validity of treatment effects from Boston charter and pilot schools with admission lotteries to schools without such lotteries. The RCT consisted of lottery students in oversubscribed charter/pilot schools; the NE consisted of lottery winners, as well as non-charter/pilot students in Boston public schools. To estimate NE treatment effects, the authors used regression models that controlled for student demographic characteristics and baseline scores.

The authors constructed a series of WSCs for sub-samples of charter and pilot schools, and for elementary and secondary grades. In cases where the WSC NE and RCT produced corresponding effects, the researchers concluded that the NE model was sufficient for addressing selection bias in an observational study of non-lottery charter/pilot schools and Boston public schools. The assumption here was that the selection process into charter/pilot schools with lotteries could be generalized to schools without lotteries. However, when the WSC NE failed to reproduce RCT benchmark results, the authors concluded that the NE model could not be used to estimate observational treatment effects. Overall, Abdulkadiroğlu et al. observed close correspondence in RCT and NE results for charter school students, and for middle school students with pilot programs. In assessing the external validity of the charter school lottery results, they found that although charter schools without lotteries produced positive and significant effects, they were smaller than effects observed from over-subscribed charter schools. The authors also found that the WSC NE model did not perform well for a sub-sample of high schools with pilot programs. As a result, they did not use the NE model to assess the external validity of treatment effects for this subsample of schools.

In a second example, Hotz, Imbens, and Klerman (2006) used a WSC design to examine treatment effect variation due to differences in program components. The researchers used RCT data from the Greater Avenues to Independence Program (GAIN) evaluation, where participants in six California counties were randomly assigned to receive job training services or to be in a control group that was denied services. Because of the local nature of treatment implementation, some county programs provided participants with general education and skills development, while other sites encouraged participants to secure immediate employment.

A goal of the evaluation was to assess treatment effect variation due to differential program components. However, because participants were not randomly assigned to sites, researchers were concerned that observed treatment effect variation may have been confounded with participants' characteristics. To address this issue, the authors constructed an WSC using *RCT control group members' outcomes*. Their goal was to examine whether NE methods and observed participant characteristics could address units' selection into sites. In places where the NE method succeeded in producing conditionally equivalent control groups, the researchers felt assured that the NE approach could be used to produce valid effect estimates of program components.

These examples illustrate how WSCs may be used to probe NE assumptions empirically. They also show how WSCs may be used to signal when NE assumptions are not well warranted in field settings. As researchers continue to use RCT and NE data to "learn more" from program and policy evaluations, WSCs provide an important method for validating NE assumptions, and for generalizing and uncovering differential treatment effects.

Conclusion

Because of increased availability of RCT data, there are now empirical evaluations of NE methods in job training, education, early childhood development, political science, international development, and public health. WSCs have also been used to evaluate more types of quasi-experimental approaches, including the regression-discontinuity design (see Cook and Wong (2008) for review) and most recently, the interrupted time series design (St.Clair, Hallberg, & Cook, 2016; St.Clair, Cook, & Hallberg, 2014). As the number of WSCs in varying contexts increases, so does the opportunity for synthesizing the literature for greater insight and external validity.

Results from WSC evaluations have had important impacts on both research practice and funding priorities in program evaluation. In most areas of the social sciences, an RCT is the preferred method for establishing causal inferences. However, WSCs have shown specific contexts and conditions where NE methods succeed in removing most if not all the bias.

Methodological advances in WSC designs, like those in presented in this special issue, will continue to improve our understanding of NE practice. As the program evaluation field turns to important policy-relevant questions such as, "*When, where, for whom, and why does it work*?", WSCs may again be instrumental in improving methodology and validating research design assumptions in field settings.

Dedication

We dedicate this two-volume special issue on within-study comparison designs to our mentor and friend, William R. Shadish. We had the honor to work with and learn from Will on the design, implementation, and analysis of several within-study comparisons. The April issue of *Evaluation Review* includes one of Will's last papers, co-authored with David Rindskopf. Will, we miss you and think of you often.

- V.W. and P.S.

Bibliography

Abdulkadiroğlu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., & Pathak, P. A. (2011). Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots. *Quarterly Journal of Economics*, *126*(2), 699–748. https://doi.org/10.1093/qje/qjr017

Agodini, R., & Dynarski, M. (2004). Are Experiments the Only Option? A Look at Dropout Prevention Programs. *Review of Economics and Statistics*, *86*(1), 180–194. https://doi.org/10.1162/003465304323023741

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a Randomized and Two Quasi-Experimental Designs in a Single Outcome Evaluation. *Evaluation Review*, *22*(2), 207–244. https://doi.org/10.1177/0193841X9802200203

Anderson, K. P., & Wolf, P. J. (2017). *Evaluating School Vouchers: Evidence from a Within-Study Comparison* (EDRE No. 2017–10). *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2952967

Anglin, K., Miller-Bains, K., Wong, V. C., & Wing, C. (2018). Methods of Reducing Bias in Time Series Designs: A Within Study Comparison. In *Society for Research on Educational Effectiveness*. Washington, DC.

Angrist, J., Autor, D., Hudson, S., & Pallais, A. (2015). Evaluating Econometric Evaluations of Post-Secondary Aid. *American Economic Review*, *105*(5), 502–507. https://doi.org/10.1257/aer.p20151025

Angrist, J. D. (2004). American education research changes tack. *Oxford Review of Economic Policy*, *20*(2), 198–212. https://doi.org/10.1093/oxrep/grh011

Angrist, J. D., & Rokkanen, M. (2015). Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff. *Journal of the American Statistical Association*, *110*(512), 1331–1344. https://doi.org/10.1080/01621459.2015.1012259

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

Arceneaux, K., Gerber, A. S., & Green, D. P. (2010). A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark. *Sociological Methods & Research*, *39*(2), 256–282. https://doi.org/10.1177/0049124110378098

Ashworth, K. E., & Pullen, P. C. (2015). Comparing Regression Discontinuity and Multivariate Analyses of Variance: Examining the Effects of a Vocabulary Intervention for Students at Risk for Reading Disability. *Learning Disability Quarterly*, *38*(3), 131–144. https://doi.org/10.1177/0731948714555020

Barrera-Osorio, F., Filmer, D., & McIntyre, J. (2014). Randomized Controlled Trials and Regression Discontinuity Estimations: An Empirical Comparison. In *Society for Research on Educational Effectiveness*.

Bell, S., Harvill, E., Moulton, S., & Peck, L. (2017). *Using Within-Site Experimental Evidence to Reduce Cross- Site Attributional Bias in Connecting Program Components to Program Impacts Using Within-Site Experimental Evidence*. Bethesda.

Bell, S., Orr, L., Blomquist, J., & Cain, G. G. (1994). *Program Applicants as a Comparison Group in Evaluating Training Programs: Theory and a Test*. Kalamazoo: Upjohn Institute for Employment Research. https://doi.org/10.17848/9780585284545

Bifulco, R. (2012). Can Nonexperimental Estimates Replicate Estimates Based on Random Assignment in Evaluations of School Choice? A Within-Study Comparison. *Journal of Policy Analysis and Management*, *31*(3), 729–751. https://doi.org/10.1002/pam.20637

Black, D., Galdo, J., & Smith, J. (2007). *Evaluating the Bias of the Regression Discontinuity Design Using Experimental Data*.

Bloom, H., Michalopoulos, C., & Hill, C. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In *Learning more from social experiments: Evolving analytic approaches* (pp. 173–235).

Bloom, H., Michalopoulos, C., Hill, C., & Lei, Y. (2002). *Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?* (MDRC Working Papers on Research Methodology).

Bratberg, E., Grasdal, A., & Risa, A. E. (2002). Evaluating Social Policy by Experimental and Nonexperimental Methods. *Scandinavian Journal of Economics*, *104*(1), 147–171. https://doi.org/10.1111/1467-9442.00276

Buddelmeyer, H., & Skoufias, E. (2004). *Evaluation of the Performance of Regression Discontinuity Design on PROGRESA*. *World Bank Policy Research Working Paper*. World Bank, Washington, D.C. Retrieved from https://search.lib.virginia.edu/catalog/u6398806

Chaplin, D., D. Cook, T., Zurovac, J., Coopersmith, J., M. Finucane, M., N. Vollmer, L., & Morris, R. (2018). The Internal and External Validity of the Regression Discontinuity Design: A Meta-Analysis of 15 Within-Study-Comparisons. *Journal of Policy Analysis and Management*, *37*. https://doi.org/10.1002/pam.22051

Clearinghouse, W. W. (2008). *What Works Clearinghouse Procedures and Standards Handbook (Version 2.0)*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v2_standards_handbook.pdf

Cook, T. D., & Wong, V. C. (2008). Empirical Tests of the Validity of the Regression Discontinuity Design. *Annales d'Economie et de Statistique*, (91/92), 127–150. https://doi.org/10.2307/27917242

Cook, T. D., & Foray, D. (2007). Building the Capacity to Experiment in Schools: A Case Study of the Institute of Educational Sciences in the US Department of Education. *Economics of Innovation and New Technology*, *16*(5), 385–402. https://doi.org/10.1080/10438590600982475

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons. *Journal of Policy Analysis and Management*, *27*(4), 724–750. https://doi.org/10.1002/pam

Dehejia, R. H., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics*, *84*(1), 151–161. https://doi.org/10.1162/003465302317331982

Diaz, J. J., & Handa, S. (2006). An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator. *Journal of Human Resources*, *XLI*(2), 319–345. https://doi.org/10.3368/jhr.XLI.2.319

Dong, N., & Lipsey, M.W. (2018). Can propensity score analysis approximate randomized experiments using pretest and demographic information in pre-K intervention research? Evaluation Review.

Dong, N., & Lipsey, M. (2014). How Well Propensity Score Methods Approximate Experiments Using Pretest and Demographic Information in Educational Research. In *Association for Public Policy Analysis and Management*.

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

Ferraro, P. J., & Miranda, J. J. (2014). The performance of non-experimental designs in the evaluation of environmental programs: A design-replication study using a large-scale randomized experiment as a benchmark. *Journal of Economic Behavior & Organization*, *107*, 344–365. https://doi.org/10.1016/j.jebo.2014.03.008

Fortson, K., Gleason, P., Kopa, E., & Verbitsky-Savitz, N. (2015). Horseshoes, hand grenades, and treatment effects? Reassessing whether nonexperimental estimators are biased. *Economics of Education Review*, *44*, 100–113. https://doi.org/10.1016/j.econedurev.2014.11.001

Fortson, K., Verbitsky-Savitz, N., Kopa, E., & Gleason, P. (2012). *Using an Experimental Evaluation of Charter Schools to Test Whether Nonexperimental Comparison Group Methods Can Replicate Experimental Impact Estimates. National Center for Education Evaluation and Regional Assistance*. Retrieved from http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED531481

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

Fretheim, A., Zhang, F., Ross-Degnan, D., Oxman, A. D., Cheyne, H., Foy, R., … Soumerai, S. B. (2015). A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation. *Journal of Clinical Epidemiology*, *68*(3), 324–333. https://doi.org/10.1016/j.jclinepi.2014.10.003

Friedlander, D., & Robins, P. K. (1995). Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods. *American Economic Review*, *85*(4), 923–937. https://doi.org/10.1016/j.jclinepi.2014.10.003

Gill, B., Furgeson, J., Chiang, H., Teh, B., Haimson, J., & Savitz, N. V. (2016). Replicating Experimental Impact Estimates with Nonexperimental Methods in the Context of Control-Group Noncompliance. *Statistics and Public Policy*, *3*(1), 1–11. https://doi.org/10.1080/2330443X.2015.1084252

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental Versus Experimental Estimates of Earnings Impacts. *Annals of the American Academy of Political & Social Science*, *589*(1), 63–93. https://doi.org/10.1177/0002716203254879

Gleason, P., Resch, A., & Berk, J. (2012). *Replicating Experimental Impact Estimates Using a Regression Discontinuity Approach*. Retrieved from https://ies.ed.gov/ncee/pubs/20124025/

Gleason, P., Resch, A., & Berk, J. (2018). RD or not RD: Using experimental studies to assess the performance of the regression discontinuity approach. Evaluation Review.

Green, D. P., Leong, T. Y., Kern, H. L., Gerber, A. S., & Larimer, C. W. (2009). Testing the Accuracy of Regression Discontinuity Analysis Using Experimental Benchmarks. *Political Analysis*, *17*(04), 400–417. https://doi.org/10.1093/pan/mpp018

Hallberg, K., Cook, T. D., Steiner, P. M., & Clark, M. H. (2016). Pretest Measures of the Study Outcome and the Elimination of Selection Bias: Evidence from Three Within Study Comparisons. *Prevention Science*, 1–10. https://doi.org/10.1007/s11121-016-0732-6

Hallberg, K., Wong, V. C., & Cook, T. D. (2016). *Evaluating Methods for Selecting School-Level Comparisons in Quasi-Experimental Designs: Results from a Within-Study Comparison* (EdPolicy Works Working Paper Series No. 47). Retrieved from https://curry.virginia.edu/uploads/resourceLibrary/47_School_Comparisons_in_Observational_Designs.pdf

Handa, S., & Maluccio, J. A. (2010). Matching the Gold Standard: Comparing Experimental and Nonexperimental Evaluation Techniques for a Geographically Targeted Program. *Economic Development and Cultural Change*, *58*(3), 415–447. https://doi.org/10.1086/650421

Heckman, J. J., & Hotz, J. V. (1989). Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. *Journal of the American Statistical Association*, *84*(408), 862. https://doi.org/10.2307/2290059

Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, *66*(5), 1017–1098. https://doi.org/10.2307/2999630

Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, *66*(5), 1017–1098. https://doi.org/10.2307/2999630

Heckman, J. J., Ichimura, H., & Todd, P. (1997). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, *64*(4), 605–654. https://doi.org/10.2307/2971733

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, *66*(5), 1017–1098.

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2005). A Comparison of Experimental and Observational Data Analyses. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family* (pp. 49–60). https://doi.org/10.1002/0470090456.ch5

Hotz, J. V., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations. *Journal of Econometrics*, *125*(1), 241–270. https://doi.org/10.1016/j.jeconom.2004.04.009

Hotz, V. J., Imbens, G. W., & Klerman, J. A. (2006). Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California GAIN Program. *Journal of Labor Economics*, *24*(3), 521–566. https://doi.org/10.1086/505050

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, *125*(1), 241–270. https://doi.org/10.1016/j.jeconom.2004.04.009

Jaciw, A. P. (2016). Applications of a Within-Study Comparison Approach for Evaluating Bias in Generalized Causal Inferences From Comparison Groups Studies. *Evaluation Review*, *40*(3), 241–276. https://doi.org/10.1177/0193841X16664457

Jaciw, A. P. (2016). Assessing the Accuracy of Generalized Inferences From Comparison Group Studies Using a Within-Study Comparison Approach. *Evaluation Review*, *40*(3), 199–240. https://doi.org/10.1177/0193841X16664456

Jacob, R., Somers, M.-A., Zhu, P., & Bloom, H. (2016). The Validity of the Comparative Interrupted Time Series Design for Evaluating the Effect of School-Level Interventions. *Evaluation Review*, *40*(3), 167–198. https://doi.org/10.1177/0193841X16663414

Keele, L. J., & Titiunik, R. (2015). Geographic Boundaries as Regression Discontinuities. *Political Analysis*, *23*(01), 127–155. https://doi.org/10.1093/pan/mpu014

Kisbu-Sakarya, Y., Cook, T.D., Tang, Y., & Clark, M.H. (2018). Comparative regression discontinuity: A stress test with small samples. Evaluation Review

L. Morgan, S., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511804564

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

Lee, W.-S. (2006). Evaluating the Effects of a Mandatory Government Program using Matched Groups within a Similar Geographic Location. *SSRN Electronic Journal*, 1–74. https://doi.org/10.2139/ssrn.936783

Leow, C., Wen, X., & Korfmacher, J. (2015). Two-Year Versus One-Year Head Start Program Impact: Addressing Selection Bias by Comparing Regression Modeling With Propensity Score Analysis. *Applied Developmental Science*, *19*(1), 31–46. https://doi.org/10.1080/10888691.2014.977995

Lottridge, S. M., Nicewander, W. A., & Mitzel, H. C. (2011). A Comparison of Paper and Online Tests Using a Within-Subjects Design and Propensity Score Matching Study. *Multivariate Behavioral Research*, *46*(3), 544–566. https://doi.org/10.1080/00273171.2011.569408

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity Scores: An Introduction and Experimental Test. *Evaluation Review*, *29*(6), 530–558. https://doi.org/10.1177/0193841X05275596

Manpower Demonstration Research Corporation. (1980). Summary and findings of the National Support Work demonstration. Cambridge, MA: Ballinger Publishing Company.

McKenzie, D., Stillman, S., & Gibson, J. (2010). How Important Is Selection? Experimental vs. Non-Experimental Measures of the Income Gains From Migration. *Journal of the European Economic Association*, *8*(4), 913–945. https://doi.org/10.1111/j.1542-4774.2010.tb00544.x

Michalopoulos, C., Bloom, H., & Hill, C. (2004). Can Propensity-Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs? *Review of Economics and Statistics*, *86*(1), 156–179. https://doi.org/10.1162/003465304323023732

Moss, B. G., Yeaton, W. H., & LIoyd, J. E. (2014). Evaluating the Effectiveness of Developmental Mathematics by Embedding a Randomized Experiment Within a Regression Discontinuity Design. *Educational Evaluation and Policy Analysis*, *36*(2), 170–185. https://doi.org/10.3102/0162373713504988

Mueller, C. E., & Gaus, H. (2015). Assessing the Performance of the "Counterfactual as Self-Estimated by Program Participants." *American Journal of Evaluation*, *36*(1), 7–24. https://doi.org/10.1177/1098214014538487

Office of Management and Budget. (2005). What Constitutes Strong Evidence of a Program's Effectiveness? Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/omb/part/2004_program_eval.pdf

Olsen, R. B., & Decker, P. T. (2001). *Testing Different Methods of Estimating the Impacts of Worker Profiling and Reemployment Services Systems*. Washington, DC: U.S. Dept. of

Labor, Employment and Training Administration. Retrieved from https://search.lib.virginia.edu/catalog/u3860217

Padgett, R. D., Salisbury, M. H., An, B. P., & Pascarella, E. T. (2010). Required, practical, or unnecessary? An examination and demonstration of propensity score matching using longitudinal secondary data. *New Directions for Institutional Research*, *2010*(S2), 29–42. https://doi.org/10.1002/ir.370

Peikes, D. N., Moreno, L., & Orzol, S. M. (2008). Propensity Score Matching: A Note of Caution for Evaluators of Social Programs. *The American Statistician*, *62*(3), 222–231. https://doi.org/10.1198/000313008X332016

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. https://doi.org/10.3102/0162373709343964

Rindskopf, D., Shadish, W.R., & Clark, M.H. (2018). Using Bayesian correspondence criteria to compare results from a randomized experiment and a quasi-experiment allowing self-selection. Evaluation Review.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. US: American Psychological Association. https://doi.org/10.1037/h0037350

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Schneeweiss, S., Maclure, M., Carleton, B., Glynn, R. J., & Avorn, J. (2004). Clinical and economic consequences of a reimbursement restriction of nebulised respiratory therapy in adults: direct comparison of randomised and observational evaluations. *BMJ*, *328*(7439), 560. https://doi.org/10.1136/bmj.38020.698194.F6

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *Journal of the American Statistical Association*, *103*(484), 1334–1344. https://doi.org/10.1198/016214508000000733

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A Randomized Experiment Comparing Random and Cutoff-Based Assignment. *Psychological Methods*, *16*(2), 179–191. https://doi.org/10.1037/a0023345

Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*(1–2), 305–353. https://doi.org/10.1016/j.jeconom.2004.04.011

Solari, C., Nisar, H., Bell, S., & Orr, L. (2017). *Quantifying the Policy Reliability of Competing Non-Experimental Methods for Measuring the Impacts of Social Programs*. *Association for Public Policy Analysis and Management*.

Somers, M.-A., 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 on Research Methodology). Retrieved from http://appam.confex.com/data/extendedabstract/appam/2012/Paper_1758_extendedabstract_156_0.pdf

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the Internal Validity and Statistical Precision of the Comparative Interrupted Time Series Design by Comparison With a

Randomized Experiment. *American Journal of Evaluation*, *35*(3), 311–327. https://doi.org/10.1177/1098214014527337

St. Clair, T., Hallberg, K., & Cook, T. D. (2016). The Validity and Precision of the Comparative Interrupted Time-Series Design. *Journal of Educational and Behavioral Statistics*, *41*(3), 269–299. https://doi.org/10.3102/1076998616636854

Steiner, P. M., Cook, T. D., Li, W., & Clark, M. H. (2015). Bias Reduction in Quasi-Experiments With Little Selection Theory but Many Covariates. *Journal of Research on Educational Effectiveness*, *8*(4), 552–576. https://doi.org/10.1080/19345747.2014.978058

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores. *Journal of Educational and Behavioral Statistics*, *36*(2), 213–236. https://doi.org/10.3102/1076998610375835

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. https://doi.org/10.1037/a0018719

Steventon, A., Grieve, R., & Sekhon, J. S. (2015). A comparison of alternative strategies for choosing control populations in observational studies. *Health Services and Outcomes Research Methodology*, *15*(3–4), 157–181. https://doi.org/10.1007/s10742-014-0135-8

Tang, Y., & Cook, T.D. (2018). Statistical Power for the comparative regression discontinuity design with a pretest no-treatment control function: Theory and evidence from the National Head Start Impact Study. Evaluation Review.

Tang, Y., Cook, T. D., Kisbu-Sakarya, Y., Hock, H., & Chiang, H. (2017). The Comparative Regression Discontinuity (CRD) Design: An Overview and Demonstration of its Performance Relative to Basic RD and the Randomized Experiment. In *Regression Discontinuity Designs* (Vol. 38, pp. 237-279 SE–6). Emerald Publishing Limited. https://doi.org/doi:10.1108/S0731-905320170000038011

What Works Clearinghouse. (2008). *What Works Clearinghouse Evidence Standards for Reviewing Studies, Version 1.0*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_version1_standards.pdf

What Works Clearinghouse. (2011). *What Works Clearinghouse TM Procedures and Standards Handbook (Version 3.0)*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_draft_standards_handbook.pdf

Wichman, C. J., & Ferraro, P. J. (2017). A cautionary tale on using panel data estimators to measure program impacts. *Economics Letters*, *151*(December), 82–90. https://doi.org/10.1016/j.econlet.2016.11.029

Wilde, E. T., & Hollister, R. (2007). How Close is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment. *Journal of Policy Analysis and Management*, *26*(3), 455–477. https://doi.org/10.1002/pam20262

Wing, C., & Cook, T. D. (2013). Strengthening the Regression Discontinuity Desgin Using Additional Design Elements: A Within-Study Comparison. *Journal of Policy Analysis and Management*, *32*(4), 853–877. https://doi.org/10.1002/pam.21721

Wong, V. C., Hallberg, K., & Cook, T. D. (2013). Intact School Matching in Education: Exploring the Relative Importance of Focal and Local Matching. In *Society for Research on Educational Effectiveness*.

Wong, V. C., Valentine, J., & Miller-Bains, K. (2017). Empirical Performance of Covariates in Education Observational Studies. *Journal of Research on Educational Effectiveness*, *10*(1), 207–236. https://doi.org/10.1080/19345747.2016.1164781

Zhou, X., & Xie, Y. (2016). Propensity Score–based Methods Versus MTE-based Methods in Causal Inference. *Sociological Methods & Research*, *45*(1), 3–40. https://doi.org/10.1177/0049124114555199

Table 1: All Known Within-Study Comparisons

Field	*Study*	WSC Design	NE Design
Consumer Science			
	Mueller and Gaus 2015	Independent	Other
Development			
	Buddelmeyer and Skoufias 2004	Dependent	RDD
	Diaz and Handa 2006	Dependent	NECG
	Handa and Maluccio 2010	Dependent	NECG
Education			
	Agodini and Dynarski 2004	Dependent	NECG
	Aiken et al. 1998	Dependent	NECG/RDD
	Anderson and Wolf 2017	Dependent	NECG
	Angrist et al. 2015	Dependent	RDD
	Ashworth and Pullen 2015	Dependent	RDD
	Barrera-Osorio, Filmer, and McIntyre 2014	Dependent	RDD
	Bifulco 2012	Dependent	NECG
	Dong and Lipsey 2014	Dependent	NECG
	Fortson, Gleason, et al. 2015	Dependent	NECG
	Fortson, Verbitsky-Savitz, et al. 2012	Dependent	NECG/ITS
	Gill et al. 2016	Dependent	NECG
	Gleason, Resch, and Berk 2012	Dependent	RDD
	Hallberg, Cook, et al. 2016	Both	NECG
	Hallberg, Wong, and Cook 2016	Dependent	NECG
	Jaciw 2016a	Dependent	NECG
	Jaciw 2016b	Dependent	NECG
	Jacob et al. 2016	Dependent	ITS
	Leow, Wen, and Korfmacher 2015	Dependent	NECG
	Lottridge, Nicewander, and Mitzel 2011	Dependent	NECG
	Luellen, Shadish, and Clark 2005	Independent	NECG
	Moss, Yeaton, and LIoyd 2014	Dependent	RDD
	Padgett et al. 2010	Dependent	NECG
	Pohl et al. 2009	Independent	NECG
	Shadish, Clark, and Steiner 2008	Independent	NECG
	Shadish, Galindo, et al. 2011	Independent	RDD
	Somers et al. 2013	Dependent	ITS
	St.Clair, Cook, and Hallberg 2014	Dependent	ITS
	St.Clair, Hallberg, and Cook 2016	Dependent	ITS
	Steiner, Cook, Li, et al. 2015	Independent	NECG
	Steiner, Cook, Shadish, and Clark 2010	Independent	NECG
	Steiner, Cook, and Shadish 2011	Independent	NECG
	Tang et al. 2017	Dependent	RDD
	Wilde and Hollister 2007	Dependent	NECG
	Zhou and Xie 2016	Dependent	NECG
Environment			
	Ferraro and Miranda 2014	Dependent	NECG/ITS
	Wichman and Ferraro 2017	Dependent	ITS
Job Training			
	Bell et al. 1994	Dependent	NECG
	Black, Galdo, and Smith 2007	Dependent	RDD
	Bloom, Michalopoulos, and C. Hill 2005	Dependent	NECG/ITS
	Bloom, Michalopoulos, C. Hill, and Lei 2002	Dependent	NECG/ITS
	Dehejia and Wahba 2002	Dependent	NECG

continued

Table 1: All Known Within-Study Comparisons

Field	*Study*	WSC Design	NE Design
	Dehejia and Wahba 1999	Dependent	NECG
	Fraker and Maynard 1987	Dependent	NECG/ITS
	Friedlander and Robins 1995	Dependent	ITS
	Heckman and Hotz 1989	Dependent	NECG/ITS
	Heckman, Ichimura, Smith, et al. 1998	Dependent	NECG/ITS
	Heckman, Ichimura, and Todd 1997	Dependent	NECG/ITS
	Lalonde 1986	Dependent	NECG/ITS
	Lee 2006	Dependent	NECG
	Michalopoulos, Bloom, and C. Hill 2004	Dependent	NECG/ITS
	Olsen and Decker 2001	Dependent	NECG
	Peikes, Moreno, and Orzol 2008	Dependent	NECG
	Smith and Todd 2005	Dependent	NECG
Health			
	Anglin et al. 2018	Dependent	ITS
	Bratberg, Grasdal, and Risa 2002	Dependent	NECG/ITS
	Fretheim et al. 2015	Dependent	ITS
	J. L. Hill, Reiter, and Zanutto 2005	Dependent	NECG
	Schneeweiss et al. 2004	Dependent	ITS
	Steventon, Grieve, and Sekhon 2015	Dependent	NECG
	Wing and Cook 2013	Dependent	RDD
Immigration			
	McKenzie, Stillman, and Gibson 2010	Dependent	NECG/IV/ITS
Political Science			
	Arceneaux, Gerber, and Green 2010	Dependent	NECG
	Green et al. 2009	Dependent	RDD
	Keele and Titiunik 2015	Dependent	RDD

This list includes all known within-study comparisons including working papers and paper presentations (where an unpublished version of the study is unavailable). We do not include simulation studies or four-arm designs where the study is intended to estimate the effect of randomization or preference rather than the performance of the non-experimental method. The WSC design column notes whether the researchers used an independent or dependent-arm design (or both if more than one study was conducted). The NE design refers to the primary research design, where NECG = Non-equivalent Comparison Group, RDD = Regression Discontinuity Design, ITS = Interrupted Time Series, and IV = Instrumental Variables. Note that we group all time series designs (including comparative interrupted time series and difference-in-differences) under the ITS label. Where authors combine non-experimental designs, we note the primary design which is tested.