## Abstract Title Page

**Title:**
Methodological Foundations for the Empirical Evaluation of Non-Experimental Methods in Field Settings

**Authors and Affiliations:**
Vivian C. Wong, University of Virginia
Peter M. Steiner, University of Wisconsin, Madison

## Abstract Body

**Background / Context:**
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 traditional WSC design, treatment effects from an RCT 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) approach 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), 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. WSCs can also address questions about the current practice of non-experimental methods. For example, given all the choices that researchers must make in constructing observational comparison groups using propensity score methods (e.g. from the estimation of the propensity score itself to using the propensity score to estimate treatment effects), can researchers – or teams of researchers – from different labs across the country reliably replicate treatment effects for the same dataset using similar matching procedures? Finally, for substantive social and behavioral science researchers, WSCs provide opportunities to check hypotheses about non-experimental methods for addressing selection bias when random assignment is not feasible. At stake is a methodology that allows program and policy evaluators to develop and refine empirically-based "best practices" for non-experimental approaches.

**Significance / Novelty of study:**
Despite the opportunities and reasons for conducting WSCs, the approach is underutilized as a method for improving research practice in social and behavioral science settings. This is because there is no coherent framework for understanding 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.

**Purpose / Objective / Research Question / Focus of Study:**
Because applications of the WSC design are published throughout the social and health sciences, important WSC methodological innovations and findings are unknown and underutilized by evaluators and researchers. This paper addresses this issue by *developing methodological foundations for within-study comparison designs that evaluate non-experimental methods*. It will present a coherent framework that addresses design and analysis issues of WSCs for evaluating non-experimental methods.

**Methodological Approach under Investigation:**
Below, we highlight three key methodological issues related to the design and analysis of WSCs, including choosing an appropriate WSC design, ensuring statistical power for assessing correspondence, and selecting a correspondence metric for comparing experimental and non-experimental results.

WSC Purposes and Designs. The paper highlights three WSC design variants. For each approach, the paper discusses the design's required assumptions, common threats to validity, benefits and limitations of the approach, and causal estimands of interest. The first class of WSC designs is the "three-arm design" because the experiment and non-experiment share either the same treatment group or some portion of the control group. This is the most common design approach and has been used to assess the performance of matching, regression, or RD designs, among others. The second design type is the "four-arm" design. It is a prospectively designed WSC that randomly assigns individuals into experimental and non-experimental conditions (Shadish et al., 2008). Units in the experimental condition are then randomized into a treatment or control condition; units in the non-experiment are allowed to select the intervention of their choice. The third design type is the "synthetic" WSC. Here, experimental data are used to construct the non-experiment, typically by deleting some portion of the experimental treatment or control group. The synthetic design has been applied to the evaluation of both matching and RD approaches. Using empirical examples to demonstrate WSC design options, the paper highlights the advantages and disadvantages of each design approach, and demonstrates the need to select appropriate design options for addressing the research question of interest. For example, synthetic designs are easily replicable, cost-effective, and well designed to address questions about the performance of analytic methods. However, synthetic designs do not involve real-world selection processes (selection is simulated) and, thus, cannot answer questions about real selection mechanisms and implementation problems that researchers are likely to encounter in the field. In this way, the synthetic design is a simulation study with experimental data.

Statistical Power in WSCs. 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 experiment and non-experiment. The paper demonstrates the unique power considerations for assessing correspondence in experimental and non-experimental results. In fact, it shows that WSCs often have greater power requirements than what is needed for detecting effects in an experiment or non-experiment alone. To see this logic, consider a scenario where the criterion for assessing correspondence in experimental and non-experimental effects is to determine whether the two study conditions reject the null hypothesis that the treatment effect is equal to zero. In other words, do the experiment and non-experiment result in the same conclusion? Figure 1 shows that in a WSC design with an independent experiment and non-experiment (e.g. four-arm WSC designs where units were randomly assigned into experimental and non-experimental conditions), the probability of rejecting the null in both study conditions depends on the statistical power in the experiment and the non-experiment. Here, the X-axis is the power in the randomized experiment, the Y-axis is the power on the non-experiment, and the vertical Z-axis is the probability of both the experiment and non-experiment arriving at the same decision of rejecting the Null hypothesis of no treatment effect. The figure shows that a well-powered experiment and non-experiment (that both have statistical power of .80) result in the same pattern of statistical significance with a probability of .68 only (= .8 x .8 + .2 x .2). In cases where the experiment and non-experiment are both underpowered for detecting effects (.20), the probability of obtaining corresponding results is again .68. This is so because both studies reject the null with a probability of .04 (=.2 x .2), but do not reject the null with a probability of .64 (=.8 x .8). This implies that when there is no significant treatment effect and both study conditions are underpowered, researchers may incorrectly interpret correspondence in statistical significance patterns as a lack of bias in the non-experiment. It also suggests that to achieve correspondence in the experiment's and non-experiment's statistical significance with a probability of .8, each arm of the WSC must be powered at .90. The paper discusses power considerations when other WSC designs are used, such as three-arm and synthetic designs, and when other correspondence measures are considered.

Assessing correspondence in experimental and non-experimental results. Finally, in most WSC designs, the researcher's question of interest is whether the non-experimental method produces an unbiased causal treatment estimate for some well-defined population. Correspondence between experimental and non-experimental effects has been assessed in a number of ways. To examine the *policy question* of whether the experiment and non-experiment produce comparable results in field settings, correspondence may be assessed by looking at the direction and magnitude of effects, as well as statistical significance patterns of treatment effects in the experiment and non-experiment. To assess the *methodological question* of whether the non-experiment produces unbiased results in field settings, researchers may look at direct measures of bias by computing the difference in non-experimental and experimental effect estimates, the percent of bias reduced from the initial naïve comparison (Shadish et al., 2008), and the effect size difference between experimental and non-experimental results (Hallberg, Wong, & Cook, under review). However, because of sampling error, even close replications of the same randomized experiment should not result in exactly identical posttest sample means and variances. Therefore, another common approach for assessing correspondence in experimental and non-experimental results is to use statistical tests of differences between non-experimental and experimental results with bootstrapped standard errors to account for covariance in the experimental and non-experimental data when appropriate (e.g. Wilde & Hollister, 2007).

Although statistical tests of difference in experimental and non-experimental results are most common in the WSC literature, a careful consideration of the typical WSC research question suggests serious weaknesses with this approach. In the standard null hypothesis significance testing (NHST) framework, if there is no evidence for a difference in effects, then the researcher fails to reject the null hypothesis. Traditionally, NHST protects against Type I error rates and there is less concern about Type 2 error rates (where the researcher concludes that there is no evidence for a difference when a difference exists). As a result, under the NHST framework, results may be inconclusive when tests of difference are used to assess whether two effect estimates are the same. The paper will propose using *statistical tests of equivalence* for assessing correspondence in WSC results. Although statistical tests of equivalence are used in public health (Barker, Luman, McCauley, & Chu, 2002), psychology (Tyron, 2001), and medicine (Munk, Hwang, & Brown, 2000), these tests are rarely applied in the analysis of WSCs. Tests of equivalence are useful for contexts where a researcher wishes to assess whether a new or an alternative approach (such as a non-experiment) performs as well as the gold standard experimental approach. This paper reviews existing methods of assessing correspondence in WSC designs, and demonstrate equivalence tests as the standard method for assessing correspondence in WSC contexts

**Usefulness / Applicability of Method:**
To demonstrate the feasibility of WSC approaches we highlight two WSCs in education. The first is a large-scale intervention that was developed to improve elementary science teaching and learning. The intervention "provid[es] fourth- and fifth-grade teachers with professional development in summer institutes and ongoing coaching and mentoring in the use of extended, inquiry-based curriculum units for elementary science" (Borman, Gamoran & Bowdon, 2008). These data are currently used in a WSC for evaluating the performance of various multi-level matching approaches. Here, the WSC is a three-arm design, where the experiment and non-experiment (matching study) share the same treatment group. We also use data from an already published WSC that examines the performance of the RD design compared to an experimental benchmark (Shadish, Galindo, Wong, Steiner, & Cook, 2011) using a four-arm design. Using these two datasets, we demonstrate design variations in WSCs, statistical power, and methods for assessing correspondence when the experimental and non-experimental data are independent and dependent.

**Conclusions:**
At stake is a high quality method for identifying real world contexts and conditions in which non-experimental methods succeed in showing "what works" in the social and behavioral sciences. Equally important is a method that alerts us to when non-experimental approaches fail to produce trustworthy results. This paper presents a coherent framework for evaluating the performance of non-experimental methods in field settings. Specifically, it examines WSC design options, highlights statistical power considerations in WSC designs, and presents methods for analyzing WSC designs that assess correspondence in experimental and non-experimental results.

## Appendix A. References


Barker, L.E, Luman, E.T., McCauley, M.M., Chu, S.Y. (2002). Assessing equivalence: An Alternative to the use of difference tests for measuring disparities in vaccination coverage. *American Journal of Epidemiology*, 156(11), 1056-61.

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects. *Journal of Research on Educational Effectiveness, 1*, 237–264.

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.

Hallberg, K., Wong, V.C., & Cook, T.D. (under review). *School Level Matching in Observational Studies.*

Shadish, W.R., Clark, M.H., & Steiner, P.M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment (with comments by Little/Long/Lin, Hill, and Rubin, and a rejoinder). *Journal of the American Statistical Association*, 103, 1334-1356.

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.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371-386.

Tryon, W. W., &Lewis, C. (2008). An Inferential Confidence Interval Method of Establishing Statistical Equivalence That Corrects Tryon's (2001) Reduction Factor. *Psychological Methods, 13*, 272-278.

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.

## Appendix B. Tables and Figures

Figure 1. Probability of achieving the same decision in a WSC design