Abstract Title Page
Not included in page count.

Title: Analyzing Empirical Evaluations of Non-experimental Methods in Field Settings

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

Abstract Body.

Background / Context:
Despite recent emphasis on the use of randomized control trials (RCTs) for evaluating education interventions, in most areas of education research, observational methods remain the dominant approach for assessing program effects. For example, among all studies reviewed by What Works Clearinghouse, only 30% of these studies were RCTs. The most common method for assessing program impacts was non-equivalent comparison group designs, followed by much smaller percentages of studies that involved regression-discontinuity and single case-study designs (Institute for Education Sciences, 2015). Given the widespread use of observational approaches for assessing the causal impact of interventions in program evaluations, there is a strong need to identify which observational approaches can produce credible impact estimates in field settings.

Over the last three decades, the within-study comparison (WSC) design has emerged as a method for evaluating the performance of non-experimental methods in field settings. In a traditional WSC design, treatment effects from a randomized experiment (RE) are compared to those produced by a non-experimental approach that shares the same target population and intervention. The non-experiment may be a quasi-experimental approach, such as a regression-discontinuity design study; it may be an observational study, where comparison units are matched to treatment cases using different statistical procedures or covariates. The goals of the WSC are to determine whether the non-experiment replicates results from a high-quality randomized experiment in field settings. Despite the potential for WSCs to improve non-experimental practice, there is no existing literature that describes the methodological foundations of the analysis for evaluating non-experimental approaches. One important methodological issue in the WSC literature is how the researcher should assess correspondence between experimental and non-experimental results. To date, there is no consensus among WSC analysts on a common method for assessing correspondence, and there has been limited discussion about the relative benefits and limitations of each approach. As a result, the existing WSCs are of heterogeneous quality, with researchers using ad hoc approaches for assessing correspondence in experimental and non-experimental results, which may or may not be appropriate for addressing the research question of interest.

Purpose / Objective / Research Question / Focus of Study:
The purpose of this paper is two-fold. First, it is to examine methods for assessing correspondence in experimental and non-experimental results in WSC designs, and their relative advantages and limitations. Second, it is to provide methods for calculating statistical power in WSC designs that evaluate the empirical performance of non-experimental methods. We apply multiple correspondence criteria to two WSC examples, highlighting contexts and conditions under which these methods converge at the same result, but also when they differ. Finally, we address correspondence criteria for WSC designs with independent experimental and non-experimental conditions, and for cases when the conditions have dependent data structures (Wong & Steiner, in progress).

Significance / Novelty of study:
Despite the growing interest in WSCs and questions regarding the internal validity of these results, there is still no coherent framework describing the analysis of WSCs for evaluating non-

experimental methods. To date, only Cook, Shadish, and Wong (2008) address theory for improving the quality of WSCs. In their guidelines for internally valid WSC designs, they urged that WSC analysts ensure that there are clear criteria for inferring correspondence in experimental and NE results. However, they did not indicate what those criteria should be, nor did they discuss the advantages and disadvantages of the approaches. The paper will examine methods for assessing correspondence, and in particular, recommend using statistical tests of equivalence for assessing correspondence in WSC results. Moreover, the current literature on empirical evaluations of non-experimental methods does not yet address the statistical power for assessing correspondence in WSC designs, nor does it provide guidelines on sample size requirements of WSC designs.

Usefulness / Applicability of Method:

The paper discusses the properties of several correspondence criteria and guides researchers in choosing the most appropriate criteria. We demonstrate the statistical performance of each method when the WSC design includes data that are dependent and independent, and at different levels of statistical power (low, medium, and high power) in both the experiment and non-experiment. The correspondence criteria and power issues discussed are relevant not only for within-comparisons but also for assessing the replication success of REs, that is, did replications of a specific RE design succeed in reproducing the same findings? Thus, the findings of this study are relevant to methodologists, meta-analysts, and practical researchers.

Statistical Approach & Findings/Results:

Aim 1: Methods for assessing correspondence in within-study comparison (WSC) designs. 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 NE effects has been assessed in a number of ways. To examine the policy question of whether the experiment and NE produce comparable results in field settings, correspondence may be assessed by looking at the (1) direction and (2) magnitude of effects, as well as (3) statistical significance patterns of treatment effects in the experiment and non-experiment. Regarding statistical significance patterns, we consider type I and II error rates, as well as S and M type error rates, where the former is concerned about errors related to incorrect conclusions about the sign of the effect and the latter is related to errors about the magnitude of the effect (Gelman & Tuerlinckx's, 2000). To assess the methodological question of whether the NE produces unbiased results in field settings, researchers may look at (4) direct measures of bias by computing the difference in NE 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 NE results (Hallberg, Wong, & Cook, under review). However, because of sampling error, even close replications of the same experiment should not result in exactly identical posttest sample means and variances. Wing and Cook (in press) address this concern by using bootstrapping the mean squared error to assess comparability of their results.

Later WSCs conducted (5) statistical tests of differences between experiment and non-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. Traditionally, standard null hypothesis significance testing (NHST) protects against Type I error rates and there is less concern about Type 2 errors (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. In cases where the experimental and NE effects are extremely different and the WSC is adequately powered, then there may be evidence to reject the null of equivalent means. But when treatment effect estimates are similar, or when power is low in the WSC, then the researcher may be uncertain how to interpret the lack of evidence to reject the Null. It could mean that there is no difference in treatment effect estimates between the experiment and non-experiment, or it could mean a Type 2 error, particularly when the WSC was underpowered for detecting effects. The paper recommends using (6) *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. Examples of such research questions include, "Can the RD design replicate experimental benchmark results?" and "Does local and focal matching of units perform as well as RCTs in producing unbiased treatment effect estimates?"

Aim 2. Statistical power for assessing correspondence in WSC designs. Another critical issue in the planning of WSCs is ensuring that the design has sufficient statistical power for assessing correspondence in experimental and non-experimental results. The paper demonstrates the unique power considerations for assessing correspondence in experimental and non-experimental results. In fact, it will show 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 both study conditions either reject or accept the null hypothesis of a zero treatment effect. In other words, do the experiment and non-experiment result in the same conclusion? In a WSC design with an independent experiment and non-experiment (e.g. 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, 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). But 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 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. Figure 1 shows the corresponding probabilities (of obtaining the same significance pattern from both studies) as a function of RE's and NE's actual power. Similar plots will be presented for all correspondence criteria.

We examine sample size requirements for all six methods of assessing correspondence in experimental and non-experimental results described above. Thus far, our work has shown sample size requirements for both independent and dependent WSC designs when the effect size differences are .2, .15, .10, and .05 standard deviations (at 80% power). We show that for an

independent WSC design to detect an effect size difference of .2 standard deviations (80% power) requires a sample size that is 33% larger than what is needed for a WSC design where the experimental and non-experimental conditions are dependent (Figure 2). The final paper will present methods for determining statistical power of WSC designs, depending on the type of WSC design (dependent and independent experimental and non-experimental conditions) and the method for assessing correspondence in results.

Moreover, the paper demonstrates methods for assessing correspondence using results from two already published WSC papers. The first example comes from an already published WSC that examines the performance of the RD design compared to an experimental benchmark (Shadish, Galindo, Wong, Steiner & Cook, 2012). In this study, students were randomly assigned to an experimental or RD arm. In the experimental arm, students were again randomly assigned to a mathematics or vocabulary intervention; in the RD arm, students were assigned to a mathematics or vocabulary intervention on the basis of a pretest score and cutoff. Achievement scores on reading and math tests were examined. The Shadish et al. dataset is an example of a WSC design where the experimental and non-experimental data are independent.

The second example takes advantage of experimental data from the Cash and Counseling Experiment, which evaluated the effects of a "consumer-directed" care program on Medicaid recipients' outcomes. The data include monthly Medicaid expenditures for 12 months prior to the intervention (pretest), and 12 months after the intervention (posttest). Medicaid participants in Arkansas, New Jersey, and Florida were randomly assigned to treatment and control conditions, where the treatment consisted of Medicaid recipients selecting their own services using a Medicaid-funded account and the control consisted of local agencies selecting services for Medicaid recipients. For the WSC, McConeghy, Steiner, Wing, and Wong (2013) used experimental data to create a simple interrupted time series (ITS) design within each of the three states by first deleting control group information, and then by estimating ITS effects for only the RCT treatment group members by looking at changes in the intercept and slope once the intervention is introduced. They also evaluated the performance of a comparative ITS design by using data from the experimental controls in other states to form two comparison groups. This study is an example of the synthetic WSC design, and we will use it to demonstrate power considerations in WSCs, and for assessing correspondence between experimental and non-experimental results.

Conclusions:

The paper demonstrates that the assessment of correspondence between RE and NE results is challenging. First, a broad variety of correspondence criteria exists which have their own strengths and weaknesses. Thus, the choice of an appropriate or multiple correspondence criteria is crucial. Second, sampling uncertainty in both the RE and NE estimates frequently results in underpowered WSCs even when both studies have a power of .8. Thus, power considerations are important for designing and conducting WSCs of high quality. This paper provides guidelines for the choice of correspondence criteria and determining required sample sizes for sufficiently power WSCs.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

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.

Berk, R., Barnes, G., Ahlman, L., & Kurtz (2010). When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology, 6*(2), 191-208.

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.

Gelman, A. & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics.* 15. 373-390.

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.

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.

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.

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.*

Appendix B. Tables and Figures
Not included in page count.

Figure 1. Probability of drawing the same conclusions from the RE and NE (conclusions are based on standard null-hypothesis significance testing, i.e., reject or accept).

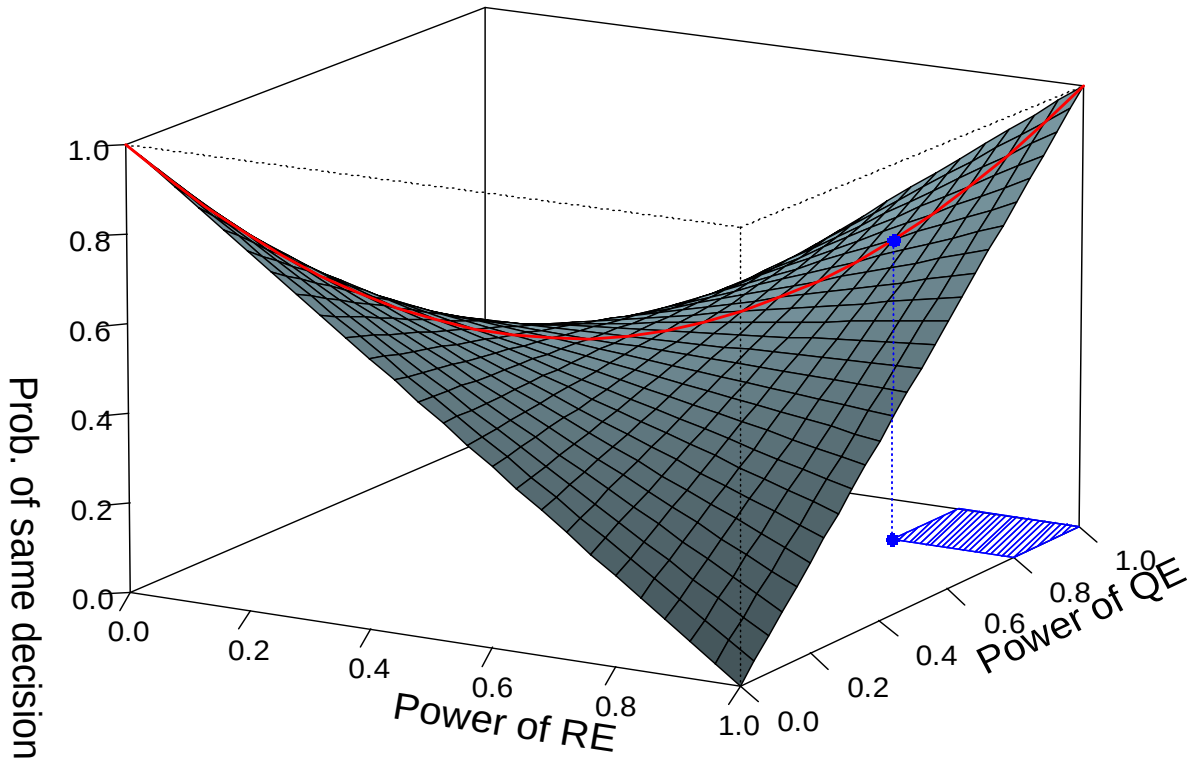

Figure 2. Sample size requirements for a WSC with independent RE and NE arms.