Working Paper:

Assessing Correspondence between Experimental and Non-Experimental Results in Within-Study-Comparisons

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

In within-study comparison (WSC) designs, treatment effects from a non-experimental design, such as an observational study or a regression-discontinuity design, are compared to results obtained from a well-designed randomized control trial (RCT) with the same target population. The goal of the WSC is to assess whether the non-experimental and experimental designs yield the same results in field settings, and the contexts and conditions under which non-experimental methods replicate benchmark RCT estimates. A common analytic challenge with WSCs, however, is identifying appropriate criteria for determining whether non-experimental and benchmark results replicate. This paper examines methods for assessing correspondence in benchmark RCT and non-experimental results in WSC designs. We examine measures for assessing correspondence in results, and the relative advantages and limitations of these approaches. We identify two classes of measures: conclusion- and distance-based correspondence approaches. Conclusion-based measures indicate correspondence in results if a researcher or policy-maker draws the same conclusion from the experiment and non-experiment. Distance-based measures investigate whether the difference in estimates is small enough to claim correspondence between methods. We use a simulation study to examine the statistical properties of correspondence measures, and recommend a new approach that combines traditional significance testing and equivalence testing in the same framework. The paper concludes with practical advice on assessing and interpreting results in WSC contexts.

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

Updated April 2016

EdPolicyWorks
University of Virginia
PO Box 400879
Charlottesville, VA 22904

EdPolicyWorks working papers are available for comment and discussion only. They have not been peer-reviewed. Do not cite or quote without author permission. Working paper retrieved from:
http://curry.virginia.edu/uploads/resourceLibrary/46_Analysis_of_Within_Study_Comparisons.pdf

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

Working Paper © 2016 Rector and Visitors of the University of Virginia. For more information please visit www.curry.virginia.edu/edpolicyworks or contact EdPolicyWorks@virginia.edu

ASSESSING CORRESPONDENCE BETWEEN EXPERIMENTAL AND NON-EXPERIMENTAL RESULTS IN WITHIN-STUDY-COMPARISONS

Peter M. Steiner & Vivian Wong

INTRODUCTION

In 1986, LaLonde introduced a research design called the within-study comparison (WSC) to empirically evaluate the performance of non-experimental methods in field settings. In his approach, he used data from a randomized control trial (RCT) of the National Supported Work Demonstration (NSW) to estimate a causal benchmark effect, and compared the result to those obtained from an observational study with the same target population. The goal was to determine whether observational methods, such as cross-sectional regression and difference-in-differences models, replicated RCT benchmark results in field settings. LaLonde concluded that observational methods failed to replicate comparable results to those obtained from the RCT benchmark. His conclusions were based on mostly heuristic approaches. He compared the direction of treatment effects from the experiment and non-experiment, the magnitude of treatment effects obtained from each study design, and the statistical significance patterns of treatment effects. Moreover, LaLonde looked at the size of the difference in non-experimental and experimental results – judging a treatment effect difference of $1600 in earnings as being substantially large, and a difference of $600 as being less consequential (p. 617).

Other WSCs soon followed, including studies by Fraker and Maynard (1987), Friedlander and Robins (1995), Bell, Orr, Bloomquist, and Cain (1995), Heckman, Ichimura, Smith, and Todd (1998), among others. Many of these studies took advantage of data from RCT job training evaluations and extant datasets that shared similar outcome measures of earnings. However, the criteria for assessing correspondence in benchmark and non-experimental results varied from study to study, with some looking at correspondence in size of treatment effects and statistical significance patterns in the RCT and non-experiment, while others constructed bias measures of the non-experiment, or tested directly statistical differences in benchmark and non-experimental results. Given the different correspondence measures used for evaluating non-experimental performance, interpreting WSC results across studies proved challenging. In an early meta-analysis of WSC results, Glazerman, Levy, and Myers' (2003) limited their analysis to include only studies that used job training data with earnings outcomes. They established a correspondence criterion of a $1,000

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

difference in benchmark and non-experimental results because a "difference of $1,000 or more can make a dramatic difference in the policy recommendation" (p. 74).

But is $1,000 an appropriate threshold for evaluating the success of non-experimental methods in all job training field settings? What about WSC cases where the outcome is not in earnings, but in test scores (Wilde & Hollister, 2007; Aiken, West, Schwalm, Carroll, & Hsuing, 1998), voter participation rates (Arceneaux, Gerber, & Green, 2010), or criminal recidivism (Berk, Barnes, Ahlman, & Kurtz, 2010)? And, how does one consider correspondence in WSC designs when even two perfectly implemented study designs will produce different effects because of sampling and measurement error? At issue are valid and consistent criteria for evaluating the performance of non-experimental methods in within-study comparison designs.

Since LaLonde's (1986) study, more than 60 WSCs have emerged evaluating the performance of non-experimental methods in field settings. These studies have spanned the disciplines of international development, job training, education, political science, environmental policy, health policy, and criminology. They have examined the performance of non-experimental and quasi-experimental approaches, including regression-adjustment, regression-discontinuity designs, matching methods, and repeated measures approaches. Results from these studies have had profound influence on funding policy and research methodology in program and policy evaluation. For example, the Department of Education (Paige, 2005), Office of Management and Budget (2004), and What Works Clearinghouse (Institute for Economic Studies, 2005) have cited results from WSC studies for guidance on methodology choice in program evaluation.

Because of strong methodological and policy interests in WSC results, recent research has focused on improving the design of these empirical tests of methods. Wong and Steiner (under review) formalize different types of WSC designs and identify the assumptions needed for WSCs to provide valid, causally interpretable results. Less clear, however, are appropriate methods for analyzing WSC designs. Thus far, WSC analysts have relied on a constellation of ad hoc approaches for judging correspondence in benchmark and non-experimental results. But this has resulted in tremendous variation in how researchers and policy-makers have interpreted WSC results, subjecting the validity of these conclusions to unconscious, or conscious, researcher bias. Moreover, there is no methodological guidance for the contexts and conditions under which different correspondence metrics should be adopted in the analysis of WSC designs. When is it desirable to just compare the direction and magnitude of benchmark and non-experimental results, as opposed to looking at the difference in benchmark and non-experimental estimates? When should statistical tests of difference

in benchmark and non-experimental results be used as the criterion for making judgments? And, how do the size of the true treatment effect, the direction of selection bias in the observational study, and statistical power of the WSC affect the performance of the different correspondence criteria?

This paper has three aims for improving the analysis and interpretation of within-study comparison designs for evaluating non-experimental methods. The first is to describe the multiple methods for assessing correspondence in WSC studies, including the purposes of each method as well as the advantages and disadvantages of the approach. The second is to present results from a Monte Carlo simulation that demonstrates the contexts and conditions under which different correspondence criteria produce interpretable results. The third is to propose a new correspondence test for evaluating whether non-experimental methods succeed in replicating benchmark results in field settings. The paper argues for the use of a correspondence measure that incorporates statistical tests of difference and equivalence within the same framework (Tyron, 2001; Tyron & Lewis, 2008). The paper concludes by offering WSC researchers with practical recommendations for the analysis and interpretation of WSCs.

Finally, in this paper, we distinguish between two types of WSC designs. The first approach is the "independent WSC design," where units are randomly assigned to a randomized experiment or a non-experimental condition (Shadish, Clark, & Steiner, 2008). In the benchmark arm, units are randomly assigned again to treatment or control conditions, while in the non-experiment, units select into a preferred treatment or control condition. Correspondence is assessed by comparing the average treatment effect from each design. The second approach is called the "dependent WSC design." This approach was implemented by LaLonde (1986), where some portion of the benchmark sample is shared between RCT and non-experimental arms of the WSC. For example, units may select into a randomized experiment, and within the RCT, units are randomly assigned to treatment and control conditions. The non-experiment is constructed by matching units who did not enter the benchmark RCT, but share the same outcome measures and were not exposed to the treatment. Figure 1 summarizes the independent and dependent arm designs, and Wong and Steiner (under review) describe these designs in further detail.

CORRESPONDENCE MEASURES IN WSC DESIGNS

Although the existing WSC literature lacks consensus on what constitutes "correspondence" in benchmark and non-experimental estimates, multiple WSC researchers have noted the challenge with interpreting results. Wilde and Hollister (2007) introduce an early discussion of correspondence

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

measures in their paper titled "How close is close enough?" In the paper, the authors use RCT data from the Tennessee Class Size experiment to examine the performance of propensity score methods. To judge "how close is close enough" between benchmark and non-experimental results, Wilde and Hollister compared the direction, size, and statistical significance patterns of treatment effects obtained from the RCT and non-experimental study, as well as conducted direct statistical tests of difference in benchmark and non-experimental results. However, Wilde and Hollister also advocated for substantive thresholds to assess correspondence in benchmark and non-experimental results. Using results from Krueger's (1999) cost-benefit analysis of the Tennessee Class Size Experiment, Wilde and Hollister defined an impact threshold of 5.4 percentiles (or .20 standard deviations) as a treatment effect large enough to warrant a policy change. They then evaluated the non-experiment's performance by examining whether both the RCT and propensity score approaches produced treatment effects as large as 5.4 percentiles to produce the same policy decision for adopting the intervention.

Cook, Shadish, and Wong (2008) also urged researchers to attend to correspondence measures in the analysis of WSC designs. They argued that for WSC designs to produce interpretable results, the approach requires consistent and clear standards for judging when non-experimental methods succeed in replicating benchmark estimates. Although their guidance did not specify which approaches WSC analysts should use for evaluating non-experimental methods, they suggested that the criteria should include a statistical threshold – that is, a formal statistical test of benchmark and non-experimental results to account for chance differences from sampling error.

In this paper, we argue that the types of correspondence measures used, and the criteria for determining correspondence, depend on the purpose of the WSC design itself. In general, WSC designs address two types of research questions. The first set of WSC questions examines the *policy issue* of whether the RCT and non-experiment produce comparable results in field settings. Here, the goal is to assess whether the policy-maker would draw the same conclusions from the RCT and non-experiment. To address this question, *conclusion-based correspondence measures* are useful. These measures include looking at the direction and magnitude of effects, as well as statistical significance patterns of treatment effects in the experiment and non-experiment. These measures require both substantive and statistical criteria for evaluating non-experimental performance.

Other WSC research questions are methodological in nature – they test whether the non-experimental method produces unbiased results in field settings. Appropriate measures for

addressing this research question estimate the size of non-experimental bias in field settings.[1] For example, the researcher may compute the difference in non-experimental and experimental effect estimates, or the percentage of bias reduced by the non-experimental method. We distinguish this class of measures as *distance-based correspondence measures*. For these measures, substantive criteria often are helpful for interpreting whether the magnitude of the bias estimates is of practical relevance for the researcher. To account for sampling error in distance-based measures, the researcher also may use statistical tests of difference or equivalence between non-experimental and experimental results.

Below, we formalize approaches for implementing conclusion-based and distance-based measures in WSC designs. We also discuss the relative advantages and limitations of each approach, and demonstrate their performance through a Monte Carlo experiment in the following section.

Conclusion-based Correspondence Measures for Addressing Policy Relevant Questions

Direction of effects. For this measure, the WSC researcher assesses correspondence by looking at whether the experiment and non-experiment produce treatment effects in the same direction, that is, whether the treatment effects have the same sign. The presumption here is that even the direction of an effect influences policy-makers' impression about the efficacy of a program. Thus, correspondence between the experimental and non-experimental estimates, $\hat{\tau}_{re}$ and $\hat{\tau}_{nx}$, is achieved if the signs of the treatment effect are identical: $\text{sgn}(\hat{\tau}_{re}) = \text{sgn}(\hat{\tau}_{nx})$, where sgn($x$) is the sign function, with sgn(x) = 1 if $x > 0$, sgn(x) = -1 if $x < 0$, and sgn(x) = 0 if $x = 0$. The conclusion-based (C in the subscript) correspondence in directions (D) is defined as:

$$C_C^D = 1[\text{sgn}(\hat{\tau}_{re}) = \text{sgn}(\hat{\tau}_{nx})] \ , \qquad (1)$$

where, the indicator function 1[.] returns 1 and indicates correspondence if the logical comparison is true, and 0 if false. Thus, $C_C^D = 1$ if the experimental and non-experimental effect estimates have the same sign, and $C_C^D = 0$ if the signs of the effect differ (or one of the signs is exactly zero).

Magnitude of effects. Researchers may also assess correspondence by comparing the size of the effect estimates produced by the experiment and non-experiment. However, one limitation of this approach is the subjectiveness of assessing "How close is close enough?" In this approach, the researcher decides *in advance* a threshold value (λ) that must be exceeded for the treatment effect to be meaningful. Here, "meaningful" may be based on results from a cost-benefit analysis (Wilde & Hollister, 2007), or an effect size that has been determined to be scientifically or programmatically

[1] Wong & Steiner (under review) identify conditions under which estimates of non-experimental bias are valid.

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

relevant for the context under investigation. Correspondence is achieved if both the experimental and non-experimental treatment effect estimates are either above or below a substantively meaningful threshold λ. Thus, correspondence in magnitude (*M*) is given when:

$$C_C^M = 1[(\hat{\tau}_{re} \geq \lambda \;\&\; \hat{\tau}_{nx} \geq \lambda) \text{ or } (\hat{\tau}_{re} < \lambda \;\&\; \hat{\tau}_{nx} < \lambda)]. \quad (2)$$

If both estimates are above a positive threshold λ, then correspondence according to C_C^M implies correspondence in signs, such that for $\lambda > 0$ and $(\hat{\tau}_{re} \geq \lambda \;\&\; \hat{\tau}_{nx} \geq \lambda)$ we have $(C_C^M = 1) \Rightarrow (C_C^D = 1)$. Correspondence in signs is not implied if both estimates fall below a positive threshold. Similarly, for a negative threshold, if $\lambda < 0$ and $(\hat{\tau}_{re} < \lambda \;\&\; \hat{\tau}_{nx} < \lambda)$ we have $(C_C^M = 1) \Rightarrow (C_C^D = 1)$.

Statistical significance patterns. Comparing the significance patterns of estimates is one of the most popular methods for assessing correspondence between experimental and non-experimental results. Here, the decision criterion about a program's efficacy is based on statistical significance rather than a threshold value for a meaningful impact. The experimental and non-experimental estimates have identical significance patterns with regard to the nil hypothesis $H_0{:}\,\tau = 0$ if (a) both estimates have the same sign and are statistically significant, or (b) both estimates are insignificant according to a two-tailed hypothesis test. Formally, correspondence in significance (*S*) is measured as:

$$C_C^S = 1[\{(\text{sgn}(\hat{\tau}_{re}) = \text{sgn}(\hat{\tau}_{nx}) \;\&\; p_{re} \leq \alpha \;\&\; p_{nx} \leq \alpha\} \text{ or } (p_{re} > \alpha \;\&\; p_{nx} > \alpha)], \quad (3)$$

where p_{re} and p_{nx} are the two-tailed *p*-values of the significance test (typically a *t*-test) for the effect estimates of the experiment and non-experiment, respectively, and *a* is the type-I error rate. Thus, correspondence in significance depends on the size and direction of the point estimates, their standard errors, the type-I error rate of rejecting the null hypothesis and each study's power for detecting the unknown true treatment effect. The power for detecting a minimum effect size is irrelevant here—what matters is the power for detecting the unknown true effect. The correspondence measure may be adapted to reflect a one-tailed hypothesis test that allows us to drop the sign condition. However, since almost all significance tests are conducted as two-tailed tests in practice, we formulated correspondence in significance in terms of a two-tailed test.

The correspondence measure C_C^S also works for null hypotheses other than the nil hypothesis. For instance, we may use the threshold value λ from the magnitude-related measure and test whether both effects, τ_{re} and τ_{nx}, are significantly different from λ $(H_0{:}\,\tau = \lambda)$. In order

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

to accommodate the more general case of hypothesis testing, we only need to change the sign condition in C_C^S to $\text{sgn}(\hat{\tau}_{re} - \lambda) = \text{sgn}(\hat{\tau}_{nx} - \lambda)$. Then, correspondence in C_C^S due to two significant effect estimates implies correspondence in magnitudes and signs: $(C_C^S = 1) \Rightarrow (C_C^M = 1) \Rightarrow (C_C^D = 1)$. However, correspondence in C_C^S due to two insignificant estimates neither implies correspondence in magnitudes nor in signs.

Combined measures. In some WSC contexts, it may be useful to combine multiple conclusion-based correspondence measures. For instance, consider the combination of magnitude- and significance-related measures, for example $C_C^{MS} = 1[C_C^M = 1 \;\&\; C_C^S = 1]$, which requires that both C_M and C_S indicate correspondence. However, this measure has the disadvantage that it may signal correspondence even in ambiguous situations, such as when both estimates exceed the threshold (suggesting a positive effect), but they are not significantly different from zero ($H_0: \tau = \lambda$). A more useful correspondence measure may be constructed if we use only parts of the C_C^S (with $H_0: \tau = 0$) and C_C^M criteria:

$$C_C^{MS'} = 1[\{(\hat{\tau}_{re} \geq \lambda \;\&\; \hat{\tau}_{nx} \geq \lambda) \;\&\; (p_{re} \leq \alpha \;\&\; p_{nx} \leq \alpha)\} \text{ or } \{(\hat{\tau}_{re} < \lambda \;\&\; \hat{\tau}_{nx} < \lambda) \;\&\; (p_{re} > \alpha \;\&\; p_{nx} > \alpha)\}]$$

Correspondence is achieved if either both estimates are significantly different from zero and exceed the threshold or both estimates are insignificant and fall below the threshold.[2]

One challenge with conclusion-based correspondence measures discussed above is that they may indicate a lack of correspondence even in cases where estimates are identical or very similar. For example, if the experimental effect estimate is slightly greater than zero and the non-experimental estimate is slightly less than zero, then the measure may indicate a lack of correspondence in the direction of the effects, even though the point estimates themselves may be considered as equivalent. In another example, the experimental and non-experimental estimates may be exactly identical in the point estimates, but the benchmark result is statistically insignificant while the non-experimental result is significant. Here, C_C^S will indicate lack of correspondence which could be due to the insufficient power (sample sizes) in the benchmark design – not because of poor performance in the non-experimental method itself. Although conclusion-based measures inform researchers

[2] It is also possible to construct asymmetric correspondence measures, but we caution against using such measures because they may result in a misleading comparison of experimental and non-experimental methods.

about whether a policy-maker would arrive at the same decision from a benchmark and non-experimental design, these measures may be less useful for assessing the performance of the non-experimental method itself. In these cases, distance-based measures may help interpret WSC results.

Distance-based Correspondence Measures for Addressing Methodological Questions

In this section, we describe correspondence measures for evaluating methodological questions about non-experimental performance in field settings. These measures all involve some variant of a distance-based metric that calculates the size of the difference in non-experimental and experimental results. We begin by introducing the most frequently used distance measures, followed by a discussion of descriptive and inferential correspondence measures.

Distance measures. A common metric for examining the discrepancy between a WSC's experimental and non-experimental estimate is the difference in effect estimates: $D_M = \hat{\tau}_{nx} - \hat{\tau}_{re}$. Given a valid benchmark estimate from the randomized experiment, the difference is interpreted as bias in the non-experimental estimate (Wong & Steiner, under review). When the treatment group is shared between the experimental and non-experimental arm of a dependent WSC design, the analyst may estimate the difference in outcomes for the independent non-experimental comparison and benchmark control groups: $D_M = \hat{\tau}_{nx} - \hat{\tau}_{re} = (\bar{Y}^c_{nx} - \bar{Y}^t_{re}) - (\bar{Y}^c_{re} - \bar{Y}^t_{re}) = \bar{Y}^c_{nx} - \bar{Y}^c_{re}$. Here, $\bar{Y}^c_{nx}$ and $\bar{Y}^c_{re}$ are the mean outcomes of the non-experimental comparison and experimental control groups, respectively, and $\bar{Y}^t_{re}$ is the mean outcome of the experimental treatment groups.

As more applications of WSCs have emerged, other distance measures that standardize the raw distance D_M have been employed. Below, we list the most important and common standardized measures of non-experimental bias.

(a) *Standardized difference in effect estimates*, $D_S = \dfrac{\hat{\tau}_{nx} - \hat{\tau}_{re}}{s_{re}}$, where s_{re} is the sample standard deviation of the outcome in the experimental control group or the pooled standard deviation of the experimental treatment and control group (Shadish et al., 2011; Hallberg et al., in progress).[3] Because the standardized difference, D_S, is in effect size units, researchers may interpret the distance on a common metric across multiple WSC studies. This is particularly useful for

[3] In estimating the standard deviation, only data from the randomized experiment should be used because the standard deviations of the non-experimental treatment and comparison groups often depend on the strength of the selection mechanisms. This is because strong selection processes in the observational study may result in relatively more homogenous treatment and comparison groups than in a randomized experiment. Thus, the standard deviation of the non-experimental control group is usually not representative of the population standard deviation of the potential control outcome.

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

qualitative and quantitative syntheses, or in fields such as education, where the effect size is a common metric for interpreting results.

(b) *Percentage of remaining bias*, $D_{RB} = \frac{\hat{\tau}_{nx} - \hat{\tau}_{re}}{\hat{\tau}_{pf} - \hat{\tau}_{re}} \times 100$, represents the bias remaining after the non-experimental adjustment in percent of the initial bias, where $\hat{\tau}_{pf}$ is the prima facie effect from the non-experimental study (i.e., the unadjusted initial effect estimate; Shadish et al., 2008). The denominator represents the initial bias (with reference to the experimental estimate) and the numerator the remaining bias after the non-experimental adjustment. A negative percentage indicates an over-adjustment relative to the benchmark estimate. If the WSC analyst is not interested in over- or under-adjustments from the non-experimental method, then she may use the absolute value of the remaining bias percentage, $|D_{RB}|$.

The WSC analyst also may compute the percentage of the bias reduced by the non-experimental adjustment, such that: 100 – D_{RB}. For individual WSC studies, D_{RB} is a useful measure for examining how strongly the non-experimental method reduces or increases initial bias relative to the experimental estimate. Across WSC studies, however, D_{RB} is challenging to interpret given the heterogeneity of the initial bias across different WSCs. That is, D_{RB} can take on extreme positive or negative values when the prima facie effect of the non-experiment is close to the experimental estimate ($\hat{\tau}_{pf} - \hat{\tau}_{re} \approx 0$). This happens if treatment selection in the non-experiment is weak, or if selection introduces positive and negative bias that approximately offset each other. Therefore, this distance measure should be used only when the initial bias is sufficiently large.

(c) *Percent difference*, $D_{DP} = \frac{\hat{\tau}_{nx} - \hat{\tau}_{re}}{|\hat{\tau}_{re}|} \times 100$, expresses the difference in experimental and non-experimental effect estimates in percent of the absolute value of the experimental estimate (Wilde & Hollister, 2007). As with the D_{RB} measure, the challenge with D_{DP} is that the distances can be extremely large if the experimental estimate is close to zero ($\hat{\tau}_{re} \approx 0$).

While all distance measures assess the discrepancy between non-experimental and experimental estimates, how "close" should these results be for a WSC researcher to judge that the two estimates replicate? Even in cases where the benchmark and non-experiment perform equally well (i.e., both designs produce unbiased estimates in the given context), one would not expect the two estimates to be identical given sampling and measurement error. Thus, without some threshold

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

values for describing meaningful differences in experimental and non-experimental results, it is often difficult to assess correspondence using distance measures alone. One solution is to define a meaningful threshold on the distance measure, another is to test the difference or equivalence in the experimental and non-experimental estimates.

Magnitude of difference. With respect to the differences discussed above, the WSC researcher may define a threshold δ_D and claim correspondence in experimental and non-experimental estimates if the absolute difference is less than or equal to the threshold, such that

$$C_D^D = 1[|D| \leq \delta_D]. \quad (4)$$

Here, D is one of the distance measures discussed above (D_M , D_S , D_{RB} , D_{DP}), and δ_D is the threshold with respect to distance measure D. For example, for the standardized distance measure D_S, one could use a threshold of $\delta_D = .1$ and claim correspondence as long as the standardized mean difference does not exceed .1 SD. Although this approach is intuitively straight-forward, it fails to account for the uncertainty due to sampling or measurement error in the outcome.

Insignificance of difference. Null hypothesis significance testing (NHST) helps in assessing whether the observed difference in effect estimates is due to systematic differences in methods or to random sampling error. For an independent WSC design where units are randomly assigned into experimental and non-experimental conditions (Wong & Steiner, under review), a researcher may assess correspondence through a two-sample t-test of the difference in the experimental and non-experimental treatment effect estimates. The standard null hypothesis is that there is no difference between experimental and non-experimental treatment effects, $H_0: \tau_{nx} - \tau_{re} = 0$. The test statistic is $t = \frac{\hat{\tau}_{nx} - \hat{\tau}_{re}}{s}$, where $s = \sqrt{s.e._{nx}^2 + s.e._{re}^2}$ is the standard deviation of the difference, $s.e._{nx}^2$ is the variance (squared standard error) of the non-experimental treatment effect and $s.e._{re}^2$ is the variance of the experimental treatment effect. The significance-related correspondence measure may be defined as:

$$C_D^S = 1[p_D > \alpha]. \quad (5)$$

Here, p_D is the two-tailed p-value from the corresponding t-test. Correspondence is achieved if the p-value is greater than the type-I error rate a such that the null hypothesis cannot be rejected. Since the correspondence measure C_D^S is a function of the p-value and the type-I error rate, the measure depends on the magnitude of the difference in effect estimates, the variance of the two treatment

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

effects (which themselves depend on the sample sizes), and the type-I error rate. Importantly, C_D^S does not depend on the experiment's and non-experiment's power to detect the true but unknown treatment effect (the true effect is differenced out by taking the difference in the experimental and non-experimental estimates). The difference in effect estimates may be significant even if both the non-experimental and experimental estimates are not significantly different from zero, that is, the conclusion-based correspondence C_C^S does not imply distance-based correspondence C_D^S (and vice versa).

For dependent WSC designs where the non-experimental arm shares some portion of the sample with the experimental arm (Wong & Steiner, under review), the statistical tests should account for the correlation in the two treatment effects. We have the same test statistic, $t = \frac{\hat{\tau}_{nx} - \hat{\tau}_{re}}{s}$, but with standard deviation $s = \sqrt{s.e._{nx}^2 + s.e._{re}^2 - 2r_{nx.re}s.e._{nx}s.e._{re}}$, where $r_{nx.re}$, denotes the correlation between experimental and non-experimental effect estimates. The third term of the standard deviation formula represents the covariance between the effect estimates and occurs because the experiment and non-experiment share the same treatment group or parts of the sample. Since the correlation or covariance between the two effect estimators is usually unknown to researchers, this test cannot directly be implemented without making some assumptions about the correlation $r_{nx.re}$ (i.e., the correlation is positive but unlikely close to one). One way of dealing with this issue is to bootstrap the *t*-statistic. Another possibility for dependent WSC designs that share a treatment group is to test whether mean outcomes in the benchmark control and non-experimental comparison groups differ. That is, we employ the *t*-test statistic $t = \frac{\bar{Y}_{nx}^c - \bar{Y}_{re}^c}{s}$, where $s = \sqrt{s.e._{\bar{Y}nx}^2 + s.e._{\bar{Y}re}^2}$ is the standard deviation of the mean difference, and $s.e._{\bar{Y}re}$ and $s.e._{\bar{Y}nx}$ are the standard errors of the experimental control and non-experimental comparison groups' mean outcomes, respectively.

We may also generalize the significance-related correspondence measure to the more general null hypothesis $H_0 : |\tau_{nx} - \tau_{re}| \leq \delta_D$, that is, we test whether the absolute difference in the two effect estimators exceeds a specific threshold δ_D. However, we do not discuss such significance tests further because (a) they are overly conservative in the sense that the test indicates correspondence though the estimates might actually be quite different and (b) they are rarely used in practice.

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Although significance tests of difference in benchmark and non-experimental results are common in the WSC literature, a more careful consideration of the typical WSC research question suggests serious weaknesses with this approach. In the standard NHST framework, the researcher does not reject the null hypothesis if the difference in effects is insignificant. Traditionally, NHST protects against Type I error rates (where the researcher concludes that there is a difference when there actually is no difference), and there is less concern about Type II error rates (where the researcher concludes that there is no evidence for a difference when a difference actually exists). As a result, under the NHST framework, WSC 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 non-experimental treatment 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 hypothesis. It could mean that there is no difference in treatment effect estimates between the experiment and non-experiment, or it could mean that the WSC is underpowered for detecting a difference. Thus, an insignificant result does not imply that the null hypothesis of a zero difference is true (even if one increases the Type-I error rate to .1 or .2). Moreover, standard NHST of a zero difference fails to address the conceptual problem of testing equivalence (as opposed to testing a difference). To overcome this conceptual problem, the WSC researcher may employ equivalence tests that use the same equivalence hypothesis as the general significance-related correspondence measure, $H_0: |\tau_{nx} - \tau_{re}| \leq \delta_D$, but formulates the equivalence as alternative hypothesis and the difference as null hypothesis.

Significance of Equivalence. Although statistical tests of equivalence have been 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. The only exceptions are by Berk et al. (2010) and more recently by Dong and Lipsey (2016), who used equivalence tests in their WSC analyses. Equivalence tests 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.

For a regular *t*-test in the standard hypothesis framework, to conclude that there is a substantial difference between two estimates, the WSC researcher must observe a difference large enough to rule out sampling error as an alternative explanation for the difference. However, in tests of equivalence, a similar approach applies, but to conclude that there is no or only a negligibly small

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

difference in estimates, one must observe a small enough difference to reject the competing explanation that the equivalence or closeness is not just due to sampling error.

Equivalence testing first requires the determination of a tolerance threshold δ_E, which is considered as a trivial or nonconsequential difference between the two estimates. For example, a difference of δ_E = .1 SD between the experimental and nonexperimental estimate might be considered as tolerable. Then, the difference in estimates is considered equivalent if the composite null hypothesis that the absolute difference in effects is larger than the threshold, $H_0: |\tau_{nx} - \tau_{re}| \geq \delta_E$, is rejected. The alternative hypothesis states the equivalence of estimates, $H_A: |\tau_{nx} - \tau_{re}| < \delta_E$, or, the absolute difference in estimates is smaller than the threshold. The composite null hypothesis of inequivalence may be formulated as two one-sided null hypotheses, $H_{01}: \tau_{nx} - \tau_{re} \geq \delta_E$ and $H_{02}: \tau_{nx} - \tau_{re} \leq -\delta_E$, which are directly testable (Schuirmann, 1987; Tryon & Lewis, 2008). If both one-sided null hypotheses are rejected, the data do not provide enough evidence for a difference that exceeds the threshold δ_E, and the WSC analyst concludes that the two methods are equivalent (within the tolerance threshold δ_E). Each of the two one-tailed tests is conducted with a nominal Type-I error rate of *a*. Thus, correspondence in equivalence is defined as:

$$C_E^S = 1[p_{E1} \leq \alpha \,\&\, p_{E2} \leq \alpha],$$

where p_{E1} and p_{E2} are the one-tailed *p*-values with respect to the two one-sided null hypotheses. The *p*-values are typically obtained from two separate *t*-tests.

One challenge with equivalence tests – and perhaps the reason why researchers have avoided these tests for assessing correspondence – is that the tolerance threshold must be defined in advance. However, the use of statistical equivalence tests may be one way to operationalize Wilde and Hollister's early advice to define meaningfully large differences in experimental and non-experimental results before a WSC analysis is conducted. Here, the substantive threshold is built within the testing framework of benchmark and non-experimental results.

Another limitation of the approach is that small tolerance thresholds of .1 or .2 SD require large samples from both the randomized experiment and the non-experiment. Otherwise, the equivalence test will not be sufficiently powered (see the simulation results below). Since a lack of evidence for equivalence does not imply that the two methods perform differently, it is useful to consider not only a dichotomous test outcome (equivalence vs. non-equivalence) but four possible

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

test outcomes of a composite correspondence test: equivalence, difference, trivial difference, and indeterminacy, which we discuss below.

Correspondence Test: Equivalence, Difference, Trivial Difference, and Indeterminacy. Tryon (2001) and Tryon and Lewis (2008) suggest a testing procedure that combines the outcomes of an equivalence test and a standard significance test for a difference in parameters. Here, we define a combined correspondence measure from the significance in equivalence, C_E^S, and the significance in difference, C_D^S, into a single framework:

$$C_{ED}^S = \begin{cases} \text{Equivalence} & \text{if } C_E^S = 1 \,\&\, C_D^S = 1 \\ \text{Difference} & \text{if } C_E^S = 0 \,\&\, C_D^S = 0 \\ \text{Trivial Difference} & \text{if } C_E^S = 1 \,\&\, C_D^S = 0 \\ \text{Indeterminacy} & \text{if } C_E^S = 0 \,\&\, C_D^S = 1 \end{cases}$$

For this measure, if correspondence is achieved in both metrics, $C_E^S = 1 \,\&\, C_D^S = 1$, the WSC analyst concludes that the two methods are equivalent. That is, the equivalence test rejects the null hypothesis of a significant difference larger than the threshold, and the significance test for the difference does not reject the nil hypothesis. If both correspondence metrics indicate non-correspondence, $C_E^S = 0 \,\&\, C_D^S = 0$, then the analyst concludes that the two estimates differ. A trivial difference is given if the estimates differ significantly (i.e., non-correspondence in difference, $C_D^S = 0$), but the equivalence test suggests correspondence ($C_E^S = 1$). Finally, if the data do not provide enough evidence for a significant equivalence ($C_E^S = 0$) or a significant difference ($C_D^S = 1$), then the WSC analyst concludes that correspondence is statistically indeterminate. Since indeterminacy results from failing to reject the null hypothesis of both tests, an indeterminate outcome is most likely when sample sizes are small (underpowered tests). We refer to the correspondence measure C_{ED}^S as a "correspondence test" because both equivalence and difference tests are used for drawing a conclusion about the methods' correspondence.

In comparison to correspondence in equivalence (C_E^S) and difference (C_D^S), the correspondence test (C_{ED}^S) has the advantage that equivalence is indicated only if the equivalence test is sufficiently powered and the difference in methods is insignificant. Since an adequately powered equivalence test (with a reasonable threshold) also implies a highly powered difference test, correspondence is achieved only if the experiment and non-experiment result in approximately the

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

same effect estimates. With small sample sizes, it is unlikely to achieve equivalence even if the estimates are nearly identical. Thus, when the benchmark and non-experimental designs are underpowered, the correspondence test will likely yield a result of statistical indeterminacy (the data do not provide sufficient evidence for a significant equivalence or significant difference). However, with large samples, or a large threshold value δ_E, it is possible to observe a significant difference in effect estimates as well as achieve correspondence in the equivalence test. In this case, the WSC analyst would conclude that a trivial difference between the methods exists. However, in the WSC context, even trivial differences should count as non-correspondence between methods unless a very small threshold value is used.

The advantages of the correspondence test, C_{ED}^{S}, are apparent. First, it provides the researcher with a statistical test that explicitly addresses the question of interest in most WSCs: Does the non-experimental method perform as well as the experiment in field settings? Second, it requires that the researcher addresses in advance what constitutes as a meaningful difference in experimental and non-experimental results. Third, it allows the researcher to conduct both statistical tests of difference and equivalence. Fourth, it provides a clear decision criterion for when experimental and non-experimental results are equivalent, different, or inconclusive.

Though we did not explicitly discuss the equivalence and correspondence tests for dependent WSCs, the very same criteria can be applied to the difference in the mean outcomes of the independent control and comparison group. Alternatively, one can also bootstrap the difference in the average treatment effects of the dependent experimental and non-experimental study arms, and then compute the correspondence measures.

SIMULATION STUDY

In this section, we examine the performance of conclusion- and distance-based correspondence measures under different WSC conditions for an independent design. For both the randomized experiment and non-experiment, we randomly sampled cases from a large joint target population and then randomly assigned the benchmark cases to the treatment or control condition, and systematically selected non-experimental cases into treatment conditions based on two covariates. To account for selection bias, we estimated the treatment effect in the non-experiment via inverse-propensity weighting and an additional covariance adjustment.

Simulation Design

To simulate the WSC design, we began by generating a target population of one million cases with two baseline covariates, X_1 and X_2, and a potential control and treatment outcome, Y_0 and Y_1. The values of the two covariates were sampled from a bivariate normal distribution with zero means, unit variances, and a correlation of .3. The pair of potential control and treatment outcomes was computed according to $Y_i^0 = \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ and $Y_i^1 = \tau SD_Y + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$, where the additive treatment effect was assumed to be τ times the standard deviation of the potential control outcome, SD_Y. Thus, τ represents the effect in terms of standard deviations. The error term ε_i was generated according to a normal distribution with zero mean and a standard deviation of .5. Note that the data-generating coefficients β_1 and β_2 as well as the error term ε_i are the same for the potential control and treatment outcome. The observed outcome, Y, was determined according to the treatment assignment status of the cases in the randomized experiment and non-experiment. In the randomized experiment, half of the cases were randomly assigned to the treatment condition, the other half to the control condition (i.e., each case had an assignment probability of .5). In the non-experiment, the assignment probabilities were determined according to the logistic model with two baseline covariates, $\text{logit}(Z_i) = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i}$.

To investigate the performance of the correspondence measures under different data-generating scenarios, we created three different target populations by varying the magnitude of the treatment effect and the direction of the selection process. Table 1 contains the parameter settings for the three populations. Here, Population 1 is characterized by a medium effect size (.3 SD) for both the benchmark and non-experiment, and a positive selection process in the non-experiment with X_1 being a stronger confounder than X_2 (X_1 has the larger coefficients in both the outcome and selection model). The selection process is of medium strength, producing a positive confounding bias of about .43 SD in the treatment effect. Population 2 has the same non-experimental selection process, but the treatment effect is only .05 SD in the benchmark and non-experimental conditions. Thus, for Population 2, it will be harder to demonstrate significant treatment effects in the experiment and non-experiment. Finally, Population 3 has the same effect size as Population 1 (.3 SD) but instead of a positive selection process, we have a negative selection that induces a negative selection bias. A negative selection bias results because the confounders' coefficients are negative for the selection model but positive for the outcome model. Since the populations' coefficients of the

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

selection model differ only in their sign but not in their magnitude, the strength of the selection process and the absolute bias is the same across all three populations.

For each population, we ran separate simulations. In each iteration of the simulation, we first drew a random sample of $n_{re} + n_{nx}$ cases, and then randomly assigned the experimental and non-experimental cases according to their assignment probabilities. Depending on the assignment status, the potential control or treatment outcome was recorded as the observed outcome. For the experimental and non-experimental data, we then used regression estimators to estimate the treatment effects.

Table 2 summarizes the regression models used to estimate treatment effects for both the experiment and non-experiment. For the experiment, we estimated treatment effects without any covariance adjustment. For the non-experimental data, we considered estimators of four models – three doubly robust estimators that combine inverse-propensity weighting with a covariance adjustment, and one model that did not control for any confounding covariates. These four models produce non-experimental treatment effects with different degrees of remaining bias. The first model controls for both confounding covariates, X_1 and X_2, that is, both covariates were used for PS estimation and covariance adjustment. Thus, the corresponding treatment effect estimator is unbiased because both confounding covariates are used and correctly modeled. The second model only controls for X_1 (in both the PS and outcome model), which results in an estimator with an expected bias of .08 SD. The third model controls only for X_2 (the weaker confounder), leading to a bias of .24 SD. Finally, the fourth model does not control for any covariates, which implies a bias of .43 SD in the treatment effect. Varying the extent of remaining bias allows us to assess the performance of the correspondence measures when the observed covariates do not succeed in removing all the selection bias in the non-experiment.

For each iteration, we computed a series of measures for assessing the correspondence between the benchmark and the four non-experimental estimates. We used the conclusion-based correspondence in direction (C_C^D), magnitude (C_C^M, with a threshold of λ = .2 SD), and significance (C_C^S) based on the regression-based *t*-test with a Type-I error of a = .05. With respect to the distance-based measures, we computed the correspondence in magnitude (C_D^M) with a threshold of δ_M = .1 SD, the correspondence in significance (C_D^S) based on a *t*-test with a = .05, as well as three correspondence measures for a significant equivalence (C_E^S) with thresholds δ_E of .1, .3, .5 SD.

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Based on the significance in difference and significance in equivalence, we then conducted the correspondence test (C_{ED}^{S}) for the three equivalence measures.

For each of the three populations we ran K = 10,000 iterations with different sample sizes for the experiment and non-experiment. For the benchmark RCT, we used sample sizes of 100, 400 and 1600 cases and for the non-experiment, we simulated 21 different sample sizes between 20 and 5000. For each simulation setting, we then computed the correspondence probabilities as the proportion of correspondence indicated by each measure. For instance, for the conclusion-based correspondence in significance, the estimated correspondence probability is given by $P(C_C^S = 1) = K^{-1} \times \sum_{k=1}^{K} C_{Ck}^S$, where C_{Ck}^S is the correspondence indicator for iteration k. For the correspondence test, which has four possible outcomes, we correspondingly computed the probabilities of equivalence, difference, trivial difference, and indeterminacy.

Results

Conclusion-based Correspondence Measures. The first three Figures (2 to 4) show the performance of the conclusion-based correspondence measures for the three different populations: positive selection with a treatment effect of .3 SD, positive selection with a very weak treatment effect of .05 SD, and negative selection with an effect of.3 SD. The performance plots for the Population 1 are depicted in Figure 2. The plots in the columns represent variations in experimental sample size and thus the experiment's power to detect the true treatment effect. The column headings indicate the RCT sample sizes of 100, 400 and 1600, which correspond to powers of .29, .81, .999, respectively. Thus, the first column of plots represents a situation where the experiment is underpowered with respect to the true effect, the second column refers to an RCT with adequate power, and the third column to an extremely well powered RCT. It is important to realize that the power figures refer to the power with respect to the unknown true effect, and not to the power of a minimum detectable effect size as used in power calculations. The rows of Figure 2 differ in the extent of remaining bias in the non-experiment. For the first row of plots, no bias is remaining because the analyses conditioned on both confounding covariates X_1 and X_2. The second row shows the results for the non-experimental analyses where X_1 has been included but X_2 has been omitted. Here, the expected bias is .08 SD. The third row refers to the non-experimental analyses with X_2 included but X_1 omitted which results in a bias of .24 SD. Finally, the last row reveals the performance of the correspondence measures when the bias is .43 SD which we obtained by not controlling for any of the confounding covariates.

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Each single plot shows the correspondence probabilities (ordinate) as a function of the non-experiment's sample size (abscissa, which is on a log scale). The abscissa shows two scales, the non-experiment's sample size and the corresponding power to detect the true effect if both confounders are conditioned (i.e., the power for the correctly specified model).[4] The three lines in the plot trace the correspondence probabilities of the three conclusion-based measures—solid lines for C_C^D, dashed lines for C_C^M, and dotted lines for C_C^S.

For the first row of plots, where the non-experiment's effect estimates are unbiased, we expect the measures to indicate correspondence because, given sufficient power, we should obtain the same results from the benchmark RCT and the non-experiment. Indeed, all three correspondence measure, C_C^D, C_C^M, and C_C^S, almost always indicate correspondence, but only if both studies have a power close to one (top right plot with an RCT sample size of 1600 and a non-experimental sample size of at least 500). If the power of one of the studies, RCT or non-experiment, is less than .8, then correspondence in magnitude and significance never exceed a probability of .8. This is not surprising because with a power of .8, the RCT or non-experiment draws the correct conclusion about the existing effect (.3 SD) only in 80% of the cases. The correspondence in magnitude (C_C^M) is rather robust and almost always indicates correspondence, except when the power of one of the studies is very low, such that the estimates might be negative (due to the relatively large sampling error). It is also worth noting that, with an underpowered RCT (power of .29, top left plot), the correspondence probability of C_C^S diminishes as non-experiment's power increases. That is, if both studies are highly underpowered, we will obtain correspondence in significance because both studies produce insignificant effect estimates. But as the sample size (and thus the power) of the non-experiment increases, the correspondence probability decreases because the non-experimental estimate will likely be significant while the RCT estimate remains insignificant. As non-experimental sample size and power increases, the correspondence probability converges to the RCT's power (in this case .29).

[4] Though we characterize the correspondence probabilities as a function of sample sizes (of both the RCT and non-experiment), the sample size requirements for obtaining sufficient power also depend on the portion of explained variance (R^2) after controlling for covariates. In our simulations, the R^2s of the analytic models are 2% for the RCT (without any covariance adjustment), 45% for the non-experimental model with both confounders X_1 and X_2, 38% with X_1, 24% with X_2, and 11% with none of the two confounders included. With higher R^2s, sample sizes can be smaller, but with lower R^2s, sample sizes must be larger to achieve the same correspondence probabilities.

The second to the fourth rows of plots display the correspondence probabilities as the bias in the non-experimental estimates increases form .08 SD to .24 SD and .43 SD. Given a biased non-experimental estimate, we expect that the conclusion-based correspondence measures should no longer indicate correspondence. However, this expectation does not bear out because the conclusion-based measures only focus on the correspondence of conclusions drawn about the treatment's impact. They are insensitive to the magnitude of biases in estimates when the bias does not result in a different policy conclusion. Since the plots in Figure 2 refer to a data-generating model that used a positive treatment effect (.3 SD) and a positive selection process (i.e., the confounders introduce a positive selection bias), failing to remove all of the selection bias does not invalidate the conclusions drawn from the non-experiment. To the contrary, the more selection bias that is left, the more likely we conclude there is a significant positive treatment effect that exceeds our magnitude threshold. Thus, it is not surprising that the correspondence probabilities increase as the bias increases (from row one to row four), provided that the RCT is sufficiently powered (at least greater than .8). In cases where the non-experimental sample is large and the power in nearly one (such that the non-experiment will almost always draw the right conclusion), the correspondence probabilities are entirely determined by the sample size (power) of the RCT. With an underpowered RCT, potential bias in the non-experimental treatment effect has almost no consequences on the correspondence probabilities.

The results discussed so far only hold for a positive treatment effect of .3 SD and a positive selection bias. The correspondence probabilities are quite different if the true treatment effect is close to zero (.05 SD), or the selection process induces negative selection bias. Figure 3 shows the results for Population 2 with a treatment effect of .05 SD. Because we used the same sample sizes as for Population 1, the power to detect a significant treatment effect is now much lower for both the RCT and non-experiment. The three columns of plots represent RCTs with a power of .06, .08 and .16. The correspondence probabilities differ strongly from the ones of Population 1. With an increasing sample size for the non-experiment, the correspondence in significance, C_C^S, tends to diminish since the non-experiment's power to detect even a small effect increases while the RCT has no power for demonstrating an effect. Thus, the conclusions often do not correspond. Due to the small true effect and the RCT's low power, correspondence in the estimates' direction (C_C^D) and magnitude (C_C^M) is also less likely. If there is remaining bias in the non-experiment's treatment effect (second to third row of Figure 3), then the conclusions based on the experimental and non-

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

experimental estimates no longer correspond, particularly when the non-experimental sample size is large. Due to the positive bias, the non-experimental analyses indicate a positive treatment effect while the underpowered RCT analyses suggest the absence of an effect.

Figure 4 shows the results for Population 3 where the selection process induces a negative selection bias. Thus, failing to control for all confounding covariates results in an underestimation of the treatment effect. The four rows of plots represent the same four bias scenarios as before except that the biases are now negative. Thus, given a true effect of .3 SD and the biases of 0, -.08, -.23, and -.41 SD, the non-experiment's average effect estimates are .3, .22, .07, and -11 SD, respectively (i.e., .3 – 0, .3 – .08, .3 – .23, and .3 – .41). Compared to Population 1, more bias now implies a smaller rather than larger effect estimate that even becomes negative if none of the two confounders is used to remove bias. Consequently, with an increasing bias in the non-experimental estimate, we now expect to see decreasing rather than increasing correspondence. This is apparent from Figure 4. The correspondence probabilities diminish as one goes from the plots in the first row (no bias) down to the plots in the fourth row (bias of -.41 SD). However, in cases where all selection bias is removed by the non-experimental analysis (first row), the correspondence measures do not depend on the direction of the selection process.

Overall, the plots in Figures 2 to 4 demonstrate that the conclusion-based correspondence measures depend on: 1) both benchmark RCT and the non-experiment's power to detect the true but unknown effect—in particular the true but unknown treatment effect and the sample size, 2) the direction of the selection process, and 3) the extent of remaining bias in the non-experimental effect estimate. Importantly, a highly biased non-experimental estimate does not always result in lack of correspondence in conclusion-based measures. In the case of a positive (negative) treatment effect and positive (negative) selection bias, more bias can imply a higher correspondence probability. The plots presented here illustrate this characteristic of conclusion-based measures. For all three populations, the characteristics of the RCT and non-experiment were held constant, but the conclusions drawn from the analyses differed dramatically.

Distance-based Correspondence Measures. Since the characteristics of the RCT and non-experiment were held constant across the three populations, the results for the distance-based correspondence measures do not differ across the three populations. That is, other than for conclusion-based correspondence measures, distance-based measures do not depend on the size of the true effect and the direction of the selection process. Only the absolute distance between the two effect estimates matters. Therefore, we discuss results for Population 1 only (treatment effect of

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

.3 SD and a positive selection process). Irrespective of the conclusions drawn from the RCT and non-experimental study (i.e., whether a treatment is effective or ineffective), distance-based correspondence measures assess whether the effect estimates are similar or different enough to conclude that the two methods perform equally well or different. Thus, if both the RCT and non-experimental estimates are unbiased, we would expect that the distance-based measures indicate correspondence. If one estimate is biased (often the non-experimental estimate, provided the experiment is valid), we expect the distance measure to indicate non-correspondence.

Figure 5 shows for Population 1 the correspondence probabilities of the correspondence in magnitude (C_D^M, with a threshold of .1 SD), the correspondence in significance (C_D^S), and the three equivalence measures C_E^S (i.e., with thresholds of .1, .3, and .5 SD). First, we consider the top row of plots where the non-experimental estimates are unbiased and where we would expect the measures to indicate correspondence in estimates. C_D^M indicates correspondence only if the sample sizes of the RCT and non-experiment are large enough such that the difference in estimates is reliably estimated (solid lines). With a larger threshold, we would obtain greater correspondence probabilities. If the sample size of one of the studies is small, the performance of C_D^M is poor because sampling error regularly produces a difference that is larger than the threshold. Correspondence in significance, C_D^S (dashed lines), almost always suggests that the RCT and non-experimental estimates do not differ. This is what we would expect from a significance test with a null hypothesis of a zero difference. With a type-I error rate of 5%, we do not reject the null hypothesis in 95% of the tests, given the null hypothesis is true (which is the case here, because both the RCT and non-experimental estimates are unbiased). For small non-experimental samples (less than about 200), the correspondence probabilities are slightly below the expected 95% because the *t*-test we employed is based on the normal approximation with identical variances instead of a *t*-distribution with Welch–Satterthwaite-corrected degrees of freedom. Both measures, C_D^M and C_D^S, are sensitive to bias in the non-experimental estimate. With increasing bias (from the second to the fourth row of plots in Figure 5), the correspondence probabilities decline, particularly if the sample sizes of both studies are large. However, with small to medium-sized biases (.24 SD) in the non-experimental estimate, and small to medium RCT and non-experimental sample sizes (up to about 500), the correspondence probabilities for the significance measure C_D^S are still high (greater than

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

.5). Due to small differences in effect estimates or because of lack of power, the C_D^S measure fails to detect true differences in RCT and non-experimental results.

As discussed above, one option for addressing the undesirable but expected liberal performance of the significance measure (C_D^S) is to use an equivalence test (C_E^S). The advantage here is that the equivalence test will not indicate correspondence in RCT and non-experimental results simply because of lack of statistical power in the WSC. However, an important consideration for equivalence tests is establishing an appropriate "tolerance" threshold for assessing correspondence in benchmark and non-experimental results. If the "tolerance" threshold is set too small, the equivalence test will indicate lack of correspondence when WSC sample sizes are not large. In our simulation, we found that if the threshold is set at .1 SD, then the equivalence test often had insufficient power for indicating correspondence, even when there was no actual bias. Here, sample sizes exceeding 5,000 in the RCT and non-experiment were required. Figure 5 shows that C_E^S with δ_E = .1 SD has a correspondence probability greater than zero only in the top right plot with 1600 RCT cases and more than 1500 non-experimental cases (dotted line). Increasing the threshold values to .3 or .5 allows for a more liberal equivalence test when WSC sample sizes are not very large. As our simulation results show, C_E^S with a tolerance threshold δ_E of .3 and .5 SD succeeded in showing equivalence when sample sizes were moderate (at least 400 in the RCT and non-experiment, second plot in the first row). However, the tradeoff is that tolerance thresholds of .3 or .5 SD may be too large for assessing non-experimental performance when there is in fact bias. Our simulation results show that even with well-powered tests, the equivalence measure regularly indicates correspondence unless the remaining bias in the non-experiment is large (.43 SD, last row of plots). However, if the bias exceeds the chosen threshold δ_E the equivalence measure always suggests non-correspondence.

The correspondence test (C_{ED}^S), which combines the C_D^S and C_E^S measures, overcomes the weaknesses of each test on its own. Figures 6 to 8 show for different threshold values (.3, .1, .5 SD), the results for the correspondence test (C_{ED}^S) with Population 1. The structure of each figure is the same as before but the ordinate now reflects the probabilities with which each outcome of the equivalence test occurs. The solid line shows the probability of equivalence, the dashed line the probability of difference, the dotted line indicates a trivial difference, and the dashed-dotted line represents indeterminacy. Figure 6 contains the plots for a threshold of .3. If there is no remaining

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

bias in the non-experimental estimate (first row of plots) and if both the RCT and non-experiment have sufficiently large sample sizes (1600 in the RCT and at least about 500 in the non-experiment; top right plot), then the correspondence test almost always indicates equivalence (solid line). With decreasing sample sizes of the RCT or non-experiment, the correspondence test suggests equivalence less likely but instead the probability of an indeterminate outcome increases (dashed-dotted line). Thus, with an unbiased non-experimental estimate, the correspondence test indicates equivalence only if the test is sufficiently powered, otherwise the test indicates indeterminacy. Only occasionally, the test suggests a difference (dashed line) or a trivial difference (dotted line).

The plots in the second to the third row of Figure 6 show that the correspondence test is also sensitive to bias in the non-experimental estimate. If the bias is large (.43 SD, last row), then the tests indicate either a difference or indeterminacy. With sufficient power (large enough sample sizes), the correspondence test almost surely indicates a significant difference. But as the power diminishes, indeterminacy is the more likely test outcome. Given a bias of .43 SD and a threshold of .3 SD, the test almost never indicates equivalence of RCT and non-experimental estimates. If the bias in the non-experimental estimate is smaller (.08 SD, second row), then the correspondence test tends to indicate equivalence, provided that sample sizes are large enough; otherwise indeterminacy is the more likely test outcome. However, given the slight bias, the probability of a difference or a trivial difference is now increased (see particularly the plot in the second row and last column). As the bias increased further (.24 SD, third column), the probability of an equivalence outcome decreases but the probability of difference increases.

The plots in Figures 7 and 8 demonstrate that the correspondence test is sensitive to the choice of the threshold. Though setting the threshold at .1 SD indicates that we are not willing to accept a difference in RCT and non-experimental estimates of .1 SD or larger as negligible, the correspondence test often lacks the power to show equivalence in this case. As the plots in the first row of Figure 7 show, the predominant test outcome is indeterminacy. Only when the sample sizes of both studies exceed 1500, the test occasionally indicates equivalence (top right plot). However, the test remains sensitive to moderate to large biases in the non-experimental estimate. For example, the correspondence test indicates a significant difference with sufficiently large sample sizes (at least 400) and a bias of .24 SD or more (last two rows of plots in Figure 6).

If the correspondence test should be more sensitive to the equivalence rather than difference in estimates, we can increase the threshold, for instance to .5 SD as shown in Figure 8. In the absence of any bias (first row of plots), the correspondence test indicates equivalence as long as both

EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

sample sizes are moderately large (at least 300). In case of biased non-experimental estimates, the correspondence test more frequently indicates a trivial difference, that is, there is a significant difference in estimates, but simultaneously the equivalence test suggest that the estimates do not differ (due to the large tolerance threshold). Given a threshold of .5 SD, one should interpret trivial differences as evidence against the equivalence of experimental and non-experimental estimates.

CONCLUSIONS

This paper examines methods for assessing correspondence in benchmark and non-experimental results, highlighting each approach's advantages and limitations. We have argued that the measures employed should reflect the purposes of the WSC design. For example, is the WSC design meant to address the policy-relevant question of whether both the benchmark and non-experimental method would result in the same conclusion by a decision-maker? If so, then the relevant measures include comparing the direction, magnitude, and statistical significance patterns of results from the experiment and non-experiment. These approaches indicate whether the policy-maker would arrive at the same decision about a program's efficacy using the benchmark or non-experimental design. However, our simulations show that results from conclusion-based measures are highly sensitive to the statistical power in the benchmark and non-experimental designs (i.e., the magnitude of the unknown true effect, the sample sizes, and error variances) and the direction of the initial selection bias in the non-experiment. Overall, we recommend that WSC analysts should interpret their conclusion-based results cautiously, within the context of their study conditions. For example, when the true treatment impact is substantially large (.3 SD) and there is positive selection bias in the non-experiment, all three conclusion-based measures are likely to indicate correspondence in benchmark and non-experimental results. This is because even if substantial bias remains in the non-experiment, the policy-maker will make the same conclusions based on the direction, size and statistical significance of results in the benchmark and non-experimental design. The only exception is in cases when one of the study designs is underpowered for detecting significant effects. In the case where the true treatment impact is near zero and there is positive selection bias in the non-experiment, then comparing the direction and magnitude of the effects from the two study designs may be informative for assessing the performance of the non-experiment (depending on the threshold established for assessing "magnitude"). However, in this case as well, patterns of statistical significance are highly sensitive to the power of the benchmark and non-experimental design. Thus, conclusion-based measures often perform poorly as an indicator of non-experimental bias; instead, these measures are informative only for addressing the

policy question of whether a decision-maker would arrive at the same conclusion using the benchmark and non-experimental study designs in the given context. A lack of correspondence does not automatically imply the failure of the non-experiment—it might be caused by a type-II error of an underpowered test in the RCT.

Given that conclusion-based correspondence measures are highly sensitive to design attributes of a specific WSC, how should one assess the magnitude of the true treatment effect, and thus the statistical power in the experiment and non-experiment, and the direction of the selection bias in the non-experiment when these attributes are not directly observable to the researcher? Often, she may be able to roughly infer this information by examining results from prior research and baseline characteristics of WSC study participants. For example, results from earlier evaluations of a similar intervention may provide the WSC analyst with guidance about the magnitude of the program's true impact. Moreover, comparing baseline characteristics between treatment and comparison units in the non-experiment may indicate the direction of the initial selection bias.

To address the methodological question of how non-experimental methods perform in field settings, distance-based measures are useful. Here, the WSC analyst compares treatment effects from two study designs to assess how well the non-experimental method is able to replicate benchmark results. Currently, the most popular approach for judging whether non-experimental and benchmark results correspond is to conduct statistical tests of difference between the two results. As we have shown, this criterion has severe limitations in the WSC context, especially when benchmark and non-experimental sample sizes are small. In these cases, a non-significant difference in benchmark and non-experimental results may indicate correspondence in the two estimates, or it may be due to lack of statistical power for rejecting the false null hypothesis. In this paper, we have argued for an alternative approach for judging correspondence in WSC contexts: the correspondence test. This test indicates equivalence or a difference in methods when there is sufficient evidence for making these conclusions, but yields a result of indeterminacy when neither of these conclusions is warranted (often due to insufficient sample sizes). As a result, the correspondence test can help guard against implicit and explicit biases from the WSC researcher or reader.

Correspondence tests also have the advantage of requiring the researcher to define in advance a tolerance threshold for when benchmark and non-experimental results are "close enough" to be equivalent. It is beyond the scope of this paper for advising exactly how one should establish these thresholds, but in general, it should be based on substantive or empirical grounds, such as results from a prior cost-benefit analysis. Our simulation results show that in a WSC context, a

relatively large tolerance threshold of at least .3 SD was needed for establishing equivalence in unbiased benchmark and non-experimental estimates. In fact, equivalence was hardly ever indicated when the threshold was as small as .1 SD. As a result, in the WSC context, the threshold should reflect the upper bound for what the analyst considers "tolerable" for equivalence in benchmark and non-experimental results.

The simulation also demonstrates that for the independent design, relatively large sample sizes are required for the correspondence test to indicate a reliable equivalence or difference in benchmark and non-experimental results. When the tolerance threshold is .3 SD, the RCT required at least 400 units, and the non-experimental arm required between 600 and 1600 participants, depending on the size of the bias. Given that the WSC researcher often plans independent designs prospectively, the approach may be feasible only in cases where the intervention is straight-forward to implement, such as a text message or informational mailer, and the outcome is obtained through administrative records. An alternative to independent arm designs is WSC approaches with dependent benchmark and non-experimental approaches. For example, in the simultaneous design, some portion of the treatment or comparison group is shared "simultaneously" in the RCT and non-experiment. Because the design is an ad hoc approach that uses existing RCT and observational data, it is often possible for the WSC researcher to include much larger sample sizes than what would be feasible in prospective WSC designs. Moreover, even with the same overall WSC sample size, the simultaneous design has improved statistical power for assessing correspondence over the independent arm design. This is because residual variance in the outcome is reduced when units are shared between the benchmark and non-experimental conditions, as opposed to the case when units in the benchmark and non-experiment are independent. Wong and Steiner (under review) discuss dependent and independent WSC designs further, highlighting validity concerns with each approach for evaluating non-experimental methods.

To assess correspondence in WSC designs, our general recommendation is that researchers begin by defining the research purpose of their evaluation, and selecting correspondence measures that address their questions directly. Second, the researcher should consider attributes of their WSC design, including the likely true impact of the intervention, the statistical power of the benchmark, non-experiment, and WSC design itself, and the direction of the selection bias in the non-experiment. These factors will likely affect the sensitivity of the correspondence measures (particularly the conclusion-based measures), and the WSC analyst should interpret the results accordingly. Finally, given that readers of WSC studies often have both policy and methodologically

related interests, it may often be useful for WSC analysts to provide a table of results that summarize all their results, including conclusion- and distance-based correspondence measures.

As standalone enterprises, individual WSC studies may have little to say about the performance of non-experimental methods in field settings. However, results from multiple WSC studies may begin to provide an empirically based "mapping" of contexts and conditions for optimal non-experimental performance in specific domains. A critical component of this endeavor is clear and consistent criteria for assessing correspondence in benchmark and non-experimental results. This paper provides researchers with guidance for analyzing and interpreting WSC results.

REFERENCES

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review, 22*, 207-244.

Arceneaux, Gerber, & Green (2010). A Cautionary Note on Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to Experimental Benchmark. *Sociological Methods & Research.* 39(2). 256-282.

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.

Bell, Orr, Bloomquist, and Cain (1995). Program Applicants as a Comparison Group in Evaluating Training Programs: Theory and a Test. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. http://dx.doi.org/10.17848/9780585284545

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.

Dong, N. & Lipsey, M. W. (2014). *How well propensity score methods approximate experiments using pretest and demographic information in educational research?* Paper presented at the 35th Annual Association for Public Policy Analysis and Management (APPAM) Research Conference, Albuquerque, NM.

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

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

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

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

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

What Works Clearinghouse (2014). *Procedures and Standards Handbook Version 3.0.* U.S. Department of Education. Retrieved February 16, 2016 from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Krueger (1999). Experimental Estimates of the Education Production Functions. *The Quarterly Journal of Economics*, 114(2), 497-532

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

Munk, Hwang, & Brown (200). Testing Average Equivalence – Finding a Compromise between Theory and Practice. Biometrical Journal. 5, 531-552.

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

Paige (2005). Scientifically Based Evaluation Methods. Federal Register, 70(15), pg. 3585. Retrieved on February 16, 2016 at http://www2.ed.gov/legislation/FedRegister/finrule/2005-1/012505a.html.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657– 680

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.

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.

Table 1. Parameter Settings for the Three Target Populations

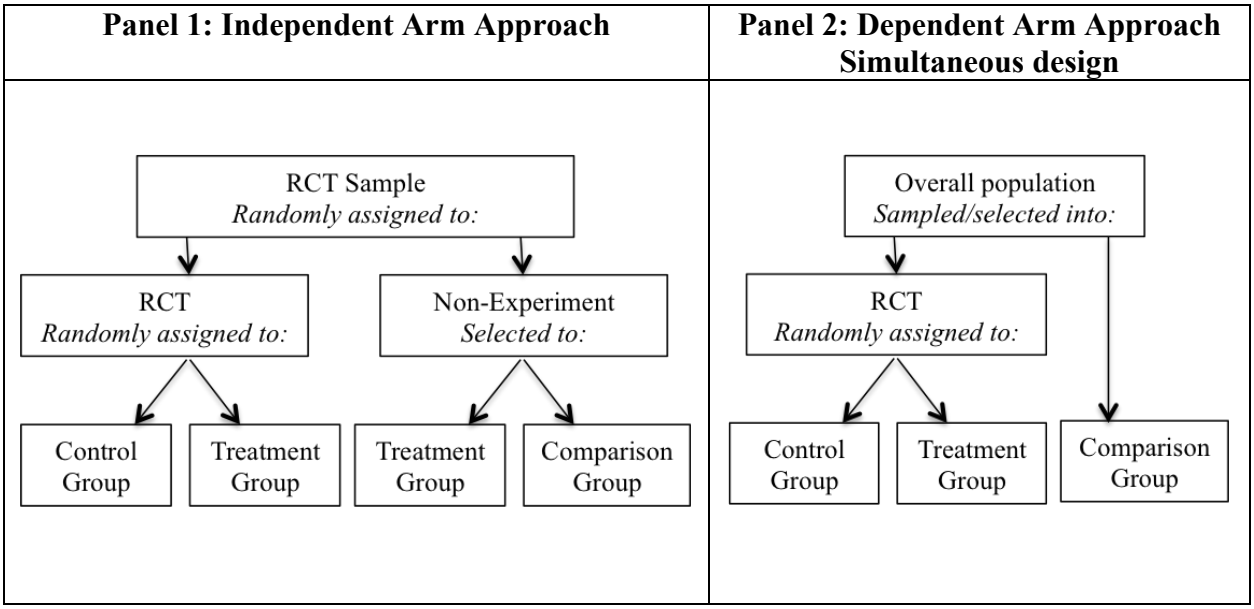
Population	Outcome Model	Treatment Effect	Selection Model Non-Experiment
Population 1: medium effect size & positive selection	$\beta_1 = .3, \quad \beta_2 = .2$	$\tau = .3SD_Y$	$\gamma_0 = -.5, \quad \gamma_1 = .5,$ $\gamma_2 = .3$
Population 2: very small effect size & positive selection	$\beta_1 = .3, \quad \beta_2 = .2$	$\tau = .05SD_Y$	$\gamma_0 = -.5, \quad \gamma_1 = .5,$ $\gamma_2 = .3$
Population 3: medium effect size & negative selection	$\beta_1 = .3, \quad \beta_2 = .2$	$\tau = .3SD_Y$	$\gamma_0 = .5, \quad \gamma_1 = -.5,$ $\gamma_2 = -.3$

Table 2. Regression Models for Estimating the Treatment Effect (τ)

Data & Analysis	Regression model
Randomized experiment	$Y = \beta_0 + \tau T + \varepsilon$
Non-experiment with both confounders (X_1 and X_2)	$Y = \beta_0 + \tau T + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ with inverse-PS weights based on X_1 and X_2
Non-experiment with confounder X_1	$Y = \beta_0 + \tau T + \beta_1 X_1 + \varepsilon$ with inverse-PS weights based on X_1
Non-experiment with confounder X_2	$Y = \beta_0 + \tau T + \beta_2 X_2 + \varepsilon$ with inverse-PS weights based on X_2
Non-experiment without any confounder	$Y = \beta_0 + \tau T + \varepsilon$

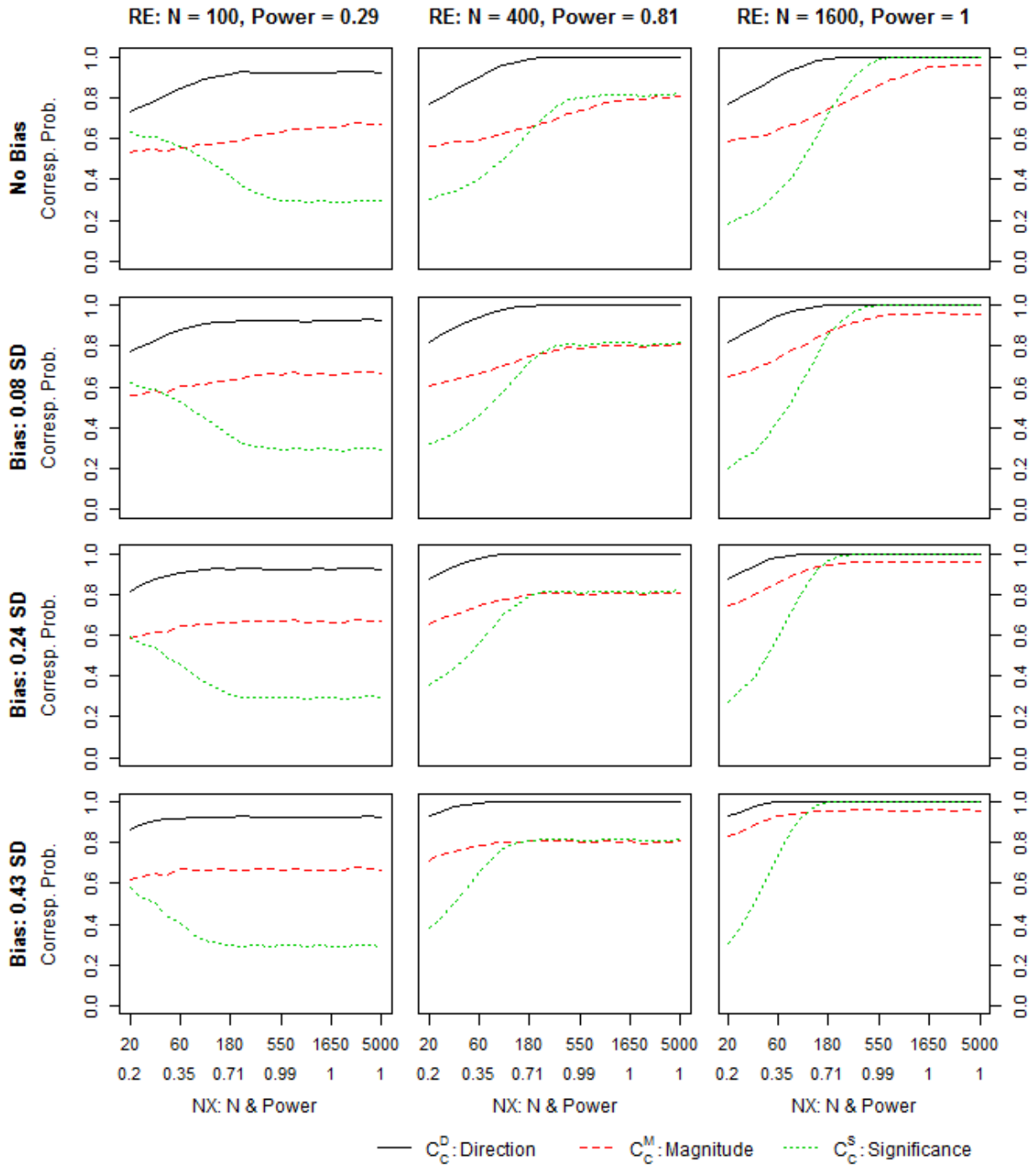
EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Figure 1: Independent versus dependent within-study comparisons



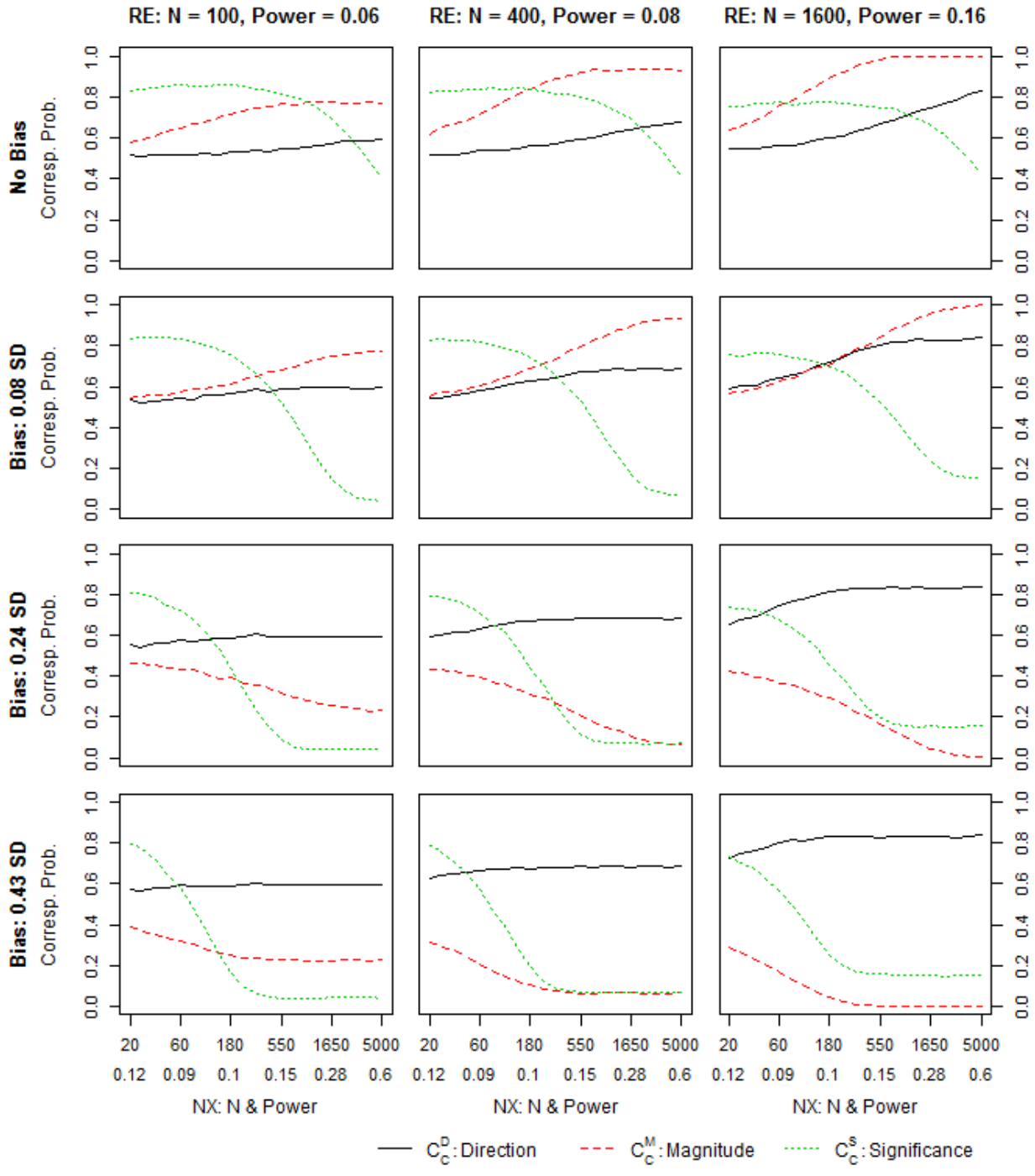
EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Figure 2. Conclusion-based correspondence for Population 1. Columns represent different sample sizes and power of the randomized experiment (RE), rows different extents of bias in the non-experiment (NX).



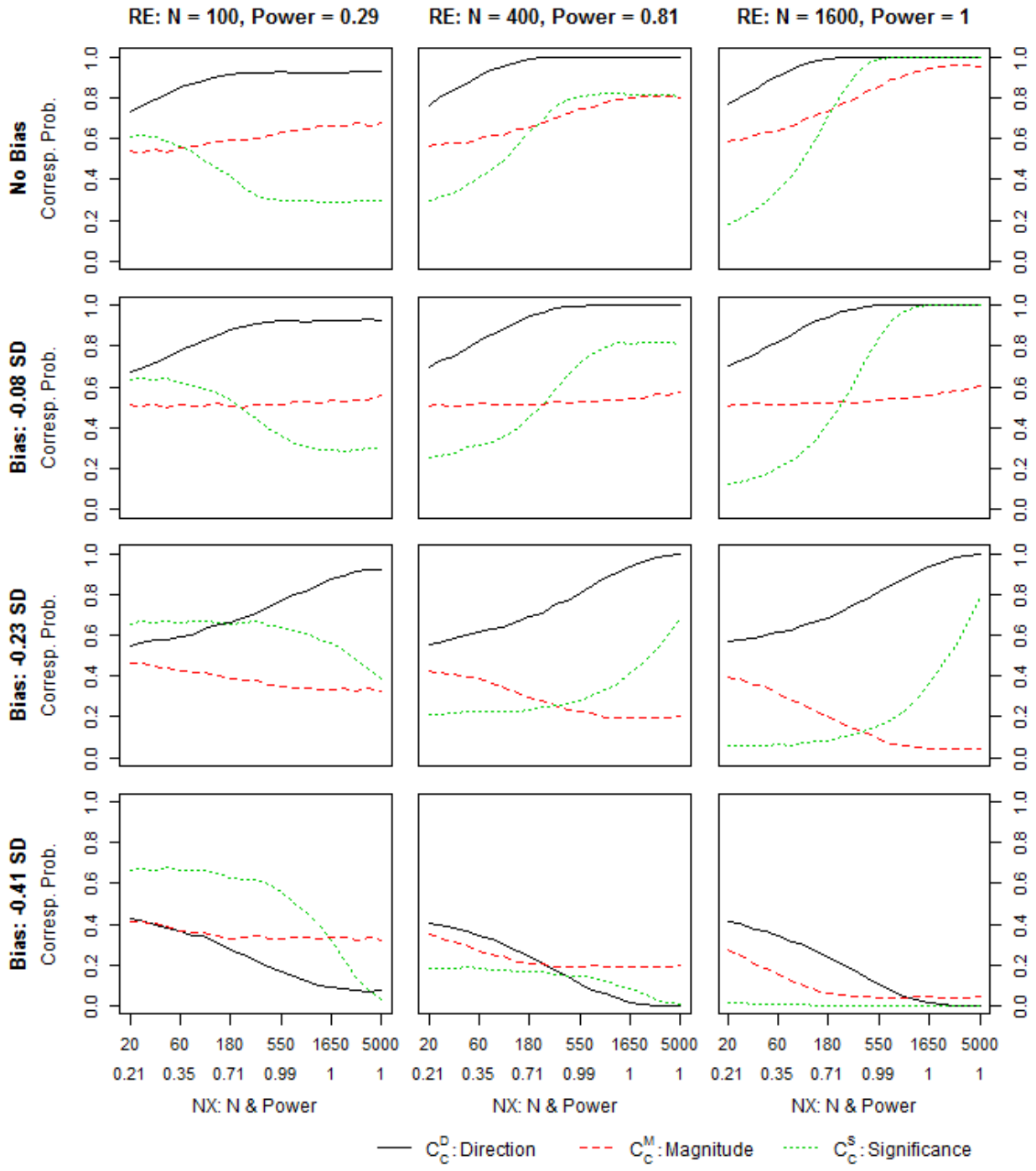
EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Figure 3. Conclusion-based correspondence for Population 2. Columns represent different sample sizes and power of the randomized experiment (RE), rows different extents of bias in the non-experiment (NX). Given the very small treatment effect of .05 SD, the regression tests' power it considerably lower than for the other Populations.



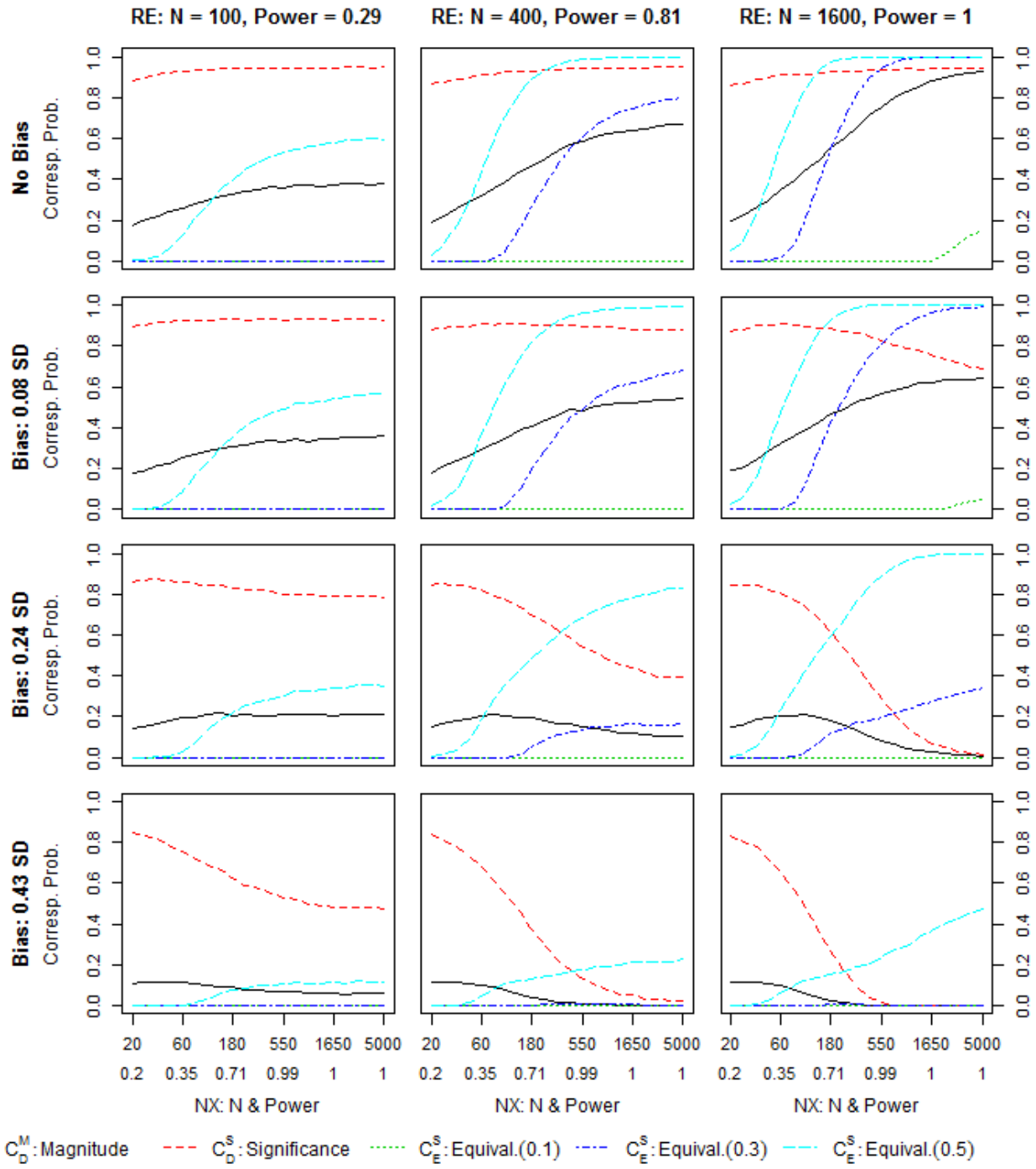
EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Figure 4. Conclusion-based correspondence for Population 3. Columns represent different sample sizes and power of the randomized experiment (RE), rows different extents of bias in the non-experiment (NX).



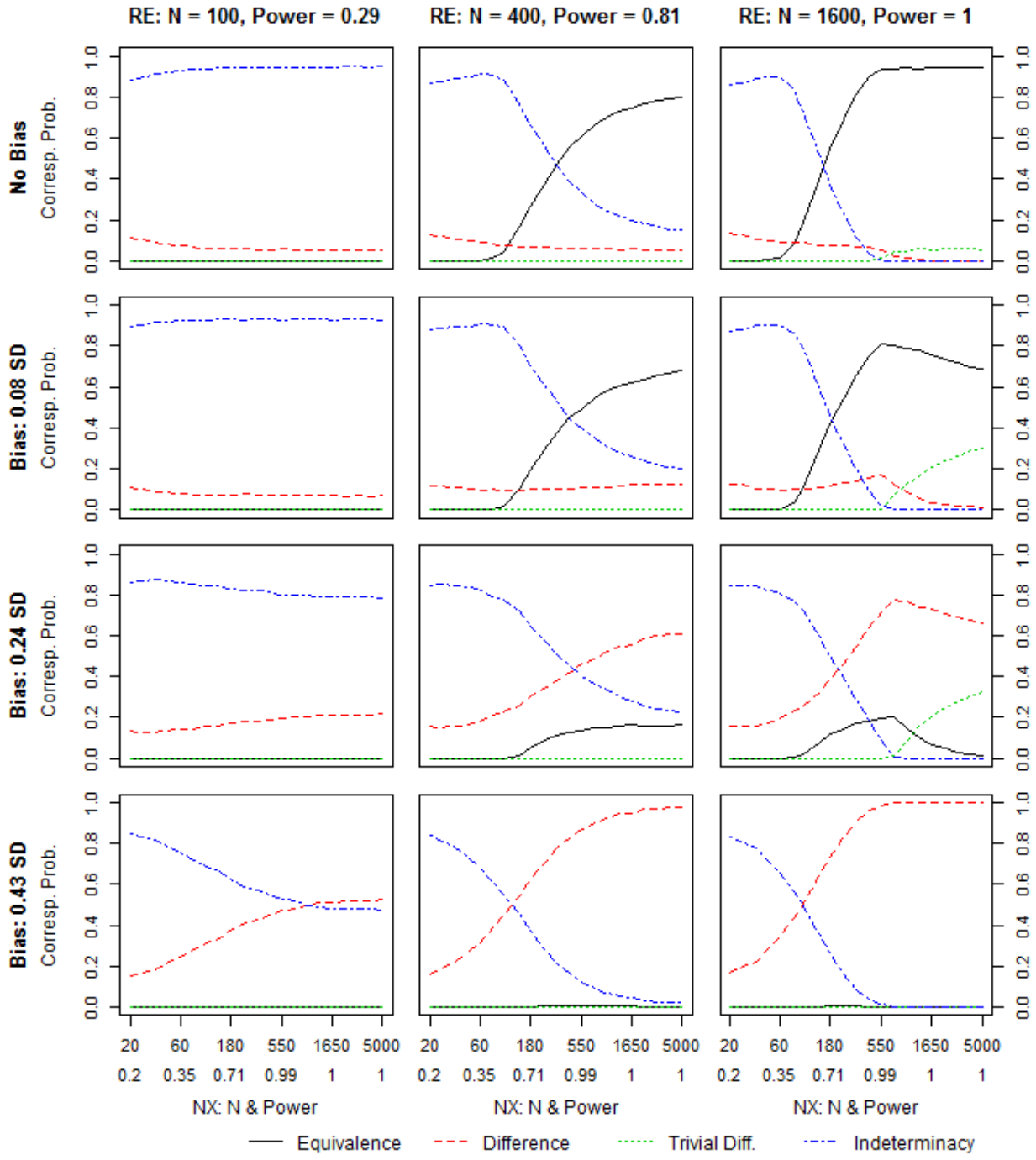
EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Figure 5. Distance-based correspondence (Population 1). Columns represent different sample sizes and power of the randomized experiment (RE), rows different extents of bias in the non-experiment (NX).



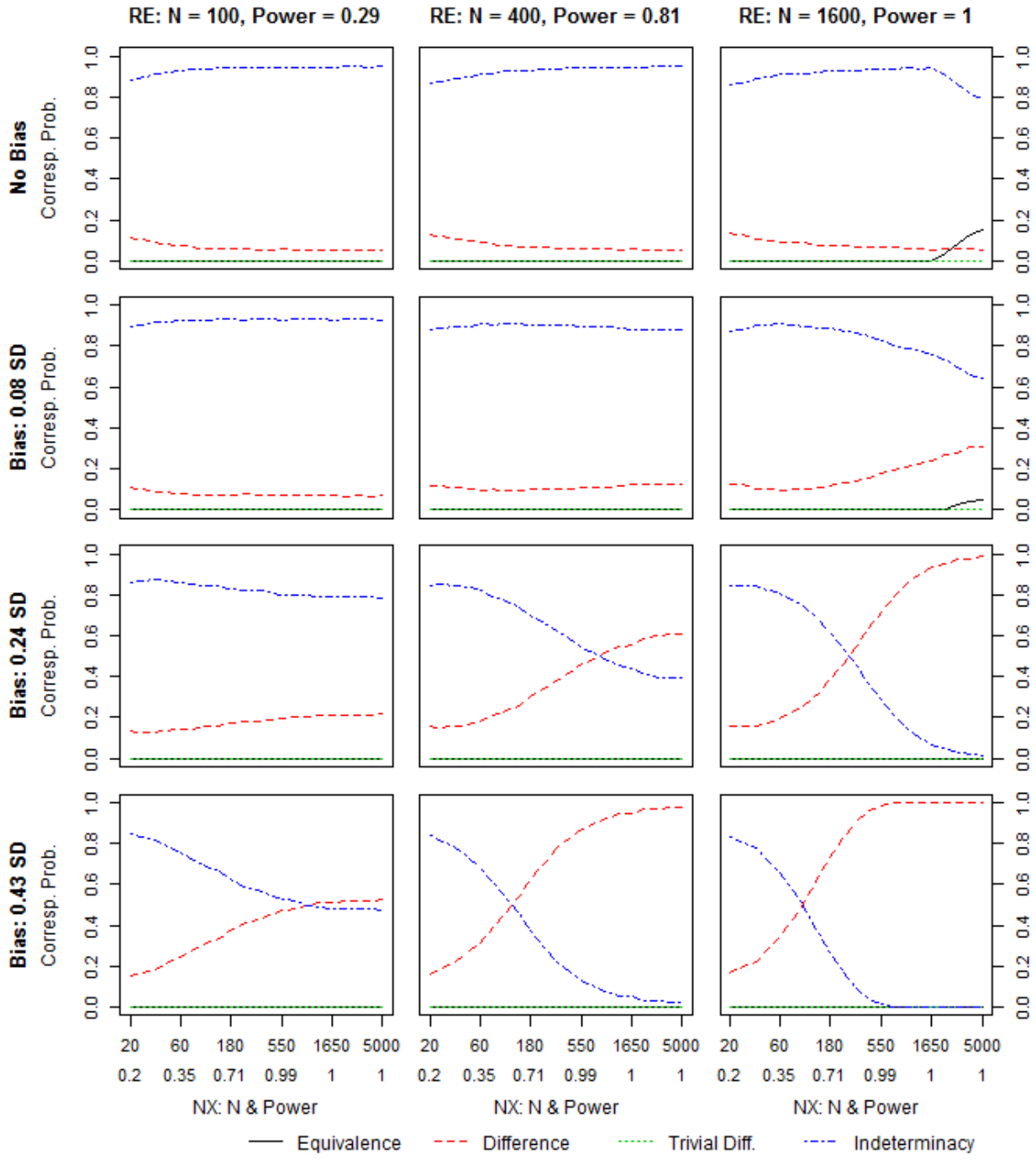
EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Figure 6. Correspondence test with δ_E = .3 SD. Columns represent different sample sizes and power of the randomized experiment (RE), rows different extents of bias in the non-experiment (NX).



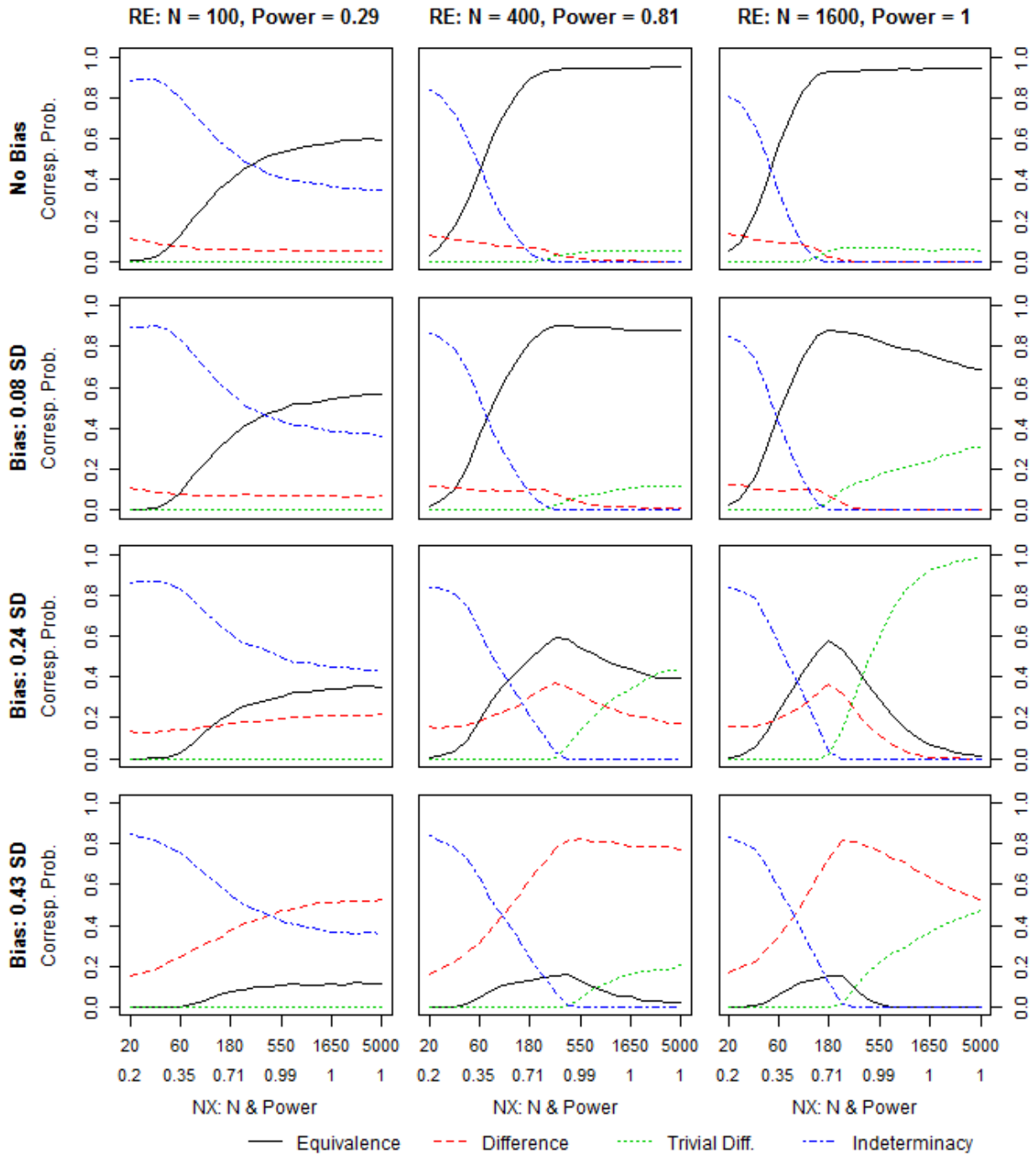
EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Figure 7. Correspondence test with δ_E = .1 SD. Columns represent different sample sizes and power of the randomized experiment (RE), rows different extents of bias in the non-experiment (NX).



EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Figure 8. Correspondence test with δ_E = .5 SD. Columns represent different sample sizes and power of the randomized experiment (RE), rows different extents of bias in the non-experiment (NX).



EdPolicyWorks Working Paper Series No. 46. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia