Empirical Education

EMPOWERING EDUCATORS THROUGH EVIDENCE AND INSIGHT

Working Paper: Are Estimates of Differential Impact from Quasi-Experiments Less Prone to Selection Bias than Average Impact Quantities?

Andrew P. Jaciw

Empirical Education Inc.

June 23, 2020

Abstract

In this work we demonstrate that estimates of differential program impact from comparison group designs that evaluate differences in outcomes for subgroups of individuals are less prone to selection bias. First, we argue for the importance of the routine evaluation of moderated impacts. Second, using formal and graphical arguments, we show that under specific conditions, cross-site comparisons of performance gradients across subgroups lead to cancelation of bias from site-specific third variable confounds. This means we can expect a reduction in standard selection bias resulting from differences between study and inference site in average performance; however, cross-site comparisons can introduce bias due to cross-site differences in the subgroup performance gradient. To examine this tradeoff in biases, we apply Within Study Comparison methods to obtain estimates of Root Mean Squared Bias from six studies, and empirically evaluate levels of each form of bias. We conclude that accuracy of estimates of subgroup differences in impact from comparison groups studies are less prone to bias overall. By yielding results with limited bias, routine analysis of moderated impacts using quasi-experiments can help broaden our understanding of the conditions under which programs are more effective.

Reference this paper: Jaciw, A.P. (2020). *Are Estimates of Differential Impact from Quasi-Experiments Less Prone to Selection Bias than Average Impact Quantities?* (Working Paper No. Empirical_AJE-WP1-2020-O.1). San Mateo, CA: Empirical Education Inc. Retrievable from https://www.empiricaleducation.com/past_research/

Introduction

The main question in an impact evaluation is whether a program achieves average positive impact for the full study sample. Studies are powered to address this question, and evidence reviews, such as by the What Works Clearinghouse in education (WWC, 2020) are based on them. In spite of their primacy, an obvious limitation to full-sample marginal impact estimates is that they do not automatically apply to subsamples. For example, the impact of a program averaged across students in the lower and upper half of the distribution of incoming achievement may not be accurate for either subgroup. To obtain more nuanced information about the robustness of a causal effect across conditions and individuals, it is necessary to evaluate differential effects as well as subgroup impacts. Tests of differential effects are especially important for indicating whether different subgroups receive the same level of benefit from a program, whether an average impact quantity generalizes to an inference population, and for understanding the potential of a program to close an achievement gap.

Study designs that are used to evaluate average impacts are also used to test for differential impacts. The Randomized Control Trial (RCT) is the preferred design for addressing both quantities. Quasi-experimental designs (QEDs) may be used when experiments are not feasible. A standard QED is the non-equivalent comparison group design (CGD) (Shadish, Cook, & Campbell, 2002), in which outcomes for a treatment group are compared to those from an ostensibly similar comparison group.

The main limitation to QEDs is the potential for results to be biased from selection, which experiments rule out by design. With CGDs, this bias occurs when we fail to account for differences between the program and comparison groups in compositional characteristics that affect outcomes. Conditions for bias in QEDs have been studied extensively, yielding design- and analysis-based strategies for addressing it (Bloom, Michalopoulos, & Hill, 2005; Cook, Shadish, & Wong, 2008).

In this work we make the case that estimates of differential impact from CGDs are less susceptible to selection bias from certain sources than estimates of average impact based on the same design. Specifically, we focus on the biasing role of confounders known as third variable confounds (Cook, Shadish, & Wong, 2008), or "macro variables" (Hotz, Imbens, & Mortimer, 2005). Such variables take different constant values in the program group and the comparison group. Unlike variables at the individual level, which allow person-to-person matching of similar cases, macro factors—such as locations across which the QE comparison is made—are wholly different between treatment and comparison groups. We argue that with CGDs, tests of differential impact across subgroups of individuals are not subject to biasing effects of macro variables in ways that estimates of average impact are.

We begin the work by arguing why it is important to evaluate differential impacts. We then discuss the role of QEDs as an alternative to randomized experiments for evaluating average and differential causal effects of programs. After this we make our main assertion of this work and develop the argument graphically and algebraically. This step leads to expressions for bias in QED-based estimates of differential impact. Focusing on outcomes in education, we use these expressions to empirically evaluate levels of bias across several studies. We end the work by drawing some conclusion.

Background

Reasons Why It Is Important to Evaluate Differential Program Impacts

The presence of effect heterogeneity has been demonstrated across multiple impact evaluations in education. Several studies have provided useful summaries of the levels of this variation, including a comprehensive meta-analysis of RCTs and QEDs of educational technology interventions (Cheung and Slavin, 2012), a synthesis of findings from a series of multisite trials of educational programs (Weiss et al., 2017), and in a review of multiple rigorously conducted experiments (Jaciw et al., 2016). Given that average impact quantities are observed to often vary with contexts and attributes of beneficiaries, it is important to study the conditions for and implications of impact heterogeneity for research and practice. We consider several motivations for investigating differential effects especially as they arise through effects of moderators.

Evaluating Differential Program Effects Helps Us To Establish Boundary Conditions For Observing Impact

Examining whether impacts vary across individuals and contexts is important because it prevents us from overgeneralizing average impact quantities to specific subgroups. If impact for boys is -1 and impact for girls is +1, then an average effect finding of zero holds little value for either subgroup. Reporting just average effect findings obscures underlying differences.

More generally, one can say that full-sample average impact findings from experiments are useful for summarizing the net benefit of a program; however, these results are vague. Grand mean positive impact findings from an experiment demonstrate that a program *can* work under *some* circumstances, but tell us little if anything about for whom and under what conditions impacts are likely to be observed (Bryk, 2014). Tests of differential effects unpack the average impact finding, allowing us to evaluate the moderating effects of individual or contextual characteristics on program impact.

Knowledge of Differential Impacts Supports Models of Generalizability

Moderator analyses are critical to several long-standing as well as emerging approaches to generalizing results from experiments. The first, "Heterogeneity of Replication" approach (Cook, 2002; Shadish, Cook, & Campbell, 2002) considers factors that moderate impact to be threats to stable main effect generalizations. The second, reweighting methods (e.g., Tipton, 2013; and described in Schochet, Puma, and Deke, 2014), adjusts grand mean impact findings from experiments for differences between study and inference populations in the distribution of moderators of impact. A third method, G-Theory (Cronbach, Rajaratnam, & Gleser, 1963; Shavelson & Webb, 2008), focuses on making explicit the interactions between the treatment and attributes of study participants and of contexts as a basis for drawing generalization. While the three methods are fundamentally different in terms of quantities supporting generalization, they all rely on moderator effects to establish the extent to which grand mean impact findings generalize.

Tests of Differential Impact Support Decisions

The correct question for evaluating if a program selectively benefits one group more than another is whether the difference in impact between subgroups is statistically significant, and not whether the impact is statistically significant for one group but the other (Gelman and Stern, 2006). This information, along with local data about subgroup baseline performance, and the proportions of individuals in each of the subgroups, can inform judgements about whether a program is increasing or closing an achievement gap. In turn, this tells us about how a program apportions benefits, and whether it is even possibly harmful for subsets of the inference population. Based on this information, stakeholders can decide whether to retain, modify or discontinue a program.

A Note About The Feasibility of Evaluating Differential Impacts

While it is important to evaluate differential impacts for several reasons, including those above, there remains the question of the feasibility of doing so. Before continuing with this work, it is important to briefly address three potential objections.

A first concern is that differential impacts are normally either very small, or too small to matter. To address this, we note that both laboratory social experiments (Cronbach, 1975) and field experiments (Jaciw & Lin, 2020; Cheung & Slavin, 2012; Edmunds et al., 2012; James-Burdumy et al., 2010; Weiss et al., 2017) have shown that in evaluations of educational programs differential effects across subgroups of students often approach, and sometimes exceed, magnitudes of average impact. A second concern is that even if differential impacts are substantive, power to detect them may be low, especially in experiments designed to detect average impacts. We address this by clarifying the assertion. Both theoretical and empirical works addressing factors influencing precision of average and differential effects estimates from RCTs (Jaciw et al., 2016; Bloom, 2005; and Spybrook, Kelcey, & Dong, 2016) show that when clusters such as school are randomized, power to detect differential impacts across subgroups of individuals within clusters (e.g., students or teachers) may be adequate with study designs powered to detect average impacts. On the other hand, it is harder to achieve adequate power to detect differences in impact across subgroups of the randomized clusters themselves. In this work, we are concerned with the former case; that is, with differences in impact across subgroups of individuals that are identified within each site. A third possible concern is that there are usually many more moderator effects to examine than main effects, and after applying multiple comparison adjustments, power for detecting any given differential effect will be low. This is a legitimate point, but the problem and its solution apply equally to any situation involving multiple comparisons, including when evaluating a series of average impacts. The goal should be to evaluate a modest number of contrasts. This concern reminds us of the need for a disciplined and economical selection of moderators. A small number of moderators selected a-priori for investigation on the basis of either theory or policy priorities addresses the problem of too many contrasts.

In sum, questions concerning differential impacts of programs for subgroups of individuals are important to address for several reasons. Further, in the context of evaluations in education, these differential impacts are often large enough that they can be evaluated with adequate power and precision using common experimental

designs. These reasons give impetus to better understand the potential of different study designs to support accurate inferences concerning differential effects. The current work addresses this point.

Use of Quasi-Experiments to Evaluate Average and Differential Impacts

RCTs are the first choice for evaluating both average and differential program impacts. Randomized experiments, when conducted well and with intact samples, allow us to estimate average and subgroup impacts with high internal validity. However, when randomized experiments are not feasible, QED's may be used instead to estimate similar quantities (Shadish, Cook, & Campbell, 2002).

As noted earlier, a principal type of QED is the CGD. With this design, the goal is to find comparison cases that represent a valid counterfactual group to individuals receiving treatment. That is, the comparison cases are chosen strategically to warrant the strong ignorability assumption (Rosenbaum & Rubin, 1983). In other words, the comparison group should be selected, and confounders of treatment identified and their effects controlled for, so that there remain no hidden factors that influence both selection into treatment and the outcome. If present, such factors would bias the estimate of the causal effect of treatment that involves a comparison of outcomes for the two groups.

While the conditions for achieving unbiased QED estimates are easy to state in principle, they are not guaranteed to be satisfied in practice. Over the past several decades, a body of work has been dedicated to empirically exploring bias in QED estimates by evaluating their capacity to replicate benchmark results from experiments. These efforts, collectively known as Within Study Comparison (WSC) studies, typically start with an experimental benchmark estimate of average impact. They then test the conditions under which QEDs yield results that approximate the benchmark quantity. Often this involves evaluating whether non-experimental comparison groups, that are selected judiciously using different strategies, can replicate how the actual controls perform in order to yield the benchmark experimental result.

WSC studies have produced detailed results and guidance for designing QEDs that yield internally valid results.[1] An important finding that is especially salient to this work is the potential for QEDs to be biased from effects of what are called "third variable confounds" (Cook, Shadish, & Wong, 2008), also described as "macros factors" (Hotz, Imbens, & Mortimer, 2005). These are "variables whose values are constant within a location, or at least within sub-locations" (Hotz, Imbens, & Mortimer, 2005, p. 248). The uniqueness of values of macro variables to individual sites presents a problem for CGD designs that involve making comparisons across sites. For instance, if the treatment group is at Location 1 and the comparison group is from Location 2, and the sites are wholly different on a basic variable that affects outcomes, then it is impossible to adjust for the biasing effect of this fundamental confound. That is, common support is lacking to permit matching individuals across

[1] We refer the reader to thorough summaries in Bloom, Michalopoulos, and Hill (2005); Cook, Shadish, and Wong (2008); Glazerman, Levy, and Myers (2003); and Wong, Valentine, and Miller-Bains (2017). More recent WSC studies include Bifulco (2012); Dong and Lipsey (2015); and Hallberg, Cook, Steiner, and Clark, (2016).

locations. In a sense, the locations *are* the individuals, and there are only two of them—one in treatment and one in control.

Consider an example from education. Schools consist of individuals at multiple levels. There are students, teachers, and the administration, often headed by one principal. If comparing two sites, we may find ranges of overlap on student pretests, and teacher experience levels. However, with only two schools, any baseline covariates affecting outcomes at the school level are fully confounded with site. An example of a macro variable is the management style of the principal. We have information for only $N = 2$ administrators, one at the treatment site and one at the untreated comparison site. If we try to establish performance for the counterfactual to treatment at one site using the performance of students who have not experienced treatment at the other site, the comparison will be completely confounded with characteristics of principals. That is, in our estimation we will not be able to de-confound the treatment effect from the "principal effect."

To address the role of macro variables, WSC studies have emphasized the need to match locally (Bloom, Michalopoulos, & Hill, 2005), which helps to ensure that the treatment and comparison groups are similar or constant in terms of possible macro variables, including unknown factors. As an analytic solution, Hotz, Imbens, and Mortimer (2005) propose using multiple locations to support model-based adjustments for effects of location level characteristics. We return to this important point later in this article.

Our Assertion

We argue in this work that differential impact quantities from QEDs that involve cross-site comparisons of performance gradients are less prone to selection bias than average impact quantities that involve the same cross-site comparisons but of average performance outcomes. Specifically, when there is a reliance on cross-site comparisons to estimate both average and differential impact, the latter estimate is less prone to selection bias attributable to macro factors.

In the next section: (a) we present our argument for why estimates of differential impact from QEDs are less prone to selection bias than estimates of average impact; (b) we apply a WSC-based approach to empirically test the assertion, and (c) we consider implications and applications of the result. As we go through these steps, we demonstrate how certain common statistics used with cluster randomized experiments may be reinterpreted for understanding bias in the context of WSC studies.

A Model for Evaluating Bias in Effect Estimates Involving Cross-Location Comparisons

We motivate our argument graphically and through an impact model.

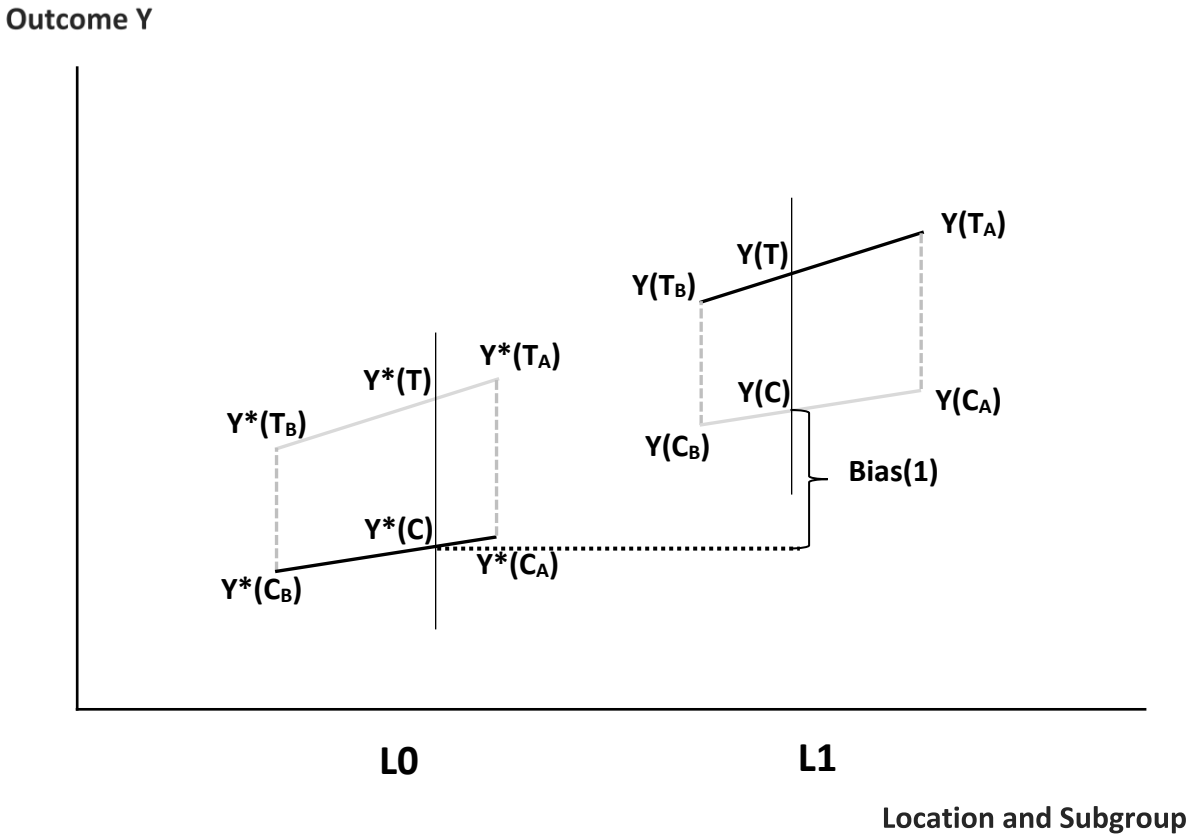

FIGURE 1. AVERAGE OUTCOMES ACROSS TWO LOCATIONS AND FOR TWO SUBGROUPS

In Figure 1, we represent impact quantities at two locations (L0 and L1). At L1 the average impact is $Y(T) - Y(C)$ and at L0 it is $Y^*(T) - Y^*(C)$. At L1, the subgroup impacts are $Y(T_A) - Y(C_A)$ and $Y(T_B) - Y(C_B)$ for subgroups A and B, respectively. At L0, they are $Y^*(T_A) - Y^*(C_A)$ and $Y^*(T_B) - Y^*(C_B)$. We represent true values at each site, and assume they could be estimated without bias through a randomized experiment.

Next, consider the scenario where we lack information about the control outcome, $Y(C)$, at L1. Assume that under a CGD, we substitute the corresponding value, $Y^*(C)$, from L0, for the counterfactual to treatment at L1. Bias in the difference estimate used to infer impact at L1 is:

$$Bias(1) = Y(T) - Y^*(C) - [Y(T) - Y(C)] = Y(C) - Y^*(C) \quad (1)$$

This is a standard expression for bias in comparison-group designs evaluated through WSC studies. That is, bias is expressed as a difference in average performance across untreated groups (Bloom, Michaolopoulos, & Hill, 2005; Heckman, Ichiumra, & Todd, 1997).

Next, assume our goal is to measure the differential impact across subgroups A and B at L1. For example, our goal may be to evaluate whether impact for subgroup A (e.g., for males) is different than impact for subgroup B (e.g., females) at that location. We would like to estimate the following quantity:

$$\Delta_{A-B} = Y(T_A) - Y(C_A) - [Y(T_B) - Y(C_B)] \quad (2)$$

Assume that, as in the previous case, we lack information about control performance (i.e., performance in the absence of treatment) at L1. We can rewrite the quantity in Equation 2, as:

$$\Delta_{A-B} = Y(T_A) - Y(T_B) - [Y(C_A) - Y(C_B)] \quad (3)$$

Under a non-equivalent comparison group design, we replace the gradient in control group performance at L1 with the corresponding quantity at L0 (as we did earlier with the average impact quantity):

$$\Delta^*_{A-B} = Y(T_A) - Y(T_B) - [Y^*(C_A) - Y^*(C_B)] \quad (4)$$

Bias in this expression is as follows:

$$Bias(2) = \Delta^*_{A-B} - \Delta_{A-B} = (Y(T_A) - Y(T_B) - \left(Y^*(C_A) - Y^*(C_B)\right) \quad (5)$$
$$-(Y(T_A) - Y(T_B) - (Y(C_A) - Y(C_B))$$
$$= (Y(C_A) - Y(C_B)) - (Y^*(C_A) - Y^*(C_B))$$

This expression, like the one in Equation 1, involves outcomes only for untreated groups at each location.

Next, we show how macro effects persist in *Bias(1)* when comparing mean outcomes across locations, but cancel in the expression for *Bias(2)* when comparing mean performance gradients between subgroups across locations. To do so, we develop our model slightly further. First, we assume that a common macro effect, Q, differentiates performance across locations for both subgroup categories:

$$Y(C_A) = Y^*(C_A) + Q \quad (6)$$

$$Y(C_B) = Y^*(C_B) + Q \quad (7)$$

If subgroups A and B are mutually exclusive and exhaustive of samples at L0, and at L1, we can express average control performance at each location as:

$$Y(C) = \pi_A Y(C_A) + (1 - \pi_A) Y(C_B) \quad (8)$$

$$Y^*(C) = \pi^*_A Y^*(C_A) + (1 - \pi^*_A) Y^*(C_B) \quad (9)$$

Next, we rewrite *Bias(1)* using these terms:

$$Bias(1) = Y(C) - Y^*(C) \quad (10)$$
$$= [\pi_A Y(C_A) + (1 - \pi_A) Y(C_B)] - [\pi^*_A Y^*(C_A) + (1 - \pi^*_A) Y^*(C_B)]$$
$$= [\pi_A (Y^*(C_A) + Q) + (1 - \pi_A)(Y^*(C_B) + Q)]$$
$$-[\pi^*_A (Y^*(C_A)) + (1 - \pi^*_A) Y^*(C_B)]$$
$$= [\pi_A Y^*(C_A) + \pi_A Q + Y^*(C_B) + Q - \pi_A Y^*(C_B) - \pi_A Q)]$$

$$-[\pi_A^* Y^*(C_A) + Y^*(C_B) - \pi_A^* Y^*(C_B)]$$
$$= [\pi_A Y^*(C_A) + Q - \pi_A Y^*(C_B))] - [\pi_A^* Y^*(C_A) - \pi_A^* Y^*(C_B)]$$
$$= Q + [(\pi_A - \pi_A^*) Y^*(C_A)] - [(\pi_A - \pi_A^*) Y^*(C_B)]$$
$$= Q + (\pi_A - \pi_A^*)(Y^*(C_A) - Y^*(C_B))$$

Bias(1) reflects the macro variable Q plus the standard subgroup confounding effect that exists if there is difference between locations in proportions of the subgroups, $(\pi_A - \pi_A^*)$, and a difference between subgroups in their average performance $Y^*(C_A) - Y^*(C_B) \neq 0$. Bias from macro effect Q persists, even if we adjust for effects arising from the imbalance between locations in the distribution of subgroups A and B.

Next, we rewrite $Bias(2)$ in similar terms (substituting quantities in Equations 6 and 7 into Equation 5):

$$Bias(2) = \Delta_{A-B}^* - \Delta_{A-B} \quad (11)$$
$$= [Y(C_A) - Y(C_B)] - [Y^*(C_A) - Y^*(C_B)]$$
$$= [Y^*(C_A) + Q - (Y^*(C_B) + Q)] - [Y^*(C_A) - Y^*(C_B)] = 0$$

In this scenario, the macro effect, Q, is differenced away. Also, the quantity is, by definition, stratified by subgroups A and B. This eliminates the additional confound due to a possible difference between conditions in the distribution of the subgroups, which was observed in Equation 10.[2]

How Do We Interpret These Results in Terms of the Graphic Representation?

$Bias(1)$ can be represented as the difference in heights of $Y(C)$ and and $Y^*(C)$ in Figure 2. We can decompose this into two components: (a) the average vertical displacement of the quadrilaterals (the macro effect), and (b) the difference in heights of $Y(C)$ and $Y^*(C)$ attributable to difference between locations L0 and L1 in the proportion of individuals belonging to subgroups A compared to B. This imbalance is represented by the solid vertical line being placed further to the right at L1 representing a larger proportion of cases in subgroup A in L1 compared to L0.

[2] There may of course be imbalance on other individual-level factors. For example, we might have a covariate with value $X_{i,L0}$ for a given individual i at L0, and $X_{j,L1}$ for a given individual j at L1. We can write each of these quantities as the per location mean of the variable plus the deviation in the individual score from its respective mean: $X_{i,L0} = (X_{i,L0} - \bar{X}_{L0}) + \bar{X}_{L0}$ for the person at L0, and $X_{j,L1} = (X_{j,L1} - \bar{X}_{L1}) + \bar{X}_{L1}$ for the person at L1. $\bar{X}_{L0}$ and $\bar{X}_{L1}$ are macro variables specific to sites and would be integrated into bias Q. Bias may still exist from differences between the shapes of distributions of $(X_{i,L0} - \bar{X}_{L0})$ centered around mean $\bar{X}_{L0}$, and of $(X_{j,L1} - \bar{X}_{L1})$ centered around mean, $\bar{X}_{L1}$. For instance, the variance or skew in the distributions of individual scores around their respective site means may be different for the two locations.

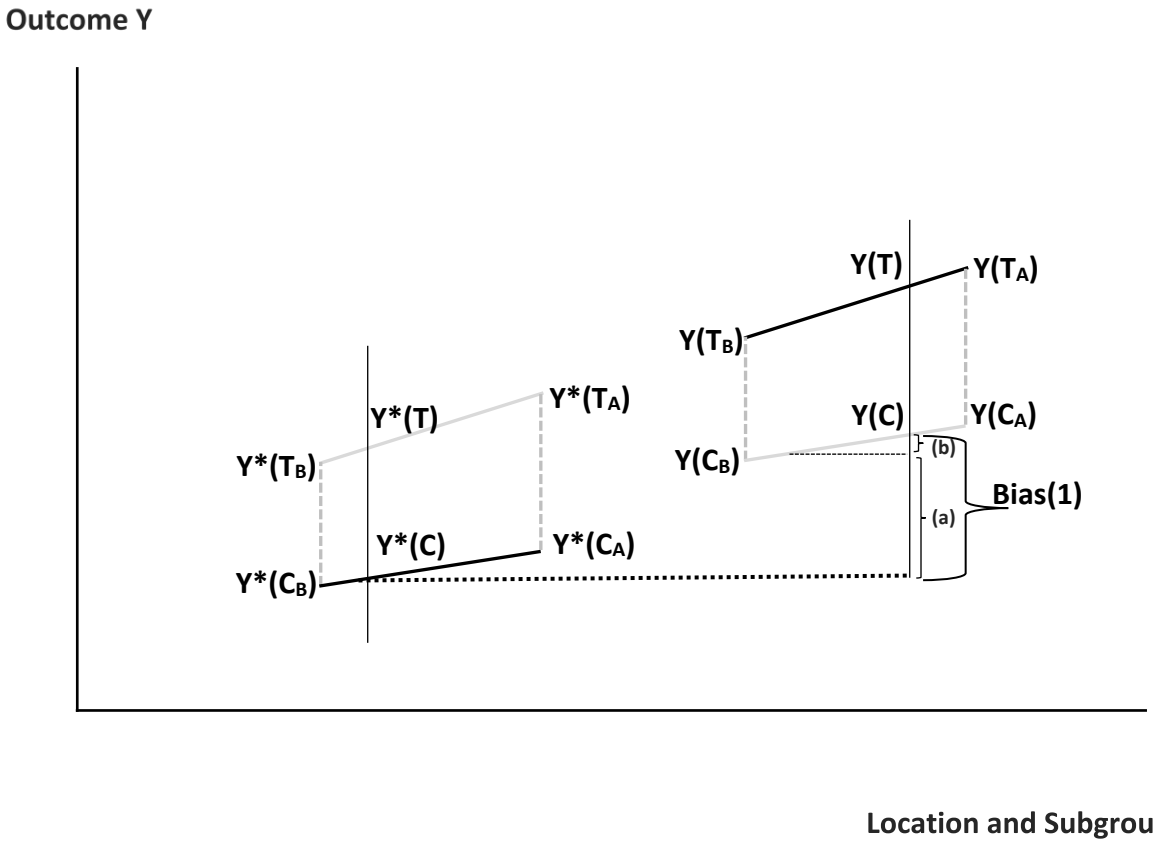

FIGURE 2. AVERAGE OUTCOMES ACROSS TWO LOCATIONS AND FOR TWO SUBGROUPS WITH COMPONENT BIASES DISPLAYED

MODELING A MORE REALISTIC SCENARIO

Under the simplified scenario considered above, *Bias(2)* does not depend on the displacement of the quadrilaterals. That is, the macro effect, which is represented as the overall vertical distance between the quadrilaterals, is differenced away.

A more realistic scenario assumes that the gradient in performance between subgroups A and B is not constant across locations. That is, we can assume the variance in the gradient in performance between subgroups across sites or clusters – referred to as the "moderator gap variance" (Spybrook, 2013) – is not zero. This is a fair assumption, with non-zero values of variance in the performance gradient across sites observed in educational experiments (Jaciw et al., 2016).

The value of the gradient is specific to the site; therefore, it is a macro effect. We model the difference between locations in the gradient in control performance across *A* and *B* as follows, with the macro value *K* determining the non-constant slope in achievement between subgroups across locations:

$$Y(C_A) = Y^*(C_A) + Q \quad (12)$$

$$Y(C_B) = Y^*(C_B) + Q + K \quad (13)$$

In Appendix A we show that in this scenario the expressions for $Bias\ (1)$ and $Bias(2)$ are as follows:

$$Bias\ (1) = (1 - \pi_A)K + Q + [(\pi_A - \pi_A^*)Y^*(C_A)] - [(\pi_A - \pi_A^*)Y^*(C_B)] \quad (14)$$

$$Bias(2) = \Delta_{A-B}^* - \Delta_{A-B} = -K \quad (15)$$

Bias(1) is the same as before, with an additional term involving macro effect *K*. *Bias(2)* is now non-zero, also reflecting the macro effect for location differences in performance gradients across subgroups. If the gradients are constant across sites, then $K = 0$ and the quantities reduce to the former ones (Equations 10 and 11).

We can interpret this in terms of the graphical representation (Figure 3.) To simplify, we assume a balanced distribution of individuals across subgroups A and B between locations (we have equal displacement of the solid vertical lines within each quadrilateral), which sets the value in square brackets in Equation 14 to zero, allowing us to focus on the macro effects *K* and *Q*.

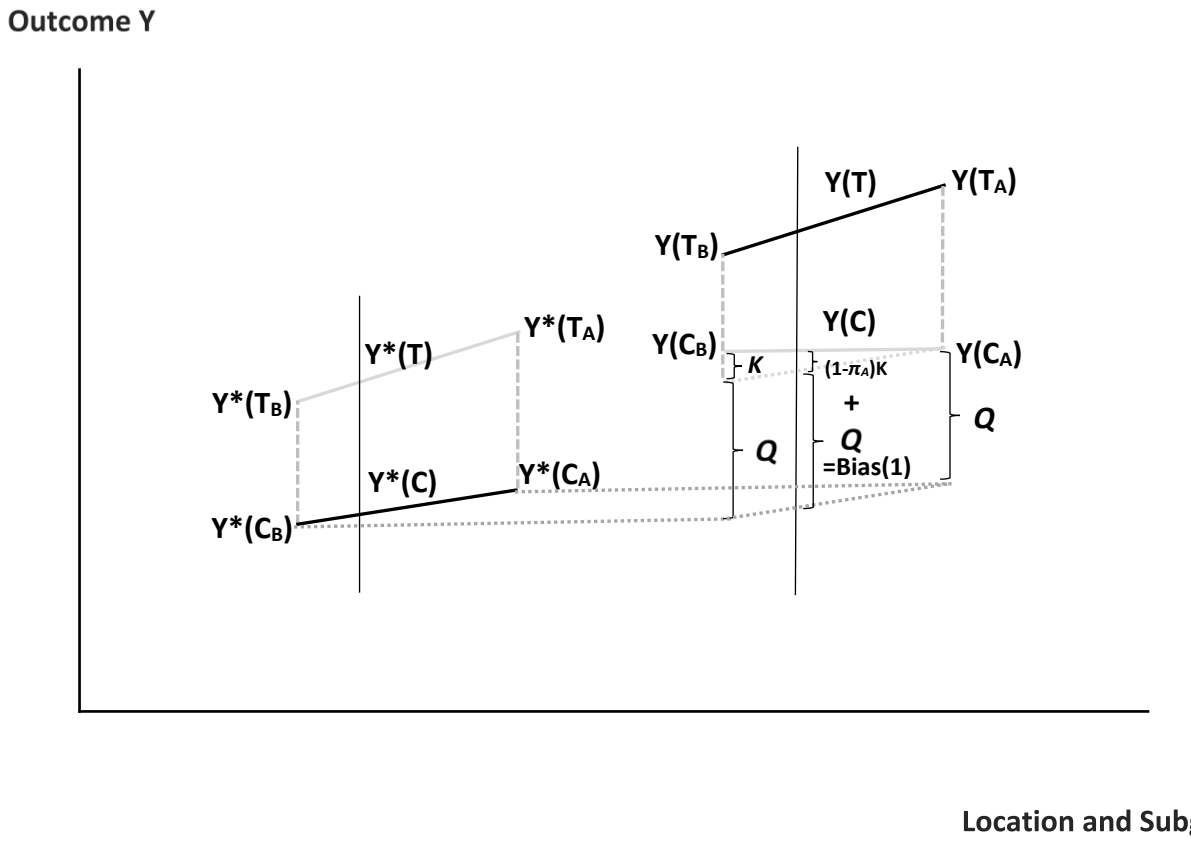

FIGURE 3. AVERAGE OUTCOMES ACROSS TWO LOCATIONS AND FOR TWO SUBGROUPS ASSUMING A DIFFERENCE IN CONTROL PERFORMANCE GRADIENT

In this scenario, the difference between subgroups in control performance at L1 is flat (i.e., zero slope). At L0 this gradient in performance is positive (as before, in Figures 1 and 2). The vertical displacement between quadrilaterals in the bottom-left vertex (i.e., average performance for controls assuming all cases at each site belong to B (i.e., $\pi_B = 1$ or equivalently $\pi_A = 0$) is $Q + K$ (consistent with Equation 14 above.)) As the proportion of cases in A increases and the proportion in B decreases, the solid vertical lines slide to the right. In the expression for $Bias(1)$, the quantity Q remains constant, but the contribution of K becomes discounted by factor $(1 - \pi_A)$. $Bias(1)$ sums to $(1 - \pi_A)K + Q$. If at both sites all members belong to A, the first terms becomes zero and $Bias(1)$ is just Q.

$Bias(2)$ is simply the difference between the slope of the control gradient at L1 (which is flat and has value 0) and the slope of the control gradient at L0, which is K. $Bias(2)$ is $0 - K = -K$.

The main idea in the examples above is that the difference in performance gradients between conditions across locations is subject to a "differencing away" of certain location-specific effects. The macro effect represented by the displacement of the quadrilaterals (Q) does not enter into the expression for $Bias(2)$. Only the macro factor influencing the performance gradient across locations (K) contributes to this bias. However, the difference in average performance between conditions across locations, is subject to the biasing effects of both macro effect Q, which is not differenced away, and a factor of macro effect K.

The potential for error terms to be differenced away is well understood in the context of difference-in-difference analyses. The current application resembles the situation in cluster-randomized experiments where cluster-level random effects are differenced away when estimating the difference in program impact between subgroups of individual, and therefore do not contribute to the standard error for the estimate of the differential effect (Jaciw et al., 2016; Bloom, 2005; Spybrook, Kelcey, & Dong, 2016).

Methods: Evaluating Bias in Cross-site Comparisons of Performance Gradients

Our goal is to evaluate levels of bias in CGS-based estimates of differential impact that depend on calculating the difference in performance gradients between conditions across locations (L0 and L1). A second goal is to compare this bias in estimates of moderated impact to bias in estimates of average impact that also involve comparing outcomes across locations.

WSC Methods

To address these goals, we use a version of the Within Study Comparison (WSCs) methodology. As described earlier, the method is normally used to evaluate discrepancies between experimental and comparison-group-based estimates of average impact.

With WSC methods, the difference between the comparison group-based average impact quantity and the experimental benchmark reduces to a difference in average performance between the comparison and control groups (Heckman, Ichimura, & Todd, 1997; Bloom, Michalopoulos, & Hill, 2005). The difference between these quantities across two locations is as follows (we repeat Equation 1 here for ease of reference):

$$Bias(1) = Y(T) - Y^*(C) - [Y(T) - Y(C)] = Y(C) - Y^*(C) \quad (16)$$

We showed earlier that bias in a differential impact quantity that relies on a comparison of performance gradients across conditions and locations also reduces to a contrast between locations in quantities measured exclusively among controls (we repeat Equation 2 here for ease of reference):

$$Bias(2) = \Delta^*_{A-B} - \Delta_{A-B} = [Y(C_A) - Y(C_B)] - [Y^*(C_A) - Y^*(C_B)] \quad (17)$$

A central feature of WSC studies is their choice of comparison group. The earliest WSC studies, which were conducted using experiments in jobs training (Fraker & Maynard, 1987; Lalonde, 1986), used survey-based outcomes from untreated individuals. Later studies used more proximal comparison cases, including eligible nonparticipants who resided in the same narrowly-defined geographic regions as the program applicants (e.g., Heckman, Ichimura and Todd, 1997). One variant of the WSC method, which we refer to as the "multisite version", takes advantage of experiments conducted at multiple sites. With this version, the benchmark experimental impact at a given inference site is compared to the impact quantity that results from substituting control performance for that site with control performance from a different site (or combination of sites) from the same trial. Controls used for the substitution effectively serve as a non-experimental comparison group for the inference site (Jaciw, 2016; Bloom, Michalopoulos, & Hill, 2005; Wilde & Hollister, 2007). As with all WSC studies, the difference between comparison-group-based and benchmark impact quantities estimates bias in the former quantity. With the multisite version of WSC, bias reflects selection into sites. Taking this approach one step further, we can also accommodate cluster randomized experiments, where units such as schools are randomized to conditions. A given control cluster serves as the unit of inference. One or more other schools assigned to control yield the non-experimental counterfactual.[3] We adopt this approach in the current work.

Selecting Benchmark and Comparison Cases and Summarizing Bias in the Multi-site Trial Version of WSC

In a given trial, we can arbitrarily select any site as the "inference site" that furnishes average control performance $Y(C)$, and designate controls from one or more other sites as constituting the comparison group, which yields the non-experimental counterfactual outcome $Y^*(C)$.

---

[3] In a multisite trial, study participants randomized to control within a site are a random sample of all study participants at that site. Therefore, their performance is the same, on average, as the average performance for all participants at that site in the absence of treatment. This means the difference between sites in average performance of individuals randomized to control (in the case of a multisite trial) is in expectation the same as the difference between the full sites in their average performance if each is randomized as a full cluster to control (i.e., in the case of a cluster randomized trial).

More formally, with cluster randomized trials, for a cluster j randomized to control (T=0), the expected value of the outcome y_{ij} over students i (i.e., the outcome averaged over all students in the control school) is as follows: $E(y_{i,j=1}|T=0, j=1)$. This quantity is the same as the expected value of control performance within that cluster where students have been randomly assigned to treatment or control within that cluster (i.e., in the case of a multisite trial with random assignment of students at the site): $E(y_{i\in T=0,j=1}|j=1)$. The expected value of the difference in control outcome between two clusters, where full clusters are randomly assigned to control (for cluster randomization), or a random subset of students within each cluster assigned to control (for within-site randomization of students) is also the same: $E(y_{i,j=1}|T=0, j=1) - E(y_{i,j=2}|T=0, j=2) = E(y_{i\in T=0,j=1}|j=1) - E(y_{i\in T=0,j=2}|j=2)$.

With N possible inference sites, we have $\binom{N}{2}$ unique pairs of sites to draw comparisons of type $Y^*(C) - Y(C)$ or $\Delta^*_{A-B}(C) - \Delta_{A-B}(C)$. Bias from unique pairings can be expressed as $Y_i(C) - Y_j(C)$, $i \neq j$ for comparisons of means and $\Delta_{i,A-B}(C) - \Delta_{j,A-B}(C)$, $i \neq j$ for comparisons of subgroup performance gradients across sites.

To summarize bias through a measure of central tendency, we can look for an average over all differences; however, if biases are centered on zero, a straight average will obscure the magnitude of bias (Bloom, Michalopoulos, & Hill, 2005). Instead, we can take an average over absolute values of differences for all unique pairs of sites across which control performance is compared: $\frac{1}{\binom{N}{2}} \sum_{i \neq j} |Y_i^*(C) - Y_j \ \ (C)|$. This expresses average absolute bias in the impact quantity that is based on cross-site comparisons. Similarly, $\frac{1}{\binom{N}{2}} \sum_{i \neq j} |\Delta^*_{i,A-B}(C) - \Delta_{j,A-B}(C)|$ is the average absolute bias in the measure of differential impact that uses between-site comparisons in performance gradients between subgroups A and B.

A somewhat different approach to summarizing bias stems from the recognition by Bloom and colleagues (2005) that in many WSC studies, the discrepancy $Y_i(C) - Y_j(C)$ over multiple pairs is centered on zero, and therefore, may be considered a form of random error. They label this "non-experimental mismatch error". The error reflects non-random selection of individuals into sites across which comparisons are drawn. We can similarly define non-experimental mismatch error for performance gradients between subgroups.

We express mismatch error from discrepancies between benchmark- and comparison-groups-based average performance as: $\sqrt{\frac{1}{\binom{N}{2}} \sum_{i \neq j} (Y_i^*(C) - Y_j(C))^2}$.

Similarly, we express mismatch error from discrepancies between benchmark and comparison-group-based performance gradients as: $\sqrt{\frac{1}{\binom{N}{2}} \sum_{i \neq j} (\Delta^*_{i,A-B}(C) - \Delta_{j,A-B}(C))^2}$.

A small modification of these expressions for mismatch error, allows us to summarize absolute levels of bias in more familiar terms. That is, we can summarize discrepancies across sites in average outcomes (or in performance gradients between subgroups) using familiar variance expressions. To do this, instead of comparing control outcomes from one site against control outcomes from just one other site, as we do in the expressions for mismatch error above, we can compare control outcomes for each site against a measure of central tendency for the outcome. Corresponding expressions to those above are as follows:

$$RMSB(Impact) = \sqrt{\frac{1}{N} \sum_i (Y_i^*(C) - \bar{Y}(C))^2} = \sqrt{\tau_{0(C)}} \tag{18}$$

for means, and

$$RMSB(Diff) = \sqrt{\frac{1}{N} \sum_i (\Delta^*_{i,A-B}(C) - \bar{\Delta}_{A-B}(C))^2} = \sqrt{\tau_{1(C)}} \tag{19}$$

for gradients.

$\tau_{0(C)}$ is the variance across clusters in average performance of controls; $\tau_{1(C)}$is the variance across clusters in the control group performance differential between the subgroups (i.e., it is the variance in the slope of the moderator variable that identifies the subgroups). *RMSB* stands for "Root Mean Squared Bias".

Expressing the Magnitude of Bias in Terms of Useful Metrics

The variance expressions above allow us to summarize the discrepancies between sites in average performance, or in performance gradients, in terms that are common to experimental evaluations. We consider three metrics.

1. Standardized Mean Squared Bias

First, we express *RMSB* quantities in the metric of the standardized effect size:

For means:

$$\frac{RMSB(Impact)}{\sqrt{\tau_{0(C)}+\sigma_C^2}} = \sqrt{\frac{\tau_{0(C)}}{\tau_{0(C)}+\sigma_C^2}} = \sqrt{ICC(C)} \tag{20}$$

For gradients:

$$\frac{RMSB(Diff)}{\sqrt{\tau_{0(C)}+\sigma_C^2}} = \sqrt{\frac{\tau_{1(C)}}{\tau_{0(C)}+\sigma_C^2}} = \sqrt{\psi(C)} \tag{21}$$

These expressions involve two quantities common in the literature on cluster randomized trials. The first is the intraclass correlation coefficient (*ICC*) indicating the proportion of total variability in outcomes attributable to between-cluster differences. The second quantity, ψ, is the random variance in the moderator gap across clusters, or "moderator gap variance ratio" (Spybrook, 2013). It is the variability in the gradient in performance between subgroups across clusters also expressed as a proportion of total variance in the outcome.[4] Here we express these quantities among controls only.

Expressing *RMSB* in the metric of a standardized effect size allows us to compare average bias in QED-based average or differential effects relative to familiar benchmarks. For instance, a standardized effect size of .20 may be considered substantial, and quantities as low as .05 can represent a meaningful difference (Bloom, Hill, Black, & Lipsey, 2008).

2. Relative Mean Squared Bias

A simple extension of (1) is to show the *RMSB* values in relation to each other through a ratio:

$$\frac{RMSB(Impact)}{RMSB(Diff)} = \sqrt{\frac{\tau_{0(C)}}{\tau_{1(C)}}} \tag{22}$$

[4] The *ICC* is normally seen as an important parameter in calculations of statistical power for multilevel experimental designs. In this work, we consider an alternative interpretation of the statistic; that is, as a metric for bias in QED-based impact estimates obtained through cross-site comparisons.

3. Root Mean Squared Bias as a Proportion of Average Impact

A third approach is to express *RMSB* relative to the magnitude of impact or differential impact. If average effects are normally larger than differential effects, then it makes sense to scale average absolute bias as a proportion of the magnitude of each effect. The expression for means is:

$$\frac{RMSB(Impact)}{(\overline{Y(T)}-\overline{Y(C)})} = \frac{\sqrt{ICC(C)}}{(\overline{Y(T)}-\overline{Y(C)})/sd} = \frac{\sqrt{ICC(C)}}{ES(avg)} \tag{23}$$

While for gradients it is:

$$\frac{RMSB(diff)}{\overline{\Delta_{A-B}}(T)-\overline{\Delta_{A-B}}(C)} = \frac{\sqrt{\psi(C)}}{[\overline{\Delta_{A-B}}(T)-\overline{\Delta_{A-B}}(C)]/sd} = \frac{\sqrt{\psi(C)}}{ES(diff)} \tag{24}$$

In the empirical part of this work, we express *RMSB* using the first and third of the three metrics considered here.

We can also relate these expressions to the models for bias from macro factors introduced earlier. Assuming that in our analysis we adjust for imbalance across locations in the distribution of characteristic *A* (which happens by necessity in calculating the achievement gradient) thereby eliminating the term in square brackets for *Bias(1)* in Equation 14, the expressions for *RMSB* can be written in terms of the macro effects *K* and *Q*. For means we have:

$$RMSB(Impact) = \sqrt{\frac{1}{N}\sum_i (Y_i^*(C) - \bar{Y}(C))^2} = \sqrt{\tau_{0(C)}} = \sqrt{Var((1-\pi_A)K + Q)} \tag{25}$$

For gradients we have:

$$RMSB(Diff) = \sqrt{\frac{1}{N}\sum_i (\Delta_{i,A-B}^*(C) - \bar{\Delta}_{A-B}(C))^2} = \sqrt{\tau_{1(C)}} = \sqrt{Var(K)} \tag{26}$$

Standardized Mean Squared Bias Adjusting for Effects of Macro Variables

A further question concerns the extent to which $RMSB(Impact)$ and $RMSB(Diff)$ are reduced by adjusting for effects of macro variables.

The strategy is based on the idea of Hotz, Imbens, and Mortimer (2005) that biasing effects of macro variables may be reduced through model-based adjustments that account for differences across locations in the distributions of those variables. This is consistent with WSC approaches that attempt to account for bias by adjusting impact estimates for effects of confounders, which in case are at the site level. In the empirical part of this work we explore the effects of such adjustments on $RMSB(Impact)$ and $RMSB(Diff)$.

Research Questions

We address the following questions:

1. Based on up to 12 outcomes across six studies, what are the values and medians of estimates of $RMSB(Impact)$ and $RMSB(Diff)$ expressed in standardized effect size units (quantities in Equations 20 and 21, above)?

2. What are the values and medians of estimates of $RMSB(Impact)$ and $RMSB(Diff)$ when expressed as proportions of average or differential impact (quantities in Equations 23 and 24, above)?
3. Our third question is concerned with a proof of concept because we are able to apply it to Study 1 only. We ask: to what extent are $RMSB(Impact)$ and $RMSB(Diff)$ reduced by adjusting for effects of macro variables?

Data

We addressed the research questions using data from six randomized experiments in education. Details of the studies are included in Table B1 of Appendix B. They include RCTs of programs addressing: reform-based math science and technology, second-language development, English language development, math skills with a focus on algebra, and language development of lower-performing readers. Outcomes were assessed using established instruments including state tests and performance measures developed by testing agencies. For several of the programs we evaluated impacts on more than one outcome, yielding up to 12 datapoints.

Estimation: Hierarchical Linear Models

To estimate the relevant quantities, we applied Hierarchical Linear (HL) models using methods described in Raudenbush and Bryk (2002) and Singer (1998). We used SAS PROC MIXED (2008) to obtain estimates of the variance components and of average and differential impact.

Our first research question is about comparing the relative magnitudes of $RMSB(Impact)$ and $RMSB(Diff)$ expressed in standardized effect size units (quantities in Equations 20 and 21). We used *limited conditional models*, in which the moderator of interest is the only covariate, to estimate the site-level variance components. We addressed separately two moderators: gender and socioeconomic status based on Free or Reduced-Price Lunch eligibility. To estimate τ_0, which is the main quantity in the expression for $RMSB(Impact)$, we used the following limited conditional model:

$$y_{ij} = \gamma_0 + \gamma_1 treatment_j + \gamma_2 M_{ij} + \upsilon_j + \varepsilon_{ij} \quad (27)$$

Here *i* indexes students, and *j* indexes the macro unit (e.g., school), y_{ij} is the outcome variable, $treatment_j$ indicates treatment assignment status (coded 1 if assigned to treatment, and 0 if assigned to control), M_{ij} is a dummy variable indicating whether student *i* in macro cluster *j* belongs to one subgroup category or the other (e.g., for the analysis involving gender, 1 for males and 0 for females), υ_j is a random effect at the macro level, and ε_{ij} is a random effect at the student level. We estimate τ_0 through $Var(\upsilon_j)$.

To estimate τ_1, which is used in the expression for $RMSB(Diff)$, we used the following limited conditional model:

$$y_{ij} = \gamma_0 + \gamma_1 treatment_j + \gamma_2 M_{ij} + \gamma_s treatment_j * M_{ij} + \upsilon_j + \nu_j M_{ij} + \varepsilon_{ij} \quad (28)$$

This model is like the previous, but with two additional terms, one for the interaction between the moderator M_{ij} and the dummy variable for treatment, and a random term, ν_j, representing the cluster-level deviation of the difference in performance between categories of M, from the grand mean of this difference. We will estimate τ_1 through $Var(\nu_j)$.

We include the moderator as the only covariate in the impact model (Equation 27) to allow a fair comparison of quantities. That is, the calculation of the differential impact in Equation 28 automatically stratifies the analysis by levels of moderator, M_{ij}; therefore, we must similarly adjust impact in Equation 27 to "net out" the effect of the moderator. If not, then the cross-site variance in M_{ij} may inflate our calculation of bias.[5]

Our second research question is about comparing the magnitudes of $RMSB(Impact)$ and $RMSB(Diff)$ with corresponding values of average and differential treatment effects (quantities in Equations 23 and 24). To estimate the average and differential impact quantities using ANCOVA we adopted *more-fully conditional models* which are identical to the limited conditional model described above, but include a series of macro-level covariates.[6,7]

Our third research question is about the extent to which $RMSB(Impact)$ and $RMSB(Diff)$ are reduced by adjusting for effects of macro variables. We used limited conditional models, with minority status as the moderator, and then examined changes from modeling effects of different sets of macro-level (i.e., site level) variables. They included site-aggregate levels of the variables listed in Table C.1 in Appendix C, in specific combinations: (1) student-based macro variables, (2) teacher-based macro variables, (3) locale-based macro variables, (4) a combination of student and teacher-based macro variables, and (5) a combination of student, teacher and locale-based macro variables. We then repeated the process with gender as the moderator. We

---

[5] Up to this point we have considered estimating variance in control performance $\tau_{0(C)}$ and in the control gradient between subgroups in performance $\tau_{1(C)}$. The differences are assessed between control groups at different sites. In the empirical part of this work, we use variance quantities available from the studies, which are based on outcomes for treatment and control members combined: τ_0 and τ_1. It is possible that variances assessed across both conditions are larger than if assessed in the control condition only (this would be the case if differences in treatment program implementation vary.) In that case, the variance components that we use will inflate bias by some amount. However, if control performance also varies depending on differences in what the counterfactual programs are, or from differences in implementation of a dominant counterfactual program, then we might expect $\tau_0 \approx \tau_{0(C)}$ and $\tau_1 \approx \tau_{1(C)}$. For the one study (Study 1) where were able to assess variation across sites by condition, we found similar values. For instance, $\hat{\tau}_0 = 237.22$, and $\hat{\tau}_{0(C)} = 223.46$, while estimates of variances in the gradient in performance across gender were , $\hat{\tau}_1 = 4.69$, and $\hat{\tau}_{0(C)} = 4.68$, while for the performance gradient across minority categories it was, $\hat{\tau}_1 = 26.76$, and $\hat{\tau}_{0(C)} = 26.39$. Still, this possible limitation should be kept in mind when interpreting results.

[6] The set of student-level covariates depend on the study. They always include the pretest, as well as other covariates, such as a dummy variable to indicate ELL status. Table B1 in Appendix B lists the covariates used in analysis.

[7] Missing values for covariates (other than the moderator) were addressed using the dummy variable method (Puma et al., 2009). This approach involves modeling a series of dummy variables, one for each covariate, that indicate whether the value of the covariate is missing. Where data are missing predominantly at the student level, as was the case in the studies examined, the dummy variable method yields effect estimates with less bias than the tolerance threshold set by the What Works Clearinghouse (described in Puma et al. 2009). Students without a posttest, and students with a missing value for the moderator being analyzed, were excluded from analysis.

explore the question with respect to Study 1 only. Evaluating this question across more studies is needed to provide a definitive answer, that is why we consider this test to be a proof of concept.[8]

Results

Research Questions 1 and 2

We address the first two research questions per moderator, first for gender and then for socioeconomic status.

With Gender as the Moderator

Figure 4 compares $RMSB(Impact)$ and $RMSB(Diff)$, with gender as the moderator. In this case we have 12 results from 6 studies that reported differential impacts for this moderator, coded 1 for male and 0 for female. The median values are .464 and .179, respectively. The result indicates that when estimates are based on cross-site comparisons, smaller levels of absolute bias are observed for differential impacts than for average impacts.

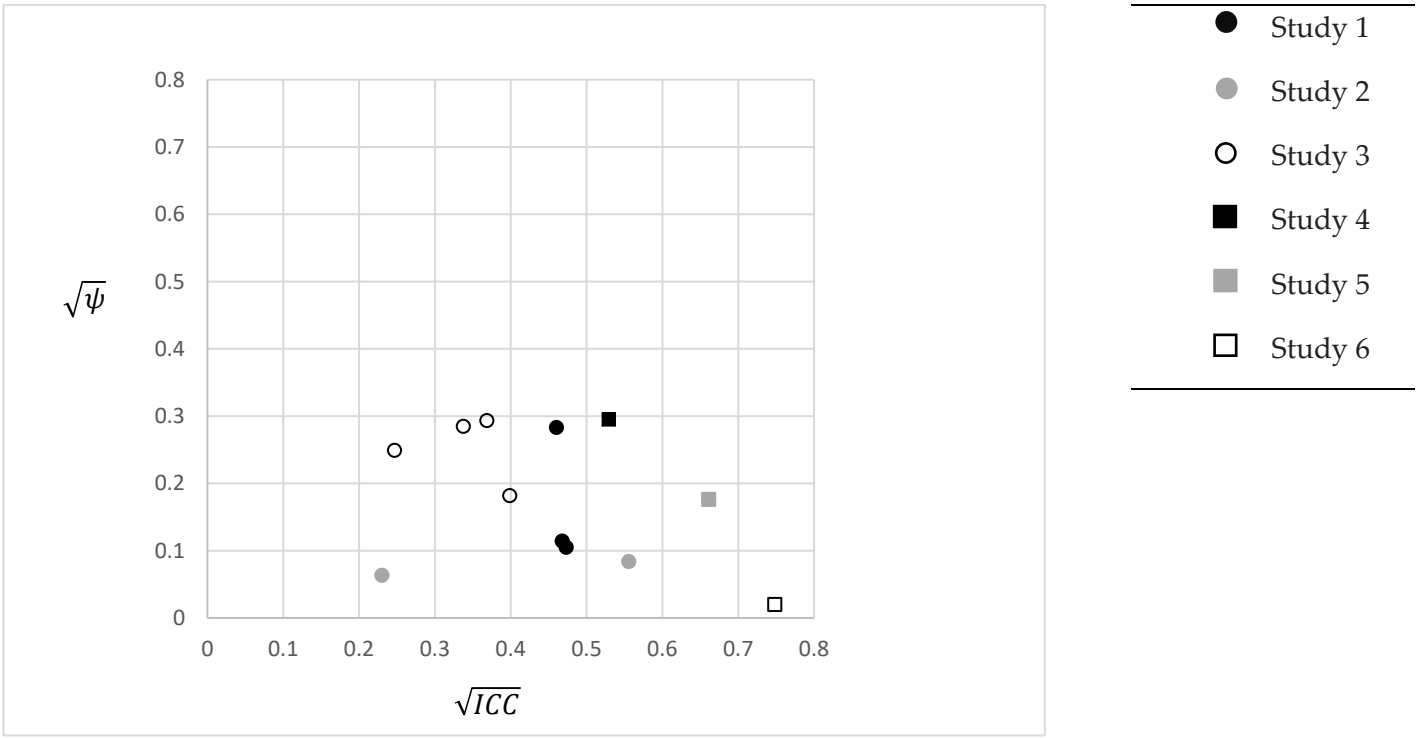

FIGURE 4. RMSB FOR AVERAGE IMPACT AND GRADIENT VALUES FOR DIFFERENTIAL IMPACT BY GENDER

[8] To address the third research question we were able to limit analysis to controls only.

In Figure 5, $RMSB(Impact)$ and $RMSB(Diff)$ are divided by magnitudes of average and differential impacts, respectively. That is, each type of Root Mean Squared Bias is expressed as a proportion of the magnitude of the average effect. The median values are 9.412 and 2.803 for average and differential effects, respectively. The level of bias as a proportion of the corresponding effect size is larger for average than for differential impacts. (While included in calculation of medians, we removed two values from the graph, where the differential impact in the denominators was very small (.001 sd) resulting in very large numbers of the ratio of 104.881 and 282.843.)

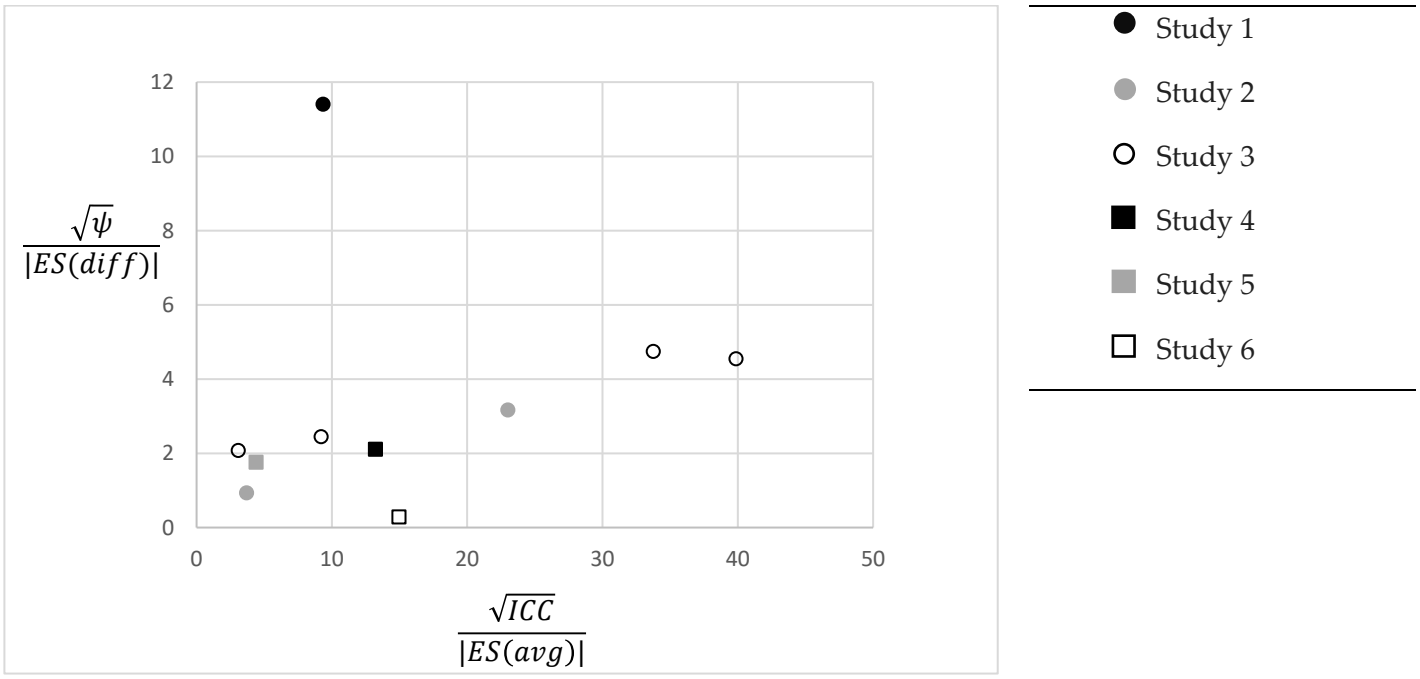

FIGURE 5. RMSB AS A PROPORTION OF AVERAGE IMPACT AND DIFFERENTIAL IMPACT BY GENDER

Note. Two values from Study 1 were removed: ordered pair (9.465, 104.881) and (7.674, 282.843). The large values along the vertical resulted from very small effect sizes for differential impact in the denominator of ~.001, each.

With Socioeconomic Status as the Moderator

Figure 6 compares $RMSB(Impact)$ and $RMSB(Diff)$ with socioeconomic status as the moderator. In this case we have 11 results from 5 studies that reported differential impacts for this moderator (in each case using student Free or Reduced Price Lunch eligibility, with eligible coded "1" and non-eligible coded "0"). The median values are .381 and .197 for the average and differential impact, respectively. As with gender, the result indicates that when estimates are based on cross-site comparisons, smaller levels of absolute bias are observed for differential impacts than for average impacts.

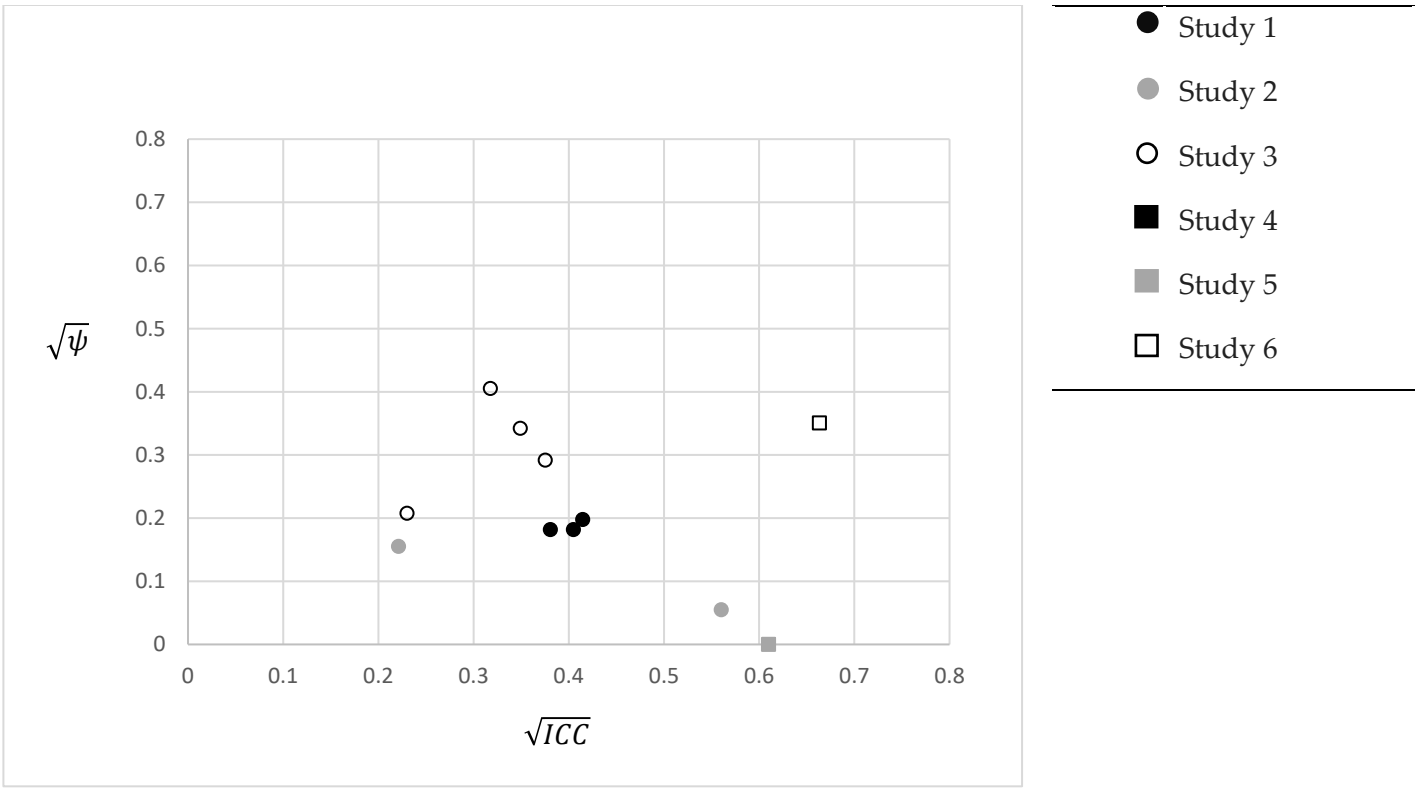

FIGURE 6. RMSB FOR AVERAGE IMPACT AND GRADIENT VALUES FOR DIFFERENTIAL IMPACT BY SOCIOECONOMIC STATUS

In Figure 7, $RMSB(Impact)$ and $RMSB(Diff)$ are divided by magnitudes of average and differential impacts, respectively. That is, each type of Root Mean Squared Bias is expressed as a proportion of the magnitude of the average effect. The median values are 8.295, 1.900 for average and differential effects respectively. As with gender, the level of bias as a proportion of the corresponding effect size is larger for average than for differential impact. While included in the calculation of medians, we removed one value from the graph, where the differential impact in the denominator was very small (.001 sd) resulting in a very large number of the ratio of 54.772.

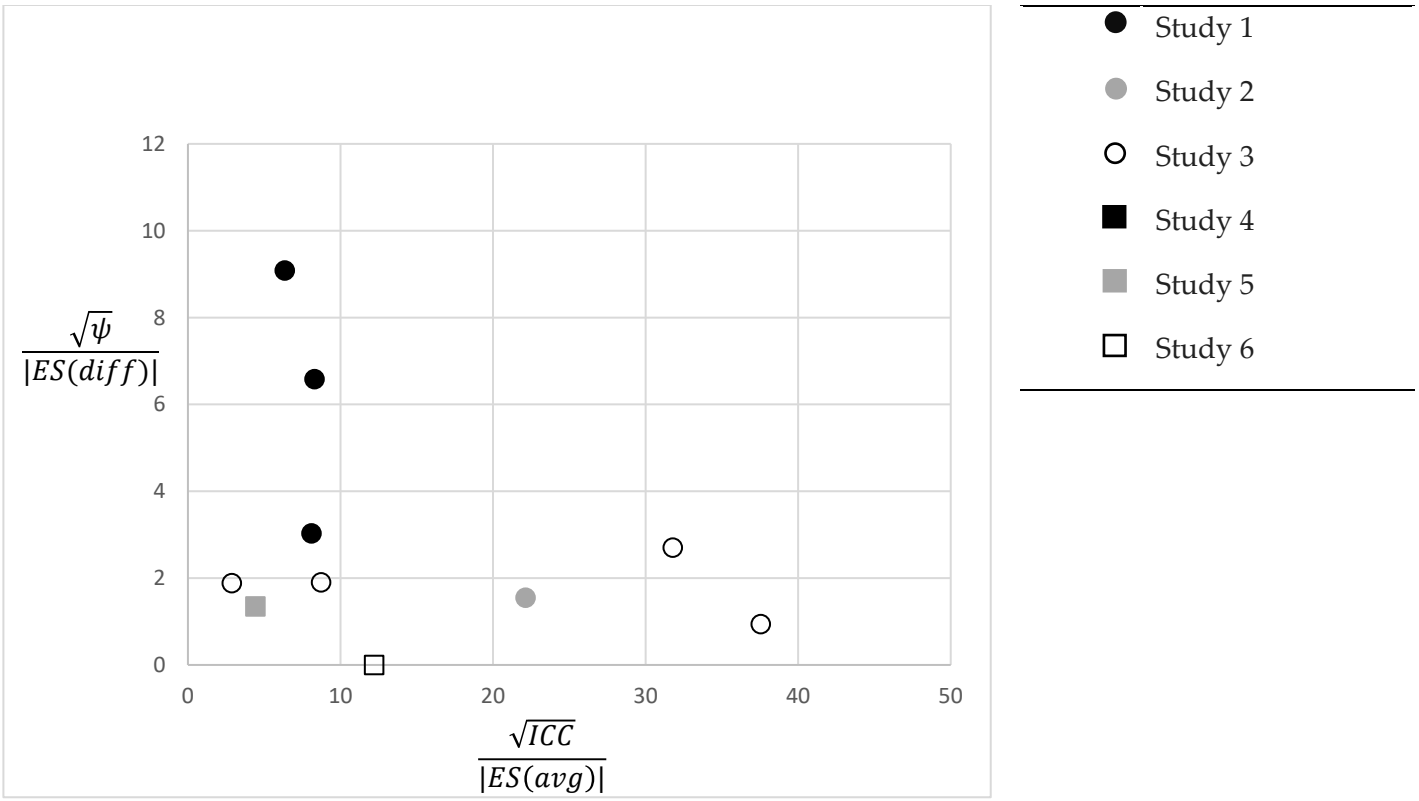

FIGURE 7. RMSB AS A PROPORTION OF AVERAGE IMPACT AND AS A PROPORTION OF DIFFERENTIAL IMPACT BY SOCIOECONOMIC STATUS

Note. Study 4 was removed because the differential impact in the denominator was very small (~.001 sd) resulting in a very large number of the ratio of 54.77.

RESEARCH QUESTION 3

The third research question addresses whether adjusting for differences in macro variables across sites reduces either $RMSB(Impact)$ or $RMSB(Diff)$. The results are based only on Study 1, and therefore should be seen as a proof of concept rather than as definitive.

Tables 1 and 2 and Figure 8, show the magnitudes of $RMSB(Impact)$ and $RMSB(Diff)$ adjusting for effects of macro variables on variation in outcomes across sites. The quantities are expressed in standard deviation units of control group performance before any covariate adjustments.

TABLE 1. ESTIMATES OF STANDARDIZED ROOT MEAN SQUARED BIAS FOR AVERAGE AND DIFFERENTIAL IMPACT CONDITIONAL ON GENDER

		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
RMSB(Impact | gender)							
1	$\frac{RMSB(Impact)}{\sqrt{\tau_{0(C)} + \sigma_C^2}} = \sqrt{\frac{\tau_{0(C)}}{\tau_{0(C)} + \sigma_C^2}} = \sqrt{ICC(C)}$	0.408	0.350	0.337	0.133	0.118	0.000
2	$H_0: \tau_{0(C)} = 0$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	.
RMSB(Diff(gender))							
3	$\frac{RMSB(Diff(gender))}{\sqrt{\tau_{0(C)} + \sigma_C^2}} = \sqrt{\frac{\tau_{1(C)}}{\tau_{0(C)} + \sigma_C^2}} = \sqrt{\psi(C(gender))}$	0.059	0.060	0.048	0.057	0.059	0.050
4	$H_0: \tau_{1(C)} = 0$	$p = .212$	$p = .207$	$p = .284$	$p = .195$	$p = .186$	$p = .226$

Note. Estimation of $\tau_{0(C)}$ for Model 6 reaches the boundary conditions for this effect (Singer & Willett, 2003) resulting in a zero value and no *p*-value given in PROC MIXED. This usually means the point estimate is trivially different from zero.

TABLE 2. ESTIMATES OF STANDARDIZED ROOT MEAN SQUARED BIAS FOR AVERAGE AND DIFFERENTIAL IMPACT CONDITIONAL ON MINORITY STATUS

		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
RMSB(Impact | minority status)							
1	$\frac{RMSB(Impact)}{\sqrt{\tau_{0(C)} + \sigma_C^2}} = \sqrt{\frac{\tau_{0(C)}}{\tau_{0(C)} + \sigma_C^2}} = \sqrt{ICC(C)}$	0.380	0.308	0.309	0.135	0.121	0.000
2	$H_0: \tau_{0(C)} = 0$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	.
RMSB(Diff(minority status))							
3	$\frac{RMSB(Diff(minority))}{\sqrt{\tau_{0(C)} + \sigma_C^2}} = \sqrt{\frac{\tau_{1(C)}}{\tau_{0(C)} + \sigma_C^2}} = \sqrt{\psi(C(minority))}$	0.140	0.136	0.147	0.133	0.126	0.114
4	$H_0: \tau_{1(C)} = 0$	$p = .031$	$p = .032$	$p = .027$	$p = .027$	$p = .034$	$p = .051$

Note. Estimation of $\tau_{0(C)}$ for Model 6 reaches the boundary conditions for this effect (Singer & Willett, 2003) resulting in a zero value and no *p*-value given in PROC MIXED. This usually means the point estimate is trivially different from zero.

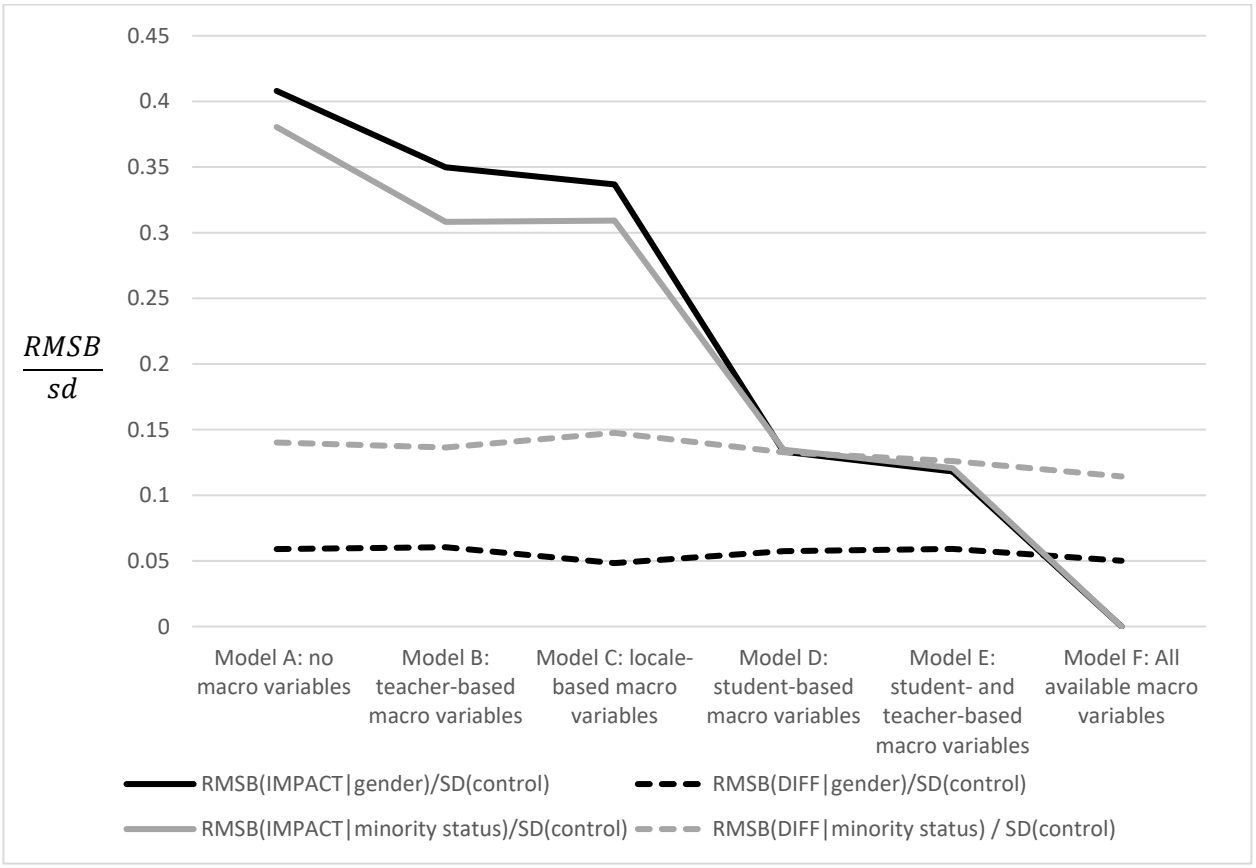

FIGURE 8. LEVELS OF RMSB EXPRESSED IN STANDARDIZED EFFECT SIZE UNITS BEFORE AND AFTER ADJUSTING FOR EFFECTS OF MACRO VARIABLES

We see that covariate adjustments lead to reductions in $RMSB(Impact)$. The estimate of cross-site variability in average performance, $\tau_{0(C)}$, is not statistically significant once effects of all macro-level variables are adjusted-for.

Consistent with findings presented above, $RMSB(Diff)$ is lower than $RMSB(Impact)$ to start with and this holds for both gender and Minority status. $RMSB(Diff)$ associated with Minority status is larger than for gender, indicating more variation across sites in the performance gradient across categories for that moderator. For the gender gradient, adjusting for effects of macro-level covariates does not produce a reduction in $RMSB(Diff)$. With Minority status as the moderator, the estimate of cross-site variability in differential performance, $\tau_{1(C)}$, is not statistically significant once effects of all macro-level variables are adjusted-for. In the case of this one study, adjusting for effects of macro variables across sites dramatically reduced *RMSB* attributable to differences across sites in average performance. In comparison, reductions in *RMSB* attributable to differences across sites in performance gradients were slight.

Conclusions

In this work we evaluated the potential of CGSs to produce unbiased estimates of differential program impact across subgroups of individuals, when comparisons are made between sites. Based on several models, we concluded that bias from macro variables that affect average achievement (our term Q) is differenced away when calculating differences across sites in the performance gradient between subgroups. The same bias term is not eliminated when estimating average impact by comparing differences in average achievement across sites. However, a macro effect associated with differences in the performance gradient across sites (our term K) can still introduce bias in estimates of differential impact based on cross-site comparisons.

We studied these points empirically using a WSC approach. We estimated $RMSB$ for average and differential impact for 12 outcomes from six studies. We found that $RMSB$ quantities associated with CGS-based estimates of differential impact were generally lower than for average impact, as our models had indicated; however, they were not trivially different from zero.

As a proof of concept, with one study we examined whether adjusting for effects of macro variables across sites using regression methods would reduce levels of $RMSB$. The strategy lowered, to near zero, the $RMSB$ associated with differences across sites in average performance, but it had a smaller effect on reducing $RMSB$ for performance gradients across sites. Given the one study used to address this question, we cannot say how typical the result is; therefore, more study of this question is necessary to draw firmer conclusions.

Implications

At the start of this work, we argued for the importance and feasibility of evaluating differential impact quantities. In this work we have demonstrated the potential for using QEDs, specifically CGDs, to produce accurate estimates of differential program impact. Given that adequate statistical power for evaluating differential impacts across subgroups of individuals is achievable (Jaciw et al., 2016; Bloom, Michalopoulos and Hill, 2005) and that CGDs can produce estimates of differential impact with limited bias, we see an opportunity to expand our knowledge base about which programs work better and for whom, by routinely building tests of differential impact into multi-site and cluster randomized experiments. While tests of moderated impact are normally seen as just exploratory, we see a much more prominent role for the evaluation of moderated impacts, especially given the potential for low bias in results.

References

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

Bloom, H. S. (2005). *Randomizing groups to evaluate place-based programs*. In H. S. Bloom (Ed.), Learning more from social experiments (pp. 115–172). New York, NY: Russell Sage Foundation.

Bloom, H. S., Hill, C. J., Black, A. R., Lipsey, M. W. (2008). *Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions* MDRC Working Papers on Research Methodology. New York: MDRC.

Bloom, H. S., Michalopoulos, C., & Hill, C. J. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effect. In H. S. Bloom (Ed.), *Learning more from social experiments* (pp. 173–235). New York, NY: Russell Sage Foundation.

Bryk, A. S. (2014). 2014 AERA Distinguished lecture: Accelerating how we learn to improve. *Educational Researcher*, *44*(9), 467–477.

Cheung, A . C. K., & Slavin, R. E. (2012). The Effectiveness of Educational Technology Applications for Enhancing Reading Achievement in K-12 Classrooms: A Meta-Analysis. A Report from the Best Evidence Encyclopedia.

Cook, T. D. (2002).Randomized experiments in educational policy research: A critical examination of the reasons the education evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*, 175–199.

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

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, 116-127.

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology, 16*, 137-163.

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

Edmunds, J. A., Bernstein, L., Unlu, F., Glennie, E., Willse, J., Smith, A., & Arshavsky, N.

(2012). Expanding the start of the college pipeline: Ninth-grade findings from an experimental study of the impact of the early college high school model. *Journal of Research on Educational Effectiveness, 5*, 136–159.

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

Gelman, A., & Stern, H. (2006). The difference between ''significant'' and ''not significant'' is not itself statistically significant. *The American Statistician, 60*, 328–331.

Glazerman, S., Levy, D., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *American Academy of Political and Social Science, 589*, 63–93.

Hallberg, K., Cook, T.D., Steiner, P.M., & Clark, M.H. (2016). Pretest Measures of the Study Outcome and the Elimination of Selection Bias: Evidence from Three Within Study Comparisons Prevention Science, 1-10.

Heckman, J. J., Ichimura, K., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*, 605–654.

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, *125*, 241–270.

Jaciw, A. P. (2016). Assessing the accuracy of generalized inferences from comparison group studies using a within-study comparison approach: The methodology. *Evaluation Review, (40)*3, 199-240. Retrieved from http://erx.sagepub.com/content/40/3/199.abstract

Jaciw, A. P. & Lin, L., (2020). The Benefits and Feasibility of Assessing Differential Program Impacts on Student Achievement in Education. **Under Review.**

Jaciw, A. P., Lin, L., & Ma, B. (2016). An Empirical Study of Design Parameters for Assessing Differential Impacts for Students in Group Randomized Trials. *Evaluation Review, 40*(5), 410–443. Retrieved from http://erx.sagepub.com/content/early/2016/10/14/0193841X16659600.abstract

James-Burdumy, S., Deke, J., Lugo-Gil, J., Carey, N., Hershey, A., Gersten, R., Faddis, B. (2010). *Effectiveness of Selected Supplemental Reading Somprehension Interventions: Findings From Two Student Cohorts* (NCEE 2010-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved March 25, 2014, from http://ies.ed.gov/ncee/pubs/20104015/pdf/20104015.pdf.

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review, 76*, 604–620.

Puma, Michael J., Robert B. Olsen, Stephen H. Bell, and Cristofer Price (2009). *What to Do*

When Data Are Missing in Group Randomized Controlled Trials (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models (2nd ed.)*. Thousand Oaks, CA: Sage.

Rosenbaum, P. R. & Rubin, B. D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

SAS/STAT Software: *Changes and Enhancements through Release 9.1* [Computer manual] (2020). Cary, NC: SAS Institute, Inc.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding Variation in Treatment Effects in Education Impact Evaluations: An Overview of Quantitative Methods (NCEE 2014–4017)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from http://ies.ed.gov/ncee/edlabs.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shavelson, R. J., & Webb, N. M. (2008). *Generalizability theory and its contributions to the discussion of the generalizability of research findings*. In K. Ercikan & W. M. Roth (Eds.), Generalizing from educational research (pp. 13–32). New York, NY: Routledge.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*, 323- 355.

Somers, M.-A., Corrin, W., Sepanik, S., Salinger, T., Levin, J., & Zmach, C. (2010). *The enhanced reading opportunities study final report: The impact of supplemental literacy courses for struggling ninth-grade readers* (NCEE 2010-4021). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved March 25, 2014, from http://ies.ed.gov/ncee/pubs/20104021/pdf/20104021.pdf.

Spybrook, J. (2013). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *The Journal of Experimental Education, 82*, 1-24.

Spybrook, J., Kelcey, B. & Dong, N. (2016). Power for detecting treatment by moderator effect in two- and three-level cluster randomized trials. *Journal of Educational and Behavioral Studies, 41*. 605 – 627.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics, 38*, 239–266.

Tipton, E. & Olsen, R.B. (2018). A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions. *Educational Researcher, 47*, 516-524

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse (2020).

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness, 10*, 843 – 876.

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management, 26*, 455–477.

Wong, V.C., Valentine, J., Miller-Bain, K. (2017). Covariate Selection in Education Observation Studies: A Review of Results from Within-study Comparisons. *Journal on Research on Educational Effectiveness, 10*, 207-236.

Appendix A. Derivation of Expressions for Bias from Macro Effects in CGSs

Express subgroup control performance at one site as a function of control performance at the other site plus bias terms Q and K:

$$Y(C_A) = Y^*(C_A) + Q \tag{A1}$$

$$Y(C_B) = Y^*(C_B) + Q + K \tag{A2}$$

Assuming subgroups A and B are mutually exclusive and exhaustive of samples at L0, and at L1, we can express average control performance at each location as:

$$Y(C) = \pi_A Y(C_A) + (1 - \pi_A) Y(C_B) \tag{A3}$$

$$Y^*(C) = \pi_A^* Y^*(C_A) + (1 - \pi_A^*) Y^*(C_B) \tag{A4}$$

Next, we write the expression for *Bias(1)*:

$$Bias(1) = Y(C) - Y^*(C) \tag{A5}$$

$$= [\pi_A Y(C_A) + (1 - \pi_A) Y(C_B)] - [\pi_A^* Y^*(C_A) + (1 - \pi_A^*) Y^*(C_B)]$$

$$= [\pi_A (Y^*(C_A) + Q) + (1 - \pi_A)(Y^*(C_B) + Q) + (1 - \pi_A) K))] - [\pi_A^* (Y^*(C_A))$$

$$+ (1 - \pi_A^*) Y^*(C_B)]$$

$$= (1 - \pi_A) K + [\pi_A (Y^*(C_A) + Q) + (1 - \pi_A)(Y^*(C_B) + Q)] - [\pi_A^* (Y^*(C_A))$$

$$+ (1 - \pi_A^*) Y^*(C_B)]$$

$$= (1 - \pi_A) K + Q + [(\pi_A - \pi_A^*) Y^*(C_A)] - [(\pi_A - \pi_A^*) Y^*(C_B)]$$

Next, we write the expression for *Bias(2)*:

$$Bias(2) = \Delta_{A-B}^* - \Delta_{A-B} \tag{A6}$$

$$= [Y(C_A) - Y(C_B)] - [Y^*(C_A) - Y^*(C_B)]$$

$$= [Y^*(C_A) + Q - (Y^*(C_B) + Q + K)] - [Y^*(C_A) - Y^*(C_B)] = -K$$

Appendix B. Details of Six Studies used in the Empirical Analysis

TABLE B1. SIX STUDIES USED TO OBTAIN SAMPLE-BASED ESTIMATES OF RMSB AND RELATED QUANTITIES

Program name	Description of program	Period of CRT	Outcome scales used	Period of trial	Covariates used in analysis	Definition of minority status	Unit of randomization	Grade levels	State(s)
Program A[a]	Reform-based math, science and technology intervention	1 year	SAT10 (math problem solving, science, reading)	2006-2008	pretest, gender, minority status, SES, ELL status, missing value indicators	Minority is defined as non-White	schools	4-8	Alabama
Program B[b]	Web-based reading and writing instruction program	1 year	State test of reading comprehension TerraNova3 reading assessments	2009-2010	pretest, gender, minority status, SES, disability status, ELL status, dummy for teacher experience > 4 years, missing value indicators	Minority is defined as Black	grade-level teams	6-10	Mississippi California
Program C[b]	Web-based English language development program for ELLs	1 year	Comprehensive English Language Learning Assessment (CELLA): Reading, Writing, and Listening/Speaking subtests and the total score	2009-2010	pretest, grade level, gender, dummy for teacher experience > 4 years, SES, missing value indicators		teachers	3-5	Florida
HMH Fuse Algebra[c]	A tablet-based version of conventional algebra text	1 year	California Standards Test (CST) Algebra	2010-2011	pretest, gender, ethnicity, disability status, ELL status, grade level, missing value indicators		sections	7,8	California
Program D[b]	An online reading intervention program designed for struggling students in grades 6-12	1 year	NWEA assessment of reading	2008-2009	pretest, gender, ethnicity, SES, disability status, ELL status, grade level, dummy for teacher experience > 4 years, missing value indicators		teachers	5-9	Texas

TABLE B1. SIX STUDIES USED TO OBTAIN SAMPLE-BASED ESTIMATES OF RMSB AND RELATED QUANTITIES

Program name	Description of program	Period of CRT	Outcome scales used	Period of trial	Covariates used in analysis	Definition of minority status	Unit of randomization	Grade levels	State(s)
Program D(II) [b]	An online reading intervention program designed for struggling students in grades 6-12	1 year	NWEA assessment of reading	2008-2009, 2010-2011	dummy for state, pretest, gender, ethnicity, SES status, disability status, ELL status, dummy for teacher experience > 4 years, missing value indicators	Minority is defined as non-White	teachers	5-9	Texas South Carolina North Carolina

Notes. SES=socioeconomic status (determined by eligibility for free or reduced price lunch in each study), ELL=English Language Learner status (based on school system designation); "missing value indicator" is a dummy variable assigned to each covariate that indicates whether the value of the covariate is missing.

[a] Results based on an impact study of a large-scale randomized experiment of a math science and technology initiative. (Analyses reported in this article were conducted during the original impact study.)

[b] Results based on an impact study that the publisher refused to release.

[c] Results based on the study of the impact of HMH Fuse Algebra Program (Jaciw, Toby, Ma, Lai, & Lin 2012).

Appendix C: Macro-Level Variables Used to Address Research Question 3

TABLE C1. MACRO-LEVEL VARIABLES FROM STUDY 1 USED TO ANALYZE RESEARCH QUESTION 3

Factors	Site-level covariates
Teacher factors	Average math degree rank
	Average science degree rank
	Average math teachers' years teaching
	Average science teachers' years teaching
	Average math teachers' years teaching math
	Average science teachers' years teaching science
	Average of self-reported level of adopting of constructivist principles in teaching math
	Average of self-reported level of adopting constructivist principles in teaching science
Student factors	Average pretest
	Proportion male
	Proportion Free or Reduced Price Lunch
	Proportion Minority
	Proportion English learner
	Proportion in 4th grade
	Proportion in 5th grade
	Proportion in 7th grade
	Proportion in 8th grade
Site factors	Site in region 1001
	Site in region 1002
	Site in region 2001
	Site in region 2002
	Is in locale 1
	Is in locale 2
	Is in locale 3
	Is in locale 4
	Number of students per site