Hold the Bets! Do Quasi- and True Experimental Evaluations Yield Equally Valid Impact Results When Effect Generalization is the Goal?

Andrew Jaciw[1]

[1]Affiliation not available

April 24, 2023

Abstract

Randomized experiments (RCTs) rule out bias from confounded selection of participants into conditions by design. Quasi-experiments (QEs) are often considered second-best because they do not share this benefit. However, when results from RCTs are used to generalize causal impacts, the benefit from unconfounded selection into conditions may be offset by confounded selection into locations. In this work we show that this tradeoff can lead to situations where estimates from QEs are less-biased from selection than are estimates from uncompromised RCTs. We establish the conditions theoretically, demonstrate the idea empirically, and discuss the implications of the results.

Hold the Bets! Do Quasi- and True Experimental Evaluations Yield Equally Valid Impact Results When Effect Generalization is the Goal?

Andrew P. Jaciw

Empirical Education Inc.

Abstract

Randomized experiments (*RCT*s) rule out bias from confounded selection of participants into conditions by design. Quasi-experiments (*QE*s) are often considered second-best because they do not share this benefit. However, when results from *RCT*s are used to generalize causal impacts, the benefit from unconfounded selection into conditions may be offset by confounded selection into locations. In this work we show that this tradeoff can lead to situations where estimates from *QE*s are less-biased from selection than are estimates from uncompromised *RCT*s. We establish the conditions theoretically, demonstrate the idea empirically, and discuss the implications of the results.

Hold the Bets! Do Quasi- and True Experimental Evaluations Yield Equally Valid Impact Results When Effect Generalization is the Goal?

It is normal for policy-makers and practitioners to ask about the potential for programs to achieve impact for an inference population that they are directly concerned with. Ideally, an uncompromised *RCT* would be conducted in their specific context to answer the question decisively. However, if an *RCT* is not possible, they may look to other sites, where the program has been used, to support an inference about what the causal impact may be for their site.

Consider two options for doing this. The first, which is the more obvious and standard one, is to use an *RCT*-based result from one or more study sites where the program has been evaluated – the "generalized from" site(s) – to infer impact for the "generalized to" site(s). This result may be adjusted for possible differences between sites in the distribution of baseline characteristics that moderate the impact. This approach has the advantage that at the "generalized from" site, the result is not biased from confounded selection into conditions; however, because it involves a cross-site comparison of outcomes, the result may be biased from confounded selection into locations in terms of characteristics that moderate the effect (Hotz et al., 2005).

A second option is to make a cross-site comparison to infer just the missing outcome at the inference site. There are two possible situations for this. In the first scenario, the treatment *has been used* at the inference site, and the goal is to compare the performance given treatment at that site, to the performance in the absence of treatment using a comparison group from another site *that has not used the program*. This is the standard observationalist's application of a non-equivalent comparison group design (CGD) (Shadish et al., 2002). In the second scenario, the treatment *has not been used* at the inference site, and the goal is to compare the performance without treatment at that site, to the performance in the presence of treatment using a group from

another site *that has used the program*. This is akin to a CGD, except it uses performance under treatment at another site to infer counterfactual performance to the group that has not received treatment at the inference site.

The second option has the advantage that is uses an estimate of half the true impact result for the inference site (e.g., achievement in the presence of treatment for the first scenario, and performance in the absence of treatment for the second scenario); however, because this option involves a cross-site comparison of outcomes, it may be biased from confounded selection into locations on characteristics that either affect average achievement or moderate program impact (Hotz et al., 2005).

The reader may well wonder why one would ever use the comparison group-based approaches (the second option) in place of the *RCT*-based one (the first option). The result based on an uncompromised *RCT* has the advantage of being unbiased from selection at the study site, and therefore seems advantageous, even if it is achieved at a different site than the one for which the generalized impact is sought. Certainly, intuition suggests that adjusting results from an *RCT* is the better option. In this work we show that, contrary to intuition, this is not always so. This finding is important because the less-biased option should be chosen whenever possible; therefore, it is critical to establish the conditions under which one approach yields less biased results than the other.

This work proceeds as follows. First, we consider three fundamental scenarios for inferring the average causal impact of a program for a study site. We use both graphical displays and formal notation to represent bias for each generalization scenario. Second, we establish the conditions under which comparison group-based generalizations are preferable to *RCT*-based solutions. We also discuss the plausibility of these conditions. Third, we develop methods for

conducting an empirical test of the question. Fourth, we provide an empirical demonstration. Fifth, we discuss limitation, draw conclusions, and consider next steps

To make this work accessible to a wider readership of program evaluators, we focus both on the intuition behind the main idea and on the formalism. To do this we use both graphical representations and more-technical notation, and the main ideas are interpretable using either approach.

Background

Recently there has been a groundswell in advances in methods of evaluation for generalizing impact findings from experiments. The starting point of these methods is an impact finding from an *RCT* conducted at one or more study sites. The generalization step involves adjusting the *RCT*-based result so that it reflects the local conditions for the inference population. Reweighting is a chief example (Schochet et al., 2014). The approach adjusts for differences between the study sample and the inference population in the distribution of moderators of impact[1]. Adjustments may also be made in terms of an index that summarizes, on one dimension, the effects of multiple moderators. Subclassification methods (e.g., Tipton, 2013) are an example of this.

Such *RCT*-based approaches reflect an internal validity-first orientation: the starting point is an estimate of average impact from a true experiment, which is assumed to be internally valid. External validity – which addresses the extent to which a causal relationship holds over variations in persons, settings, outcomes or treatment variants (Shadish, et al., 2002 p. 256) – is achieved after, and is based on the impact finding from a completed *RCT*. This is consistent with a specific orientation in program impact evaluations that internal validity is the "sine qua non",

[1] Moderators of impact are baseline characteristics of persons, or other units, that are associated with changes in the effect of treatment.

that is, “the basic minimum without which any experiment is uninterpretable” (Campbell et al., 1963, p. 5, in Shadish, et al., p. 97). The logic is that we should first establish the causal relationship between variables, and only then address questions about the reach of the causal inference to contexts beyond the study.

The *RCT*-based approach, however, is not the only one. Consider the case where a principal has implemented a program school-wide and would like to know if the program had an average positive impact on student achievement at her school. With the *RCT*-based approach to generalization described above, she would look for an average impact finding from an *RCT* conducted elsewhere, preferably from a locale similar to hers. She might then reweight the impact result from the remote site to more-closely reflect the distribution of characteristics of individuals for her site. Alternatively, using a quasi-experimental (*QE*-based) approach, she may estimate impact for her site by comparing the average performance at her site, in which everyone has received treatment, to contemporaneous performance from one or more similar locales where the program has not been used. This is a standard non-equivalent comparison design (Shadish, et al., 2002), with the inference locale being *the treated site* (i.e., the comparison is with a sample that provides a plausible counterfactual value for what the performance at the inference site would have been in the absence of treatment.)

The comparison group-based approach accommodates also an alternative “reversed” scenario, where the program has not been used at the principal’s (inference) site, and where she would like to know about the potential impact that could be achieved for her site. In this situation, the principal can again use the *RCT*-based result, or, applying the comparison group-based strategy, she can infer impact to her site by comparing average performance at other similar locales, where the program has been used, against the average performance (in the

absence of treatment) at her site. This approach is a version of the non-equivalent comparison design (Shadish, et al., 2002), where the comparison is made with a treated group to support a causal inference for *the untreated site* (i.e., the comparison is with a sample that provides a plausible counterfactual value for what the performance at the inference site would have been if the treatment had been used at that site.)

For either of these scenarios, the *RCT*-based result has the advantage that it is unbiased for the source sample (assuming the *RCT* has not been compromised in some way.) The success of generalization depends on completely identifying and adjusting for the effects of factors that produce a difference in impact between the experimental and inference sites (Cole & Stuart, 2010; Hotz et al., 2005; Imai et al., 2008). On the other hand, the comparison-group based approach has the advantage that it uses information from the actual inference site; however, the comparison with another site puts the result at risk of being biased from selection on confounders that affect average achievement, or on moderators that affect achievement by way of their interactions with the treatment. Put another way, the *RCT*-based option uses an estimate of the true (unbiased) result, but it is from someplace other than the inference site; whereas the comparison-group based approach makes use of an estimate of the true outcome for the inference site itself; however, it is observed for just one condition (i.e., it is half of the unbiased solution for the inference location.)

It is not a-priori clear that for these scenarios, where causal generalization of the average impact is being sought, the *RCT*-based result always yields a better (more accurate) generalization than the *QE*-based alternative. A related question is whether there are certain conditions under which the *QE* strategy is the better option. In this work we address the following questions: (1) What are the expressions for bias for the *RCT*- and *QE*-based

generalizations described above? (2) Under what conditions is net bias in one of these quantities lower than in the other? (3) In a single empirical application, what are estimated levels of each type of bias, prior to and after applying site-level covariate adjustments?

The question we raise here, which we explore in this work, reflects a different perspective on the relationship between internal and external validity. Rather than seeing internal validity as the sine qua non, it treats the two forms of validity as interdependent. There is precedent for this in the program evaluation literature. Shadish et al. (2002), although accepting that internal validity is the "sine qua non", also stressed that it is inseparable from external validity. They considered the latter as "the desideratum" (the purpose or objective) of educational research (Shadish, et al., p. 97). Thus, the two kind of validity – internal and external – are deeply complementary (Shadish et al., 2002). Taking a stronger position, some evaluators outright rejected the precedence of one form of validity over the other, instead emphasizing contextual factors as co-causes of the results (e.g., Cronbach, 1975; Cronbach 1982; Scriven, 2008). Cronbach, for example, stressed the "limited reach" of internal validity. For him identification of "the cause" of an effect had limited value without understanding the conditions and scope of that effect. Put another way, Cronbach considered that "if observations are not generalizable, then causal validity is irrelevant" (Albright et al., 2000, p. 338). According to this interpretation, what we observe as the marginal impact in an *RCT*-based evaluation, is the product of both the experimental manipulation plus the interactions of treatment with observed and unobserved moderators of the effect for that specific context. Any generalization requires making additional strong assumptions about the role and impact of those factors in the "generalized-to" context.

In this work, we provide an additional perspective on the positioning of internal and external validity. We show that under certain plausible conditions, quasi-experimental

comparisons can yield generalized inferences that are less-biased than *RCT*-based ones. We develop the idea that both the *RCT*- and *QE*-based inferences have potential for bias – either from confounders that affect average performance across sites, or from moderators that produce differential impacts across sites. Net bias depends on whether effects of confounders and moderators compound or offset one another. Therefore, when generalization is the concern, the role of factors affecting internal validity (e.g., confounders) and external validity (e.g., moderators of impact) must be considered simultaneously. We develop this idea formally in the next section[2].

Three Scenarios to Motivate the Main Ideas of this Work

[2] Four points require clarification before proceeding:

1. The quantities considered in the following sections represent true values. Later, in the empirical section of this work we will address estimation with observed values that figure-in sampling error at different levels of analysis.

2. Consistent with standard WSCs, the term "bias" is used to describe departures of values of parameters of interest from their true benchmark values. An alternative would be to consider these differences as between estimands for average impact, where a different estimand is associated with each site. The main ideas of this work do not depend on which of these interpretations we choose, and we follow the standard WSC usage.

3. In this work *RCT*s and *QE*s both are considered "experiments". The notation designates results from *RCT*s as "*XP*", and from quasi-experiments as "*QE*", where the *QE*s in this work are Non-Equivalent Comparison Group Designs (CGDs), in which comparisons are made between sites, and with bias potentially arising from differences between sites in the distribution of factors that affect average achievement or impact. (The convention and terminology reflects a preference for considering *QE*s as experimental, and it contrasts with the usage of these terms in other sources [e.g., Glazerman et al., 2003; and Imai et al., 2008, who differentiate "experimentalists" from "observationalists".])

4. To help bring out the main idea, this work assumes a certain type of program participation among sites. First, in sites where an experiment of the program is conducted, all eligible subjects are randomly assigned to conditions, and there is compliance with assignment to conditions at those sites. Second, in sites where an experiment is not conducted, all eligible cases either receive the treatment, or none of them do. As a result, a non-experimental comparison between conditions involves a comparison between sites. Selection is of individuals into sites. The focus is on one selection mechanism – that of individuals into sites – as it affects average performance and impact, similar to certain established WSC designs (e.g., Bloom et al., 2005 and Wilde & Hollister, 2007). Further, we assume no missing outcomes for individuals. (Future extensions of this work may integrate additional selection mechanisms that determine, for experimental sites, whether individuals choose to participate in the experiment or not, and for sites without an experiment, whether individuals select into the program or not within the site.)

Scenario 1: Establishing the Experimental Benchmark

The goal is to estimate the impact of a program T relative to counterfactual C on outcome Y at a given site N. Express the true average impact quantity for this site as follows (*XP* stands for "experimental" to denote a quantity that would be estimated without bias through an uncompromised randomized experiment):

$$\Delta_{XP|N} = Y_N(T) - Y_N(C) \quad (1)$$

This work assumes that there is a true benchmark average impact quantity for each site. If it was possible to randomize individuals to T or C at an inference site, then the unconfounded assignment of individuals into conditions would allow the true value to be estimated without bias. For instance, if the site is a school, this quantity would be estimable if students and teachers were randomly assigned to conditions and if their outcomes were observed. The true value average impact is represented pictorially in Figure 1.

The problem addressed in this work arises when information is missing about performance in one of the conditions at the inference site, N, which requires the use of information from elsewhere (i.e., from other sites) to generalize impact to the inference site. These cases are explored in the next two scenarios.

Insert Figure 1 here

Scenario 2: Inferring Average Impact at Site N When Treatment is Provided to Everyone at the Site.

In Scenario 2 the goal is to infer average impact for site N, $\Delta_{XP|N}$, in the situation where an experiment is not possible at the site because everyone is provided with the program. This means average performance of individuals in the absence of treatment at the site is unknown.

This situation would arise if a program is being implemented school-wide. The administrator may want to know if the program achieves positive average impact on student outcomes, compared to if it had not been used.

For this scenario, two options are considered to infer average impact at N.

Option 1. The first option under Scenario 2 is to infer impact for N using the result of a randomized experiment conducted at a different location M[3]. That is, the following impact quantity can be generalized from the source site M to the inference site N:

$$\Delta_{XP|M} = Y_M(T) - Y_M(C) \quad (2)$$

Figure 2 displays both average impact at the inference site N, $\Delta_{XP|N}$, and the average impact from the comparison site M, $\Delta_{XP|M}$. For Scenario 2, it is assumed that performance in the absence of treatment is not observed at N (represented by the empty horizontal bar), which prompts the use of the result from M.

Insert Figure 2 here

The difference between the true average impact at M, and the one at inference site N, is the bias in the former when used to generalize impact to the latter site. This is shown as $Bias1$ in Figure 2, and is expressed as follows:

$$Bias1 = \Delta_{XP|M} - \Delta_{XP|N} \quad (3)$$

[3] A specific case of this scenario would happen if an administrator of a site where a program is being used wants to know if the program is working; that is s/he wants to draw an inference about whether the program is achieving a positive impact at his/her site. In the context of educational research in the U.S., s/he may look to a compendium of impact findings from randomized experiments of the program done in other settings, such as through the What Works Clearinghouse, to support a conclusion about whether the program is working at his/her site.

$Bias1$ is represented in Figure 2 as the difference in the length of the vertical bars. $Bias1$ is present if there is a difference between the sites in the average impact of treatment[4].

What are Possible Sources of $Bias1$? $Bias1$ is attributable to factors that result in systematic differences in impact between sites (i.e., not counting random sampling error). Among possible sources, most commonly considered are "unit characteristics" – person attributes that moderate the program's impact. That is, $Bias1$ will be present if there is imbalance between sites *N* and *M* in the distribution of characteristics of persons that interact with treatment (Jaciw, 2016[a]; Cole & Stuart, 2010). For example, if the program achieves greater impact for students with higher incoming achievement, and if inference site N has a systematically lower proportion of students with higher incoming achievement than site M, then, unless this difference is adjusted for analytically, the average impact at M will have positive bias if generalized to N. Random assignment of individuals to locales would eliminate this confound; however, such assignment rarely occurs in practice (Hotz, et al., 2005). A common strategy to limit $Bias1$ is to adjust for effects for imbalance on moderators through reweighting or subclassification methods (Tipton, 2013). An assumption of these methods is that the adjustment strategies result in conditionally unconfounded selection into locations on variables that interact with treatment – a necessary condition to mitigate $Bias1$.

Lack of $Bias1$ implies that the selection of individuals into locations across which outcomes are being compared are unconfounded (or is conditionally unconfounded) by factors moderating impact; however, the converse of this is not necessarily true. Macro effects (Hotz, et al., 2005) are another possible source of $Bias1$. These are site characteristics measured through

[4] In Figure 2, the short horizontal bars represent performance levels. The solid-filled black bars represent observable quantities. The white-filled horizontal end-bar (in this case representing average performance in the absence of treatment for N) represents a quantity that is not observed.

variables that have values specific to sites. For example, if N and M are different in terms of principal leadership style or school technological capacity, and if the impact of the treatment depends on (interact with) these variables, then they will induce $Bias1$ if $\Delta_{XP|M}$ if used to infer impact at N. Macro variables may also be site-averages of unit characteristics measured at each site. If a site is a school, then this may include the school average of teachers' years of experience, or the school average of student performance assessed before the study. When the comparison involves just two sites (in our case, N and M) the effects of macro variables on outcomes cannot be analytically de-confounded from the effect of treatment. (With more than two sites, model-based adjustments may be used to try to limit imbalance on macro variables [Hotz et al., 2005]. Our empirical example will illustrate this idea further.)

$Bias1$ may be attributable to other sources also, including if the treatment itself varies across locations either in its offering (program practices) or uptake (participation in activities) (Bloom, et al., 2003), if fidelity of implementation is different (Hulleman et al., 2009), or if the control condition or programs vary across sites (Weiss, et al., 2014). An additional potential source of $Bias1$ is a difference between the sites either in the measures or the comparability of the scales used to assess outcomes.

Option 2 The second option is to make a quasi-experimental (*QE*) comparison between N, where we assume everyone is receiving treatment, and individuals at a different site M, where we assume the treatment is not implemented. The resulting difference quantity may be represented as follows:

$$\Delta_{QE1} = Y_N(T) - Y_M(C) \qquad (4)$$

Figure 3 displays both the benchmark impacts quantity at site *N*, $\Delta_{XP|N}$, and the average impact using the *QE* comparison with site M, Δ_{QE1}. As before, the short horizontal bars represent

performance levels. The solid-filled black bars represent observable quantities. The white-filled horizontal end-bar (in this case for average performance in the absence of treatment for N) represents a quantity that is not observed.

Insert Figure 3 here

The difference between the *QE* result, Δ_{QE1}, which involves a comparison of outcomes from M, and the *RCT*-based benchmark impact at N, is the bias in the former quantity when used to infer impact to N. This is shown as $Bias2$ in Figure 3. It is expressed as follows:

$$Bias2 = \Delta_{QE1} - \Delta_{XP|N}$$
$$= Y_N(T) - Y_M(C) - [Y_N(T) - Y_N(C)]$$
$$= Y_N(C) - Y_M(C) \quad (5)$$

$Bias2$ is represented in Figure 3 as the difference between the inference site N and the comparison site M in the vertical heights corresponding to average performance in the absence of treatment. That is, $Bias2$ is present if there is a difference between the sites in their average performance without the program. This is the standard expressions for bias in a CGD-based difference quantity relative to an experimental benchmark for an inference site. (It is usually the starting point in WSC empirical studies that evaluate bias in impact estimates from CGD's relative to empirical *RCT*-benchmarks, and in these studies treatment group performance typically is differenced away [Heckman et al., 1997; Weidmann & Miratrix, 2021]).

What are Possible Sources of $Bias2$? $Bias2$ can result from imbalance between the locations being compared on any factor that produces a difference in outcomes in the absence of treatment. As with $Bias1$, this includes "unit characteristics", in this case person attributes that affect the outcome. They include student baseline achievement in impact evaluations in

education (Unlu et al., 2021), and individual prior earnings in impact evaluations of employment training programs (Glazerman et al., 2003). Such variables are important because they are highly predictive of later outcomes. Stated more formally, an important potential source of $Bias2$ in Scenario 2 is confounded assignment of persons to location (Hotz et al., 2005) on factors that affect performance in the absence of treatment. As is the case with $Bias1$, random assignment of individuals to locales would eliminate this confound; however, such assignment rarely occurs in practice (Hotz, et al., 2005). A common strategy to limit $Bias2$ is to adjust for effects of confounders, or a corresponding balancing score (Rosenbaum et al., 1983) to achieve conditionally unconfounded assignment to locations. It is notable that "unit characteristics" (the person attributes) that result in $Bias1$ may or may not be the same as those that produce $Bias2$.

Lack of $Bias2$ implies that the selection of individuals into locations that are being compared is unconfounded (or is conditionally unconfounded) by factors affecting performance in the absence of treatment; however, the converse is not necessarily true. $Bias2$ may be present for reasons other than person-level selection. For example, outcomes may be affected by site-level characteristics - the "macro variables" (Hotz, et al., 2005) discussed under Option 1. Such variables have values specific to sites. Using the example from the last section, if locales N and M are different in terms of principal leadership style or school technological capacity, and if these variables have a bearing on the outcome, for example achievement, then they will induce $Bias2$ in Δ_{QE1}. As was discussed under Option 1, if the comparison involves just two sites (in our case, N and M) the effects of macro variables on outcomes cannot be analytically de-confounded from the effect of treatment. Additional possible sources of $Bias2$ is a difference between the "business as usual" (BAU) program being used at M and one that would be used in

the absence of treatment at N, and a difference between the sites either in the measures or the comparability of the scales used to assess outcomes

Recap of Scenario 2. Scenario 2 assumes treatment *is* implemented across site N, and the goal is to infer average impact of the treatment for the site. The unbiased "benchmark" average impact for site N is $\Delta_{XP|N}$.

Option 1 uses the uncompromised *RCT*-based impact quantity from site M, $\Delta_{XP|M}$, to infer impact at N. The result, used to infer average impact for site N is biased by the amount $Bias1 = \Delta_{XP|M} - \Delta_{XP|N}$.

Option 2 makes use of a quasi-experimental comparison of average performance between the treated sample at inference site N, and average performance in the absence of treatment at the comparison site M: $\Delta_{QE1} = Y_N(T) - Y_M(C)$. The result, when used to infer average impact for site N is biased by the amount $Bias2 = Y_N(C) - Y_M(C)$.

The Two Options in Scenario 2 Raise the Following Question: Given Scenario 2, if the goal is to achieve lower net bias, what is the better choice for inferring impact at N? We explore this point empirically later.

Scenario 3: Inferring Average Impact at Site N when the Treatment has Not Been Implemented at that Site

This is like Scenario 2, except no one at the inference site, N, has received the treatment. Scenario 3 may seem to be more-obviously about generalization, in the sense that an externally valid causal inference is being sought for a population that has not yet received the program or been involved in an experiment of the program. We assume that a randomized experiment cannot be conducted at N in the short term, and a plausible value for impact at N is needed immediately

– perhaps to guide programming decisions in the present – and before a randomized experiment can happen.

As with Scenario 2, we consider two options for estimating impact at N.

Option 1. The first option is to use the experimental quantity from location M, as we did in Case 1 under Scenario 2. With this option, bias is the same as in Equation 3 above, reflecting the difference in impact between the two sites.

Option 2. The second option is less obvious. It consists of a quasi-experimental comparison between treated individuals at a different site M, and untreated individuals (everyone) at the inference site N. The resulting quasi-experimental difference quantity is represented as follows:

$$\Delta_{QE2} = Y_M(T) - Y_N(C) \quad (6)$$

This is the difference between the treated at comparison site M, and the non-treated (everyone) at the inference site N.

Figure 4 displays both the quantity of interest, which is the average impact for the inference site N, $\Delta_{XP|N}$, and the average impact based on a comparison of outcomes with site M, Δ_{QE2}. As in the prior figures, the solid-filled horizontal black bars represent observed values, and the white-filled bar (in this case representing average performance under assignment to treatment at N) represents a value that is not observed. The absence of this value prompts the comparison with the treated group at M.

Δ_{QE2} in this scenario is different from Δ_{QE1} in Scenario 2. In Scenario 2, we inferred the average counterfactual performance to T at N using outcomes in condition C at M. Under Scenario 3, we infer average counterfactual performance to C at N using outcomes in condition T at M. The comparison is inverted between the two scenarios, reflecting the information that is

available or needed to infer the average causal effect for site N . This is evident in Figures 3 and 4: under Scenario 2 (Figure 3) the known and unknown quantities (the solid black bars and white bars) at N are for the treated and untreated groups, respectively. Under Scenario 4, the coloring of the bars is reversed for the two groups at that site[5].

Insert Figure 4 here

To formulate bias in Δ_{QE2}, first we expand the expression in Equation 6:

$$\Delta_{QE2} = Y_M(T) - Y_N(C) = [Y_M(T) - Y_M(C)] + [Y_M(C) - Y_N(C)] \quad (7)$$

The difference between the *QE* result obtained through a comparison of outcomes at M, Δ_{QE2}, and the experiment-based benchmark impact at N, $\Delta_{XP|N}$, is the bias in the former quantity when used to generalize impact at the inference site N. We express this as follows:

$$\begin{aligned}
&\Delta_{QE2} - \Delta_{XP|N} \\
&\quad = \{Y_M(T) - Y_M(C) + [Y_M(C) - Y_N(C)]\} - \{Y_N(T) - Y_N(C)\} \\
&\quad = [Y_M(T) - Y_M(C)] - [Y_N(T) - Y_N(C)] + [Y_M(C) - Y_N(C)] \\
&\quad\quad = \{[Y_M(T) - Y_M(C)] - [Y_N(T) - Y_N(C)]\} - [Y_N(C) - Y_M(C)] \\
&\quad = \Delta_{XP|M} - \Delta_{XP|N} - (Y_N(C) - Y_M(C)) \\
&\quad = Bias1 - Bias2 \qquad (8)
\end{aligned}$$

[5] During a presentation of this work at a conference, one attendee pointed out that, in research, one would never start with an untreated group and look for a counterfactual among treated individuals. This may be true in research in which one seeks to understand impact for an established treated group. However, it is not routinely true in program evaluations where decision-makers, for instance administrators for individual sites, such as schools or districts, have a well-defined inference population that has not used the program (i.e., the students at the site), and want to draw an inference about how those students would have performed *had the program been used*. This information has value for such decision-makers, and getting the answer right is important before deciding whether or not to buy that program (as opposed to a potentially better alternative) and implement it for the students in their charge.

$Bias1$ and $Bias2$ are represented in Figure 4. The figure is essentially a composite of Figures 2 and 3. The sources of bias are the same as before. We see that Δ_{QE2} has two bias terms when used to infer impact at N: due to the difference between the sites in average achievement in the absence of treatment ($Bias2$) and due to the difference between the sites in their impact ($Bias1$). This implies that the difference quantity, Δ_{QE2}, potentially reflects selection into sites both on factors affecting average achievement in the absence of treatment, and on factors that moderate impact.

Recap of Scenario 3: Scenario 3 assumes treatment *is not* implemented across site N, and the goal is to infer average impact of the treatment for the site. The unbiased "benchmark" average impact for site N is $\Delta_{XP|N}$.

Option 1 uses the uncompromised *RCT*-based impact quantity from site M, $\Delta_{XP|M}$, to infer impact at N. As with Scenario 2, the result is biased by the amount $Bias1 = \Delta_{XP|M} - \Delta_{XP|N}$.

Option 2 makes use of a quasi-experimental comparison of average performance between the treated sample at site M, and average performance in the absence of treatment at the inference site N:, $\Delta_{QE2} = Y_M(T) - Y_N(C)$. The result, used to infer average impact for site N is biased by the amount $Bias1 - Bias2$

The Two Options in Scenario 3 Raise the Following Question: For Scenario 3, what is the better choice for inferring impact at N to achieve lower net bias: the comparison group-based result, Δ_{QE2}, which is susceptible to both $Bias1$ and $Bias2$, or the experiment-based quantity $\Delta_{XP|M}$, susceptible to $Bias1$ only? On an intuitive level, the second option seems preferable because it is potentially affected by only one form of bias. But is the second option always better?

We address this question in the next section. We summarize the alternatives discussed above in Table 1.

Insert Table 1 here

Under Which Conditions should a Quasi-Experimental Generalization be Preferred to an Experiment-Based One?

Scenarios 2 and 3 each explores two options for inferring impact at N, and the two scenarios suggest different rules for deciding the preferred alternative.

Scenario 2

When everyone at N (the inference site) receives treatment, impact may be inferred using $\Delta_{XP|M}$ which is susceptible to $Bias1$, or Δ_{QE1}, which is susceptible to $Bias2$. The better choice is the one with a lower magnitude of bias (assuming we use the level of bias as the criterion for the better choice). Below we explore this question empirically by comparing the magnitudes of the estimates of each type of bias.

Scenario 3

When no one at N (the inference site) has received treatment, impact may be inferred using $\Delta_{XP|M}$, which is susceptible to $Bias1$, or Δ_{QE2}, which is susceptible to both biases, specifically, net bias in the difference: $Bias1 - Bias2$. As with Scenario 2, we assume the better choice is the one with a lower magnitude of net bias. To evaluate this, we can compare $|Bias2|$ to $|Bias1 - Bias2|$, or alternatively, we compare $Bias2^2$ to $(Bias1 - Bias2)^2$.

Before exploring this question empirically below, we consider the two options numerically. Applying the criterion of lower net bias, we should prefer Δ_{QE2} when:

$$(Bias1 - Bias2)^2 < Bias1^2 \quad (9)$$

We can rewrite this as:

$$(Bias1^2 - 2Bias1Bias2 + Bias2^2) < Bias1^2 \quad (10)$$

Cancelling terms on both sides:

$$(Bias2^2 - 2Bias1Bias2) < 0 \quad (11)$$

For this to be true, $2Bias1Bias2$ has to be positive, which implies the biases have the same sign, and (a) if both biases are positive, $Bias1 > 1/2\ Bias2$, or (b) if both biases are negative, $Bias1 < 1/2\ Bias2$.[6] Before further discussing the alternatives, we should consider if satisfying these conditions is even plausible.

First, can we expect the conditions in Equation 11 to be satisfied under certain plausible values for bias? Past empirical work shows that without adjustment for effects of covariates, $Bias2$ can achieve levels of .20 standard deviations of the outcome variable (Unlu et al., 2021), and $Bias1$ can achieve similar levels (Jaciw et al., 2021; Orr et al., 2019). However, there is also variability in magnitudes of each type of bias, which suggests that the inequality in Equation 11 is not unexpected (e.g., based on past empirical results one can imagine situations in which $Bias1 > 1/2\ Bias2$ [when both biases are positive] or $Bias1 < 1/2\ Bias2$ [when both biases are negative]).

Second, given the point made above, that Δ_{QE2} is preferable to Δ_{XP} when the two biases affecting the former quantity have the same sign, is this condition plausible in terms of experimental evaluation scenarios that arise in the real world? Presence of $Bias2$ implies a difference in average achievement between N and M in the absence of treatment. Presence of

[6] This issue is considered further in Appendix A. The appendix includes a graphical display of the regions in the space of $Bias1$ and $Bias2$ in which Δ_{QE2} is preferable to Δ_{XP}, given the criterion of lower net bias. A hypothetical possible shape of the joint distribution of $Bias1$ and $Bias2$ is overlaid. Various distributions are imaginable; however, true empirical distributions will depend on collecting datapoints of estimated bias across multiple studies including multisite trials of different program types.

$Bias1$ implies average impact varies between N and M. For the biases to have the same signs, the differences have to be in the same direction; that is, average performance in the absence of treatment must be lower, and impact higher, at the comparison site, M, relative to the inference site (see Scenario A in Figure B1 in Appendix B) or average performance in the absence of treatment must be higher, and impact lower, at the comparison site, M, relative to the inference site (see Scenario D in Figure B1 in Appendix B). These situations are plausible under implementation of more-equitable programs that focus improvements for sites in which students on average perform on the lower end of the incoming achievement scale. With the use of such programs, it is reasonable to expect greater impacts for lower achieving students. Under these conditions, same-sign biases are necessary to achieve lower overall bias with Δ_{QE2} compared to $\Delta_{XP|M}$ when inferring impact at N[7].

Having established the condition under which Δ_{QE2} is preferable to $\Delta_{XP|M}$ for assessing impact at N through generalization of outcomes at M, and having shown that the level of each form of bias to support a preference for Δ_{QE2} is plausible both numerically, and can be imagined for real-life scenarios, we move to an empirical example.

Method

Setting up the Empirical Example

We adopt standard Within Study Comparison (WSC) methods to empirically evaluate the levels of each type of bias. WSC studies have traditionally been used to evaluate bias in non-experimental estimates against experimental benchmarks (pioneering studies are by Lalonde, [1986], and Fraker &Maynard, [1987]).

[7] The alternative scenarios B and C in Appendix B are situation in which impact is more positive at the site where students perform higher in the absence of treatment.

The starting point for a WSC study is typically the impact finding from an uncompromised randomized experiment. This result serves as an unbiased benchmark quantity. Control group outcomes are then substituted with those from a different comparison group. The difference in the estimated impact that results from this substitution reflects bias in the impact based on the quasi-experimental comparison (corresponding to $Bias2$ in the current study) as well as random sampling error. WSC studies include use of multisite trials in which controls at certain sites serve as the non-experimental comparison group for the inference site(s) (e.g., Bloom, et al., 2005; Wilde and Hollister, 2007). After bias is estimated, design and analysis strategies are then applied to see if they reduce bias. There are many examples of WSC studies (Wong, et al., [2018] cite 66, and there are more recent ones, such as by Unlu et al., [2021]).

Recently, the method (or methods similar to WSC) has been extended to empirically assess discrepancies in generalized causal inferences from benchmark experimental impacts (Jaciw et al., 2021; Dehejia et al., 2021; Kern et al., 2016; Orr et al., 2019). The approach parallels the standard one described above – the benchmark impact estimate from an uncompromised *RCT* at a site is replaced by a generalized quantity based on *RCT*-based findings from one or more of the other sites. The difference between the generalized impact and the benchmark impact for the site reflects bias in the former quantity (corresponding to $Bias1$ in the current work). As with standard applications of WSC, design and analytic strategies are then evaluated in terms of their capacity to reduce bias.

As noted above, in this work we use a "multisite variant" of the WSC method (Bloom, et al., 2005; Michalopoulos, et al., 2004; Wilde & Hollister, 2007). With this approach, each site in a multisite experiment has a benchmark true value that may be estimated without bias through an uncompromised *RCT* (i.e., each can play the role of inference site N), and controls at some or all

of the remaining sites serve as the comparison group that yields an alternative quasi-experimental result (i.e., they play the role of M). Our empirical study is an application of WSC where we compare estimates of the alternative generalized impact quantities $\Delta_{XP|M}$, Δ_{QE1}, Δ_{QE2} relative to the estimated benchmark impacts for each inference site, M, and evaluate estimated bias in each ($Bias1$, $Bias2$, and $Bias1 - Bias2$, respectively.) Below we discuss the steps in this application of WSC.

In going forward with this application, it is important to keep in mind that each WSC study is just one of potentially many of the question. It takes multiple WSC studies to establish the empirical distributions of bias. That is, no single WSC study yields definitive results for understanding conditions for bias. As an empirical procedure (as opposed to a theoretical demonstration of an idea) it requires the accumulation of results from many studies before consistent rules for avoiding bias can be established through research summaries of the findings, such as by Bloom et al., (2005), Cook et al., (2008) or Glazerman et al., (2003) for traditional WSC studies. Therefore, the results from this study should be considered as providing one piece of the evidence for evaluating how discrepant *RCT*-based and *QE*-based generalizations for individual sites are from the *RCT* benchmarks for those sites. Multiple similar studies can help to establish empirical generalities about this. (The current work should also be considered a proof of concept, because it is a novel application of WSC methods.[8])

Steps of the Method

[8] We emphasize that for sake of brevity, the current description of WSC studies does not address some of the more recent developments. Included among them is a Causal Replication Framework (Steiner et al., 2019) that comprehensively addresses the many reasons why *QE*-based estimates in WSC studies may fail to replicate *RCT*-based benchmarks. We addressed some of these earlier in discussion of possible sources of *Bias1* and *Bias2*. Another development is work by Steiner and Wong (2018) for assess correspondence between quantities compared in replication efforts, including WSCs, using both difference and equivalence tests. Additionally, Orr et al., (2019) discuss approaches to helping policy-makers draw decisions from the results using a Bayesian decision theory framework (Bell, et al., 1995).

Before describing the three steps of the method, we make one modification. Up to this point, we have discussed assessing the accuracy of a generalized causal impact quantity that involves a comparison between just two sites, the inference (generalized to) site N, and the comparison (generalized from) site M. Going forward, instead of drawing comparisons with just one other site (M) we do so with the average of outcomes across all other sites. This has four advantages: (1) it allows testing the accuracy of "large to small" generalized inferences – corresponding to the meaningful and policy-relevant question of whether findings assessed on a larger scale apply on the smaller scale (Jaciw et al., 2021; Orr et al., 2019; Shadish et al., 2002), (2) with a larger sample, the effects of sampling error at the student and intermediate (teacher) levels are reduced in estimates of differences between sites in their average outcomes, (3) it allows discrepancies from benchmark site-specific impacts to be summarized in a convenient way, specifically, as variance expressions (a point we develop below), and (4) with multiple such comparisons, it is possible to evaluate the reduction in the average magnitude of bias conditional on the effects of site-level (macro) variables (i.e., through application of "model-based adjustments" noted by Hotz et al. [2005]).

Step 1. **Express $Bias1$ and $Bias2$ for a Given Site j.**

$Bias1$ with respect to inference site $S = j$ is impact averaged across all sites except j, $\bar{\Delta}_{XP|S\neq j}$, minus impact at site j, $\Delta_{XP|D=j}$:

$$Bias1_j = \bar{\Delta}_{XP|S\neq j} - \Delta_{XP|S=j} \quad (12)$$

$Bias2$ with respect to in inference site j is the control performance at site $S = j$, $Y(C)_{|S=j}$, minus the average of control performance across all sites except j, $\bar{Y}(C)_{|S\neq j}$:

$$Bias2_j = Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j} \quad (13)$$

Step 2. Express Bias in Δ_{XP} Δ_{QE1}, and Δ_{QE2} for each of N sites:

Site-specific biases in Δ_{XP} are as follows:

$$Bias1_{|S=1} = (\bar{\Delta}_{XP|S\neq 1} - \Delta_{XP|S=1}) \quad (14)$$

$$Bias1_{|S=2} = (\bar{\Delta}_{XP|S\neq 2} - \Delta_{XP|S=2}) \quad (15)$$

…

$$Bias1_{|S=N} = (\bar{\Delta}_{XP|S\neq N} - \Delta_{XP|S=N}) \quad (16)$$

Site-specific biases in Δ_{QE1} are as follows:

$$Bias2_{|S=1} = \left(Y(C)_{|S=1} - \bar{Y}(C)_{|S\neq 1}\right) \quad (17)$$

$$Bias2_{|S=2} = (Y(C)_{|S=2} - \bar{Y}(C)_{|S\neq 2}) \quad (18)$$

…

$$Bias2_{|S=N} = (Y(C)_{|S=N} - \bar{Y}(C)_{|S\neq N}) \quad (19)$$

Site-specific biases in Δ_{QE2} are as follows:

$$Bias1_{|S=1} - Bias2_{|S=1} = \left(\bar{\Delta}_{XP|S\neq 1} - \Delta_{XP|S=1}\right) - \left(Y(C)_{|S=1} - \bar{Y}(C)_{|S\neq 1}\right) \quad (20)$$

$$Bias1_{|S=2} - Bias2_{|S=2} = (\bar{\Delta}_{XP|S\neq 2} - \Delta_{XP|S=2}) - (Y(C)_{|S=2} - \bar{Y}(C)_{|S\neq 2}) \quad (21)$$

…

$$Bias1_{|S=N} - Bias2_{|S=N} = (\bar{\Delta}_{XP|S\neq N} - \Delta_{|S=N}) - (Y(C)_{|S=N} - \bar{Y}(C)_{|S\neq N}) \quad (22)$$

Step 3. Summarize Bias for Each Alternative (Δ_{XP}, Δ_{QE1} and Δ_{QE2}) using Means of Squared Differences (i.e., Mean Squared Bias (MSB)).

If overall bias was summarized by simply averaging over site-specific biases, then cancellation of positive and negative values would result in the underestimation of the average magnitude of bias (Bloom et al., 2005). Summarizing average levels of bias using the mean squared bias is one way to avoid this problem:

For Δ_{XP}:

$$MSB_{XP} = \frac{1}{N}\sum_{j=1}^{N}(\bar{\Delta}_{XP|S\neq j} - \Delta_{XP|S=j})^2 \quad (23)$$

For Δ_{QE1}:

$$MSB_{QE1} = \frac{1}{N}\sum_{j=1}^{N}(Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j})^2 \quad (24)$$

For Δ_{QE2}:

$$MSB_{QE2} = \frac{1}{N}\sum_{j=1}^{N}[(\bar{\Delta}_{XP|S\neq j} - \Delta_{XP|S=j}) - (Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j})]^2,$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left(Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j}\right)^2 + \frac{1}{N}\sum_{j=1}^{N}\left(\bar{\Delta}_{XP|S\neq j} - \Delta_{XP|S=j}\right)^2$$

$$- \frac{2}{N}\sum_{j=1}^{N}(Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j})(\bar{\Delta}_{XP|S\neq j} - \Delta_{XP|S=j})$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left(Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j}\right)^2 + \frac{1}{N}\sum_{j=1}^{N}\left(\Delta_{XP|S=j} - \bar{\Delta}_{XP|S\neq j}\right)^2$$

$$+ \frac{2}{N}\sum_{j=1}^{N}(Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j})(\Delta_{XP|S=j} - \bar{\Delta}_{XP|S\neq j}) \quad (25)$$

A point to emphasize here is that MSB for Δ_{XP} (Equation 23) is an expression for the cross-site variance in impact. Similarly, MSB for Δ_{QE1} (Equation 24) is an expression for the cross-site variance of average performance in the absence of treatment, and MSB for Δ_{QE2} (Equation 25), is the sum of these quantities plus twice the covariance between terms for $-Bias1$ and $Bias2$.

These expressions allow us to address the main questions of this work, including about the degree of bias in comparison-group-based and *RCT*-based generalizations when compared to experimental benchmarks. We can expect that on average the magnitude of bias in Δ_{QE2} is less than the magnitude of bias in Δ_{XP}, when the following condition is satisfied:

$$MSB_{QE2} < MSB_{XP}$$

$$\Leftrightarrow \frac{1}{N}\sum_{j=1}^{N}(Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j})^2 + \frac{2}{N}\sum_{j=1}^{N}(Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j})(\Delta_{XP|S=j} - \bar{\Delta}_{XP|S\neq j}) < 0 \quad (26)$$

This requires that each of the following conditions is met:

$$\frac{2}{N}\sum_{j=1}^{N}(Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j})(\Delta_{XP|S=j} - \bar{\Delta}_{XP|S\neq j}) < 0 \quad (27)$$

$$\left|\frac{2}{N}\sum_{j=1}^{N}\left(Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j}\right)\left(\Delta_{XP|S=j} - \bar{\Delta}_{XP|S\neq j}\right)\right| > \left|\frac{1}{N}\sum_{j=1}^{N}\left(Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j}\right)^2\right| \quad (28)$$

Equation 27 show the necessity of a negative covariance between site deviations in average achievement, and site deviation in impact, and therefore that $Bias1$ and $Bias2$ have the same sign, in expectation [9]. (The conditions described here capture, on the aggregate, the criteria discussed earlier for the two-site case [i.e., the condition in Equations 9 – 11]).

Step 4. Adjust for Effects of Confounders and Moderators.

As with standard WSC studies, a salient question is whether quantities that summarize bias, in this case the values of MSB, are reduced in magnitude after conditioning on effects of covariates. More specifically, the question is whether adjusting for effects of covariates lowers (1) MSB_{XP} by reducing the influence of moderators of impact, (2) MSB_{QE1} by limiting the influence of confounders that affect average achievement, and (3) MSB_{QE2} by reducing the influence of either confounders or moderators. To do this, quantities in Equations 23 – 25 above will be estimated conditioning on a series of covariates, $\boldsymbol{X}$: $MSB_{XP|\boldsymbol{X}}$; $MSB_{QE1|\boldsymbol{X}}$, and $MSB_{QE2|\boldsymbol{X}}$.

For the empirical example below, estimates of $\widehat{MSB}_{XP}$, $\widehat{MSB}_{QE1}$, and $\widehat{MSB}_{QE2}$ will be obtained prior to and after adjusting for effects of covariates. To allow a direct comparison of the results on the same scale that average program impacts are commonly reported (i.e., in the metric of the standardized effect size), estimates will be reported as the square root of these quantities divided by the pooled standard deviation of the outcome measure.

Identifying Sources of Random Sampling Error

[9] The covariance is between $Bias2_{|S=j}$, $\left(Y(C)_{|S=j} - \bar{Y}(C)_{|S\neq j}\right)$, and the negative of $Bias1$, $-\left(\bar{\Delta}_{|S\neq j} - \Delta_{|S=j}\right) = \left(\Delta_{|D=N} - \bar{\Delta}_{|D\neq N}\right)$. Therefore, a negative covariance implies that $Bias1$ and $Bias2$ have the same sign, in expectation.

Estimates of MSB reflect sources of random sampling error. A secondary question of this work is whether, in the case of multisite designs with randomization of both students and teachers within sites, results are sensitive to modeling the intermediate teacher level. The question is whether statistically significant and substantively important differences in estimates of the MSB terms result from ignoring this level. The generalizations discussed in this work are concerned with differences in impact across sites, and not with random class-level sampling error within sites; therefore, it is important to make sure that measurements of the former are not conflated with effects of the latter.

Research Questions

1. In the context of a multisite WSC design, what are the estimated average magnitudes of the discrepancies between site (benchmark) impacts and the three different version of generalized impact corresponding to Δ_{XP}, Δ_{QE1} and Δ_{QE2}, and are the estimated average magnitudes statistically significant?
2. Are the estimates of the discrepancies in (1) different from each other?

These questions will be addressed prior to and after covariate adjustments, and with and without adjusting for effects of class-level random sampling error. Given the focus of this work, the contrast of main interest is between the experiment-based results (e.g., $\widehat{MSB}_{XP}$) and each of the comparison-group based results ($\widehat{MSB}_{QE1}$ and $\widehat{MSB}_{QE2}$).

Estimation

Hierarchical Linear Models (HLM) (Raudenbush & Bryk, 2002) are used to produce maximum likelihood estimates of the following quantities: $\widehat{MSB}_{XP}$, $\widehat{MSB}_{QE1}$, $\widehat{MSB}_{QE2}$, $\widehat{MSB}_{XP|\boldsymbol{X}}$, $\widehat{MSB}_{QE1|\boldsymbol{X}}$ and $\widehat{MSB}_{QE2|\boldsymbol{X}}$. Random Intercept Random Coefficient (*RIRC*) models are used in estimation (Jaciw et al, 2021; Miratrix et al., 2021). The precedent for using of HLM in WSC

applications is in Jaciw (2016[b]))[10]. (The full details of the models used are provided in Appendix C.)

Data

The application uses data from the Tennessee STAR (Student-Teacher Achievement Ratio) class size reduction multisite trial. Results from the study are reported in Finn and Achilles, (1990), Mosteller, (1995), Nye, Hedges and Konstantopoulos, (1999), Nye, Hedges and Konstantopoulos, (2000), Nye, Hedges and Konstantopoulos, (2001) and Nye, Hedges and Konstantopoulos, (2002). The multisite trial started in 1985 and lasted 4 years. In the original study, 6,400 kindergarten students and their teachers were randomly assigned to (1) small classes

[10] HL models provide several advantages to summarizing discrepancies between the benchmark *RCT*-based impacts for individual sites and each of the three alternatives: *XP*, *QE*1 and *QE*2. First, the use of variances components from HLM allow us to summarize average *absolute* discrepancies between benchmark impacts for individual sites and generalized impacts that draw on outcomes from the remaining sites. In standard WSC methods it was recognized that summarizing biases through a straight average is misleading because positive and negative biases cancel, resulting in a mean close to zero; therefore, "average absolute biases" were used which consisted of taking the mean of absolute values of bias. Root Mean Squared biases (RMSB) are equally adequate for this purpose (with precedent in Bloom (2005), through conceptualization of bias in WSC studies as a form of random error labelled "non-experimental mismatch error", and in more recent applications such as by Kern et al (2016), for example). In the WSC application of the current work, the bias for each site is the difference between the *RCT*-based impact for each site and an *RCT* or *QE*-based impacts that uses information from across all sites. As shown in Equations 23-25 the MSBs in this application conveniently take the form of expressions for variances and covariances. HLM provides direct estimation of these quantities, including of the covariance term that is important for understanding the conditions under which *QE*2 should be preferred to *XP*. A second advantage to using the HL framework, it that it allows efficiently parsing-out potential sources of sampling error that it would be a mistake to ignore. In this work, we specifically net-out teacher-level sampling error through inclusion of a random effect at the intermediate (teacher) level. This would be much harder to do outside of HLM on a site-by-site basis. Third, a reason to summarize bias over multiple sites simultaneously is that it allows examining the role of site-level (macro) variables, both as confounders that influence average achievement across sites (and thereby lead to $Bias2$) and as moderators of impact that lead to effect heterogeneity (and thereby lead to $Bias1$). The role of macro variables may be tested by including them as main effects, and as variables that are interacted with treatment in the HL models, and examining the changes in RMSB that are measured in terms of the variance components. Finally, the use of variance components for summarizing bias my be more-easily appreciated if we think of variance components in a way that is different from how we usually think about them. A variance is usually conceived of as a measure of how distant individual observations are from a true average, and how erroneous the individual observations are – think of measurement error around the "true score" value in classical measurement theory. We recommend an alternative orientation to and interpretation of variance components in the context of this work. Instead of thinking of the deviations as degrees of distance and error of individual observations relative to the average, we can think of *the average* as a deviated value from the true (benchmark) impacts for individual sites. That is, in this work we pay attention to how wrong the average impact is, in expectation, as an alternative generalized quantity for individual sites, when compared to *RCT*-based experimental benchmark values for sites. The variance quantities in this sense suitably summarize *wrongness of the average*, instead of *wrongness from the average*.

(13–17 students), (2) regular classes (22–25 students), or (3) regular classes with an aide within each of 79 schools (sites). Approximately 100 classes were allotted to each arm of the trial. The intervention continued through third grade. Teachers were randomly assigned to conditions within grades as the student cohort moved from kindergarten through third grade. The design aimed to retain students in the condition to which they were originally assigned over that period. Students who joined the study in intervening years were randomly assigned to conditions. In previous studies, regular classes, with or without an aide, are considered the control group, and we adopt this approach. Having multiple control classes per school, allows estimation of sampling error for intermediate (class) units, which helps to address the question of the effect of ignoring the teacher level in estimates of MSB[11].

In the original study, outcomes were assessed on reading and math in kindergarten through third grade using the SAT-7 assessment. For the demonstration of this work, impacts are evaluated on second grade reading outcomes. The sample consists of students with posttests, who joined the study either in kindergarten or in first grade, and who remained in their assigned condition through second grade. To maintain a hierarchically structured dataset (i.e., with students nested in schools) analysis is limited to students who remained in the same school through second grade. Applying these criteria, the final dataset consists of 3,452 students among 314 second grade classrooms, among 73 schools[12].

[11] Results from past studies include the following: by the end of the second year, students in small classes had an advantage of more than .20 standard deviations in achievement over students in regular-sized classes (Finn & Achilles, 1990; Nye, Hedges & Konstantopoulos, 2000); impacts were cumulative, increasing from kindergarten through third grade (Nye, Hedges & Konstantopoulos, 2001); small class sizes did not have above-average impact for low performing students (Nye, Hedges & Konstantopoulos, 2002); impacts persisted at least five years after re-entry into regular-sized classes (Nye, Hedges & Konstantopoulos, 1999).

[12] Student attrition in this study is notable (Nye, Hedges & Konstantopoulos, 2000). The reader is referred to prior work by Nye, Hedges and Konstantopoulos, (2000) who examined potential for bias from attrition. They examined differences between stayers and leavers in terms of impacts of small classes in prior grades. They conclude: "it seems implausible that attrition substantially biased the treatment effects in the following year" (p. 131). This gives

We examine results (a) prior to covariate adjustments; (b) after adjusting for effects of student-level school-centered variables (gender, eligibility for Free or Reduced Price Lunch, minority [non-White] status, years of experience teaching by the student's teacher, whether a student's teacher holds a Master's degree or higher, and end-of-kindergarten scores in math and reading, and the interactions of these covariates with treatment); and (c) and after adjusting for the effects of both the variables in (b) and site-level macro variables, including site averages of uncentered student-level variables, and school urbanicity (i.e., whether a school is inner-city, suburban, rural or urban) and their interactions with treatment.[13]

Results

The main results are displayed in Table 2 and Figures 5a and 5b. Each triplet of bars shows, from left-to right: $\sqrt{\widehat{MSB_{XP}}}$ (light gray), $\sqrt{\widehat{MSB_{QE2}}}$ (dark gray), and $\sqrt{\widehat{MSB_{QE1}}}$ (black), expressed in units of the standard deviation of the outcome variable. (Estimates of corresponding variance components are shown in Table E1 in Appendix E[14].)

Insert Table 2 here

Insert Figure 5a here
---

assurance that results from the current study, which use samples that are similar to those in the prior works noted above, are not substantively biased by attrition.

[13] All variables used as moderators were assessed at the start of the trial with the exception of post-kindergarten achievement. Baseline performance was not assessed in kindergarten. Therefore, based on precedent (Nye, Hedges & Konstantopoulos, 2002), as an exploratory strategy, the earliest achievement results available that are not likely to be affected by treatment are used (i.e., school averages of achievement at the end of kindergarten for controls only.)

[14] The expressions that we developed for Mean Squared Bias (Equations 23 – 25) involve removing the inference site from each average against which it is compared (this "one out" approach was first discussed in Author (2016[a])). It's inclusion, however, has minimal effect on the overall result. (This is demonstrated in Appendix D); therefore, the one-out approach is not used in estimation.

Insert Figure 5b here

What Is Observed?

How Large on Average are the Discrepancies of Generalized Quantities from Experimental Benchmarks?

Estimates of *XP*. Before adjusting for class-level random sampling error, the estimates of $RMSB$ for the experiment-based generalization, in the metric of the standardized effect size are .31, .31 and .29, without covariate adjustment, with adjustment for student-level site-centered covariates, and additionally with adjustment for school-level covariates, respectively. After adjusting for class-level random sampling error, the corresponding quantities are .18, .17 and .14. All estimates are statistically significant before adjusting for the class-level random sampling error (at level α=.05), but none are after this adjustment.

Estimates of *QE2*. The estimates of $RMSB$ for the comparison group-based generalization that involves substituting outcomes for cases assigned *to treatment* are .46, .47 and .34 in the metric of the standardized effect size, before controlling for the class level random sampling error, and .40, .42 and .27, after. (They are reported in the same order as for *XP* above.) All estimates are statistically significant (at level α=.05) with the exception of the model that adjusts for class-level sampling error and includes all covariates.

Estimates of *QE1*. The estimates of $RMSB$ for the comparison group-based generalization that involves substituting outcomes for cases assigned *to control*, in the metric of the standardized effect size and reported in the same order as for *XP* above, are .45, .44 and .25, before controlling for the class level random sampling error, and .43, .42 and .19, after. All estimates are statistically significant (at level α=.05).

How Different are the Generalized Quantities from Each Other?

We also tested whether the generalized estimates are different from each other given that it is preferable to examine the statistical significance of the difference between quantities being compared, than the difference between them in whether they reached statistical significance (Gelman, 2006).

Comparing Estimates of *XP* and *QE2*. Two main trends are evident. First, adjusting for the class-level random effect marginally increases the difference in MSB, favoring the experiment-based solution. This reflects that modeling the intermediate (class) level decreases $\sqrt{\widehat{MSB_{XP}}}$ proportionately more than $\sqrt{\widehat{MSB_{QE2}}}$. That is, the experiment-based generalization benefits more from properly accounting for sampling error at the intermediate level.

Second, adjusting for the effects of school-level covariates reduces the discrepancy in outcomes between the two generalization approaches. After adjustment, the difference in MSB quantities is not statistically significant.

Comparing Estimates of *XP* and *QE1*. The first main trend described above is less obvious in this case. The second persists – adjusting for the effects of school-level covariates leads to a small and non-statistically significant differences between methods of generalization, as indexed by the difference between them in *MSB*.

Comparing *QE1* and *QE2*. The differences between the estimates never reach statistical significance.

What are the Main Take-Aways from the Results?

There are three main take-aways. First, in almost all instances examined, the experiment-based generalization (*XP*) shows a smaller magnitude of bias than the generalization involving a non-experimental comparison (*QE*). Second, after adjusting for effect of school-level covariates,

the advantage of *XP* over *QE* ceases to reach statistical significance. Third, adjusting for the class-level random effect leads to a statistically significant reduction in the site-level variance components based on the change in the deviance statistic (the results are shown in Appendix E), with implications for generalizability. Notably, the experiment-based generalization is the most sensitive to adjustment for class-level effects, with MSB being close to halved, which is more-favorable to the experiment-based alternative.

Some Conclusion about the Results

The Need for Additional WSC Studies of this Type.

Replication is needed. Knowledge from WSC studies about conditions for bias is cumulative, and requires replication. As noted earlier, Wong et al. (2018) cite 66 WSC studies that have led to more-general understanding about conditions for $Bias2$. These efforts have led to overarching strategies in the design and analysis of quasi-experiments to reduce selection bias that corresponds to $Bias2$ in this work. The current work extends use of WSCs to address problems of bias in causal generalization. The study should be considered as an N of 1 of potentially many studies of similar questions that over time will lead to more global rules about conditions for bias. More specifically, given the importance of cumulative knowledge about conditions for bias from multiple WSC studies, the results from the single STAR data set explored in this work should not be considered as definitive. For example, adjusting for pretest often performs quite well in reducing $Bias2$, including the finding from recent works that adjusting for the pretest allows near replication of experimental benchmarks (Unlu et al., 2021). In this sense, the current work may be an exception.

This study alerts us to certain things. Replication of this work is needed; however, as an early study comparing experiment- and comparison-group based approaches to

generalizability, the results of this work provoke an important question for subsequent replication efforts: Is it the case that, normally, differences between experiment-based and comparison-group-based generalizations are diminished, after adjusting for macro effects, and are resulting differences small enough to not matter? If so, then the implications are significant. If comparison group-based generalizations are as reliable as experiment-based ones, at least under specific conditions, then we can use both designs in more-targeted applications, potentially yielding more abundant evidence form evaluations of program effectiveness. A specific question to focus on in subsequent studies, based on what the derivations of this work predict, is whether usually the *QE*2 alternative comes closer than *XP* to experimental benchmarks, specifically for programs that achieve greater positive impacts for persons who perform lower on pre-intervention measures of the outcome variable.

Conventional WSC studies at providing direction for future work. Continuing work may focus on some of the standard questions typically addressed with traditional WSC studies, including about the benefits of local matching, the effects of adjusting for standard demographic variables as opposed to productive and theory-driven covariates (Shadish, et al. 2006) the role of different types of adjustments, including matching on propensity scores (Heckman et al., 1997; Rosenbaum et al., 1983). Recent work has also evaluated different metrics for capturing correspondence between *QE* and *XP* results (Steiner et al., 2018), and focused on quantification of discrepancies from benchmarks in terms of policy-relevant metrics (Kruger, 1999; Kruger, 2000; Orr et al., 2019; Wilde & Hollister, 2007). These extensions may be adapted to the problems studied in this work, and we earmark them for the future. It is also important to understand the role of other factors limiting replication in efforts to generalize (see Steiner et al., 2019), because a search for effective moderators to account for impact heterogeneity may be

fruitless if the variation is attributable to other sources. Separately from WSC studies, the current work also may be developed further through application of a framework for separating estimation error into components due to sample selection and treatment imbalance (Imai et al., 2008), especially in situations where we assume the components are not additive, and in the context of generalization from large-to-small as we have been considering in this work.

Possible Threats to the Validity of Estimates in this Study.

Sampling Error. Estimates of MSB used to index generalizability reflect between-school differences in average performance and impact, as well as random sampling error attributable to classes within schools, and to students within classes. An important result from this study is that estimates of MSB_{XP}, MSB_{QE1} and MSB_{QE2} are sensitive to the inclusion of class-level random sampling error. Ignoring the variance attributable to class-level differences within schools leads to its absorption at the school level, resulting in upward-bias in variance at that level. This is important because estimates of variances at the school-level, both in average achievement and impact, are used to measure the proximity of generalizations to benchmark values. Ignoring class-level sampling variation biases these results. This effect may be more apparent if there are few classes per school, as was the case in the STAR experiment.

In the STAR experiment, both students and teachers were randomly assigned to conditions, making the inclusion of the class random effect sensible; however, with multisite trials, modeling the intermediate level may be important even under different randomization schemes. For instance, if students are randomly assigned to conditions, but teachers are not, then bias may result from teachers' selecting into conditions (e.g., certain teachers may jockey for the position to teacher students assigned to treatment). Even if randomization is blocked on classes, with students balanced across conditions within classes, if characteristics of classes and teachers

interact with treatment this will be reflected in heterogeneity in impact among classes within sites. Unless these within-school effects are modeled explicitly, they may be misinterpreted as reflecting variability across schools.

Additionally, we can consider the stability of the results assuming a hypothetical re-sampling from an imaginary "super universe" of schools (of which the study sample is one realization). The standard errors of estimates of the school-level variance components (and of MSB_{XP}, MSB_{QE1} and MSB_{QE2}), reflect uncertainty in this scenario, assuming a theoretical sampling distribution. Ultimately, the current study with its results may be seen as one from among many potential study replications, where each study is a replicate from the "super universe" of studies. This calls for conducting actual replications to assess if the main results of this study represent the norm or are an aberration.

Potential for Bias from Overfitting. A possible concern is that the covariate adjustments at the site level induce bias from overspecification, which would raise doubt about the effects of those adjustments. To address this possibility, results were examined from 22 models that included specific subsets of main effects of school-level covariates and their interactions with treatment. The first set of models adjusted for class-level random sampling error. The second set excluded the class-level effects. This yielded 44 results.

Results showed that the relative changes in estimates in school-level variance components were very similar when adding the same sets of school-level covariates to less- or more-parameterized models. That is, inclusion of specific sets of school-level covariates produced similar relative changes in estimated variance components regardless of how many school-level covariates were already in the model. If the models were overfitted, one would expect instead that changes in estimates of variance components from inclusion of additional

school-level covariates to depend greatly on how saturated the model already is. The details of these tests are provided in Supplement A.

Discussion

This work applied a WSC approach to evaluate the accuracy of experiment-based and comparison-group-based causal generalizations to individual sites. It was posited at the outset that the former type of generalization should not be automatically accepted as less-biased just because it stems from an experiment conducted elsewhere – each approach can yield biased estimates. Experiment-based solutions ($\Delta_{XP|M}$) are susceptible to bias from imbalance on moderators of impact, whereas comparison group-based solutions (Δ_{QE1}, Δ_{QE2}) can be biased from imbalance on confounders that affect average achievement, on moderators of impact, or on both types of variables. Susceptibility to both forms of bias does not mean greater net bias, because the two types of bias may cancel-out each other[15]. The idea was tested empirically, and in the case of a single multisite trial it was found that experiment-based results, for the most part, were less biased than comparison-group based approaches; however, after adjusting for effects of site-level covariates, the difference between them in the average magnitude of bias was not statistically significant. Further study of the questions addressed in this work, and replication of results is recommended, possibly with focus on conditions that in this work were posited in theory to produce lower net bias in Δ_{QE2} – in cases where impacts are being evaluated for programs that are expected to achieve more-positive impacts for persons who perform lower on the pre-intervention measure of the outcome variable.

[15] An attendee at a conference where this work was presented was concerned with the idea that bias from Δ_{QE2} could be less than for $\Delta_{XP|M}$. S/he summarized this intuition by saying that "two wrongs don't make a right". However, we have shown that because the biases may offset each other, sometimes "two wrongs" when represented in relation to each other, may be closer to the benchmark solution.

Relatedly, it is important to evaluate situations in which adjustment for certain covariates reduces both forms of bias, versus just one but not the other. At least intuitively, it seems that it would be harder to find covariates that routinely reduce $Bias2$ regardless of the program (in the sense that a pretest does for $Bias1$). This is because moderators of impact are treatment-specific and therefore depend on the joint working of the program and the moderator by way of their interaction.

An important idea of this work is that when causal generalization is the goal, bias from confounded selection on factors that affect average achievement in the absence of treatment (factors resulting in $Bias2$), cannot be considered separately from bias due to confounded selection on factors that affect achievement by way of their interactions with treatment (i.e., moderators leading to $Bias1$). As noted earlier, traditional WSC studies are almost exclusively concerned with $Bias2$. Calculation of this bias "differences away" performance in the treatment condition (Bloom et al., 2005; Weidmann & Miratrix, 2020). Consequently, in standard applications of WSC studies, the role of treatment is immaterial to the discussion of bias. In my view, this misses half the problem, one that is unavoidable when causal generalization is the goal.

At the start of this work, it was noted that the theoretical results of this work have implications for how we view the relationship between internal and external validity. This topic deserves a longer discussion, one that is beyond the scope of this work; therefore, we just briefly elaborate on the idea, with the hope that it generates further discussion.

Causal generalization requires taking into account at least imbalance on moderators (sources of $Bias1$ if the generalized inference is based on $\Delta_{XP|M}$) and possibly, additionally, imbalance on factors affecting average achievement absent treatment (i.e., $Bias1 - Bias2$, if the

generalized inference is based on Δ_{QE2}). The implication is that, to establish the external validity of a causal inference, one does not first demonstrate internal validity through ruling out $Bias2$ (the usual concern of standard WSC studies) and then proceed to establish external validity by ruling out $Bias1$. Establishing the external validity of a causal inference requires ruling out either $Bias1$ alone (in the case where the generalization is from an experiment conducted elsewhere), or $Bias1$ and $Bias2$ simultaneously (in the case of a *QE*2). Ruling out just $Bias2$ is trivial to establishing external validity[16].

Based on this, our interpretation of the assertion that *internal validity comes first* (i.e., is the "sine qua non" [without which there is nothing]), is that it applies only in the circumscribed case where, by design, external validity is not in question. Internal validity comes first in the same sense that one places first in a race by being the only competitor – you cannot lose! (But are you really winning? And can you even call it a competition?)

[16] When is $Bias2$ the sole concern? I would argue it is so only when the study sample is the inference sample, and when external validity is not at issue. For example, $Bias2$ is the primary concern of standard WSC studies that measure bias that compromises internal validity in CGDs when the causal inference concerns the study sample. Internal validity, and its potential to be compromised through $Bias2$, are insufficient for addressing problems of external validity.

References:

Albright, L. & Malloy, T. E. (2000) Experimental validity: Brunswik, Campbell, Cronbach and enduring Issues. *Review of General Psychology, 4*, 337-353.

Bell, S.H., Orr, L.L., Blomquist, J.D., and Cain, G.G. (1995). Program applicants as a comparison group in evaluating training programs. Kalamazoo, MI: Upjohn Institute for Employment Research.

Bloom, H. S., Hill, C. J. & Riccio, J. A. (2003). Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments. *Journal of Policy Analysis and Management 22*(4): 551-575. doi: 10.1002/pam.10154.

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.

Campbell, D. T. & Stanley, J. C. (1963) *Experimental and quasi-experimental designs for research*. Chicago: RandMcNally.

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG-320 trial. American Journal of Epidemiology, 172, 107 – 115.

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*, *30*(2), 116–127.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.

Dehejia, R., Pop-Eleches, C. & Samii, C. (2021). From local to global: External validity in a fertility natural experiment. *Journal of Business and Economic Statistics*, *39*, 217 – 243.

Finn, J. D., & Achilles, C. M., (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27,* 557-577.

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*(4), 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.

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.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child development perspectives, 2*(3), 172-177.

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.

Hulleman, C. S. & Cordray, David S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2*(1), 88–110.

Imai, K., Kong, G. & Stuart. E. A. (2008). Misunderstandings between experimentalists and \ observationalists about causal inference. *J. R. Statist. Soc. A, 171, Part 2*, 481–502.

Jaciw, A. P. (2016a). 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. (2016b). Applications of a within-study comparison approach for evaluating bias in generalized causal inferences from comparison group studies. *Evaluation Review, (40)*3, *(40)*3, 241-276. http://erx.sagepub.com/content/40/3/241.abstract

Jaciw, A. P., Unlu, F. & Nguyen, T. (2021). A Within-Study Approach to Evaluating the Role of Moderators of Impact in Generalizations from 'Large to Small'. *The American Journal of Evaluation,* https://doi.org/10.1177/10982140211030552

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness, 9*, 103 – 127.

Krueger, A. (1999). "Experimental Estimates of Education Production Functions." Quarterly Journal of Economics 114 (2): 497–532.

Krueger, A. (2000). "The Class Size Policy Debate." Economic Policy Institute Working Paper No.121.

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

Michalopolous, C., Bloom, H. S., & Hill, C. J., (2004). Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *The Review of Economics and Statistics, 86,* 156-179.

Miratrix, L. W., Weiss, M. J. & Henderson, B. (2021). An applied researcher's guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *Journal of Research on Educational Effectiveness, 14,* 270-308. DOI: 10.1080/19345747.2020.1831115

Nye, N., Hedges, L. V., Konstantopoulos, S., (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Educational Evaluation and Policy Analysis, 21,* 127-142.

Nye, B., Hedges, L. V., & Konstantopoulos, (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal, 37,* 123-151.

Nye, Hedges and Konstantopoulos (2001). Are effects of small classes cumulative? Evidence from a Tennessee experiment. *The Journal of Educational Research, 94,* 336-345.

Nye, N., Hedges, L. V., Konstantopoulos, S., (2002). Do low-achieving students benefit more from small classes? Evidence from the Tennessee class size experiment. *Educational Evaluation and Policy Analysis, 24,* 201-217.

Orr, L.L., Olsen, R.B., Bell, S.H., Schmid, I., Shivji, A., and Stuart, E.A. (2019). Using the results from rigorous multi-site evaluations to inform local policy decisions. *Journal of Policy Analysis and Management, 38,* 978 – 1003.

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

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

SAS Institute Inc. (2011). *SAS/STAT®9.3* [Computer Software]. SAS Institute Inc.

Scriven, M. (2008). A summative evaluation of RCT methodology: & an alternative approach to causal research. *Journal of Multidisciplinary Evaluation, 5,* 11-24.

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

Shadish, W. R., Luellen, J. K., & Clark, M. H., (2006) Propensity scores and quasi-

experiments: A testimony to the practical side of Lee Sechrest. In R.R. Bootzin & P.E., McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 143-157). Washington DC: American Psychological Association.

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.

Smith, J. A., & Todd, P. E., (2005). Does matching overcome Lalonde's critique of non-experimental estimators? *Journal of Econometrics, 125,* 305-353.

Steiner, P. M., & Wong, V. C. (2018). Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons. *Evaluation review*, *42*(2), 214-247.

Steiner, P. M., Wong, V. C. & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift fur Psychologie*. *227*, 280 – 292.

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

Unlu, F., Lauen, D., Fuller, S. C., Berglund, T., and Estera E. (2021). Can Quasi-Experimental Evaluations that Rely on State Longitudinal Data Systems Replicate Experimental Results: Findings from a Within-Study Comparison. Forthcoming at *Journal of Policy Analysis and Management, 40(*2*),* 572-613.

Weidmann, B. and Miratrix, L. (2021). Lurking inferential monsters? Quantifying selection bias in evaluations of school programs. *JPAM, 40,* 964-986.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management, 33*(3), 778–808.

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., Steiner, P. M. & Anglin, K. L. (2018). What can be learned from empirical evaluations of nonexperimental methods? *Evaluation Review, 42,* 147-175.

Appendix A: Additional Theoretical Considerations Under Which we can Expect $Bias2$ to Exceed $Bias1$

We explore graphically the conditions under which the impact based on a comparison of outcomes between the remote and inference sites (Δ_{QE2}) is less biased than the experiment-based result from the remote sites ($\Delta_{NX|M}$). This condition is satisfied when the following relation holds (repeating Equation 10).

Using the criterion of lower net bias, we should prefer Δ_{QE2} to $\Delta_{XP|M}$ when:

$$(Bias1^2 - 2Bias1Bias2 + Bias2^2) < Bias1^2$$

$$\Leftrightarrow (Bias2)^2 - 2Bias2Bias1 < 0$$

$$\Leftrightarrow (Y_N(C) - Y_M(C))^2 - 2(Y_N(C) - Y_M(C))(\Delta_{XP|M} - \Delta_{XP|N}) \quad \text{(A1)}$$

We represent the terms for bias on a standard coordinate system:

$$(Bias2 = x, Bias1 = y)$$

Net bias for Δ_{QE2} is less than for $\Delta_{XP|M}$ under the following condition:

$$x^2 - 2xy < 0$$

$$\Leftrightarrow x^2 < 2xy \quad \text{(A2)}$$

When $x > 0$ this inequality is satisfied when the following condition is met:

$$x/2 < y \quad \text{(A3)}$$

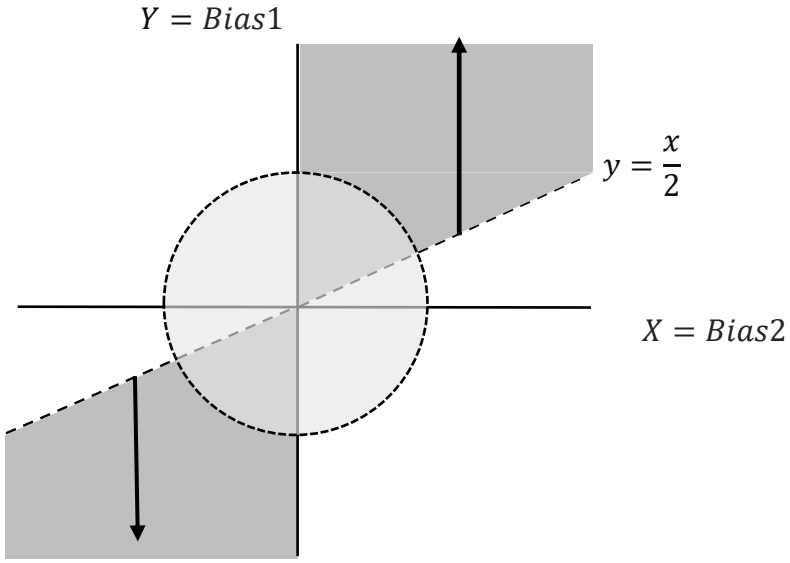
When $x < 0$ this inequality is satisfied when the following condition is met:

$$x/2 > y \quad \text{(A4)}$$

We observe that Δ_{QE2} is less than for $\Delta_{XP|M}$ in the regions indicated by the arrows in Figure A4. The proportion of occasions on which this condition is satisfied depends on the distribution of the two biases across multiple studies. Assuming no correlation, with the distribution of biases being roughly circular (schematically shown as the dashed circle centered

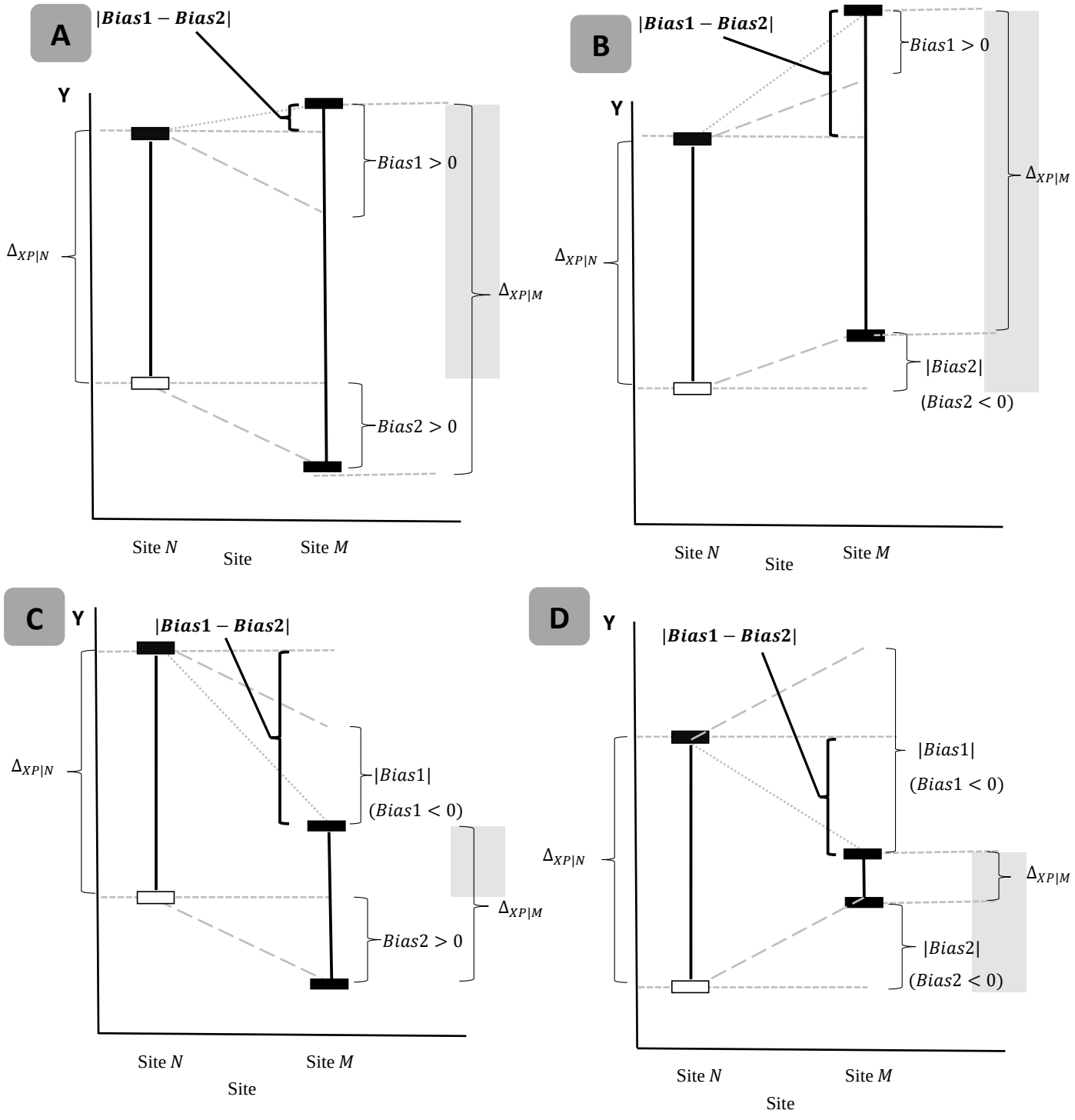
on the axes), on average $\Delta_{XP|M}$ would be preferred (more of the area of the circle is in the white region than the gray region). Given a positive correlation, with an ellipse (instead of the circle) tilted to the upper right, we may expect a more even preference for Δ_{NX2} or Δ_{XP}.

Figure A1. Mapping the space over which Δ_{QE2} has less net bias than $\Delta_{XP|M}$.



Appendix B: Four Main Scenarios for Comparing Bias

Figure B1. Four Prototypical Scenarios Involving $Bias1$ and $Bias2$



Appendix C: HL Models Used in Estimation

The Base Model

The HL model reflects the study design, including the unit of random assignment and sources of random sampling error:

The student-level (level-1) model is as follows:

$$y_{ijk} = \alpha_{0jk} + \alpha_{1jk}T_{ijk} + \varepsilon_{ijk} \quad \text{(C1)}$$

y_{ijk} is the achievement of student *i* in class *j*, in school *k*. Student random assignment to conditions is indicated by T_{ijk} (with value 1 for treatment, and 0 for control). The term ε_{ijk} is the random deviation in student performance from the class mean (i.e., the source of student-level random sampling error).

The class-level (level-2) model is as follows:

$$\alpha_{0jk} = \beta_{00k} + e_{0jk} \quad \text{(C2)}$$

$$\alpha_{1jk} = \beta_{10k} \quad \text{(C3)}$$

The term e_{0j} is the random deviation in class performance from the school mean (i.e., the source of class-level random sampling error).

The school-level (level-3) model is as follows:

$$\beta_{00k} = \gamma_{000} + r_{0k} \quad \text{(C4)}$$

$$\beta_{10k} = \gamma_{100} + r_{1k} \quad \text{(C5)}$$

The term r_{0j} is the random deviation in school average achievement from the grand mean of achievement, and r_{1k} is the random deviation in school average impact from the grand mean of the impact.

We can also summarize the model using a single mixed model formulation:

$$y_{ijk} = \gamma_{000} + \gamma_{100}T_{ijk} + [r_{0k} + r_{1k}T_{ijk} + e_{0jk} + \varepsilon_{ijk}] \quad \text{(C6)}$$

We assume the following distributions of the random effects:

$$r_{0k} \sim N(0, \tau_0) \quad \text{(C7)}$$

$$r_{1k} \sim N(0, \tau_1) \quad \text{(C8)}$$

$$e_{0jk} \sim N(0, \nu) \quad \text{(C9)}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2) \quad \text{(C10)}$$

$$\begin{pmatrix} r_{0k} \\ r_{1k} \\ e_{0jk} \\ \varepsilon_{ijk} \end{pmatrix} \sim N\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0 & \tau_{01} & 0 & 0 \\ \tau_{01} & \tau_1 & 0 & 0 \\ 0 & 0 & \nu & 0 \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix} \quad \text{(C11)}$$

Reflecting the point that Mean Squared Biases are variances of site-level averages of performance and of impact, and the covariance between site-level averages and impact, corresponding to quantities in equations 23 – 25, the variance and covariance components estimated at the site-level in the HL models are used to summarize each form of Mean Squared Bias:

$$\widehat{MSB}_{NX(a)} = \widehat{\tau_0}, \quad \text{(C12)}$$

$$\widehat{MSB}_{NX(b)} = \widehat{\tau_0} + \widehat{\tau_1} + 2\widehat{\tau_{01}} \quad \text{(C13)}$$

$$\widehat{MSB}_{XP|D=1)} = \widehat{\tau_1}, \quad \text{(C14)}$$

In addition to the base model described above, results from two additional types of models were examined:

1. With site-centered student-level covariates and their interactions with the treatment variable (modeled at Level-1) (gender [1=male, 0=female], eligibility for Free or Reduced Price Lunch [1=eligible, 0=non-eligible], and minority [non-White] status [1=minority, 0=non-minority], the years of teaching experience of a student's teacher,

whether the student's teacher holds a Master's degree or higher [1=year, 0=no], and end of kindergarten scores on tests of math and reading.)

2. With the effects in (1) and with the main effects of site-level uncentered covariates modeled at level-3 and their interactions with treatment. (Covariates at the school level are school averages of uncentered student-level covariates and variables indicating school urbanicity [whether a school is inner-city, suburban, rural or urban.])

Appendix D: Deriving the Relationship Between In-Sample and Out-of-Sample Large-to-Small Generalizations

We derive an expression for the difference between two variance quantities. The first includes the impact for the inference site in the grand mean, and supports an in-sample generalization:

$$\tau_1^* = \frac{1}{N}\sum_{i=1}^{N}(\Delta_i - \frac{\sum_i \Delta_j}{N})^2 \quad \text{(D1)}$$

The second quantity excludes the inference site from the grand mean and supports an out-of-sample generalization:

$$\tau_1^{**} = \frac{1}{N}\sum_{i=1}^{N}(\Delta_i - \frac{\sum_{j\neq i} \Delta_j}{N-1})^2 \quad \text{(D2)}$$

We rewrite Equation D1 as follows:

$$\tau_1^* = \frac{1}{N}[(\Delta_1 - \frac{(\Delta_1+\cdots+\Delta_N)}{N})^2 + \cdots + (\Delta_N - \frac{(\Delta_1+\cdots+\Delta_N)}{N})^2] \quad \text{(D3)}$$

We rewrite Equation D2 as follows:

$$\tau_1^{**} = \frac{1}{N}[(\Delta_1 - \frac{(\Delta_2+\cdots+\Delta_N)}{N-1})^2 + \cdots + (\Delta_N - \frac{(\Delta_1+\cdots+\Delta_{N-1})}{N-1})^2] \quad \text{(D4)}$$

Next, we express τ_1^* in terms of τ_1^{**}.

$$\tau_1^* = \frac{1}{N}[(\Delta_1 - \frac{(\Delta_1+\cdots+\Delta_N)}{N})^2 + \cdots + (\Delta_N - \frac{(\Delta_1+\cdots+\Delta_N)}{N})^2]$$

$$= \frac{1}{N}\{\left(\Delta_1(\frac{N-1}{N}) - \frac{(\Delta_2+\cdots+\Delta_N)}{N}\right)^2 + \cdots + \left(\Delta_N(\frac{N-1}{N}) - \frac{(\Delta_1+\cdots+\Delta_{N-1})}{N}\right)^2\}$$

$$= \frac{1}{N}\{\frac{1}{N^2}\left((N-1)\Delta_1 - (\Delta_2+\cdots+\Delta_N)\right)^2 + \cdots + \frac{1}{N^2}\left((N-1)\Delta_N - (\Delta_1+\cdots+\Delta_{N-1})\right)^2\}$$

$$= \frac{1}{N}\{\frac{(N-1)^2}{N^2}\left(\Delta_1 - \frac{(\Delta_2+\cdots+\Delta_N)}{N-1}\right)^2 + \cdots + \frac{(N-1)^2}{N^2}\left(\Delta_N - \frac{(\Delta_1+\cdots+\Delta_{N-1})}{N-1}\right)^2\}$$

$$= \frac{(N-1)^2}{N^2}\{\frac{1}{N}\left(\Delta_1 - \frac{(\Delta_2+\cdots+\Delta_N)}{N-1}\right)^2 + \cdots + \frac{1}{N}\left(\Delta_N - \frac{(\Delta_1+\cdots+\Delta_{N-1})}{N-1}\right)^2\}$$

$$= \frac{(N-1)^2}{N^2}\{\frac{1}{N}\sum_{i=1}^{N}(\Delta_i - \frac{\sum_{j\neq i}\Delta_j}{N-1})^2\}$$

$$= \frac{(N-1)^2}{N^2}\tau_1^{**}$$

$$= \left(\frac{N-1}{N}\right)^2 \tau_1^{**} \quad \text{(D5)}$$

The quantity supporting the in-sample generalization is smaller compared to the quantity supporting an out-of-sample generalization. This is consistent with the intuition that for an in-sample generalization, the quantity being generalized is an average across all sites including the inference site; therefore, the difference between it and the inference site is slightly reduced because the inference site figures into both quantities.

Appendix E: Variance Components Estimates and Deviance Statistics for the Main Models

Table E1. Variance Components Estimates and Deviance Statistics for the Main Models

	Without class effects			With class effects		
Models	1	2	3	4	5	6
Variance components estimates	No covariates (base model)	School-centered student-level variables only	School-centered student-level variables and school-level variables	No covariates (base model)	School-centered student-level variables only	School-centered student-level variables and school-level variables
Variance in school deviations in average achievement from grand mean achievement: $\widehat{\tau_0}$	406.07****	393.79****	122.66****	364.84****	349.96****	70.7429**
Variance in school deviations in average impact from grand mean of impact: $\widehat{\tau_1}$	186.51****	190.28****	166.7***	66.8598	59.7474	39.2772
Covariation between school deviations in average achievement and average impact from their grand mean impacts: $\widehat{\tau_{01}}$	-86.017	-71.951	-29.357	-52.978	-32.237	15.1589
Variance in class deviations in average achievement within schools: $\hat{v}$				147.94****	148.21****	147.37****
Variance across students within schools $\widehat{\sigma^2}$	1651.73****	1536.04****	1535.59****	1562.6****	1448.83****	1448.36****
-2 log likelihood statistic	35598.5	35355.7	35280.6	35538.3	35290.6	35214.9
Number of effects estimated	4 random 2 fixed	4 random 12 fixed (L1)	4 random 12 fixed (L1), 20 fixed (L3)	5 random 2 fixed	5 random 12 fixed (L3)	5 random 12 fixed (L1), 20 fixed (L3)

****p<.001, ***p<.01, **p<.05, *p<.10

n(students)=3452, n(schools)=73;

L1=Level 1 (student), L3 = Level 3 (school); All fixed effects include main and interaction effects with treatment (e.g., 12 fixed effects at L1 are main effects of six covariates and their interactions with treatment.); The average impacts of small classes are .25, .24, .25 and .24 standard deviation units for Models 1, 2, 4 and 5 (i.e., for models without interactions of treatment with site-level characteristics), each significant at α=.05.

Figure 1. Representation of the Experimental (*XP*) Average Treatment Effect of Assignment to *T* relative to *C* at site *N*

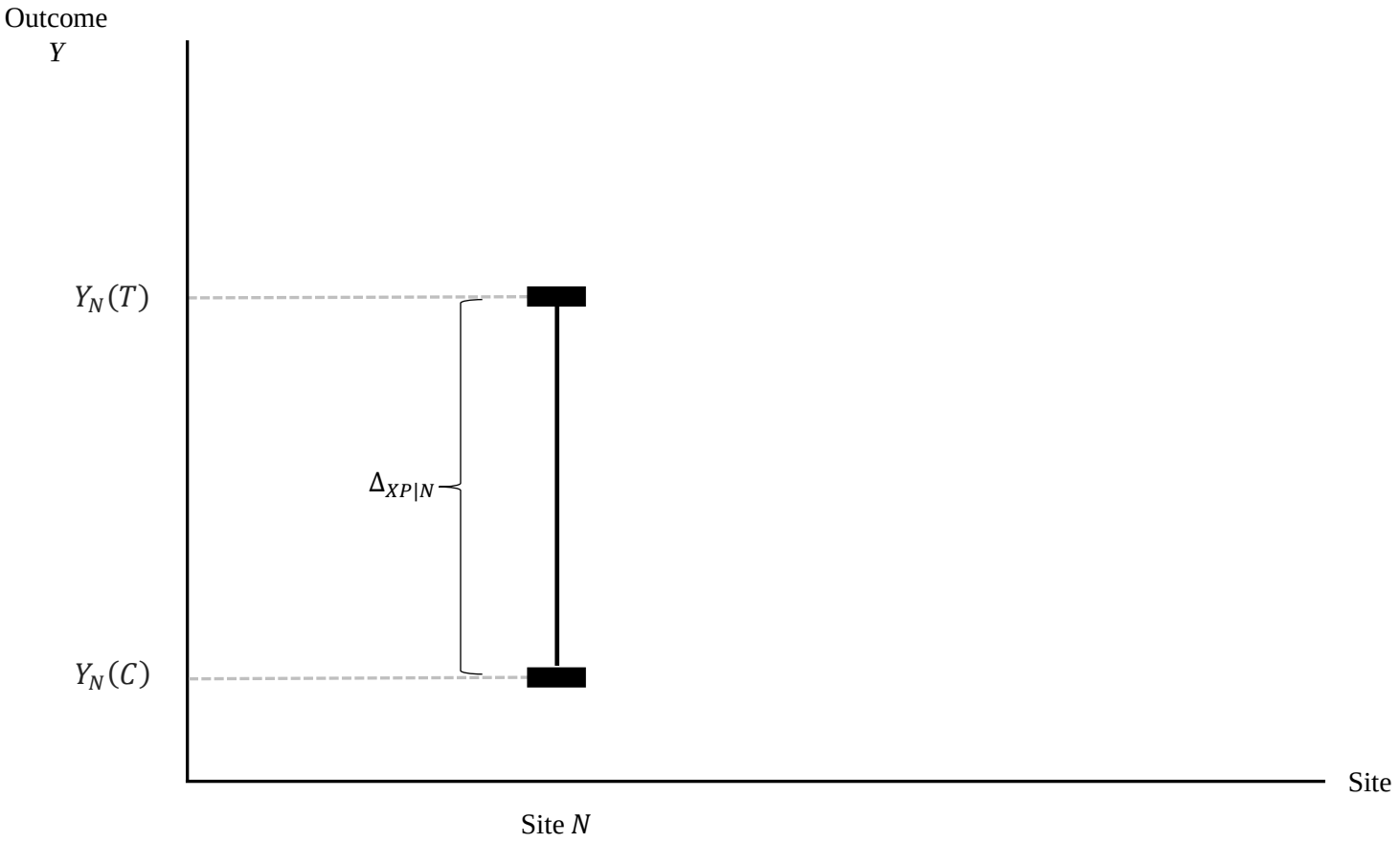


Figure 2. Representation of the Experimental (*XP*) Impact Obtained at Site *M*, $\Delta_{XP|M}$, as Used to Infer Impact at *N*.

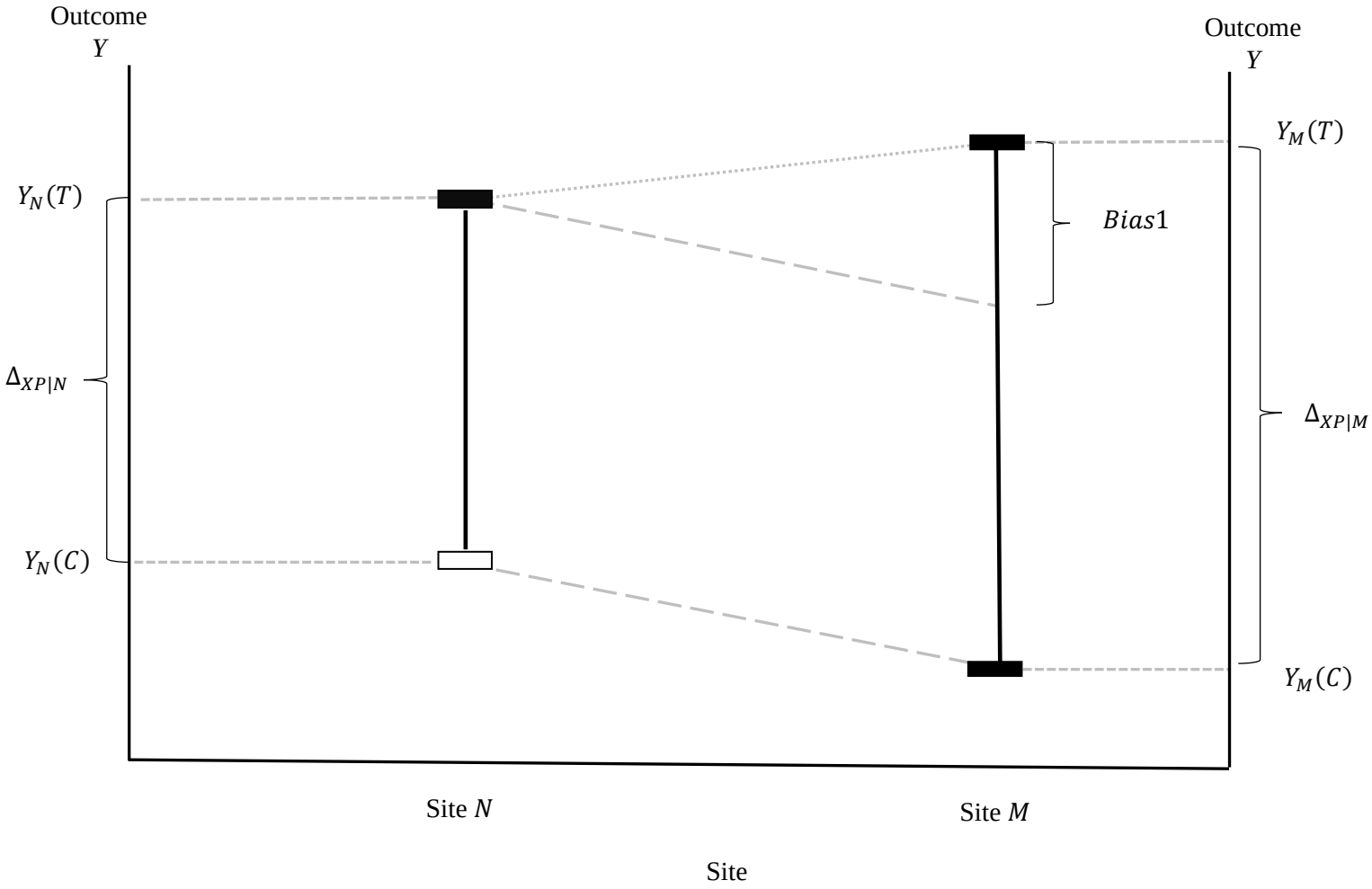


Figure 3. Representation of the Quasi-Experimental Average Treatment Effect (*QE1*) Inferred to Site *N* Through Comparison with Site *M*

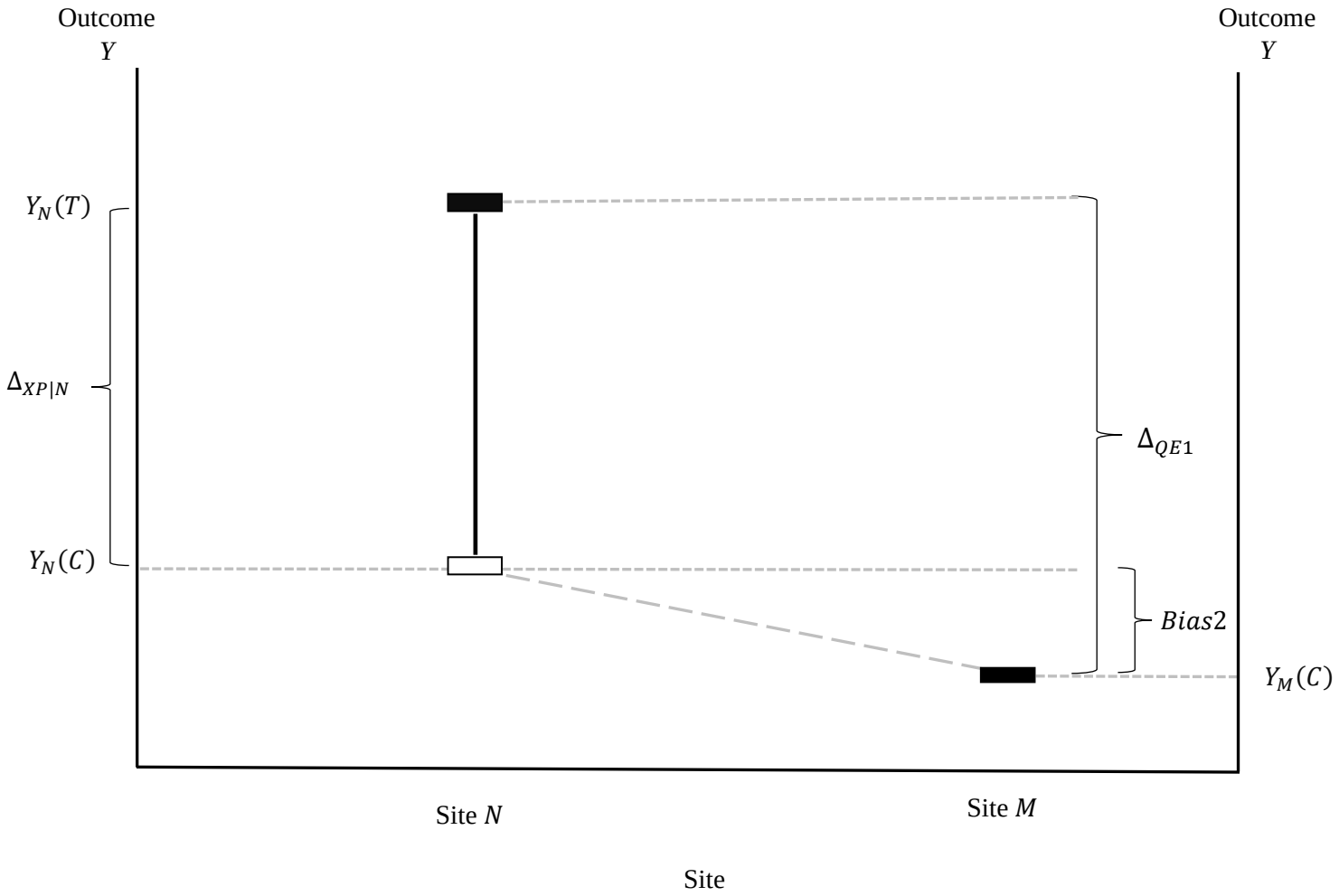


Figure 4. Representation of the Quasi-Experimental Average Treatment Effect (*QE2*) Inferred to Site *N* Through a Comparison with Site *M*.

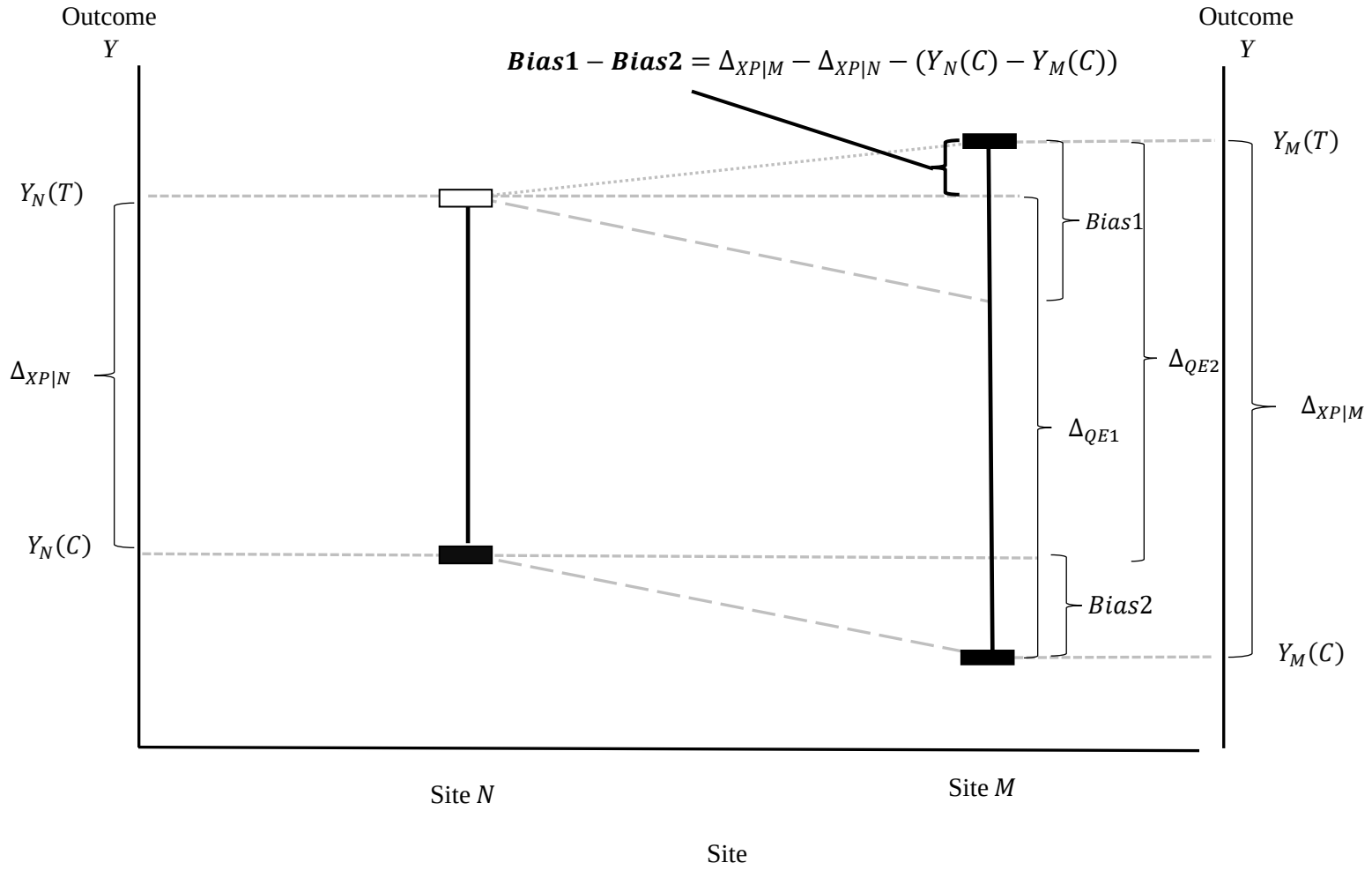


Figure 5a. Estimates of Root Mean Squared Bias *Without* Adjustment for Class-level Random Effects (Expressed in Units of the Standard Deviation of the Outcome Distribution)

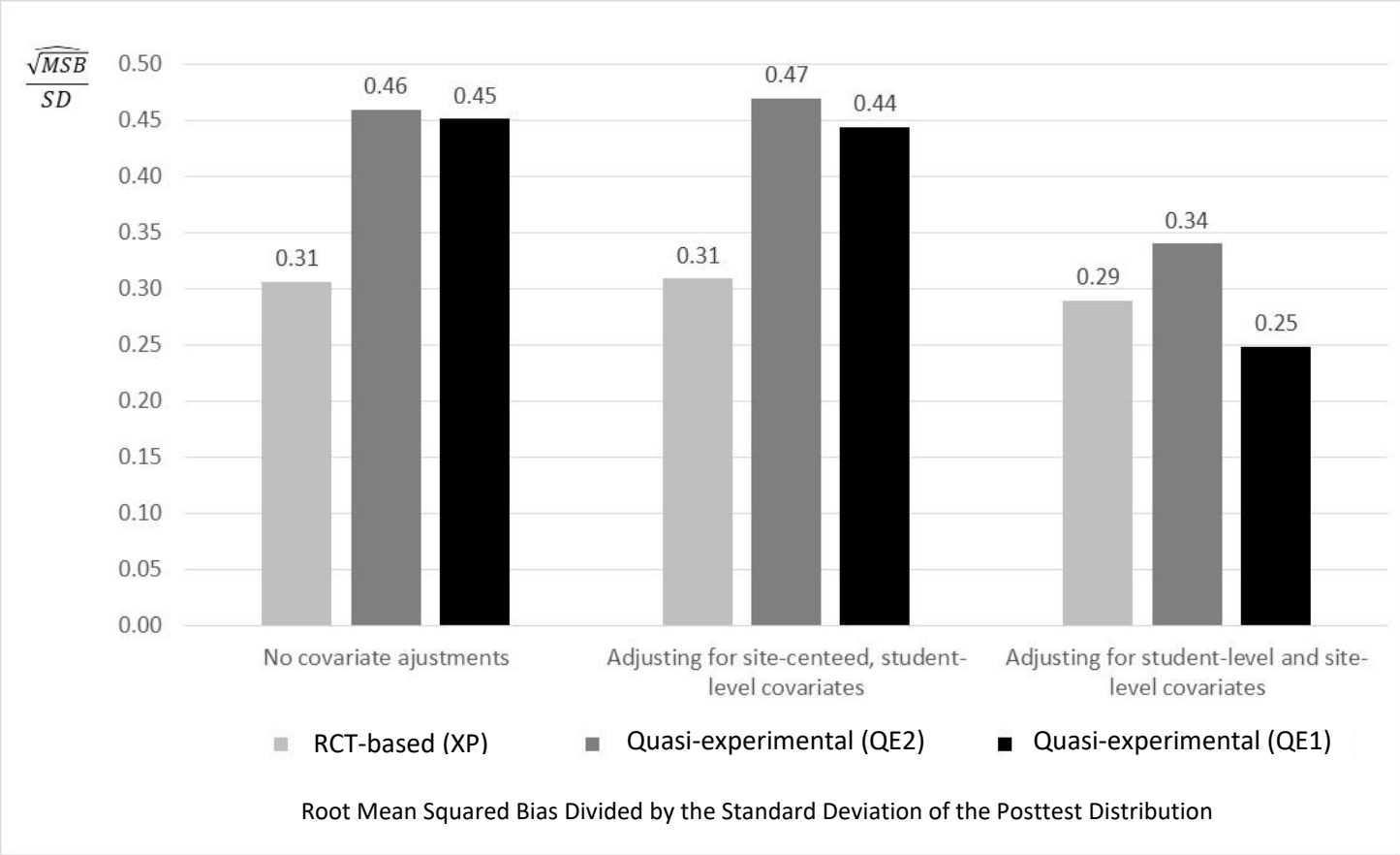


Figure 5b. Estimates of Root Mean Squared Bias *With* Adjustment for Class-level Random Effects (Expressed in Units of the Standard Deviation of the Outcome Distribution)

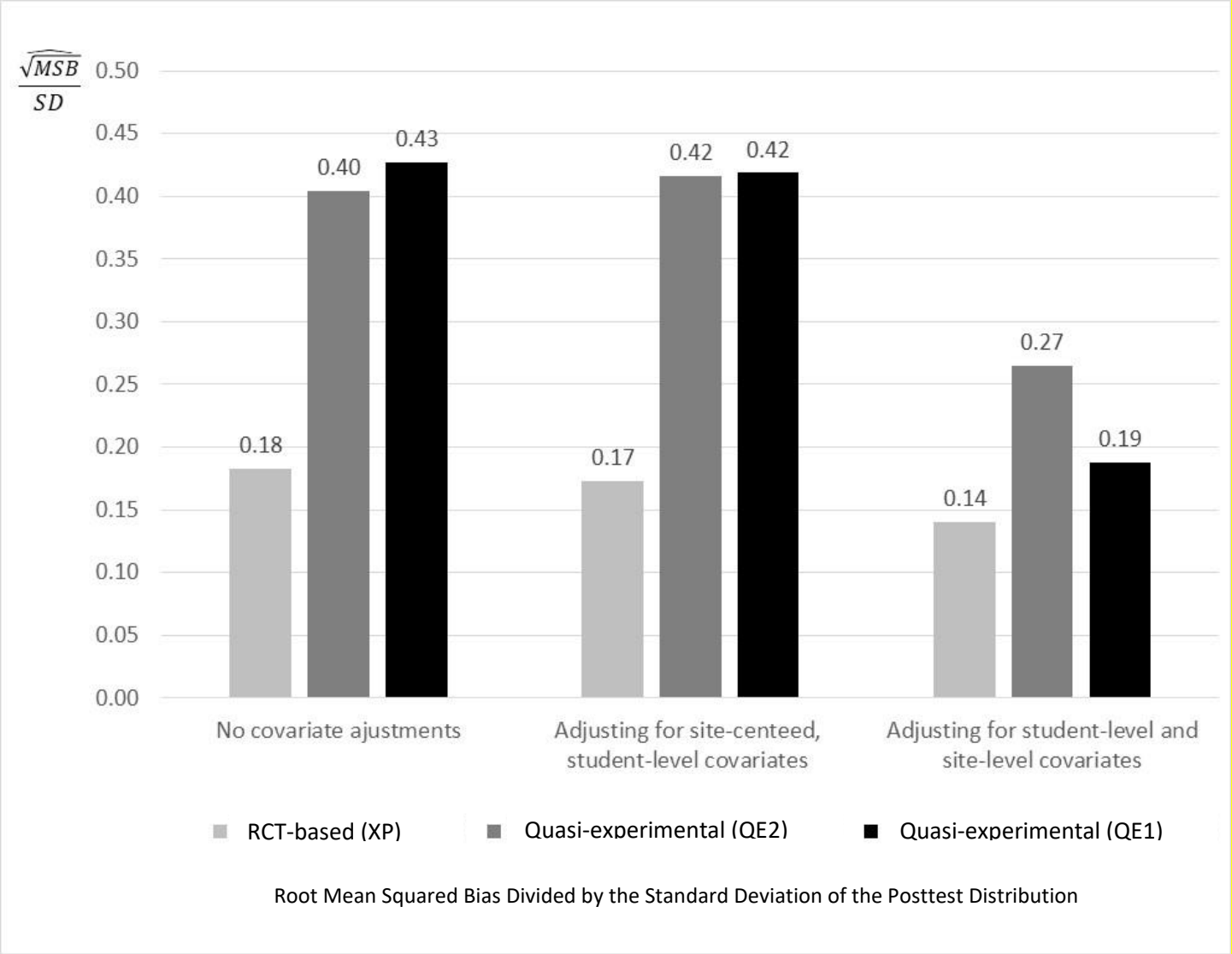


Table 1. Scenarios for Inferring the Average Causal Impact of Program T to Site N Through a Comparison of Outcomes from Site M.

Causal quantity of interest	Missing Component	Quantity used to infer impact at N	Bias
$\Delta_{XP\|N} = Y_N(T) - Y_N(C)$	None (Scenario 1)	$\Delta_{XP\|N}$	0
	$Y_N(C)$ (Scenario 2)	Option 1: $\Delta_{XP\|M} = Y_M(T) - Y_M(C)$	$Bias1 = \Delta_{XP\|M} - \Delta_{XP\|N}$
		Option 2: $\Delta_{QE1} = Y_N(T) - Y_M(C)$	$Bias2 = Y_N(C) - Y_M(C)$
	$Y_N(T)$ (Scenario 3)	Option 1: $\Delta_{XP\|M} = Y_M(T) - Y_M(C)$	$Bias1 = \Delta_{XP\|M} - \Delta_{XP\|N}$
		Option 2: $\Delta_{QE2} = Y_M(T) - Y_N(C)$	$Bias1 - Bias2 = \Delta_{XP\|M} - \Delta_{XP\|N} - [Y_N(C) - Y_M(C)]$

Table 2. Estimates of Root Mean Squared Biases for Evaluating Several Approaches to Generalization.

Research question:	On average, is there bias in the experiment-based (*XP*) generalization ?	On average, is there bias in the comparison group-based generalization that involves **imputing average achievement with treatment** for the inference sites (i.e., *QE2*)?	On average, is there bias in the comparison group-based generalization that involves **imputing average achievement in absence of treatment** for the inference sites (i.e, *QE1*)?	Is the average magnitude of bias different between the experiment-based generalization (*XP*) and the one based on the non-experimental comparison *QE2*? (Scenario 3)	Is average magnitude of bias different between the experiment-based generalization (*XP*) and the one based on the non-experimental comparison *QE1*? (Scenario 2)	Is average magnitude of bias different between the comparison group-based generalizations QE1 and *QE2*?
Null hypothesis:	$H_0: MSB_{XP} = 0$	$H_0: MSB_{QE2} = 0$	$H_0: MSB_{QE1} = 0$	$H_0: MSB_{QE2} - MSB_{XP} = 0$	$H_0: MSB_{QE1} - MSB_{XP} = 0$	$H_0: MSB_{QE1} - MSB_{QE2} = 0$
Corresponding estimates of interest in Figures 5a and 5b	Height of light gray bar	Height of dark gray bar	Height of black bar	Difference: dark gray bar – light gray bar	Difference: black bar – light gray bar	Difference: Black bar – dark gray bar
	Estimates of RMSB (or differences in RMSB) expressed in SD units (*not controlling for* class-level random sampling error)					
No covariates	.31****	.46***	.45****	.15*	.14**	-.01
[a]Student-level (site-centered) covariates only	.31****	.47***	.44****	.16*	.13**	-.03
[a] Student-level (site-centered) and site-level covariates	.29***	.34***	.25****	.05	-.04	-.09
	Estimates of RMSB (or differences in RMSB) expressed in SD units (*controlling for* class-level random sampling error)					
No covariates	.18	.40**	.43****	.22*	.14**	.03
[a]Site-centered student-level covariates only	.17	.42**	.42****	.25**	.25**	0.00
[a] Student-level (site-centered) and site-level covariates	.14	.27	.19**	.13	.05	-.08

****p<.001, ***p<.01, **p<.05, *p<.10

[a]Models include the main effects of the covariates and their interactions with treatment.

Note: the values displayed are the square roots of estimates of corresponding *MSB* divided by the standard deviation of the outcome variable. Expressing results in the metric of the standardized effect size allows comparison with values for average impact and yearly expected growth. For example, the average impact of small classes on second grade reading performance in this experiment is .24-.25 (similar to results in Nye, Hedges and Konstantopoulos, 2000). For reference, annual expected growth in second grade reading scores is approximately .60 standard deviations (Hill et al., 2008).

Covariates at the student level (all school centered) are: gender (1=male, 0=female), eligibility for Free or Reduced Price Lunch (1=eligible, 0=non-eligible), and minority (non-White) status (1=minority, 0=non-minority), the years of teaching experience of a student's teacher, whether the student's teacher holds a Master's degree or higher (1=year, 0=no), and end of kindergarten scores on tests of math and reading. Covariates at the school level are school averages of uncentered student-level covariates and variables indicating school urbanicity (whether a school is inner-city, suburban, rural or urban.)

Supplement 1: Demonstrating that the Results do not Reflect Model Overspecification

A possible concern is that the results of analysis that involves adjustment for school-level covariates reflect model overspecification. More specifically, our data included 73 schools. The most saturated model included 7 continuous variable school-level variables, and a 4-level categorical variable indicating school urbanicity. The most saturated model included both main effects of these covariates and their interactions with treatment.

Model overspecification may bias the estimates of variance components, on which the results of this work rely.

To assess whether over-specification is occurring we examined results for 44 models where we gradually introduced predetermined sets of school-level covariates, and their interactions with treatment, into the models.

We observed that inclusion of specific sets of school-level covariates produced similar relative changes in estimated variance components regardless of how many school-level covariates were already in the model. If the models were overfitted, we would instead expect estimates of variance components to be less robust when including school-level covariates for more-saturated models.

The results from the 44 models are summarized in Figure SB1 below (The sets of covariates and school main effects used with each model are displayed in Table SB1). It shows the estimates of τ_0, τ_1, τ_{01} and ν for different combinations of covariates.

The results on the left half of the display (models A0 – C7), show parameter estimates without a class-level random effect. The results on the right half (models D0 – F7) , show corresponding estimates after inclusion of a class-level random effect. (The line with longer black dashes is the estimate of the variance component for classes, and it appears only on the right half of the figure.)

The six vertical bands in different shades show a similar progression of models:

Vertical band 1 (white): includes models with no teacher random effects, no interactions between site-level covariates with treatment, and gradual introduction of main effects of site-level covariates:

A0: no site-level covariates (no main effects of school-level covariates).
A1: two teacher-based school-level covariates: proportion of teachers with advanced degree, average years teaching (2 main effects of school-level covariates).
A2: three student-based school-level covariates: proportion low-SES, proportion male, proportion minority (3 main effects of school-level covariates).
A3: two student-based school-level covariates: end-of-K school average math and reading achievement scores (2 main effects school-level covariates).
A4: three indicators for four levels of urbanicity (3 main effects of school-level covariates).
A5: all student-based covariates (in A2 and A3) (5 main effects school-level covariates).

A6: all student-based and teacher-based covariates (in A1 - A3) (7 main effects of school-level covariates).
A7: all covariates (10 main effects of school-level covariates).

Relatively few school-level covariates (at most 10 with a sample size of 73 schools) are introduced with this set of models, thereby limiting risk of bias from over-specification.

We observe that between-school variation in average performance (solid black line) is differentially reduced depending on which main effects of school-level covariates are modeled. The lowest estimated between-school variability in average achievement is achieved when all 10 covariates are included (model A7). The school variation in impact (gray line) is stable, which we expect given that moderating effects of the school-level variables on the impact are not included in these models. (The number of terms involving school-level covariates across the eight models A0 – A7 is 0, 2, 3, 2, 3, 5, 7 and 10)

Vertical band 2 (light gray): includes models like in the first vertical band, except we double the number of terms involving school-level covariates in each model by introducing the corresponding interactions with treatment. For example, model B3 includes the same three student-based school-level covariates as model A2, but also includes the interactions of those variables with treatment, resulting in six terms with school-level covariates.

A larger number of covariates is used with this set of models, introducing more risk of bias due to model overspecification. However, if we compare results of models A1 – A7 (vertical band 1) with corresponding models B1 – B7 (vertical band 2) we see that the school variation in average performance (solid black line) is almost parallel across band 1 and band 2. We also see that introducing interactions of treatment with school-level variables leads to a small degree of fluctuation in estimates of school variation in impact (gray line) across the models B1 – B7 (as compared to the nearly flat line in A1 – A7).

Increasing the number of effects involving school-level variables leads to minor departures in estimates of variance component from corresponding less-parameterized models. (The number of terms involving school-level covariates across the seven models B1 – B7 is 4, 6, 4, 6, 10, 14 and 20.)

Vertical band 3 (darker gray): the models include the main effects of all 10 school-level covariates, and then introduce interactions between site-level covariates and treatment corresponding to those used with models in vertical band 2. Therefore, the number of terms involving school-level covariates is 10+2 for Model C1, 10+3 for Model C2, 10+2 for Model C3, etc.

We observe that the school variation in average performance across models (solid black line) is almost flat, which reflects that we are consistently including all main effects of school-level covariates across models C1 – C7. Also, inclusion of interactions of treatment with school-level variables leads to a relatively small fluctuation in school variation in impact (gray line) across the models, similar to what we observed with corresponding models in vertical band 2.

Increasing the number of effects involving school-level variables leads to minor departures in estimates of the variation in the treatment effect across schools compared to corresponding less-parameterized models. (The number of terms involving school-level covariates across the seven models C1 – C7 is 12, 13, 12, 13, 15, 17 and 20.)

Vertical bands 4, 5 and 6: The models correspond exactly to those in vertical bands 1, 2 and 3, respectively, except a class random effect has been included.

The pattern almost exactly parallels the one across vertical bands 1, 2 and 3 except that there is a uniform downward shift in the variance across schools in average performance, and a proportionately larger downward shift in the variance across schools in average impact. The within-school variation between-classes is now added (represented by the solid black line with longer dashes.)

As above, if the results were becoming biased with the inclusion of a larger number of terms involving site-level covariates, we would expect greater instability of the variance components estimates, potentially with a large reduction in their values as the number of covariates increases. We do not observe this trend. First, we observe the same relative reductions in variability in average outcomes across schools with the addition of main effects of site-level covariates, regardless of how many covariates are already in the model. Second, we observe little fluctuation in the variance component for impact across schools across the models. The only factor that makes an across-the-board difference in estimates of the variance components is the inclusion of the class-level random effect.

Table SB1. School-level Variables Included in Tests of Sensitivity of Variance Components Estimates to Model Overspecification

	Main effects of:				Interactions of treatment with:				Random effects:				
Model	TB	SB	PRE	URB	TB	SB	PRE	URB	τ_0	τ_1	τ_{10}	ν	σ^2
A0									*	*	*		*
A1	*								*	*	*		*
A2		*							*	*	*		*
A3			*						*	*	*		*
A4				*					*	*	*		*
A5		*	*						*	*	*		*
A6	*	*	*						*	*	*		*
A7	*	*	*	*					*	*	*		*
B1	*				*				*	*	*		*
B2		*				*			*	*	*		*
B3			*				*		*	*	*		*
B4				*				*	*	*	*		*
B5		*	*			*	*		*	*	*		*
B6	*	*	*		*	*	*		*	*	*		*
B7	*	*	*	*	*	*	*	*	*	*	*		*
C1	*	*	*	*	*				*	*	*		*
C2	*	*	*	*		*			*	*	*		*
C3	*	*	*	*			*		*	*	*		*
C4	*	*	*	*				*	*	*	*		*
C5	*	*	*	*		*	*		*	*	*		*
C6	*	*	*	*	*	*	*		*	*	*		*
C7	*	*	*	*	*	*	*	*	*	*	*		*
D0									*	*	*	*	*
D1	*								*	*	*	*	*
D2		*							*	*	*	*	*
D3			*						*	*	*	*	*
D4				*					*	*	*	*	*
D5		*	*						*	*	*	*	*
D6	*	*	*						*	*	*	*	*
D7	*	*	*	*					*	*	*	*	*
E1	*				*				*	*	*	*	*
E2		*				*			*	*	*	*	*
E3			*				*		*	*	*	*	*
E4				*				*	*	*	*	*	*
E5		*	*			*	*		*	*	*	*	*
E6	*	*	*		*	*	*		*	*	*	*	*
E7	*	*	*	*	*	*	*	*	*	*	*	*	*
F1	*	*	*	*	*				*	*	*	*	*
F2	*	*	*	*		*			*	*	*	*	*
F3	*	*	*	*			*		*	*	*	*	*
F4	*	*	*	*				*	*	*	*	*	*
F5	*	*	*	*		*	*		*	*	*	*	*
F6	*	*	*	*	*	*	*		*	*	*	*	*
F7	*	*	*	*	*	*	*	*	*	*	*	*	*

All effects displayed are of school level variables. (Student-level site-centered covariates and their interactions with treatment are included in each model)

TB=teacher-based, SB=student-base (except pretests), PRE=pretests, URB=urbanicity

Figure SB1. Variance and Covariance Estimates for Multiple Models with and Without Class-Level Random Effets and Adjusting for Effects of Specific Covariates.