Towards the Unification of External and Internal Validity: An Empirical Approach

Andrew P. Jaciw

Empirical Education Inc.

Abstract

It is commonly accepted that internal validity comes before external validity. The former is concerned with establishing the causal relationship between variables, and the latter addresses the extrapolation of a causal relationship across conditions. In this work we challenge the long-standing notion of the primacy of internal validity. We argue that when generalization of a causal inference is the goal, the two forms of validity achieve parity in their logical priority and importance. The role of selection, which results in a difference in the composition between the "generalized from" and the "generalized to" samples, poses a concurrent threat to both forms of validity. As a result, neither takes precedence, and strategies for eliminating bias from selection simultaneously affect both forms of validity. We present the argument graphically and algebraically, and include an empirical example. An implication of this result is a challenge to the legitimacy of Randomized Control Trials (RCTs) as a gold standard method for causal inference, because this prioritization stems from the assumption of the primacy of internal validity. RCTs may be considered gold standard only under the special circumstances where generalization is not an objective. The results of this work have the potential to expand what is regarded as an admissible range of study designs, as well as the evidence base, for evaluating which programs work for whom and under what conditions.

Towards the Unification of External and Internal Validity: An Empirical Approach

In research and evaluation across many fields it has become conventional to think of internal validity of causal inferences as coming before external validity. The argument is that we first establish as best as we can whether the relationship between a treatment or exposure and an outcome is causal (internal validity), and only then do we address the problem of demonstrating that the relationship generalizes across changes of conditions (external validity).

In the current work, we introduce a methodological innovation that runs contrary to the conventional ordering of internal and external validity. Specifically, we argue that when causal generalization is the aim, and when there is a systematic difference between the "generalized from" and the "generalized to" groups in terms of the composition of their samples, then establishing internal and external validity should be considered as part of the same problem.

This innovation is important because it has the potential to elevate the status of non-experimental methods (such as quasi-experimental designs, including comparison group design) and increase their use for drawing generalizations about causal effects. This, in turn, may potentially increase the pool of allowable evidence for informing generalizations, serving a larger number of stakeholders and decision-makers who need more-ample evidence more quickly.

This work proceeds as follows. First, we summarize the standard view concerning internal and external validity and their prioritization. Second, we develop an alternative view that addresses the important role of selection bias for both forms of validity, and has implications for drawing causal generalizations. We develop the argument (1) graphically, (2) algebraically, and (3) empirically. We end the work by drawing conclusions and discussing implications.

The Standard Model: The Primacy of Internal Validity

In this work, we provide an alternative interpretation of the relationship between internal and external validity, with potential for rebalancing criteria for causal validity, and reprioritizing what are considered "gold standard" methods. Before we make our case, we review standard definitions of internal and external validity, and the accepted model for their prioritization.

Internal validity. The internal validity of a causal inference is normally defined as addressing whether the observed covariation between two variables reflects a causal relationship (Shadish et al. 2002). Primary threats to internal validity include factors that compete with treatment in the explanation of outcomes (Shadish et al., 2002), with selection being the main concern. Selection happens when the treatment and comparison groups differ in the distribution of attributes of their members that affect the outcome. For example, the internal validity of the inference that an educational program causes improvement in students' reading skills may be challenged if the conclusion is reached by comparing program users to non-users, and if the two groups differ in terms of home environments that have a bearing on reading, independently of treatment. When this happens the effect of the program is conflated with effects of individual characteristics (quality of home environment). Stated more formally, one threat to internal validity is a confounded selection of persons to conditions (Hotz et al., 2005), which results in a biased causal inference.

The best way to limit the threat to internal validity from selection for a study sample is to randomly assign subjects to conditions. This random selection mechanism, by definition, eliminates bias from confounded selection into conditions (Hotz et al., 2005). A quasi-experiment (QED) often supports weaker claims of internal validity because the mechanism for selection into conditions is not fully understood.

External validity. A common definition of external validity is the extent to which a causal relationship holds across differences in "persons, settings, outcomes or treatment variants" (Shadish et al., 2002, p. 256). Threats to the external validity of a causal inference include factors that moderate the relationship between a treatment and an outcome, including characteristics of persons that interact with treatment and thereby produce variation in its impact.

The primacy of internal validity. These definitions of internal and external validity are consistent with the idea that the internal validity of a causal inference is established before making claims concerning its external validity. According to this view, one first establishes the validity of the statement that the relation between treatment (T) and the outcome (O), represented as T→O, is causal, and only then does one discuss the role of additional factors that impinge on the relationship and moderate the impact, thereby limiting its generalizability. In practice, this prioritization of validity means using designs and methods that help to rule out plausible rival explanations for the covariation of treatment and outcome, and then evaluating the robustness of that relationship across conditions that might moderate it. Failure to replicate a causal relationship implies a limitation of external validity (Cook, 2002; Shadish et al., 2002). From this perspective, if the internal validity of a causal inference is dubious, then claims about its external validity are even more uncertain because there is a high risk that they rest on a false causal claim.

That internal validity comes first is seen in much of standard writing on the topic. Shadish et al., (2002) describe it as the "sine qua non", that is, the "without which not" (Shadish et al., 2002) or as " 'the basic minimum without which any experiment is uninterpretable' " (Campbell and Stanley, 1963, p. 5, in Shadish, et al., p. 97). However, they are also careful to note that internal validity is inseparable from external validity. The latter is considered "the desideratum" (the purpose or objective) of educational research (Shadish et al., 2002, p. 97).

External validity essentially is a close second. The focus on "internal validity first" is also observed in economics (and more recently many other fields) with the development of the Within Study Comparison (WSC) research tradition (Lalonde, 1986; Fraker et al., 1987). This research lineage is dedicated to understanding the role of selection bias as it affects internal validity separately from the problem of external validity. In a recent count, more than 60 such studies have been conducted (Wong et al., 2018). The works evaluate the capacity of comparison group designs, and other non-experimental methods, to replicate benchmark impacts from experiments. A few exceptions that apply WSC methods to empirically evaluate the accuracy of generalized causal inferences include works by Dehejia et al., (2021), Hotz et al., (2005), Hotz et al, (2006), Jaciw (2016[a],2016[b]), Jaciw et al. (2021) and Orr et al., (2019).

The primacy of internal validity is also reflected in how evidence is prioritized in technical standards for conducting impact research and evaluation such as by the What Works Clearinghouse (WWC) (U.S. Department of Education) and Evidence for ESSA (https://education.jhu.edu/2020/02/evidence-for-essa/) in the U.S., and the document *Measuring Impact by Design* (Privy Council Office, 2019) in Canada. For example, in standards of evidence by WWC, only a handful of designs can achieve the coveted status of meeting standards "without reservations", with Randomized Control Trials (RCTs) chief among them. Comparison group designs, which are quasi-experimental, can at best meet standards with reservations[1]. In *Measuring Impact by Design* the primacy of internal validity is absolutely clear: "Once internal validity is assured to a good degree … external validity will then become an important issue for future roll out and design. (p. 49)."

[1] In this work, among QEDs we consider primarily comparison group designs in which a group that receives treatment is compared with one that does not, and where selection into groups is not determined through random assignment (i.e., the "non-equivalent comparison group design" [Shadish et al., 2002]). There are many types of QEDs (Shadish et al., 2002), but for this work we use the term interchangably with comparison group designs.

Implication of putting internal validity first. If we accept the premise that internal validity is prior to external validity, then it follows that generalizations that are based on evidence from QEDs are on a shakier foundation than ones that are based on true experiments. Generalizations from non-degraded experiments are subject only to external validity bias ($EVbias$) whereas generalized causal inferences from quasi-experiments are subject to both internal validity bias ($IVbias$) and $EVbias$, so that the situation can only get worse. (In the section that follows we mount a challenge to this standard view.)

Before stating the main ideas of this work, we note that not all traditions in research and evaluation consider internal validity to come before other forms of validity, or even define internal validity as we have described above. Lee Cronbach and his colleagues are one example, (e.g., Albright et al., 2000; Cronbach, 1975; Cronbach 1982). Also, not everyone regards randomized experiments as the one best method for establishing causal relationships between variables (e.g., Patton, 2015; Scriven, 2008).

The Important role of Selection as a Threat to External Validity

In this work we assert that if the generalizability of a causal relation is the goal, then the primacy of internal validity is unfounded, and that there is a principled way to demonstrate this. That is, when the goal is to establish the external validity of a causal inference, then there is no reason to accept that internal validity must *always* precede external validity. A better rendering of their relationship, is that in most contexts, and especially in evaluations in the field, internal and external validity of causal inferences are co-occurring and are related or complementary.

Selection. Our argument rests on the foundational idea that selection plays a role in both internal and external validity bias. As described above, a principal source of $IVbias$ is

imbalance between conditions on factors affecting average performance; that is, confounded selection into conditions. However, selection is also an important source of $EVbias$, specifically from selection into the "generalized from" and "generalized to" samples, that produces imbalance on moderators of impact (Cole et al., 2010, Hotz et al., 2005, Jaciw, 2016[a]). Moderators of the treatment impact are characteristics of individuals or settings that interact with the treatment, thereby increasing or decreasing the magnitude of the effect. Imbalance on moderators between the "generalized from" and "generalized to" samples results in heterogeneity in impact between them, which limits the generalizability of the average causal impact. Selection resulting in $EVbias$ is described as due to confounded selection into locations (Hotz, Imbens and Mortimer, 2005).

Why is it important that selection plays a role in both types of bias? If selection underlies both $IVbias$ and $EVbias$, it unifies them in a way that undermines the prioritization of one over the other. There are two basic situations where we seek to draw a causal inference:

Case 1: If the study sample is the inference population, then internal validity is first and an RCT is the best study design for eliminating possible bias from selection into conditions. However, internal validity comes first only because external validity is not in question. In other words, internal validity is first only because it is the sole concern.

Case 2: If the goal is to evaluate program impact for a sample that is not identical to the study sample (it may be overlapping with or mutually exclusive to the study group) – that is, if the full study sample is not the inference population (or a random sample of the inference population) – then selection may be a source of $IVbias$ and $EVbias$, and we should consider how it can simultaneously lead to either form of bias.

$EVbias$ is immaterial to Case 1, therefore the question of the primacy of either form of validity applies only to Case 2. We consider these situations in the next section.

The Implication of Parity of $IVbias$ and $EVbias$ for Study Design

We noted above that for Case 1 – where the study sample is the inference population – an RCTs is an optimal design because it rules out biasing effects from confounded selection into conditions within the study sample. What is the optimal design choice for Case 2; that is, when bias in impact may happen from confounded selection into conditions or locales?

In what follows, we demonstrate that with Case 2 an RCT does not always produce the less-biased result. *In other words, under specific conditions, a QED is better for avoiding bias from selection.*

To illustrate the main argument of this work, we use graphical, algebraic and empirical arguments. We start with four graphical scenarios. We draw on examples from education.

Scenario 1.

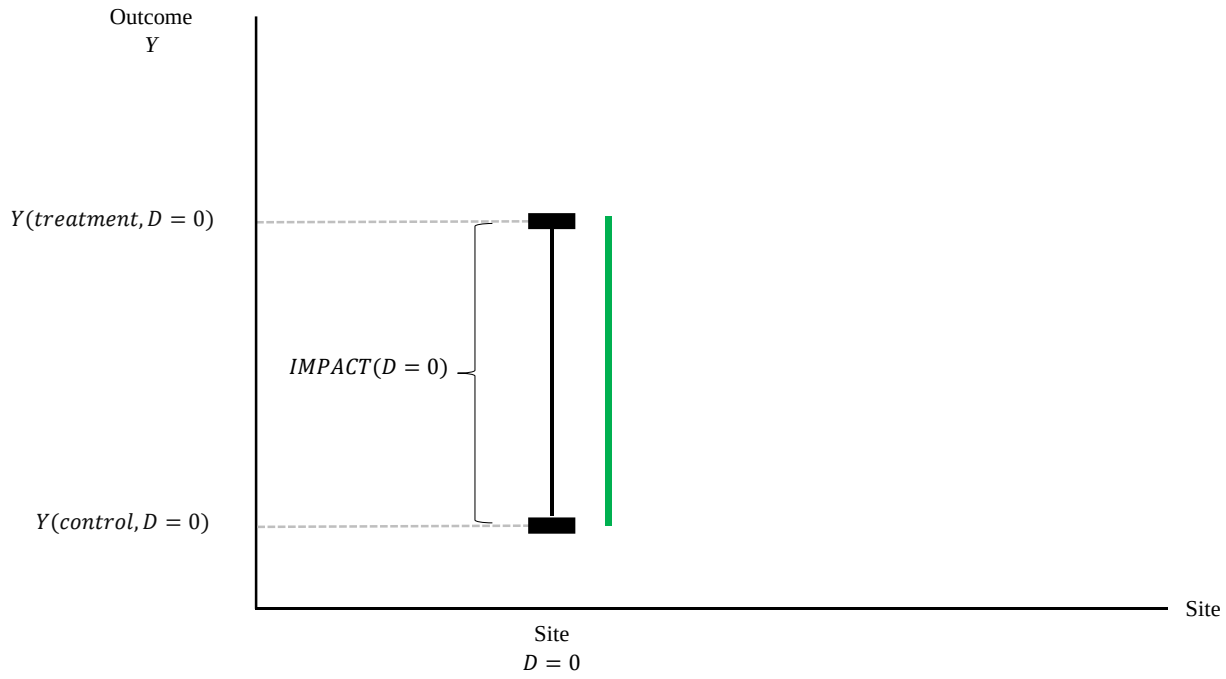
Referring to Figure 1, the goal is to evaluate the average program impact for the full sample at inference site $D = 0$. An experiment is conducted at a site $D = 0$. The study sample consists of the whole student population at $D = 0$ at the start of the study. External validity bias is not a concern, because the study sample is the inference sample (i.e., Case 1 in the previous section).

We express average impact at $D = 0$ as the difference in average outcome Y between individuals randomly assigned to treatment, and those randomly assigned to control[2]:

$$IMPACT(D = 0) = Y(Treatment, D = 0) - Y(Control, D = 0) \qquad (1)$$

The average impact quantity at $D = 0$ that is estimated through an RCT at that site is represented in terms of the length of the green line in Figure 1 ($IMPACT(D = 0)$). This is the benchmark "true" average impact for the site[3].

Figure 1. Average impact of $T = 1$ relative to $T = 0$ for inference site $D = 0$



[2] We adopt a basic notation to avoid excessive formalism. A more formal representation of a similar approach using the "potential outcomes" framework (Rubin, 1974) is given in Author, (XXXX).

[3] For the scenarios considered here, we focus on true values for impact and bias. We discuss the role of random sampling error later.

Scenario 2.

Referring to Figure 2, the goal is to evaluate the average program impact for the full sample at inference site $D = 0$. The whole student population at $D = 0$ is assigned to treatment T, which precludes an experiment at the inference site.

External validity bias is not a concern, because the study sample – all students at $D = 0$ – constitute the inference sample, (like Scenario 1, this scenario corresponds to Case 1 described earlier.)

Given that a counterfactual to average performance in the other conditions (i.e., without the program) is not available for $D = 0$, a comparison group is formed using individuals from one or more other sites ($D = 1$) who have not used the program. The average impact quantity at $D = 0$ based on this quasi-experimental (QED) comparison is expressed as follows:

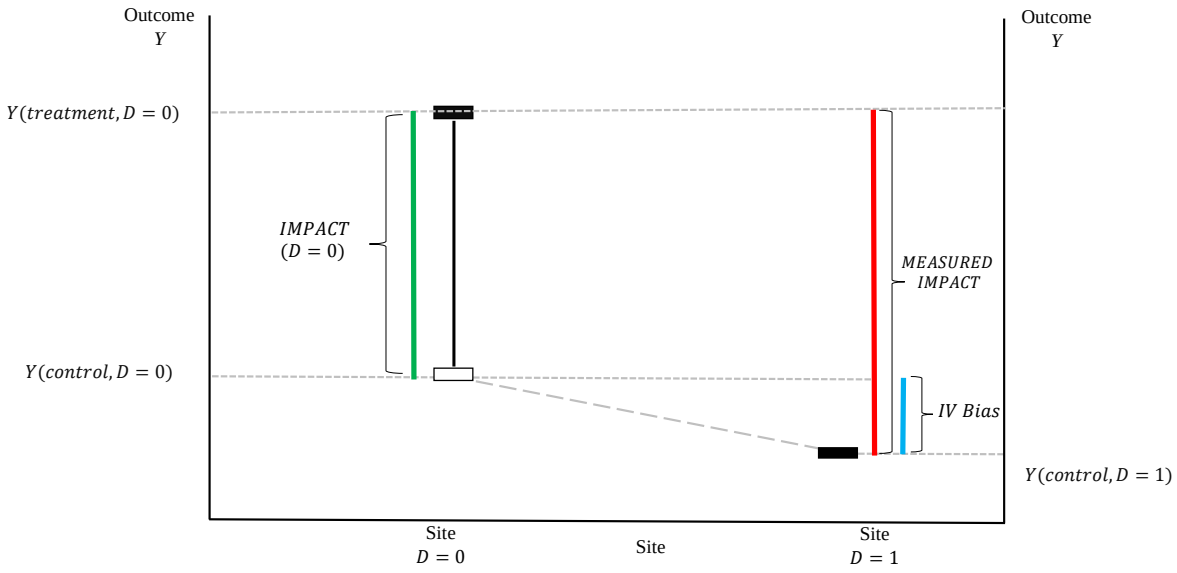
$$QED1(D = 0) = Y(Treatment, D = 0) - Y(Control, D = 1) \quad (2)$$

This is represented as the length of the red line (*MEASURED IMPACT*) in Figure 2. Bias in this causal quantity is the difference between the inferred and true (benchmark) impact quantities for inference site $D = 0$. For $QED1$ this is:

$$IVbias = Measured\ Impact - True\ Impact = QED1(D = 0) - IMPACT(D = 0)$$
$$= Y(Control, D = 0) - Y(Control, D = 1) \quad (3)$$

$IVbias$ is represented in Figure 2 as the length of the blue line, which is the difference in lengths of the red and green lines. This is the difference between the sites in their average performance in the absence of treatment.

Figure 2. Average impact inferred for site *D=0* through a comparison with controls at site *D=1*



Note: short horizontal bars representing average performance are black if the value is assumed measured, and empty (white) if the value is assumed unknown and therefore must be obtained from a comparison site.

Scenario 3:

The goal is to infer what the average program impact is for all students at site $D = 0$.

No students have received the program at that site, and an experiment has not been conducted there. This requires generalizing a value of program impact to the inference site $D = 0$ using information from another locale.

The impact finding from an uncompromised experiment at a comparison site $D = 1$ is used to infer impact at the site of interest, $D = 0$. That is, locale $D = 1$ supplies an estimate of the potential difference in performance between treatment and control for site $D = 0$[4].

$$RCT_{D1}(D = 0) = Y(Treatment, D = 1) - Y(Control, D = 1) \quad (4)$$

[4] $RCT_{D1}(D = 0)$ is experimental and identified for the study site, $D = 1$; however, it is not identified for the inference site, $D = 0$. The sample at $D = 1$ is effectively a non-experimental comparison group for generalizing impact to $D = 0$.

(We denote this "$RCT_{D1}(D = 0)$" to emphasize that the inferred impact is from an uncompromised experiment conducted at a comparison site $(D = 1)$ and applied to the inference site $(D = 0)$.) The impact inferred from $D = 1$ is represented as the red line in Figure 3.

External validity *is* a concern in this case. When applied to $D = 0$, the impact for $D = 1$ is subject to $EVbias$ from confounded selection into locations. Bias in $RCT_{D1}(D = 0)$ is represented as the difference between the inferred and benchmark (true) impact for $D = 0$:

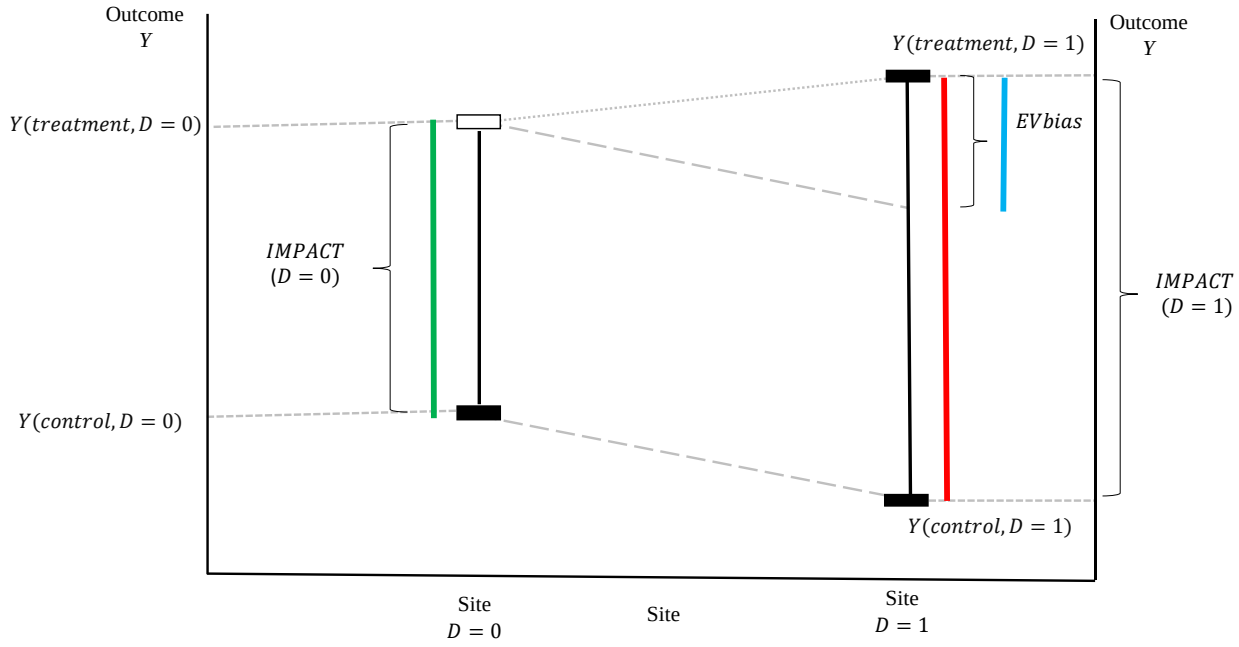
$$EVbias = RCT_{D1}(D = 0) - IMPACT(D = 0)$$

$$= Y(Treatment, D = 1) - Y(Control, D = 1)$$

$$-[Y(Treatment, D = 0) - Y(Control, D = 0)] \quad (5)$$

$EVbias$ is represented in Figure 3 as the length of the blue line, which is the difference in lengths of the red and green lines. This is the difference between the sites in average impact.

Figure 3. Average impact inferred for site *D=0* using the impact at site *D=1*



Note: short horizontal bars representing average performance are black if the value is assumed measured, and empty (white) if the value is assumed unknown and therefore must be obtained from a comparison site.

Scenario 4:

The situation is the same as in Scenario 3: the goal is to generalize impact to inference site $D = 0$ where no one at that site has been assigned to the program.

In this case, the treatment effect at $D = 0$ is inferred using the difference in average performance between individuals assigned to treatment $T = 1$ at $D = 1$ and those not receiving treatment, $T = 0$, which is everyone at the inference site $D = 0$. This contrasts with Scenario 3 because in this case we use information about performance at the inference site (i.e., in the absence of treatment) to arrive at a generalized causal impact quantity for that site[5].

The QED-based impact quantity is represented as the length of the red line in Figure 4. We represent this quantity as follows:

$$QED2(D = 0) = Y(Treatment, D = 1) - Y(Control, D = 0) \quad (6)$$

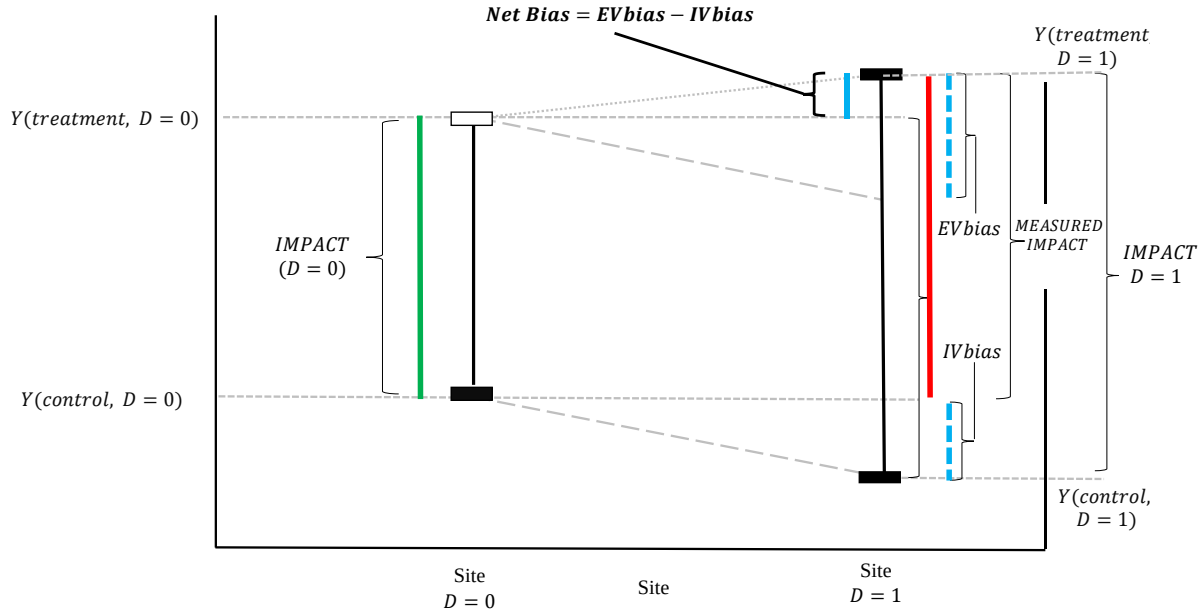
By how much is this quantity systematically different from the true impact at $D = 0$? Bias in $QED2$ is the difference between the inferred and benchmark (true) impact for $D = 0$ (we shortened "$Treatment$" to "T" and "$Control$" to "C" to allow a more compact representation):

$$\begin{aligned} Bias = {} & QED2(D = 0) - IMPACT(D = 0) \\ = {} & Y(T, D = 1) - Y(C, D = 0) - [Y(T, D = 0) - Y(C, D = 0)] \\ = {} & \{[Y(T, D = 1) - Y(C, D = 1)] + [Y(C, D = 1) - Y(C, D = 0)]\} \\ & - [Y(T, D = 0) - Y(C, D = 0)] \\ = {} & \{[Y(T, D = 1) - Y(C, D = 1)] - [Y(T, D = 0) - Y(C, D = 0)]\} \\ & - [Y(C, D = 0) - Y(C, D = 1)] \\ = {} & EVbias - IVbias \end{aligned} \quad (7)$$

[5] In our examples, we assume that at a site without an RCT everyone at a site either receives treatment or does not, and at a site with an RCT, everyone participates in the condition they are assigned to (everyone is a complier). This rules out additional selection effects within sites, which require more complex scenarios beyond the scope of this work.

Total (net) bias is represented in Figure 4 as the length of the red line minus the length of the green like. This is the same as the difference between $EVbias$ (the top blue dashed line) and $IVbias$ (the bottom blue dashed line).

Figure 4. Average impact inferred for site *D=0* through a comparison with treated at site *D=1*



Note: short horizontal bars representing average performance are black if the value is assumed measured, and empty (white) if the value is assumed unknown and therefore must be obtained from a comparison site.

Implications of the Four Scenarios for Bias in Generalized Causal Inferences

What do the results based on the four scenarios reveal? The goal is the same across all the scenarios: to arrive at an accurate value for average impact for the inference site $D = 0$. The information that is available and that gets used across the four scenarios is summarized in Table 1.

In Situation 1 an uncompromised experiment at the site yields the unbiased impact for the inference site, which is optimal. In the situations 2 – 4, incomplete information about

performance in one or both conditions at $D = 0$ requires using outcome data from the comparison site, $D = 1$, to infer impact for $D = 0$.

Table 1. A summary of main quantities for scenarios1 - 4

Scenario	Quantity used to infer impact at $D = 0$	Information missing at inference site ($D = 0$)	Information used from inference site ($D = 0$)	Information used from comparison site ($D = 1$)	Is External Validity a concern for causal inference at $D = 0$?	Susceptible to $EVbias$ or $IVbias$?
1	$IMPACT(D = 0) = Y(T, D = 0) - Y(C, D = 0)$	None	$Y(T, D = 0)$, $Y(C, D = 0)$	None	No	Neither
2	$QED1(D = 0) = Y(T, D = 0) - Y(C, D = 1)$	$Y(C, D = 0)$	$Y(T, D = 0)$	$Y(C, D = 1)$	No	$IVbias$
3	$RCT_{D1}(D = 0) = Y(T, D = 1) - Y(C, D = 1)$	$Y(T, D = 0)$	None	$Y(T, D = 1)$, $Y(C, D = 1)$,	Yes	$EVbias$
4	$QED2(D = 0) = Y(T, D = 1) - Y(C, D = 0)$	$Y(T, D = 0)$	$Y(C, D = 0)$	$Y(T, D = 1)$	Yes	$IVbias$, $EVbias$

In situation 2, a QED-based estimate that uses control outcomes from $D = 1$ potentially introduces *IVbias*. In both Situations 1 and 2, performance conditional on assignment to treatment is known for the inference site. The external validity of the treatment effect is not a concern, and the focus is on finding an accurate counterfactual to the treated group.

In contrast to this, in Scenarios 3 and 4, for the inference site, $D = 0$, we know only about performance in the absence of treatment. To infer impact there requires drawing on performance outcomes from a site where the treatment has been used ($D = 1$). This raises the question about the external validity of the inference, because performance under assignment to treatment must be generalized from someplace else. If we use the experiment-based result from $D = 1$ to infer impact at $D = 0$ (i.e., $RCT_{D1}(D = 0)$) the only source of bias is $EVbias$. If instead we infer impact to $D = 0$ by comparing performance for the treated group at $D = 1$ with performance of the non-treated group at $D = 0$, (i.e., using $QED2$) net bias is $EVbias - IVbias$.

These results raise the following question: *When external validity is a concern (Situations 3 and 4), which is the less-biased option for inferring impact for* $D = 0$*, an RCT-based result from* $D = 1$ *(i.e.,* $RCT_{D1}(D = 0)$*) in Situation 3 that is subject to* $EVbias$*, or a comparison group-based result that contrasts performance under treatment at* $D = 1$ *with performance in the absence of treatment at* $D = 0$ *(i.e.,* $QED2(D = 0)$ *in Situation 4, which is subject to net bias* $EVbias - IV\ bias$*)*? When external validity is the concern, we have to ask, when is magnitude of *net bias* from $QED2$ less than magnitude of bias for $RCT_{D1}(D = 0)$? In other words, under what conditions is it the case that:

$$\sqrt{(EVbias - IVbias)^2} < \sqrt{EVbias^2} \qquad (8)$$

When this inequality holds, an experimental impact finding from elsewhere (RCT_{D1}) is *less preferred* than one that uses a cross-site comparison ($QED2$). It may see counterintuitive that when generalization of a causal impact is the goal, a comparison-group-based result may be less biased than an experiment-based one, and therefore, preferable. We can make this idea more intuitive by considering that there is a tradeoff between RCT_{D1} and $QED2$. The former is experimental, but is from outside the inference sample, whereas the latter is non-experimental, but half the solution for the impact uses data from the inference site (i.e., performance in the absence of treatment), which is an unbiased solution for performance in just one condition (i.e., we have the control-half of an unbiased impact quantity for the inference site $D = 0$). There are pros and cons to each alternative.

Additional specification of conditions under which a comparison group design is preferable. We explore graphically the conditions under which the impact based the cross-site comparison ($QED2$) is less biased than the experiment-based result from the comparison site (RCT_{D1}). This condition is satisfied when the following relation holds.

$$(EVbias - IVbias)^2 < EVbias^2 \quad (9)$$

Representing the terms on a standard coordinate system ($IVbias = x, EVbias = y$) we have:

$$(y - x)^2 < y^2$$

$$\Leftrightarrow y^2 - 2xy + x^2 < y^2$$

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

$$\Leftrightarrow x^2 < 2xy \quad (10)$$

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

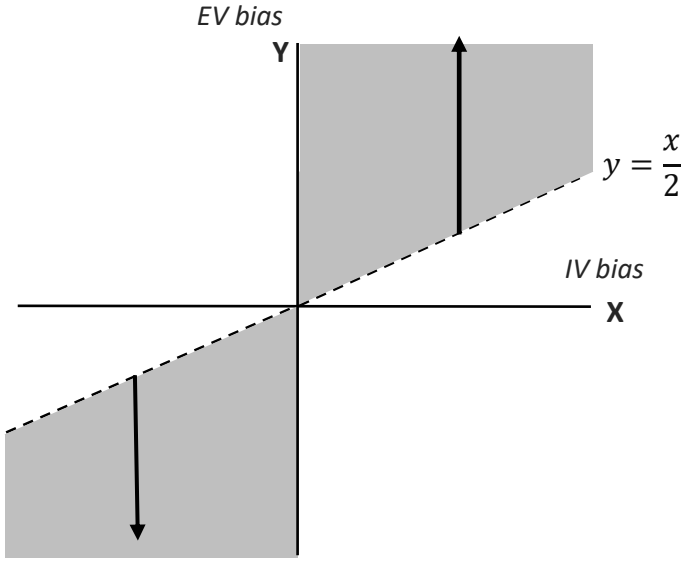
$$x/2 < y \quad (11)$$

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

$$x/2 > y \quad (12)$$

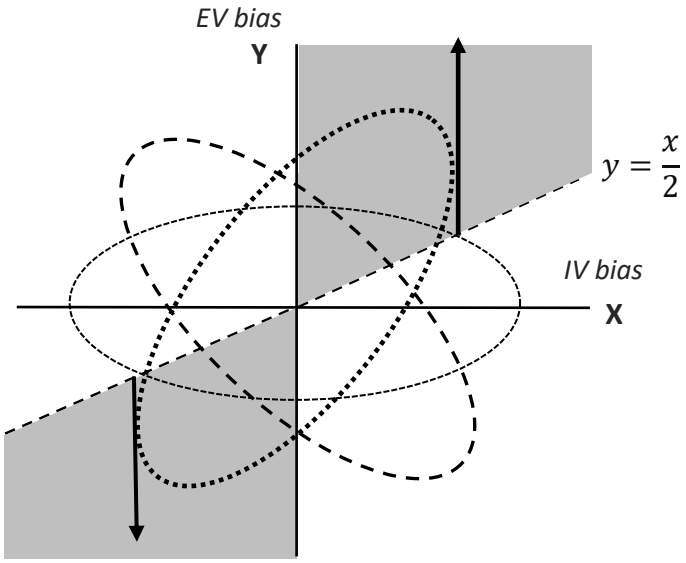
We observe that net bias in $QED2$ is less than for RCT_{D1} in the regions indicated by the arrows in Figure 5. (This result is confirmed in Appendix A).

Figure 5: Comparison of internal and external validity bias and space over which RCT_{D1} has less net bias than $QED2$.



The result raises the question of what are the empirical distributions of $EVbias$ and $IVbias$ across many studies. Figure 6 displays three hypothetical scenarios for the bivariate distribution, each demonstrating different coverage of the gray exclusion region.

Figure 6: Three hypothetical distributions for $EVbias$ and $IVbias$:



An empirical evaluation of levels of internal and external validity biases in generalized causal inferences based on RCT_{D1} and $QED2$

Thus far, we have derived expressions for bias in RCT_{D1} (i.e., when generalizing an RCT-based impact quantity from $D = 1$ to the inference site $D = 0$), and for $QED2$ (i.e., when using a comparison group based result that contrasts performance under treatment at $D = 1$ with performance in the absence of treatment at inference site $D = 0$). We have established the relationship between RCT_{D1} and $QED2$, and the conditions under which one exceeds the other in principle.

Next, we discuss the potential to empirically study the properties of RCT_{D1} and $QED2$ and their respective biases when used for causal generalization. Relevant questions include the following:

1. What are average magnitudes of bias in RCT_{D1} and $QED2$?

2. What is the joint distribution of bias in RCT_{D1} and $QED2$?
3. Does adjusting for effect of covariates reduce either form of bias?

Further questions are:

4. Do specific baseline variables that are imbalanced between sites affect both average achievement (confounders) and interact with treatment (moderators); If yes, does this imply that adjusting for their effects reduces both $EVbias$ and $IVbias$, and correspondingly bias in both RCT_{D1} and $QED2$?
5. How should we group studies when examining the distribution of the biases across them? For example, in education, when compiling results to produce empirical distributions of the type in Figure 6, should we aggregate results for all interventions and outcomes, or is it more informative to disaggregate studies based on the type of treatment (e.g., depending on whether curricula stress balanced literacy versus a phonics-first approach?), or outcome (e.g., math as opposed to English Language Arts) or the subgroups of sites or students in the studies?

Studying the questions empirically: An example of WSC

Background to WSC

In this section we demonstrate an approach to addressing the first three of the questions posed above (the latter two are left for future study). Specifically, we apply WSC methods to the problem of evaluating external validity bias in generalized causal inferences when using RCT_{D1} or $QED2$. This contrasts with the aims of standard WSC studies, which examine levels of and conditions for reducing internal validity bias in QEDs (i.e., $IVbias$ in comparison group designs of type $QED1$). Our extension of WSC methods to the question of external validity relates to other works by Hotz, et al., (2005), Hotz, et al., (2006) and Jaciw, (2016[a]).

We build on the standard rationale for and procedures used with WSC studies, which we briefly review here. Typically, in WSC studies, first, an experimental estimate of the causal impact of a program is obtained. This serves as the benchmark impact quantity to be replicated (Scenario 1 considered earlier). Second, a quasi-experimental estimate is constructed by replacing the outcome for the experimental control with one from a matched comparison group, (corresponding to $QED1$ in Scenario 2.) Third, $IVbias$ is estimated using the observed difference between the QED-based and experimental benchmark estimates. Fourth, different design and analytic strategies are applied to evaluate if the bias can be reduced or eliminated. If yes, then the QED-based estimate effectively replicates the experimental benchmark. (Of the more than 60 traditional WSC studies [Wong et al., 2018], notable ones include Lalonde (1986); Bloom et al., (2005), in jobs training, and Unlu et al., (2021), and Wilde et al., (2007), in education.[6])

In our application we use a version of WSC that evaluates bias in the context of cross-site comparisons in multisite trials (e.g., Bloom et al, 2005; Wilde et al., 2007). The approach assumes an uncompromised multisite trial involving *N* sites, in which each site supplies an unbiased impact estimate. That is, the trial yields *N* experimental benchmark values. In this version of WSC, a non-experimental estimate is constructed for a given site by replacing the outcomes in one condition – usually the control – with a result for the same condition from one or more of the other sites. This gives *N* estimates of both non-experimental estimate $QED1$ and corresponding $IVbias$. Following the standard WSC approach, the QED-based result for each site is compared to its experimental benchmark, and the difference between them is summarized across sites using average absolute bias. In this application, bias is attributable to selection into sites within the trial (Bloom et al., 2005).

[6] Recently Steiner et al., (2019) have advanced WSC methods through a framework that takes into account of all factors potentially affecting causal replication.

Extension of WSC to evaluating bias in RCT_{D1} and $QED2$

Our application evaluates bias in RCT_{D1} and $QED2$ relative to the experimental benchmark impact for each site. For RCT_{D1}, the impact estimate that is generalized to a given site out of *N* is the average of impacts across the remaining *N*-1 sites. (Each of the component *N*-1 impacts is experimental for the site from which it is obtained, but involves a non-experimental comparison when generalized to the remaining inference site(s).) The resulting estimate corresponds to RCT_{D1}, susceptible to $EVbias$. For $QED2$, the impact estimate that is generalized to a given site out of *N* is the difference between the average of performance of the treatment groups across all other (*N*-1) sites, and control performance at the inference site. This corresponds to $QED2$, which is susceptible to net bias $EVbias - IVbias$.

Each of the *N* sites yields an estimate for each of the two biases. Using the idea that $IVbias$ is distributed as "non-experimental mismatch error" (Bloom, 2005), Jaciw (2016[a]) and Jaciw et al (2021) show that the average level of bias across the sites – that is the average discrepancies of the QED estimates from corresponding benchmark impacts – can be summarized in terms of the cross-site variability in outcomes and impacts. In our application the average of magnitude of bias in RCT_{D1} is expressed as "Root Mean Squared Bias":

$$RMSB(RCT_{D1})/SD = \frac{1}{SD}\sqrt{Variance(Impact)} \quad (13)$$

(This is analogous to the expression in Equation 5, which is $EVbias$ associated with RCT_{D1} for the two-site case).

The average of magnitude of bias in $QED2$ is expressed as:

$$\frac{RMSB(QED2)}{SD}\frac{1}{SD}\sqrt{Variance(Impact) + Variance(Control) + 2Cov(Impact, Control))} \quad (14)$$

(This is analogous to Equation 7, which is $EVbias - IVbias$ associated with $QED2$ for the two-site case). "Control" denotes average site performance for the control group.

We standardize the values by dividing them by the standard deviation of the outcome variable (SD). This allows a comparison of $RMSB$ within and across studies and with the overall average impact for a study, as well as with meaningful benchmarks such as expected annual growth (Hill et al., 2008) which are often expressed in standardized effect size units.

Data and Results.

As a demonstration, and proof of concept, we summarize $RMSB$ from two multisite trials in education. The first is a multisite trial of Alabama Math Science and Technology Initiative (AMSTI) (Newman et al., 2012), the other of Tennessee Class Size Reduction Experiment (Project STAR) (Finn et al., 1990)[7]. STAR has been used in a prior WSC study by Wilde et al. (2007). Table 2 shows, in bold, estimates of $RMSB(RCT_{D1})/SD$ and $RMSB(QED2)/SD$ for two outcomes in STAR (reading and math) and one outcome in AMSTI (reading). We also display these results in Figure 7. (For completeness, in Table 2 we also include $RMSB(QED1)/SD$, with the result not bolded).

Table 2 displays results prior to and after adjusting for effects of site-level covariates (listed in Appendix B). (Detailed results for AMSTI and STAR reading have been provided elsewhere (Jaciw et al., 2021); Our application of WSC to the STAR math data is new, and we report detailed results in Supplement A.)

Table 2. Estimates of $RMSB$ for $QED1$, RCT_{D1} and $QED2$.

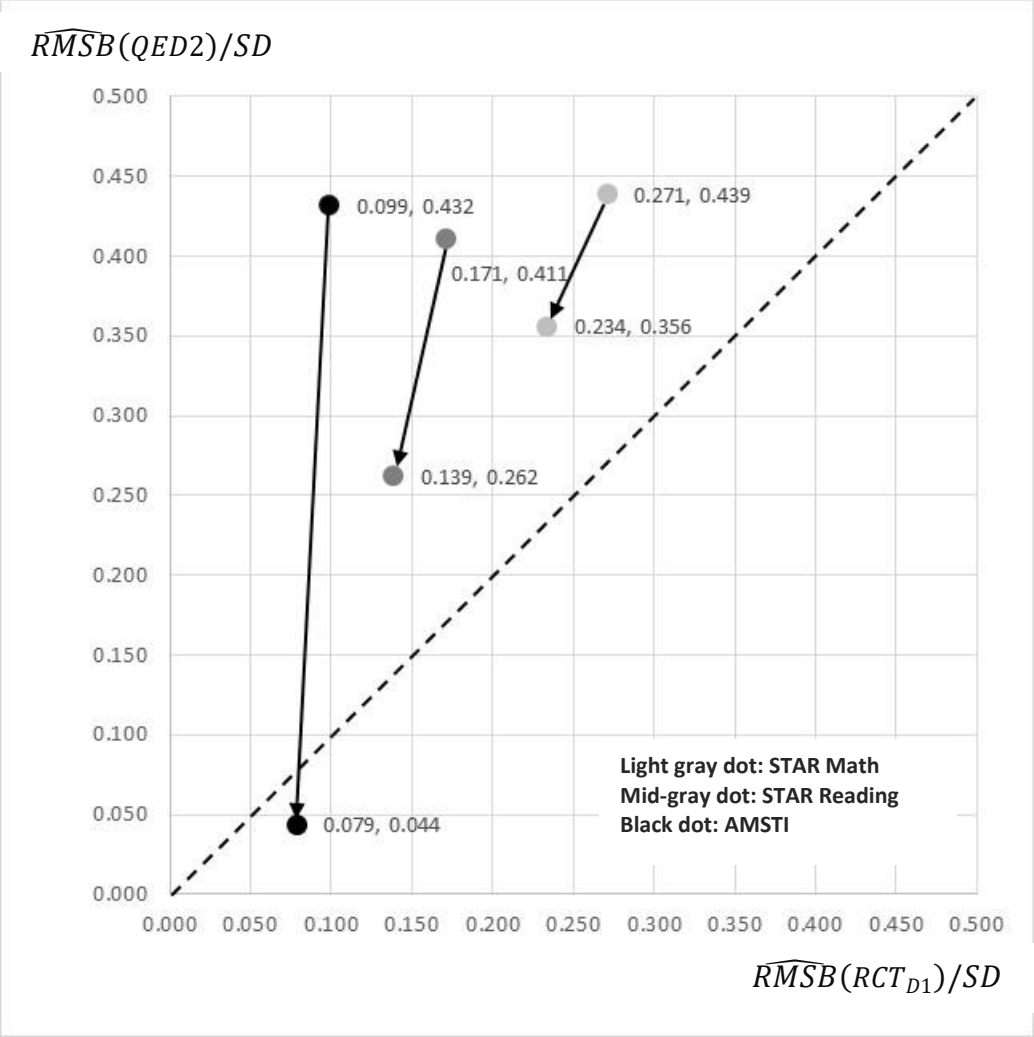
[7] For AMSTI we use results from analysis conducted during the original study.

	Site-level Covariate adjustments	$RMSB(QED1)/SD$	$\boldsymbol{RMSB(RCT_{D1})/SD}$	$\boldsymbol{RMSB(QED2)/SD}$
AMSTI Reading (N=40 sites, n=17922 students)	Without	0.420****	**0.099***	**0.432****
	With	0.074****	**0.079****	**0.044***
STAR Reading (N=73 sites, n=3452 students)	Without	0.414***	**0.171***	**0.411****
	With	0.186*	**0.139**	**0.262***
STAR Math (N=73 sites, n=3452 students)	Without	0.438****	**0.271****	**0.439 *****
	With	0.296****	**0.234***	**0.356****
	With	0.186*	**0.139**	**0.262**

****p<.01, ***p<.05, **p<.10, *p<.20

Note: Estimates of *RMSB* are expressed in standardized effect size units for the distribution of the outcome variable.

Figure 7. Estimates of $RMSB$ /SD for RCT_{D1} and $QED2$ prior to and with covariate adjustments



The observed values of $RMSB(RCT_{D1})/SD$ are .099 (p=.172), .171 (p=.173), and .271 (p=.067) for AMSTI(reading), STAR(reading) and STAR(Math), respectively, before covariate adjustments and .079 (p=.033), .139 (p=.252), and .234 (p=.106), after. For $RMSB(QED2)/SD$ they are .432 (p<.001), .411 (p=.010), and .439 (p=.013), respectively, before covariate adjustment, and .044 (p=.131), .262 (p=.163), and .356 (p=.057) with covariate adjustments.

Two noteworthy findings are: (1) the levels of bias are substantively important prior to and after covariate adjustments when compared to empirical benchmarks, for example, in terms of average annual expected growth in achievement (Hill et al., 2008), and achieved sample-wide average impacts for the studies (.07 SD for AMSTI and .24 for STAR); (2) importantly for the current work, RCT_{D1} has less bias than $QED2$ prior to covariate adjustments; however, the values of the two types of estimates are closer to each other after conditioning on effects of covariates, and $QED2$ is less than RCT_{D1} for AMSTI. (We include p values in reporting of results in Table 2, however, we recommend caution when using them as indicators of remaining bias because they reflect sampling error in the estimates, which reflects the number of degrees of freedom available in estimation. We advise examining p values with a view to the magnitudes of the estimates. A thorough review of metrics for evaluating levels of bias in the context of WSC studies is given in Wong et al. (2019)[8]).

Conclusions about the results

[8] In this work we have not elaborated on our approach to estimating the quantities in Table 2. Details of the approach are provided in (Author, XXXX). In brief, we used Hierarchical Linear Models (HLM) (Raudenbush and Bryk, 2002) with Random Intercept Random Coefficient (RIRC) models (Jaciw et al., 2021; Miratrix et al., 2021) to estimate the variance and covariance components in Equations 12 and 13. HLM has the advantage of removing withing-site between-student variability (at level-1) from estimates of variation in impact and achievement and covariance between impact and achievement across sites. In this work we also have not discussed the issue of exclusion of the inference site from the cross-site average against which it is compared. Systematically removing each site from the average, a "one-out" approach, is discussed in Jaciw (2016[a]) and Orr et al., (2019), and it is shown to have limited effect when the number of sites is sufficiently large (Author, XXXX) as is the case in this study.

Earlier in this work we established that bias in $QED2$ can be less than for RCT_{D1} in principle. Our empirical study above is a demonstration of an application of WSC methods to evaluate the question of the comparative magnitudes of bias for RCT_{D1} and $QED2$. With the six datapoints of Figure 7 we observe that bias in $QED2$ is larger than in RCT_{D1} except in one case (AMSTI with covariates); however, covariate adjustments result in convergence between RCT- and QED-based estimates in the accuracy of causal generalizations.

As with any application of WSC methodology, a given study yields a limited set of datapoints, and replication with multiple studies is necessary to address the question decisively. Therefore, more results from a collection of multisite trials are needed to determine levels of bias in RCT_{D1} and $QED2$ with greater confidence. We remind the reader that between 1985 and 2018, over 66 WSC studies have been conducted to evaluate levels of $IVbias$ (Wong et al., 2018), which has led to valuable empirical summaries and general conclusions about levels of and conditions for $IVbias$, as well as about design and analytic solutions for reducing this bias (Bloom et al., 2005; Cook et al., 2008; Glazerman et al., 2003; Wong et al., 2018). The current work sets in motion similar empirical research into conditions for $EVbias$, and the closely related question about when it is advisable to use RCT_{D1} compared to $QED2$. That is, the questions posed in this work may be answered through a similar cumulative acquisition of evidence.

Conclusions

In this work we have established that when generalization of causal effect estimates is the goal, internal and external validity must be considered together because neither takes precedence in a logical sense. The two forms of validity are unified because both are susceptible to threats from selection. Whether an RCT-based result from elsewhere (RCT_{D1}) is less biased than one

involving a comparison group design that involving outcomes from inference and comparison locations ($QED2$) depends on a selection mechanism that addresses the role of confounders and moderators of impact operating simultaneously. The question of whether the same variables act as confounders and moderators, and therefore account both for $EVbias$ and $IVbias$, is earmarked for future study.

We believe that the unification of $EVbias$ and $IVbias$ through effects of selection gives justification to reject the idea that RCTs are the gold standard, or provide gold standard results. An RCT is a gold standard only if the causal inference is intended for the sample itself (or with a probability sample, which is rare and hard to achieve in educational research [Tipton and Olsen, 2018]). When selection effects that influence both average achievement and impact heterogeneity are present, the idea of a gold standard study design loses its potency. A possible consequence of this work is the reduction of the authoritativeness of results based on RCTs and elevation of QEDs, and serious consideration of QEDs *as equally authoritative under certain conditions*. Understanding those conditions is an important aspiration.

Aside from running many WSC studies to build an empirical joint distribution of $IVbias$ and $EVbias$, can we say anything else about when, in principle, we might expect $QED2$ to be less biased than RCT_{D1}? The role of the covariance term in ($RMSB(QED2)$) is important in this respect. If negative, it reduces net bias. (The WSC result for AMSTI with covariate adjustments is one example.) Under what conditions do we expect a negative correlation between site deviations in average achievement and site deviations in average impact? This will happen when impact increases with decreasing incoming achievement within sites across which comparisons are made. For such interventions, generalizations that use comparison group designs ($QED2$) may be more-valid.

We end this work by emphasizing the need for WSC replication efforts to build cumulative knowledge of the relationship between $IVbias$ and $EVbias$. Standard WSC studies come with caveats and tests that apply here also. They include the following considerations:

a. Results from WSC studies may underestimate bias because they are limited to samples that have selected to join a study, such as a multisite trial (Jaciw, 2016[a], Orr et al., 2019).
b. Results, including bias, should be reported using policy-relevant metrics (Wong et al. [2018] reviews several metrics, with recent additional development in Orr et al., 2019).
c. Choice of confounders and moderators must be made smartly, and preferably reflects theory of selection (Smith et al., 2005); Each moderator uses up a degree of freedom, which may affect statistical power, and possibly conclusions about impact heterogeneity.
d. Incoming levels of the outcome measure are considered important determinants of bias in standard WSC studies (Glazerman et al., 2003; Unlu et al., 2021). It should not be automatically assumed that the pretest is as important in accounting for $EVbias$ – it is obvious that the pretest is predictive of average posttest performance, but it is not obvious that it routinely interacts with treatment to produce impact heterogeneity. That will depend on the program itself (as discussed in Jaciw [2016[a], 2016[b]]).
e. Sensitivity analyses are recommended. As evidence accumulates from WSC studies that jointly address $IVbias$ and $EVbias$, it is important to establish that the findings are not being driven through a particular or arbitrary approach to analysis (e.g., choice of estimator or estimand).
f. Related to the previous point, replication generally should take into account factors producing heterogeneity in average achievement and impact (Steiner et al., 2019). Adjusting for effects of confounders and moderators may be counterproductive if bias

reflects other sources of variation, ranging from researcher biases to changes of “effect generating processes” (Steiner et al., 2019) that moderators do not account for.

References:

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

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

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

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

Evidence for ESSA https://education.jhu.edu/2020/02/evidence-for-essa/

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.

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

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.

Hotz, V. J., Imbens, G. W., & Klerman, J. A. (2006). Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the California GAIN Program. *Journal of Labor Economics, 24*, 521–566.

Jaciw, A. P. (2016[a]). 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. (2016[b]). Applications of a within-study comparison approach for evaluating bias in generalized causal inferences from comparison group studies. *Evaluation Review, (40)*3, 241-276. Retrieved from 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

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

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

Newman, D., Finney, P.B., Bell, S., Turner, H., Jaciw, A.P., Zacamy, J.L., & Feagans Gould, L. (2012). Evaluation of the Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI). (NCEE 2012–4008). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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.

Patton, M. Q. (2015). *Qualitative Research and Evaluation*. Sage Publications: NY.

Privy Council Office / Impact and Innovation Unit. (2019). *Measuring impact by design: A guide to methods for impact measurement*. Author. ISBN: 978-0-660-29539-8.

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

Rubin, D. B., (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701.

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.

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. (2019). Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons. *Evaluation review*, *42*(2), 214-247.

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

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.

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

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. Verification of exclusion regions in Figure 6

Impact based the cross-site comparison ($QED2$) is less biased than the experiment-based result from the comparison site (RCT_{D1}) when the following relation holds.

$$(EVbias - IVbias)^2 < EVbias^2 \quad \text{(A1)}$$

Representing the terms on a standard coordinate system ($IVbias = x, EVbias = y$) we have:

$$(y - x)^2 < y^2 \quad \text{(A2)}$$

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

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

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

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

Case 1: $X = a > 0$

a. Show that when $\Delta > 0$, $y = \frac{a}{2} + \Delta$ satisfies the condition in (A4).

$$L.S. = (\tfrac{a}{2} + \Delta - a)^2 = (\Delta - \tfrac{a}{2})^2 = \left(\Delta^2 - a\Delta + \tfrac{a^2}{4}\right) < (\tfrac{a}{2} + \Delta)^2 = (\tfrac{a^2}{4} + a\Delta + \Delta^2) = R.S.$$

$$\Leftrightarrow \left(\Delta^2 - a\Delta + \frac{a^2}{4}\right) < (\frac{a^2}{4} + a\Delta + \Delta^2)$$

$$\Leftrightarrow -a\Delta < a\Delta$$

Because $a > 0$ and $\Delta > 0$, this is true.

b. Show that when $\Delta < 0$, $y = \frac{a}{2} + \Delta$ does not satisfy the condition in (A4).

The derivation is the same as above:

$-a\Delta < a\Delta$

Since $a > 0$ and $\Delta < 0$, this is never true.

Case 2: $X = a < 0$

a. Show that when $\Delta > 0$, $y = \frac{a}{2} + \Delta$ does not satisfy the condition in (A4).

The derivation is the same as above:

$-a\Delta < a\Delta$

Since $a < 0$ and $\Delta > 0$, this is never true.

b. Show that when $\Delta < 0$, $y = \frac{a}{2} + \Delta$ satisfies the condition in (A4).

The derivation is the same as above

$-a\Delta < a\Delta$

Since $a < 0$ and $\Delta < 0$, this is true.

Appendix B. Covariates used in analysis

STAR Experiment

Covariates at the school level are school averages of uncentered student-level covariates (gender, eligibility for Free or Reduced Price Lunch, minority [non-White] status, the years of teaching experience of a student's teacher, whether the student's teacher holds a Master's degree or higher, and end of kindergarten scores on tests of math and reading) and variables indicating school urbanicity (whether a school is inner-city, suburban, rural or urban.)

AMSTI Experiment

Covariates at the school level are baseline achievement in reading, proportion male, proportion eligible for Free or Reduced Price Lunch, proportion of students who are non-White, and proportion of English Learners.

Supplement A: More detailed results for analysis of STAR reading outcomes in second grade

A1. Variance components estimates prior to covariate adjustments (J=73 sites (schools), N=314 teachers, n=3,452 students).

Covariance Parameter Estimates					
Covariance Parameter	**Level**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
$Variance(Control)$	*School (Site)*	349.96	77.719	4.5	<.0001
$Cov(Impact, Control)$	*School (Site)*	-32.237	52.4885	-0.61	0.5391
$Variance(Impact)$	*School (Site)*	59.7474	63.2987	0.94	0.1726
$Variance$	*Teacher nested in school*	148.21	30.1673	4.91	<.0001
$Variance$	*Student nested in teacher*	1448.83	36.5533	39.64	<.0001

A2. Variance components estimates with covariate adjustments (J=73 sites (schools), N=314 teachers, n=3,452 students).

Covariance Parameter Estimates					
Covariance Parameter	**Level**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
$Variance(Control)$	*School (Site)*	70.7429	30.7587	2.3	0.0107
$Cov(Impact, Control)$	*School (Site)*	15.1589	30.1399	0.5	0.615
$Variance(Impact)$	*School (Site)*	39.2772	58.8632	0.67	0.2523
$Variance$	*Teacher nested in school*	147.37	29.6705	4.97	<.0001
$Variance$	*Student nested in teacher*	1448.36	36.5277	39.65	<.0001

Note. Covariates at the school level are school averages of student-level covariates (gender, eligibility for Free or Reduced Price Lunch, minority [non-White] status, the years of teaching experience of a student's teacher, whether the student's teacher holds a Master's degree or higher, and end of kindergarten scores on tests of math and reading) and variables indicating school urbanicity (whether a school is inner-city, suburban, rural or urban.)