An Applied Researcher’s Guide to Estimating Effects From Multisite Individually Randomized Trials: Estimands, Estimators, and Estimates

An Applied Researcher's Guide to Estimating Effects from Multisite Individually Randomized Trials: Estimands, Estimators, and Estimates

Luke W. Miratrix*
lmiratrix@g.harvard.edu
Larsen 603
Harvard School of Education
14 Appian Way
Cambridge MA 02138
510-735-7635

Michael J. Weiss*
michael.weiss@mdrc.org
MDRC
200 Vesey Street (23rd floor)
New York, NY 10281
212-340-8651

Brit Henderson
brit.henderson@mdrc.org
212-340-8892

*Corresponding authors and co-lead authors.

Researchers face many choices when conducting large-scale multisite individually randomized control trials. One of the most common quantities of interest in multisite RCTs is the overall average effect. Even this quantity is non-trivial to define and estimate. The researcher can target the average effect across individuals or sites. Furthermore, the researcher can target the effect for the experimental sample or a larger population. If treatment effects vary across sites, these estimands can differ. Once an estimand is selected, an estimator must be chosen. Standard estimators, such as fixed-effects regression, can be biased. We describe 15 estimators, consider which estimands they are appropriate for, and discuss their properties in the face of cross-site effect heterogeneity. Using data from 12 large multisite RCTs, we estimate the effect (and standard error) using each estimator and compare the results. We assess the extent that these decisions matter in practice and provide guidance for applied researchers.

ACKNOWLEDGEMENTS
We would like to thank Natalya Verbitsky-Savitz, Steven Raudenbush, Daniel Schwartz, and Howard Bloom, all of whom contributed to early discussions about this work.

Introduction

A postsecondary "Learning Community" is a semester-long program where students are grouped into cohorts that co-enroll in two or more courses to help them succeed. But do learning communities work? To find out, researchers randomized students into receiving this intervention or not, at 11 different- college campuses. Such *multisite experiments* can help us learn about a program or policy's effectiveness. Researchers typically characterize such effectiveness with estimates of, for example, the average treatment effect (ATE) for a specified population of people. There are a variety of such summary quantities, called estimands, that one might wish to estimate, with ATEs being the most common (see, for example Raudenbush and Bloom, 2015; Schochet, 2016, or Raudenbush and Schwartz, 2020).

But even the simple concept of an ATE has important nuances when faced with a multisite experiment. For example, do we care about the impact of learning communities for the average *college student* or the average *college*? These nuances, which can have important implications for the estimation and presentation of study results, is what we examine in this work. In particular, we examine four different estimands that are all versions of the ATE, and which vary along two dimensions:

1. *Estimating for individuals vs. sites*: Whether the target of inference is the effect for the average individual or the average effect for the average site.
2. *Making inference to a finite vs. super population*: Whether the target of inference is the average effect for the individuals and sites in the evaluation sample (i.e., those in the study itself) or the average effect for individuals and sites from a larger population.

Combined, the above gives four possible estimands—finite population individual, finite population site, super population individual, and super population site.[1] All four of these are ATEs. They differ in how the averaging is done and how uncertainty is assessed.

To make matters more complicated, there are also many different *estimators*, the methods used to calculate an intervention effect estimate, used in practice. Confusingly, some estimators do not obviously connect to an individual average or site average effect. It can even be unclear whether an estimator is finite sample or super population targeting. If treatment effects vary across sites, then these various estimators could yield different overall ATE estimates and estimated standard errors, potentially changing the interpretation of a study's findings. This concern may be more than theoretical—*applied* research demonstrates that in some instances there is substantial cross-site variation in treatment effects (Weiss et al., 2017).

Some estimators have a potential bias-precision tradeoff in the face of cross-site variation in treatment effects. For example, fixed effect regression (a commonly used approach) is best understood as a biased estimator targeting a person-weighted estimand, where the bias is in exchange for smaller standard errors. For multilevel modeling (aka, random effects or mixed modeling), Raudenbush and Bloom (2015) explore their *theoretical* properties, showing that, in the face of cross-site variation in treatment effects, these estimators can be biased. Here again the bias comes from a tradeoff to ideally achieve precision gains. Given this tension, researchers must decide if they are willing to risk possible bias for improved precision in their estimator selection.

[1] There are actually multiple possible super population models; to streamline our discussion, we refer to them as a general class and note salient differences when they arise. Furthermore, as we will see, there can be additional variation within each of the four classes due to concerns such as nonresponse or nonrandom sampling.

The main motivations behind this paper are to unpack the above themes and tensions and to explore their implications in practice. We clarify that researchers conducting multisite trials must make two key decisions: the choice of estimand and the choice of estimator. To help ensure these decisions are made intentionally, rather than implicitly and perhaps inadvertently, we explicitly connect estimators to their appropriate estimands. To achieve this, we first unpack our two axes for defining an ATE estimand, emphasizing the importance of defining an evaluation's estimand(s) and discussing reasons researchers might choose each one. We then identify, for a series of estimators advocated for in the literature and/or commonly used in practice, which estimand is the most logical inferential target, what conditions could lead to bias, and what factors impact precision. We finally provide empirical evidence on the extent that these decisions matter in practice so that appropriate time and energy can be focused on consequential decisions.

To keep the foundational themes of our investigation maximally salient, we focus on Intent to Treat (ITT) effects and follow the classic superpopulation sampling frames implied by the empirical methods most commonly used in practice. These frames do bring several strong assumptions, such as individuals within sites being representative of their sites, and the sites being representative of their superpopulation. Attrition or, most importantly, the convenience sampling of sites, can easily undermine such claims. To address such concerns one would need to use methods from, e.g., the generalizability literature. While we do not engage with these tools in this work, we do discuss these connections further in the Section titled *Further Perils of the Superpopulation Frameworks*, below.

To inspect how relevant the choice of estimand and estimator are in practice, we turn to a collection of 12 large multisite RCTs. For each trial we estimate the average effect, β, and its associated standard error, $SE(\beta)$, using 15 different estimators. We then compare the resulting effect estimates along with their estimated standard errors. The goal is to inform researchers regarding the extent that the choice of estimand and estimator "matter." In particular, we seek to address the following research questions:

(1) To what extent can the choice of estimator of the overall average treatment effect (β) result in a different impact estimate?

   a. To what extent do effect estimates vary across all estimators?

   b. For a given estimand, to what extent do effect estimates vary across reasonable estimators?

(2) To what extent can the choice of estimator of the standard error of the overall average treatment effect ($\widehat{SE}(\beta)$) result in a different estimated standard error?

   a. To what extent does the estimated standard error (SE) vary across all estimators?

   b. For a given estimand, to what extent does the estimated standard error (SE) vary across reasonable estimators?

   c. To what extent are differences in standard error estimates due to estimation error of the standard error or due to the estimators actually being differently precise?

Finally, estimators that plausibly target the same estimand may still yield differing estimates because some estimators are potentially biased to increase precision. This motivates our third question:

(3) How do the theoretically possible bias-precision tradeoffs play out in real data?

This work offers several important contributions to the literature. First, in a single document, we provide an overview of the estimands and estimators commonly used in multisite randomized controlled trials. Next, we apply these estimators to a series of large-scale multisite trials so that researchers can appreciate whether and how much the choice of estimand and estimator does (or does not) matter. We supplement this empirical evaluation with two substantial supplementary documents: a technical appendix (Online Appendix A) that gives more detail for the different estimators, and an extensive multifactor simulation study (Online Appendix B) that further illustrates and verifies the trends in the empirical findings. We have also posted the complete code for this simulation. Understanding the practical implications of these decisions can help inform researchers as they plan studies, analyze data, and interpret results from large scale multisite trials.

Overall, we find that estimand selection indeed matters. Interest in a super population inference or site averaged effect often comes with substantial cost in terms of precision. Appreciating this from the outset may help ensure studies that target these estimands are adequately powered. We also find that site- and person-average estimators occasionally give different effect estimates, although usually not by a lot.

With one exception, once an estimand is selected, choice of estimator, among estimators that align with the estimand, tends to be inconsequential (see remarks throughout this manuscript for caveats). The exception is the super population site estimand. Here, the two primary estimators can yield different treatment effect estimates and frequently result in different estimated standard errors as well; this is partially driven by the bias-precision tradeoff.

The Estimands

Within an evaluation sample, consider people nested within site, with person i in site j. Throughout this paper we use the potential outcomes framework (for an overview see Imbens and Rubin, 2015 or Rosenbaum, 2010), which allows for precise description and discussion of the various estimands of interest. Under this framework each unit (e.g., person) has two potential outcomes, $Y_{ij}(1)$ for if it gets treated, and $Y_{ij}(0)$ for if it does not. If we treat, we observe $Y_{ij}(1)$

and if we do not treat, we observe $Y_{ij}(0)$. The *individual* intention-to-treat (ITT) effect[2] for person i at site j is then $B_{ij} = Y_{ij}(1) - Y_{ij}(0)$.[3] It follows that B_j is the average of the individual effects for all persons in the study at site j:

$$B_j = \frac{1}{N_j}\sum_{i=1}^{N_j} B_{ij}, \qquad (1)$$

where N_j is the number of persons in the study at site j. Each site has its own B_j, and they can differ. We are interested in summary measures of these B_j across sites.

Individual vs. Site Average Estimands

We start by focusing on the individuals and sites in the evaluation sample only, extending to a super population afterwards. Given a collection of sites with different sizes, we might ask what the person-average effect is (i.e., weighting each person equally), or what is the site-average effect (i.e., weighting each site equally, regardless of its size). These quantities can be different. For example, if larger sites have larger treatment effects, then the person-average effect will be larger than the site-average effect.

We therefore have two estimands, one for the average effect across individuals, and one for the average of the site-average effects. The average effect across individuals in the evaluation sample is defined as:

$$\beta_{FP-Persons} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j} B_{ij} = \sum_{j=1}^{J}\frac{N_j}{N}B_j \qquad \text{with } N = \sum_{j=1}^{J} N_j \qquad (2)$$

[2] In this paper we focus on the intention-to-treat effect. For ease of exposition we refer to units as “treated” or “not treated” rather than “offered a treatment” or “not offered a treatment.”

[3] For this model to be sensible, we must assume no spillover, and a well-defined treatment. I.e., we cannot have the outcome of person i changing if person k gets a different treatment. This is often called the Stable Unit Treatment Value Assumption, or SUTVA. See Rosenbaum (2010) for further discussion and a good overview.

where J is the total number of sites in the evaluation sample. This is the average of the individual casual effects across all individuals in the sample.

The average effect across sites in the evaluation sample is defined as:

$$\beta_{FP-Sites} = \frac{1}{J}\sum_{j=1}^{J} B_j \ . \tag{3}$$

In order for $\beta_{FP-Persons}$ and $\beta_{FP-Sites}$ to differ, the N_j and B_j must both vary and be related.

Finite vs. Super population Estimands

A second consideration is whether to target the average effect for the units in the experimental sample, as we did above, or to view the experimental units as a sample from a larger population and target the average effect in that larger population. These do not have to be the same quantity. The estimands in the prior section are **finite-population (FP)** estimands, residing in the finite population framework.

Under the **super population (SP)** framework, by contrast, we assume that the study sites are a random sample of a much larger population of sites. We discuss alternate superpopulation models below, but what follows generally holds in substance. Index our sites in the super population with $1,2,\dots,J^*$ with J^* being the total number of sites in the super population. J^* is assumed to be very large, $J^* \gg J$. We assume that when we sample a site we sample all units at that site. Under this regime we have two estimands analogous to those for the finite-population. First, there is the average effect across all individuals in the super population:

$$\beta_{SP-persons} = \sum_{j=1}^{J^*} \frac{N_j}{N^*} B_j \text{ with } N^* = \sum_{j=1}^{J^*} N_j. \tag{4}$$

The average effect across sites in the super population, similarly, is:

$$\beta_{SP-sites} = \frac{1}{J^*}\sum_{j=1}^{J^*} B_j. \tag{5}$$

When there is cross-site treatment effect variation, a given random sample of sites may have a different average treatment effect than the population. In this case the finite and superpopulation estimands will not be the same. For more on how they are connected, please see Appendix A.

A technical (ignorable) remark on superpopulation frameworks. There are actually many super population frameworks that are possible, and the different frameworks tie to different standard error estimates, as we will see below. First, researchers often assume some hypothetical *infinite superpopulation*, where the sites are sampled i.i.d. from some distribution. We instead use a large but finite superpopulation to avoid the use of integrals in the definition of the estimands, and assume J^* is substantially larger than J so that any finite population correction would be negligible. For more on finite population corrections, as well as the tensions between superpopulation and finite population estimation, and the different formulations of superpopulations, see, e.g., Lohr (2019), Sarndal et al. (2003), or Cochran (1977).

Second, we generally consider that all relevant individuals within each site are sampled when we sample the site. For example, for a school-level tutoring intervention targeting low-achieving students, we would take the students in the evaluation sample at that school as the population of students targeted. An alternate super population model would have the observed sample of individuals within a site coming from a second sampling step of individuals from a super populations of individuals at that site. To directly extend to this context, we would need to assume unbiased sampling of individual within site. For the person-weighted results, we would generally also need to assume that the sample size at each site is proportional to the site size.[4]

[4] This is necessary to prevent, e.g., a case where sites with generally low impacts have more people sampled, leading to a higher proportion of low-impact units in the overall sample than in the population. Site population numbers

Alternatively, we can assume individuals are randomly sampled across a fixed collection of sites, with site as a categorical covariate of these individuals; this is a classic econometric view, and aligns more with finite population estimation. See both Schochet (2016) and Pashley and Miratrix (2020) for further discussion of sampling regimes with regards to the blocked or multisite experiments perspective. Pashley and Miratrix (2020) specifically compares finite vs. superpopulation settings.

Estimand Selection

Taking the individual vs. site-average and finite- vs. super-population as two dimensions, we are left with four common estimands of interest, laid out in Figure 1. When faced with these four estimands, the question naturally arises of which one to select for a given evaluation. Ideally, we would like to know the *true* effects for all individuals in an evaluation, B_{ij}, or the *true* site average effects, B_j, rendering such a decision unnecessary. In practice, however, there is typically no way to obtain these quantities with any degree of accuracy, and thus we turn to summary measures. We generally advocate selecting based on context and the goal of the study, acknowledging that the practical limitations of a real-world evaluation may also play into decision-making.

could instead be included as weights, rather than site sample size numbers, to correct for this. Considering such situation is beyond the scope of this paper.

Figure 1. The Four Common Estimands.

Target Population

		Finite	Super
Site-weighting	Person	$\beta_{FP-persons} = \sum_{j=1}^{J} \frac{N_j}{N} B_j$	$\beta_{SP-persons} = \sum_{j=1}^{J^*} \frac{N_j}{N^*} B_j$
	Site	$\beta_{FP-sites} = \sum_{j=1}^{J} \frac{1}{J} B_j$	$\beta_{SP-sites} = \sum_{j=1}^{J^*} \frac{1}{J^*} B_j$

Person or Site. From a utilitarian perspective, e.g., where the benefit of policy is thought of in terms of total individual impact, the estimands that target the effect for the average person make the most sense since they consider the impact of all persons equally. From this perspective, in an evaluation with one or a few sites that are exceptionally large, these outlier sites should drive the overall findings since they represent a large proportion of persons in the population.

In contrast, a site-level decision-maker (e.g., a school principal or district superintendent) *may* be most interested in the effect for the average site. In the absence of better information, this decision-maker may find it most useful to know how effective a program was at the average site, so that sites are given equal weight regardless of their size. (A site-level decision-maker often would prefer to know the effect of an intervention in those sites that are most like theirs. While such site-level subgroup analyses are in principle more informative; such analyses are frequently underpowered, which can bring us back to the effect for the average site.) This is especially true if it were thought that site impacts were especially variable due to, for example, variation in implementation.

It is noteworthy that when intervention effects vary (across individuals and/or sites), an overall average impact, whether person or site, could be misleading. In this case, studying the distribution of treatment effects defined over sites could be quite valuable, and can be done for large-scale multisite trials. Here the overall site ATE would be important as an indicator of the center of that distribution (Raudenbush and Bloom, 2015).

Finite or Super population? There are many reasons researchers design studies with the goal of making a finite population inference – that is, focusing on the experimental units at hand. For example, in efficacy trials we are often interested in knowing whether an intervention *can* work in a site or set of sites, as a proof of potential. Relatedly, we may be interested in an intervention's effects in less than ideal circumstances, to assess its resilience in a worst-case test. We may study an intervention's effects in a sample of sites that is purposefully selected to include the most plausible moderators. Testing for theorized moderation in this way may facilitate causal generalization, even if it is not "backed by a statistical logic that justifies formal generalizations" (Shadish, Cook, Campbell, 2002, p. 24).

In some instances, the sample at hand may even be the immediate and only inferential target. For example, for a social impact bond evaluation[5] (e.g., Parsons et al., 2016) where an investor will be paid based on the impact a program had on service recipients, the finite inference approach seems most appropriate. Similarly, if an organization is evaluating itself and includes nearly all its sites in their evaluation, there may be little interest beyond the people and sites in the study.

On the other hand, the super population framework, i.e., considering the evaluation sample of sites a representation of a larger whole, is often the preferred target of inference. Given

[5] From Parsons et al. (2016): "A social impact bond is an innovative form of pay-for-success contracting that leverages private funding to finance public services. In a social impact bond, private investors fund an intervention through an intermediary organization—and the government repays the funder only if the program achieves certain goals, which are specified at the outset of the initiative and assessed by an independent evaluator."

the time and expense associated with conducting most large-scale multisite trials, it is rare that the true goal of a study is to estimate the average effects of a program for a specific set of people and sites at a point in time. Rather, generalizing beyond is inherent to the process of policy formulation. Such generalization typically involves an argument that the results of the study carry information about what would plausibly happen if this intervention were tried elsewhere. The theoretical justification is often that the settings, participants, and historical period of the study represent conditions that are plausibly of broad interest. Under these conditions, we want our standard errors to be big enough to reflect some uncertainty about the proposition regarding how the results would apply (on average) elsewhere.

That said, multisite trials are rarely conducted on a *random* sample of sites (or individuals) from a larger population (Olsen, Orr, Bell, & Stuart, 2013). (There are exceptions, such as the National Growth Mindset evaluation (Yeager et al., 2017), explicitly designed to be representative although even here nonresponse makes this connection less immediate.) When the sampling process is not probabilistic, classic super population inference involves generalizing to a more vaguely defined set of units from which the study sample could reasonably have been randomly drawn, rather than to a super population that can be enumerated and described with well-defined characteristics. (We discuss repairing this using tools from the generalization literature below).)

The question of finite or super population can therefore come down to a preference for using logical arguments, with little assessment of uncertainty, to generalize beyond a study sample (as would be required under the finite sample) vs. using statistical inferences to acknowledge, even if imperfectly, some of the uncertainty when generalizing to a difficult to define population.

Pragmatically, there are trade-offs in the precision of the estimators of the four estimands we have discussed. If sites are radically different sizes, then the estimated effect for the average site will be much less stable than the estimated effect for the average person.[6] Additionally, if intervention effects vary across sites, then finite sample inference will be more precise, at the cost of any ability to generalize statistically. This tension is particularly of concern in trials with few sites; in these trials our sample of sites may easily not be representative of the super population (if site average impacts varied substantially) due to an unlucky draw, and this uncertainty would have to be included in the overall uncertainty estimate of a superpopulation average impact. In fact, for very few sites it may not be feasible to generalize what may already be a low-powered experiment due to this additional uncertainty. Assessing the relative costs of these decisions is one purpose of our empirical investigation.

Further Perils of the Superpopulation Frameworks

As discussed above, in current impact evaluation practice the superpopulation framework is usually left implicitly defined and there is no effort to adjust for the impact of attrition on treatment variation or nonrandom selection of sites. But what does one do when the sites are not sampled at random, such as with convenience sampling? Or what if there is attrition related to responsiveness to treatment? In this section (entirely optional reading for those interested) we discuss how these concerns tie in with the above themes, and also a bit on how they could be addressed.

First, if the treatment effect in the sample of sites is systematically different than the treatment effect in the target population, then the evaluation sample, without adjustment, does

[6] Recall we assume each site's evaluation sample constitutes the entire site.

not represent the larger population. This context directly links to the generalizability literature, which proposes estimating the sampling mechanism of how sites came to be evaluated based on observed covariates in order to reweight the sample to match baseline characteristics of a specified target population to make the sample representative of a larger whole, in terms of observed characteristics. For two overviews of the generalization approach see Tipton & Olson (2018) or Kern et al. (2016). To generalize, one typically fits a propensity score model of being in the sample vs. not as a first step. Then one reweights the sample in an attempt to make it represent the target population; by contrast, the methods we examine in this work implicitly assume the representativeness of the sample.

Unfortunately, when weights are included in the subsequent inference, they tend to increase uncertainty even beyond what would naturally occur with simple random sampling. See Miratrix et al. (2018) for an examination of this, and specific discussion of sampling weights, in the case of a single site. We therefore view the differences between the estimation approaches we investigate in this work as, roughly speaking, providing a lower bound on how much uncertainty estimates can change moving from a finite to superpopulation framing. We do not focus on any bias due to systematic sample nonrepresentativeness.

Multisite trials have further complications: within sites, the number and characteristics of a site's evaluation sample (i.e., those units used in estimation) may not correspond to the number or characteristics of the site's inferential target population. Within-site nonrepresentativeness could be due to the sampling procedures used to select individuals for a study (e.g., if they are not probabilistic) or attrition (differential or otherwise) due to, e.g., survey nonresponse. If the evaluation sample at a site has a different response to treatment than the full population of interest at the site, then the true average impact for the evaluation sample may not be representative of the true average impact for the target population at that site. In this case individuals within each site would require associated sampling weights that would have to be propagated throughout the estimation process.

In the case of differences only in relative numbers of individuals, site-specific estimands and estimates could be reweighted to correspond to the *target* site size rather than the *realized* size of the evaluation sample for the person-weighted estimands. This repair is most directly seen with design-based estimators, where weights are explicitly calculated and used in the estimation process. See Schochet (2016) for more on using general weights in estimation.

The Estimators of $\boldsymbol{\beta}$

Once an estimand is identified, a researcher must decide how to estimate it, given the observed data, and obtain uncertainty estimates (standard errors) for those estimates. We call the estimators "effect estimators" and the uncertainty estimators "standard error estimators." In this section, we systematically go through three general classes of effect estimators, discussing variants of implementing each. We attempt to identify what weighting each estimator appears to be targeting (i.e., site vs. individual effects), and how we might interpret them in a finite- or super- population context.[7] To simplify exposition and formulas, we do not include covariates (e.g., pretests and the demographic covariates) in our discussion or analyses, but the overall intuition carries over.

Our estimators are all built on unadjusted ITT effect estimates for the individual sites:

$$\hat{B}_j = \hat{Y}_j(1) - \hat{Y}_j(0) \qquad (7)$$

[7] In our technical appendix we elaborate on these estimtaors more fully, providing reference to the core ideas and concepts tied to each estimator.

In words, an unadjusted estimate of the average ITT effect at site j, $\hat{B}_j$, can be calculated as the difference in the average observed outcomes of all treatment group members at site j, $\hat{Y}_j(1)$, and the average observed outcomes of all control group members at site j, $\hat{Y}_j(0)$.

Common estimators of the overall ATE can then be described as a weighted average of the $\hat{B}_j$:

$$\hat{\beta} = \sum_{j=1}^{J} w_j \hat{B}_j \qquad \text{with } \sum_{j=1}^{J} w_j = 1 \tag{8}$$

The w_j are determined by which estimator is used. We discuss the different implied weights as we discuss the estimators, below. With covariate adjustment, the individually estimated $\hat{B}_j$ would be adjusted, and the weighting less clear, but the above holds in substance.

Along with impact estimates, we need uncertainty estimates. This is where the finite-sample vs. superpopulation tension is really critical. In particular, when we are estimating a superpopulation parameter, we need to account for two sources of uncertainty: the uncertainty in estimating the impact on our sample, and the uncertainty of how representative our sample is with respect to the population it came from. For finite-sample inference, by contrast, we only need to focus on the first source of uncertainty. An important point here is the sampling uncertainty is due to how much treatment effects vary across sites. If effects vary across sites, then finite-sample standard error estimators will grossly underestimate the true super-population uncertainty. If effects do not vary across sites, then they will not. Our technical appendix, section 2, gives further discussion of this point.

The different standard error estimates differ for two reasons: they focus on different kinds of variation, and they rely on different types of assumptions. In particular we can estimate variation in our sample by looking at how individuals vary within site, or by looking at how site-level estimates vary across site. Generally, for example, finite-population approaches estimate uncertainty within site and then average this uncertainty across sites to get overall uncertainty estimates. We can impose assumptions such as homoskedasticity of residuals across site or across treatment arm to stabilize this process. Super-population uncertainty estimators, by contrast, tend to rely on how the individual site-level impact estimates vary among sites to capture both any within-site estimation error along with the uncertainty associated with sampling sites from a larger population. Modeling assumptions, such as homoskedasticity, can stabilize this just as with the finite population approaches. We highlight these themes as we discuss the estimators.

Table 1. Estimator names, notation, and target estimands

Estimator Name	**Notation**	**Estimand**
Design Based		
(1) Finite Population, person weighted	$\hat{\beta}_{DB-FP-person}$	$\beta_{FP-person}$
(2) Finite Population, site weighted	$\hat{\beta}_{DB-FP-site}$	$\beta_{FP-site}$
(3) Super Population, person weighted	$\hat{\beta}_{DB-SP-person}$	$\beta_{SP-person}$
(4) Super Population, site weighted	$\hat{\beta}_{DB-SP-site}$	$\beta_{SP-site}$
Linear Regression		
(5) Fixed effects	$\hat{\beta}_{FE}$	$\beta_{FP-person}$
(6) Fixed effects, heteroscedasticity robust SEs	$\hat{\beta}_{FE-Het}$	$\beta_{FP-person}$
(7) Fixed effects, cluster robust SEs	$\hat{\beta}_{FE-CR}$	$\beta_{SP-person}$
(8) Fixed effects, ClubSandwich SEs	$\hat{\beta}_{FE-Club}$	$\beta_{SP-person}$
(9) Fixed effects, weighted – person	$\hat{\beta}_{FE-weight-person}$	$\beta_{FP-person}$
(10) Fixed effects, weighted – site	$\hat{\beta}_{FE-weight-site}$	$\beta_{FP-site}$
(11) Fixed effects w/ interactions, person weighted	$\hat{\beta}_{FE-inter-person}$	$\beta_{FP-person}$
(12) Fixed effects w/ interactions, site weighted	$\hat{\beta}_{FE-inter-site}$	$\beta_{FP-site}$
Multi-level Modeling		
(13) Random Intercept, Constant Treatment Coefficient	$\hat{\beta}_{ML-RICC}$	$\beta_{FP-person}$
(14) Random Intercept, Random Treatment Coefficient	$\hat{\beta}_{ML-RIRC}$	$\beta_{SP-site}$
(15) Fixed Intercept, Random Treatment Coefficient	$\hat{\beta}_{ML-FIRC}$	$\beta_{SP-site}$

We now investigate three common classes of estimators: design-based estimators, linear regression approaches, and multi-level modeling. As we discuss the estimators, we also discuss how standard errors are typically generated, and the choices available there. Table 1 (above) lists all 15 estimators we consider. For each estimator, we note the estimand that we believe is the most logical target of inference.

Design-Based Estimators

We begin with design-based estimators. Although their use may not be currently as common as linear regression or multi-level modeling, the effect estimators are easier to describe and understand, and seem to be gaining traction (Schochet, 2016). Design-based estimators, in contrast to several of the estimators below, are explicitly designed to target the four estimands. Design-based estimators focus on the random assignment mechanisms and the sampling mechanism as the primary sources of uncertainty, rather than specifying a linear model with a random residual component. They are therefore generally considered to rely on weaker assumptions than other approaches.

For multisite trials, design-based estimators generally use the difference in means estimator within each site (see equation (7)) to obtain site-level effect estimates. They then take a weighted average of these site-level effect estimates to obtain either the individual-weighted average (if one weights by site size) or site-weighted average (if one takes the average of the site level estimates) effect.

When the target of inference is the effect for the average person ($\beta_{FP-person}$ or $\beta_{SP-person}$), the design based estimator of β sets $w_j = \frac{N_j}{N}$, yielding:

$$\hat{\beta}_{DB-person} = \sum_{j=1}^{J} \frac{N_j}{N} \hat{B}_j. \tag{9}$$

This estimator is unbiased with respect to $\beta_{FP-person}$ and $\beta_{SP-person}$ and will generally be precise relative to reasonable alternatives. However, if the proportion assigned to treatment (p_j) vary dramatically across site, this estimator can be less precise than some precision-weighted estimators discussed below, especially if some very large sites have very lopsided treatment proportions. In this case, these large sites will have proportionally greater uncertainty despite their large size, and this uncertainty gets propagated to the overall estimate.

When the target of inference is the effect for the average site ($\beta_{FP-site}$ or $\beta_{SP-site}$), the design-based estimator of β sets $w_j = \frac{1}{J}$, yielding:

$$\hat{\beta}_{DB-site} = \sum_{j=1}^{J} \frac{1}{J} \hat{B}_j. \tag{10}$$

This estimator is unbiased with respect to $\beta_{FP-site}$ and $\beta_{SP-site}$. While unbiased, site weighting can increase uncertainty: with this estimator, a small site (where effects are estimated imprecisely) will have the same influence on the overall average effect estimate as a large site (where effects are estimated precisely).

Standard error estimators for $\hat{\beta}_{DB-person}$ and $\hat{\beta}_{DB-site}$ can be calculated for either the finite or super-population estimand. In the finite case, design-based estimators traditionally build an overall variance estimate by first estimating uncertainty of the site-specific estimates, and then averaging them across the sites, with weights dictated by the target estimand. This gives a weighted average of the site-level uncertainties of

$$\widehat{Var}\left[\hat{\beta}_{DB-w-finite}\right] = \sum_{j=1}^{J} w_j^2\, \widehat{Var\left(\hat{\mathrm{B}}_{\mathrm{j}}\right)} \tag{11}$$

where the weight w$_j$ are either n$_j$/n (person weighted) or 1/J (site weighted), and where the $\widehat{Var\left(\hat{\mathrm{B}}_{\mathrm{j}}\right)}$ can be estimated using Neyman's classic formula:

$$\widehat{Var\left(\hat{\mathrm{B}}_{\mathrm{j}}\right)} = \frac{1}{\mathrm{p_j N_j}} \mathrm{s}_{\mathrm{jT}}^2 + \frac{1}{(1-\mathrm{p_j})\mathrm{N_j}} \mathrm{s}_{\mathrm{jC}}^2 = \frac{1}{\mathrm{N_{jT}}} \mathrm{s}_{\mathrm{jT}}^2 + \frac{1}{\mathrm{N_{jC}}} \mathrm{s}_{\mathrm{jC}}^2 \ . \tag{12}$$

Here p_j is the proportion of units treated in site j (allowed to vary across site), N_{jT} is the number of treatment group members at site j, s_{jT}^2 is the variance of the outcome across treatment group members at site j, and N_{jC} and s_{jC}^2 are the corresponding values for the control group. Under a finite population framework this expression is conservative due to the unidentifiable correlation of potential outcomes across individuals. There are some corrections, although their impact is typically small. See, e.g., Aronow, Green, and Lee (2014) or Schochet (2016).

For super-population inference, design-based uncertainty estimates of β_{SP} rely on the variation in the site-level impact estimates, $\hat{B}_j$, to incorporate in the additional uncertainty due to site sampling. In particular, the individual site effect estimates are dispersed around the overall average effect estimate both due to site variation in average impact as well as within-site estimation error of site-level average impact. We can thus leverage this dispersion to get our overall uncertainty estimates. The formula for standard error estimation, with some additional discussion, are included in Appendix A. For detailed discussion of these estimators and their associated standard error estimators, see Schochet (2016).

Linear Regression Approaches

Linear regression is probably the most common tool for analyzing data from randomized experiments. We consider three common versions of regression modeling that are used in multi-site contexts: fixed effect models, fixed effect models with individual unit weights, and fixed effect models with treatment by site interaction terms. In the third case the site-level impact estimates are then combined in a second step to estimate the overall ATE. We first discuss these three different variants of linear regression, and then discuss uncertainty estimation at the end.

Fixed Effects with a constant treatment (FE). With fixed effects, the researcher fits a linear regression with site fixed effects and a single parameter for the overall ATE. The traditional model is,

$$Y_{ij} = \sum_{k=1}^{J} \alpha_k Site_{k,ij} + \beta T_{ij} + e_{ij}, \tag{13}$$

with $Site_{k,ij}$ an indicator variable for unit *ij* being in site k (out of J sites), and the e_{ij} classically considered to be independent, identically distributed (iid) draws from some normal distribution with 0 mean and unknown variance. Taken literally, the model assumes a constant treatment effect with no variation in effects across sites. For estimation, we find parameters values that minimize the sum of squared residuals and use $\hat{\beta}$ as our estimate of average impact.

This overall estimator is not, however, generally unbiased for either an individual or site average effect when there actually is cross site effect variability, depending on how this variability is associated with site size and treatment rates. Given the single impact parameter, the least squares process gives a *precision weighted* estimate of average impact, weighing each site average impact ($\hat{B}_j$) proportionally to the estimated overall precision in being able to estimate that site-average impact (Raudenbush & Bloom, 2015):

$$\hat{\beta}_{FE} = \sum_{j=1}^{J} \frac{N_j p_j(1-p_j)}{Z} \hat{B}_j, \tag{14}$$

with normalizing constant $Z = \sum_{j=1}^{J} N_j p_j\left(1 - p_j\right)$. The precision weights $w_j \ = N_j p_j\left(1 - p_j\right)/Z$ are the (normalized) inverses of $Var\left[\hat{B}_j\right]$ under assumed homoskedasticty with

$$\mathrm{Var}\left[\hat{B}_j\right] = \left(\frac{1}{N_j p_j} + \frac{1}{N_j(1-p_j)}\right)\sigma^2 = \frac{\sigma^2}{N_j p_j(1-p_j)}. \tag{15}$$

The σ^2 is the standard deviation of the residuals across sites and experimental conditions (assumed to be constant). This expression comes directly from the more general Neyman formulation above, if we assume the variances are equal, and ignore the possible correlation of potential outcomes (this term will naturally drop out under some assumed sampling models; see, e.g., the discussion of superpopulation inference of Miratrix et al. (2013)).

As Equation 14 shows, precision weighting weights larger sites and sites where the proportion assigned to treatment is closest to 0.50 more heavily. Precision weighting yields a standard error $SE\left[\hat{\beta}_{FE}\right]$ that will generally be smaller than any of the alternative standard errors. This is by design: fixed effect regression estimates the estimand that is, in principle, the most easy to estimate.[8] While an attractive feature, it comes at a price – the fixed effects estimator is potentially biased with respect to all four of our estimands. This is a bias-precision tradeoff.

To illustrate the potential bias, consider this estimator's potential bias with respect to the person-weighted mean estimands (β_{person}). If B_j is related to $p_j\left(1 - p_j\right)$, this estimator will be biased. This situation may be plausible if, for example, p_j is an indication of demand for program services and more effective sites experience greater demand. This could be true for an evaluation of charter schools that relies on natural lotteries to identify the effects of each school. Nonetheless, since N_j plays a major role in the implied weighting of the fixed-effect regression, we believe the most reasonably implied estimand for this estimator is $\beta_{FP-person}$. Indeed, if the proportion treated does not vary across sites (or $p_j\left(1 - p_j\right)$ is not correlated with B_j), then this estimator targets the individual average impact without bias.

With respect to the site-weighted mean estimands (β_{site}), this estimator will be biased if B_j is related to $N_j p_j\left(1 - p_j\right)$. In addition to the potential relationship between B_j and p_j, it will often be plausible that site size (N_j) is related to site ATE (B_j) – larger and smaller sites often serve different clientele, implement programs differently, are situated in different contexts, and offer different counterfactual services – each of which could cause treatment effects to systematically vary with size. For example, in the *Encouraging Additional Summer Enrollment* Study at 10 community colleges, site size, which varied substantially, was related to the number

[8] The "in principle" refers to the assumption of homoskedasticity that gives these precision weights. If different sites have different degrees of variation, then the precision weights may not correspond to the actual precisions of the sites and this estimand may not in fact be the easiest estimand to estimate. The fixed effect estimator would no longer be truly optimal in this case.

of summer course offerings, a factor that could moderate the intervention's effectiveness at increasing summer enrollment rates.

The estimand for which the fixed effects estimator is unbiased is one that weights each site average impact by $N_j p_j\left(1 - p_j\right)$. This estimand does not lend itself to a natural or policy-relevant description. Its definition is an artifact of a statistical property. If it is believed there is little to no cross-site effect variation, then this estimator may be entirely appropriate, owing to its improved precision. Otherwise, users of this estimator should be making an intentional decision regarding the potential bias precision tradeoff.

Fixed effects with a constant treatment and weighted units (FE-weight). The bias of the general fixed effect model, above, can easily be corrected by weighting individuals proportional to their inverse chance of treatment, as done in Edmunds et al. (2017). Such inverse probability of treatment weights can be calculated as:

$$\omega_{ij}^{person} = T_{ij}\left(\frac{p}{p_j}\right) + \left(1 - T_{ij}\right)\left(\frac{1-p}{1-p_j}\right), \tag{16}$$

where p is the proportion of the evaluation sample assigned to treatment. While this weighting can hurt precision relative to the fixed effects estimator ($\hat{\beta}_{FE}$), one can show that the expected value of the treatment effect estimate is the individual average effect (see derivation in Appendix A). Consequently, this estimator is unbiased with respect to the person-weighted average effect.

This weighting approach forces the weighted proportion assigned to treatment to be the same at each site and equal to the overall proportion assigned to treatment, while keeping each site's weighted size equal to its unweighted size. Due to this, the overall average impact estimate, $\hat{\beta}_{FE-weight-person}$, will be a site-size weighted average of the site impacts (see equation 9).[9]

A related weighting approach can be used to create an unbiased estimator of the effect for the average site (β_{site}). (See Mayer, Patel, Rudd, and Ratledge (2015) for an example that is very similar to this approach). This approach uses inverse probability of treatment weights in combination with forcing each site to contribute equally to the overall average effect estimates, resulting in the same impact estimate as the design-based estimator in Equation 10. See the technical appendix for the expression for these weights, with further discussion.

These weighting methods have strong ties to the design-based estimators. In fact, the effect estimates will be identical, although the methods for estimating uncertainty differs.

Fixed Effects with Interactions (FE-inter). One regression option that explicitly models cross site variation is the "fully interacted" model – that is, a model that includes both site indicators and site by treatment interaction terms. This model produces an ATE estimate for each site ($\hat{B}_j^{FE-inter}$) along with associated standard errors. The overall model can be written as

$$Y_{ij} = \sum_{j=k}^{J} \alpha_k Site_{k,ij} + \sum_{k=1}^{J} B_k Site_{k,ij}\, T_{ij} + e_{ij}. \qquad (17)$$

The residual structure is usually the same as with the usual fixed effect model above. The lack of intercept makes α_j the control mean at site j, and B_j the effect at site j.

Fitting this model gives effect estimates for each site. These J site effect estimates ($\hat{B}_j^{FE-inter}$) are then averaged in either of two ways. The first is to weight each site effect

[9] Interestingly, the weights make the site fixed effects not strictly necessary for estimation. The regression of outcome onto treatment, with these weights and without further covariates such as site fixed effect indicators, will give an unbiased estimate of the individual average impact. The site fixed effects, however, can absorb outcome variation across sites to increase precision.

estimate by the site size to obtain the individual average impact estimate, as in Equation 9. The second is to average across sites (see, for example Clark, Gleason, Tuttle, and Silverberg (2011)); this estimates the site level average impacts and is equivalent to Equation 10. These estimates will give identical values to the corresponding design-based estimators. Standard errors, if using the classic regression framework, are calculated by taking a weighted average of the standard errors associated with the $\hat{B}_j^{FE-inter}$ from the model described by Equation 17, which will give different measures of uncertainty as compared to design-based approaches. The regression approach, however, naturally allows for including covariate for adjustment (covariate-adjusted design-based approaches generally fit this model, in fact).

Uncertainty Estimation for linear models. There are three general approaches for getting standard errors for these linear regression models. The first is to specify a residual variance structure, i.e., how the individual units vary around the values predicted by the model. This is the classic Linear Regression approach. This approach usually makes assumptions about how the sites are similar, e.g., by assuming the residual variances are the same across sites. Two alternatives to imposing a residual model are heteroscedasticity robust (Huber-White) standard errors (for an example used in a multisite trial, see Weiss, Ratledge, Sommo, and Gupta, 2019) and cluster robust standard errors (for an example used in a trial, see Richburg-Hayes, Visher, and Bloom (2008)). These choices correspond to finite vs. super population estimation. The model-based standard errors are finite sample. Of the two types of “robust” standard error estimators – heteroskedastic and cluster robust – the former is appropriate for finite population inference and the latter for super population. We next make this reasoning more explicit.

Traditional presentations of the regression framework consider the covariates fixed and the residuals random. For a fixed effect regression, therefore, the sites are all considered fixed and the only uncertainty is the i.i.d. residuals for the individual outcomes. This implies finite sample inference and associated uncertainty estimates. For the models without interactions, any treatment effect variation leads to heteroskedastic residuals, a violation of the modeling assumptions.

Huber-White errors are primarily viewed as allowing for heteroscedasticity in the residual structure.[10] That being said, traditional theory typically assumes a sampling framework where individual observations (with an observation being the triplet of outcome, treatment assignment, and site) are sampled i.i.d. from some larger population. Critically, in this case the sites are considered to be fixed categories for an individual-level covariate akin to the categories of gender or race. This aspect of sites being fixed is what makes these standard errors still finite-sample targeting: we are estimating an impact while adjusting for the covariate of site membership. We are not allowing for variation due to our sites being a random draw of a broader population of sites.

Lin (2013) connects Huber-White standard errors to finite-sample, design-based inference. In this work, Huber-White standard errors are shown to be an appropriate (albeit conservative) choice for estimating finite-sample uncertainty in the context of individually randomized trials adjusted by covariates. Also see Schochet (2016) for related discussion and derivations.

Cluster robust standard errors, by contrast, assume entire clusters (sites) are sampled i.i.d. from a super population, and therefore treat the sites as integral clusters and give super-

[10] There are several variants of Huber-White, all of which differently adjust for concerns such as degrees of freedom issues; we use “HC1.” HC1 is the default of STATA’s “robust” option and is commonly used in the literature. All these variants are generally quite similar for large experiments.

population estimates of uncertainty. The potential correlation of individuals within site, as well as heteroskedasticity, is all accounted for by aggregating the site-specific patterns of residuals within site. These site-level aggregates are then averaged in a way that mimics sampling sites from a super population of sites. Here cross-site impact variation causes the within-site residuals to generally be correlated, giving larger aggregated amounts and larger standard errors; this is similar to design-based approaches that look at the dispersion of individual site effect estimates that directly incorporate cross-site impact variation. If applied to weighted regression, the weights will impact standard errors by effectively weighting the sites differently in the aggregation step.

For few sites, classic cluster-robust SEs are known to be unstable and biased downward.[11] There are refinements such as the "club sandwich" (Pustejovsky & Tipton, 2018) or bootstrapping (Cameron, Gelbach, & Miller, 2008). Also see Abadie, Athey, Imbens, & Wooldridge (2017) for further discussion of using cluster-robust vs. Huber-White standard errors in practice.

Multilevel Modeling

Multilevel modeling is another common approach for evaluating multisite trials in the social sciences. The traditional presentation of this framework posits a super population of sites, with the ATE of each site being a draw from a random effects distribution. The individual outcomes within each site are then thought of as extra residuals added on top of the site effects. This view strongly suggests multilevel modeling as targeting the super-population site average impact. MLM can also be viewed from a Bayesian perspective (see, e.g., Gelman & Hill, 2006);

[11] Similar to Huber-White's HC0, HC1, and so forth, there are several different proposed degrees-of-freedom corrections to account for this.

we do not investigate that here. Also see Hodges & Clayton (2011) for reflections on MLMs as a smoothing approach.

There are several proposed multilevel modeling strategies currently in the field, some which have been recently developed with cross-site impact variation in mind. We focus on one of these, the fixed intercept, random (treatment) coefficient (FIRC) estimator as described in Raudenbush and Bloom (2015) and Bloom, Raudenbush, Weiss, and Porter (2017). We do not examine Bayesian approaches in this work but use maximum likelihood estimation as is currently standard. The FIRC model has an indicator variable for each site, analogous to classic fixed effect estimator above, and a random effect for the average treatment across sites.

Level 1:

$$Y_{ij} = \sum_{k=1}^{J} \alpha_k \, Site_{k,ij} + B_j T_{ij} + e_{ij} \qquad (18)$$

Level 2:

$$B_j = \beta + b_j$$

Here the e_{ij} are independent normal with some variance structure, as with fixed effect regression. The b_j are typically modeled as i.i.d. normal, one draw per site, with an unknown variance, τ^2, which represents the variance of the distribution of site-level impacts. The β is the average site impact and is usually taken as the estimated ATE when reporting results.

Even though the parameter is clearly a site-average impact, this estimator is not, in general, unbiased for either an individual or site average effect. Instead it is adaptive. Similar to the fixed effects estimator, this estimator weights each site average impact ($\hat{B}_j$) proportionally to the precision in being able to estimate that site average impact (i.e. $\widehat{SE}\left[\hat{B}_j\right]$; see Raudenbush and Bloom (2015)). Importantly, this includes (estimated) uncertainty due to cross-site effect variation. The less cross-site effect variation is detected, the more similar this estimator is to the

FE, precision weighted one. In the face of greater cross-site effect variation this estimator adapts and comes closer to taking an unweighted average of the individual site estimates to avoid bias.

That is, weights are not only a function of the proportion of units treated and the size of the sample at that site, but they also depend on how much treatment effects are estimated to vary across sites. These weights, if the cross-site treatment effect variation (τ^2) were known, would yield the multi-level model fixed-intercept random treatment coefficient estimator:

$$\hat{\beta}_{ML-FIRC*} = \sum_{j=1}^{J} \frac{1}{Z}\left(\frac{\sigma^2}{N_j p_j(1-p_j)} + \tau^2\right)^{-1} \hat{B}_j \text{ with } Z = \sum_{j=1}^{J} \left(\frac{\sigma^2}{N_j p_j(1-p_j)} + \tau^2\right)^{-1}. \quad (19)$$

Notice how as the cross-site treatment effect variation diminishes, this estimator converges to the fixed effects estimator ($\hat{\beta}_{FE}$). As τ^2 increases, this estimator moves toward weighting the sites more and more equally. This is the adaptive aspect of this approach. The actual $\hat{\beta}_{ML-FIRC}$, roughly speaking, uses an estimated τ^2, but the logic is similar.

The two other common multilevel model specifications are the random-intercept, random-coefficient (RIRC) model and a random intercept model with a single (constant) treatment coefficient (RICC, analogous to fixed effect regression, above). Formally, RIRC can be written as

Level 1:

$$Y_{ij} = A_j + B_j T_{ij} + e_{ij} \quad (20)$$

Level 2:

$$A_j = \alpha + a_j$$
$$B_j = \beta + b_j.$$

The e_{ij} are again independent normal with some variance structure. The a_j and b_j are modeled as i.i.d. from a joint normal distribution, with unknown variances τ_a^2, τ_b^2 and an unknown covariance τ_{ab}. The τ_a^2 represents the variance of the distribution of site-level control group average outcome levels, τ_b^2 the variance of the distribution of site-level average impacts, and τ_{ab} their covariance.

Finally, RICC can be written as

Level 1

$$Y_{ij} = A_j + BT_{ij} + e_{ij} \tag{21}$$

Level 2

$$A_j = \alpha + a_j$$

The a_j are typically modeled as i.i.d. normal, one draw per site, with an unknown variance, τ^2, which, assuming a true constant treatment impact, represents the variance of the distribution of site average control-side outcome levels. Interestingly, because the RICC model does not allow treatment effects to vary by site, it is essentially the precision-weighted FE model from linear regression, and we thus we do not consider it as targeting a site average effect and instead consider it to be targeting the finite population person average effect.

For estimating uncertainty, multilevel modeling traditionally uses maximum likelihood estimation theory (or restricted maximum likelihood), which requires a complete model for both the random effects and the residual variances.

Multilevel modeling could potentially be made to explicitly target the average individual impact rather than site-average by weighting the regression as with the weighted fixed effect regression approaches discussed above. We do not investigate this approach in this work. See Raudenbush and Schwartz (2020).

A Unified Representation of the Effect Estimators

The effect estimators across the three general classes described above can be all viewed as different weighted averages of the site-specific effect estimates ($\hat{B}_j$), as described in Equation (8). Under this view, for the average effect estimates the different estimators differ only in how they weight the site-specific effect estimates. Table 2 (below) summarizes the four weighting schemes, borrowing from Raudenbush and Bloom (2015).

Table 2. Summary of implied weights of common effect estimators:

Weight name	**Weight**	**Estimators using this weight**
Unbiased person-weighting	$w_j \propto N_j$	$\hat{\beta}_{DB-FP-person}$ $\hat{\beta}_{DB-SP-person}$ $\hat{\beta}_{FE-weight-person}$ $\hat{\beta}_{FE-inter-person}$
Fixed-effect precision weighting	$w_j \propto N_j p_j (1 - p_j)$	$\hat{\beta}_{FE}$ $\hat{\beta}_{FE-HW}$ $\hat{\beta}_{FE-CR}$ $\hat{\beta}_{FE-Club}$ $\hat{\beta}_{ML-RICC}$ (approximately)
Random-effect precision weighting	$w_j \propto \left[\hat{\tau}^2 + \frac{\sigma^2}{N_j p_j (1 - p_j)}\right]^{-1}$ (approximately)	$\hat{\beta}_{ML-FIRC}$ $\hat{\beta}_{ML-RIRC}$ (approximately)
Unbiased site-weighting	$w_j \propto 1$	$\hat{\beta}_{DB-FP-site}$ $\hat{\beta}_{DB-SP-site}$ $\hat{\beta}_{FE-weight-site}$ $\hat{\beta}_{FE-inter-site}$

Notes: We have not derived a closed form expression for the implied weight for $\hat{\beta}_{ML-RIRC}$, but include it under random effect precision weighting for completeness (it differs slightly from FIRC in practice). While we have no closed form expression for the implied weight for $\hat{\beta}_{ML-RICC}$, empirical and simulation results find it to be extremely similar to fixed effect precision weighting.

Table 2 illustrates several things. First, if the weights are independent of the site average effects, or if the impacts do not vary, then all the estimators will coincide, up to estimation error, although they may differ in terms of their actual and estimated precision. Second, by examining the formula we can see how different estimators up- or down-weight sites based on site-level characteristics (sample size and proportion of units treated) which aids in determining what quantities the different estimators are targeting. Subsequently, we will refer to the weighting itself as an effect estimator, e.g., “the unbiased person-weighted effect estimator” denotes all four of DB-FP-person, DB-SP-person, FE-weight-person, and FE-inter-person.

Remark. In practice, there are two further concerns with these estimators, especially those that are variants of regression (fixed effects or multilevel modeling). First, treatment is often assigned within randomization subblocks (e.g., cohorts) within a site. These subblocks are not independent and need to be aggregated to get site ATEs. Second, if there is believed to be variation in treatment effect within site, then the usual homoscedasticity assumption for the residuals can be viewed as implausible, for the parameteric models (the multilevel models in particular). The most flexible option would be to therefore allow for different variances for each treatment arm for each site, but this can introduce instability in estimation, so something between complete homoskedasticity and this extreme may be preferred. Following Bloom, Raudenbush, Weiss, and Porter (2017), we advocate for two variance parameters, one for the control side and one for the treatment side, each pooled across site. This is discussed further in Supplementary Appendix A.

A Taxonomy of the Standard Error Estimators

The effect estimators can all be expressed as weighted sums of the individual site-level effect estimates, as discussed above. The standard error estimators can also be roughly grouped into families. The design-based approach is to use the known random assignment (and assumed known sampling of sites, in the case of super population inference) to generate standard errors that do not depend on parametric modeling assumptions. The modeling-based approach posits a model for how outcomes deviate from a structural form (along with, for super population estimators, a parameterized sampling framework such as normally distributed site level random effects), and uncertainty estimation is based on that. Importantly, the robust estimators (Huber-White and Cluster Robust methods) lie in the former, design-based camp.[12]

The modeling vs. design-based tension raises questions of the benefits and costs of the modeling assumptions. On one hand, we might expect the models to provide more precisely estimated standard errors, and on the other we might get biased estimates of the standard errors under misspecification. This is another bias-precision tradeoff, but now for estimating uncertainty rather than the estimand itself. In our simulation study we examine the gains of modeling under correct model specification, but leave the consequence of misspecification to future work. We also see how model-based vs. design-based approaches work empirically as well.

Empirical Examples

We examine how the above estimators perform across a convenience sample of 12 multisite field trials, all but two of which were also used in Weiss et al. (2017). These trials were selected based on several criteria. The most important criterion was that each study could provide internally valid estimates of average ITT effects *by site*. Datasets were restricted to multisite studies that randomly assigned individuals within sites, for as many sites as possible, with as many individuals per site as possible. The total number of individual sample members ranges across studies from roughly 1,400 to 69,000, and the total number of sites ranges from 9 to over 300. Consequently, our findings are most applicable to large scale RCTs. Our findings could be quite different in smaller scale studies. To facilitate implementation of our analyses, we

[12] This classification coupled with the biased impact estimators gives even more estimation approaches. For example, one could use cluster robust standard errors on top of multilevel modeling (this is what the HLM7 software and the "robust" option in STATA does) to keep a sampling framework of sites from a superpopulation while also allowing for complex, unknown correlation structures of individuals' residuals within site.

restricted data sets to those that were readily available to our team. The selected studies represent a broad spectrum of educational levels, ranging from preschool through primary and secondary school to postsecondary education, along with one welfare to work program. The outcomes include test scores, measures of progress toward and completion of higher school and college, and earnings. Interventions vary substantially, including, for example, two elementary school after school programs, a supplemental reading course for high school students and financial incentives for community college students.

For each empirical example, we estimate average effects and associated standard errors using all the methods (save one) described above and listed on Table 1. All of our examples are multisite trials but in most cases treatment was assigned within random assignment block within site, which creates technical issues discussed in online Appendix A.[13] Most of these examples have multiple outcomes, and we estimate impacts for all outcomes. For comparability and simplicity, we standardized all estimates and report results in effect size units; for a more nuanced discussion of how standardization works (and other options), see Weiss et al. (2017). Here, we use estimators without covariates; as a sensitivity check, we also conducted all analyses including relevant covariates and found that our substantive results hold (see Online Appendix D for discussion and covariate adjusted results).

Online Appendix C summarizes the 12 studies, describing the program and target population examined, the research design used, and the primary outcomes measured. Citations are included to help readers learn more about each study. Table 3 provides statistics about each study, including information about the site-level distribution of sample sizes and proportions

[13] Due to these concerns, we excluded $\hat{\beta}_{ML-RIRC}$. The other estimators have reasonable adjustments; with $\hat{\beta}_{ML-RIRC}$ the alternatives did not seem sensible.

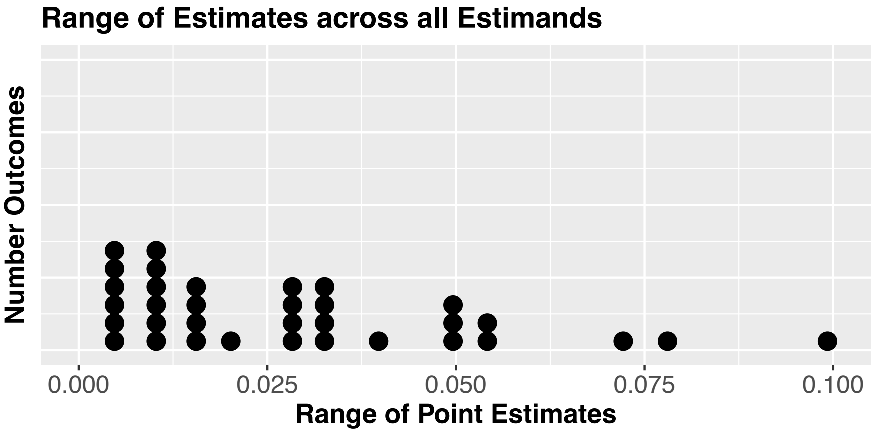
assigned to treatment, as well as an estimate of how much effects varied across sites. See Table 4 for all estimated effects and estimated standard errors using our final set of 14 estimators.

Results

To what extent can the choice of estimator of the overall average treatment effect (β) result in a different impact estimate?

To assess the greatest degree that estimator (or estimand) can matter, for each outcome we first consider the range of effect estimates across all our estimators (i.e., $max(\hat{\beta}) - min(\hat{\beta})$). Take, for example, the study of Learning Communities for community college students and the outcome of total cumulative credits earned through three semesters. For this specific outcome and study, the minimum effect estimate is -0.001σ (from the unbiased site weighted estimator) and maximum effect estimate is 0.031σ (from the fixed effect estimators and FIRC), for a range of 0.032σ. In most studies where a decent sized effect is required to deem a program a success, a 0.032σ difference would be considered small (substantively), suggesting that in this instance the choice of estimator does not matter much. The range of effect estimates for each of the 34 outcomes is on Figure 2 (below).

Figure 2: The Range (across all implemented estimators) of Estimates of β from 12 Studies (34 outcomes)



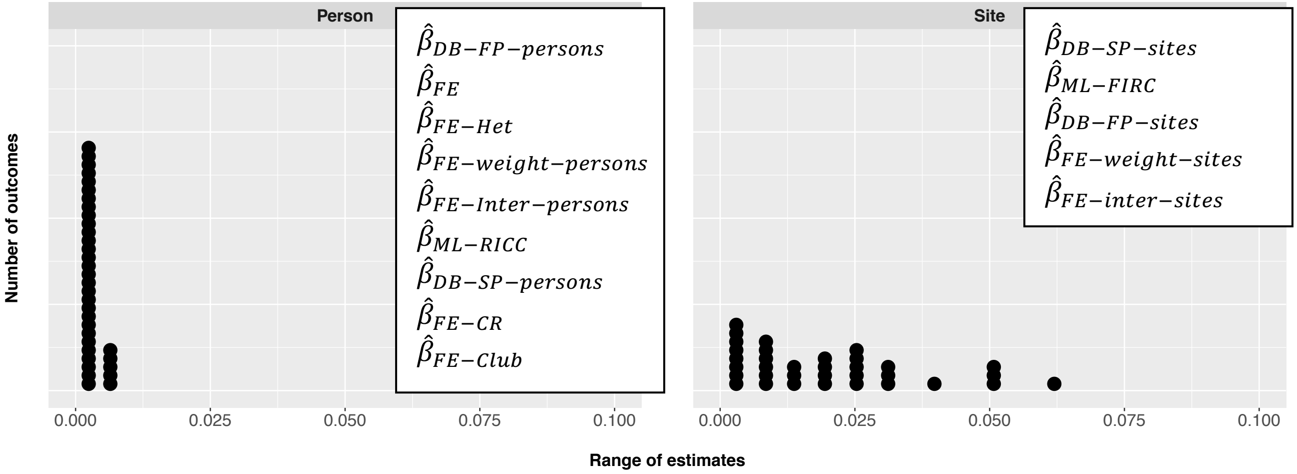
Notes: Each dot represents a single outcome for a single study. The x-axis is the range of effect estimates ($max[\hat{\beta}] - min[\hat{\beta}]$) across the final 14 estimators, in effect size units.

We see that on occasion, the choice of estimator can result in non-negligible differences in estimates of the overall ATE. The range of estimates is less than 0.02σ in about half the cases, which is small by most standards. In the three most extreme cases, the range is 0.10σ, 0.08σ, and 0.07σ. The extent that such differences are practically meaningful likely depends on the study and outcome. For a very low-cost intervention where an effect of size 0.05σ might be deemed "worth it," differences of this magnitude could be considered substantial. In contexts where the target is an effect of 0.25σ or above, only these most extreme differences may be considered clinically meaningful.

There are two theoretical reasons the estimators could produce different effect estimates for a given outcome. The first is if the person and site average estimands differ; in this case the estimators targeting the person average effects would systematically differ with the estimators targeting the site average effects. The second is that the estimators themselves could differ due to random variation or, for some, the bias-precision trade-off.

To investigate these reasons, for each outcome we calculate the range of effect estimates among the estimators targeting site estimands and person estimands. Results are shown in Figure 3; we do not further subdivide by finite vs. superpopulation as any effect estimator could in principle be used for either population framework.

Figure 3: The Range of Estimates of β for all estimators targeting a given estimand from 12 Studies (34 outcomes)



Notes: Each dot represents a single outcome for a single study. The x-axis is the range of effect estimates ($max[\hat{\beta}] - min[\hat{\beta}]$) across the estimators listed in the legend, in effect size units.

First, if a person-weighted estimand is chosen (left hand side), then the choice of $\boldsymbol{\beta}$ estimator generally does <u>not</u> matter. For these estimands, there are essentially two main effect estimators possible – the unbiased person-weighted estimators and the precision-weighted fixed effect estimators. The largest range among these estimators is 0.007σ, with an average range of 0.001σ. This implies that the potential bias in the bias-precision trade-off of the fixed effect estimators is negligible in practice. In many of these studies the proportion assigned to treatment do not vary substantially across sites (see Table 3), so this finding may be unsurprising.

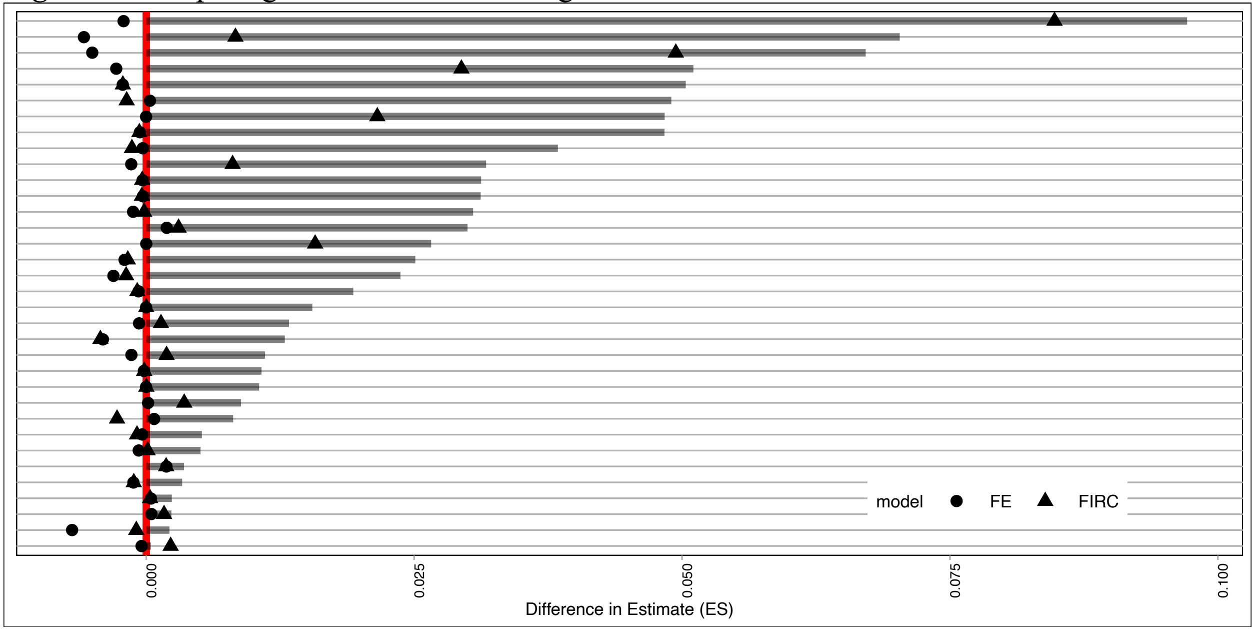
Second, if a site-weighted estimand is chosen (right hand side), then the choice of $\boldsymbol{\beta}$ estimator can matter. For this estimand, there are two different effect estimators considered – the unbiased site-weighted estimator and FIRC. The largest range in estimates is 0.062σ, with an average range of 0.020σ. In 15 cases (44 percent) the range of estimates is greater than 0.02σ, although there's only three instances where the range is greater than 0.05σ. Thus, even after identifying the site-average impact as one's target, the choice of estimator, within this class, can still matter. This appears to be due, in large part, to the adaptive nature of FIRC: when it does not estimate substantial variation in site-average effects, it aligns more with the person-weighted estimators, while the unbiased site-weighted estimator does not.

Finally, with respect to estimating $\boldsymbol{\beta}$, the choice of estimand can matter. The choice to target the effect for the average person vs. the effect for the average site can influence the estimated effect. In fact, the average difference in estimated effects between the unbiased person- and site-weighted estimators is 0.026σ. The average range of estimated effects among all distinct estimators is 0.028σ, only 9 percent larger. The difference between these two unbiased estimators explains 99 percent of the variation in the range across all estimators. The empirically observed differences in effect estimates do not necessarily mean the underlying estimands are different (the differences could all be dictated by estimation error). This does, however, illustrate that the driver behind any substantial differences in estimated effects across estimators is the correlation between $\hat{\beta}_j$ and N_j.

To reinforce these points, consider Figure 4, which shows the relative similarity of the different estimators across all 34 outcomes. The x-axis is the magnitude of the difference in effect estimates using various estimators. The grey bars present the difference in estimated effects between the unbiased person- and site-weighted estimators. We orient each estimator so the person-weighted estimate is always at zero and site weighted is always positive.[14] The fixed effect and FIRC estimators are then plotted as colored dots on these lines.

[14] We flip the sign of the outcome as needed to make the site-weighted estimate larger, just for this visualization.

Figure 4. Comparing Effect Estimates using Various Estimators



Notes: Gray bars represent the range of unbiased person weighted and unbiased site weighted effect estimates. Fixed effects estimates are circles and FIRC estimates are triangles. All bars centered so 0 corresponds to the unbiased person-weighted effect estimate.

First, notice that the fixed effect precision weighted estimator (red dots) are always near zero: it is strikingly similar to the person-weighted estimator. This shows that a correlation between $\hat{\beta}_j$ and $p_j(1-p_j)$ is not the reason behind differences in estimated effects across estimators.

Second, notice that the range of estimates across all estimators is rarely meaningfully larger than the range between the person- and site-weighted estimates alone (the gray bars). The FE and, to some extent, the FIRC deviations from the person weighted estimate are generally small in comparison to the grey bar magnitude. The range between the person- and site-weighted estimates is the primary driver of the variation in effect estimates. The weighting formula show that these differences stem from an empirical correlation between $\hat{\beta}_j$ and N_j.

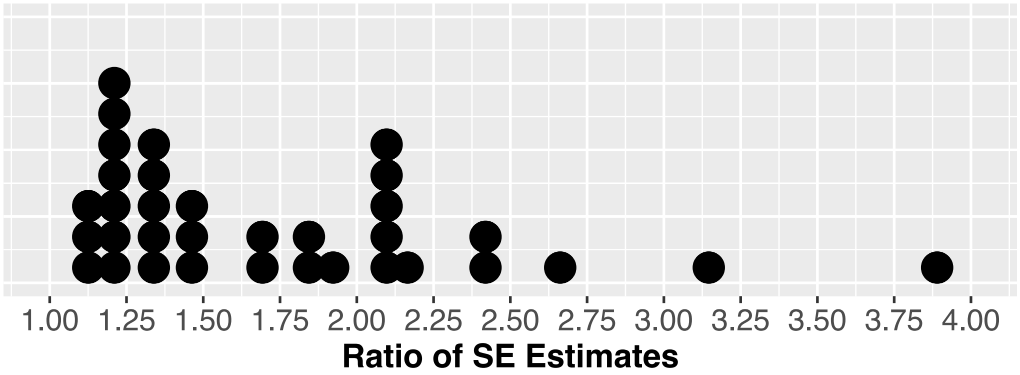
Finally, the FIRC effect estimates tend to lie between the person- and site-weighted estimates. This is the bias-precision tradeoff. In our examples, the FIRC estimates generally tended more toward the person-weighted estimates than the site-weighted estimates. This explains why, when the super-population site estimand is selected, the choice between FIRC and DB-Sites can matter. The source of difference is again the correlation between $\hat{\beta}_j$ and N_j.

To what extent can the choice of estimator of the standard error of the overall average treatment effect ($SE(\beta)$) result in a different estimated standard error?

For each outcome, we consider the ratio of the largest estimated standard error to the smallest estimated standard error (i.e., the $max\left[\widehat{SE}(\hat{\beta})\right]/min\left[\widehat{SE}(\hat{\beta})\right]$) across all estimates of the ITT effect. This provides an indicator of the greatest extent that the choice of estimator (or estimand) can matter when estimating precision. Take, for example, the Communities in Schools study of for high school students and the chronic absenteeism outcome. The minimum $\widehat{SE}(\hat{\beta})$ is 0.046σ and the maximum $\widehat{SE}(\hat{\beta})$ is 0.080σ, an increase of 173 percent. In this instance, the choice of estimator with respect to $\widehat{SE}(\hat{\beta})$ can matter a lot.

Figure 5 (below) plots the ratio of the largest estimated standard error to the smallest estimated standard error for each outcome. As shown, the ratio from the Communities in Schools example is not unusual. For half of the outcomes, the ratio of the largest to smallest standard error estimate is greater than 150%, large by most standards. For some studies, the largest estimated standard error is more than three times the smallest estimated standard error. With respect to the estimated standard error of the overall ATE, choice of estimator can matter a lot in practice.

Figure 5: The Ratio of the Largest to Smallest Estimates of $SE(\beta)$ across our final 14 estimators from 12 Studies (34 outcomes)

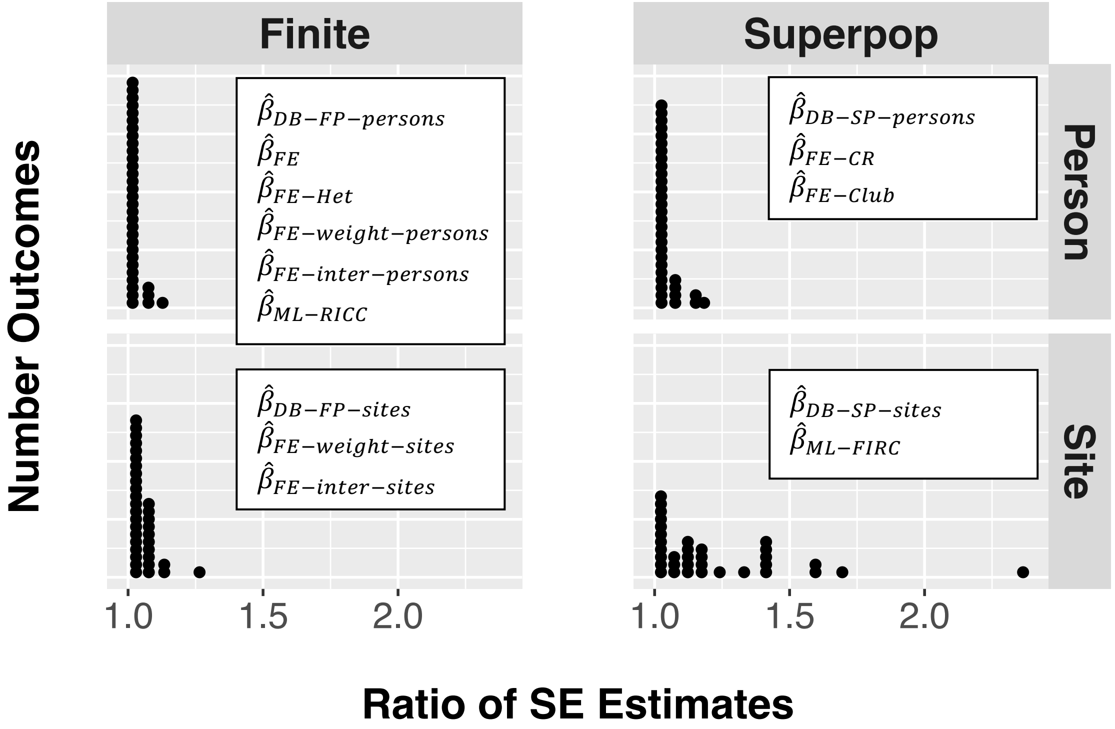


Notes: Each dot represents a single outcome for a single study. The x-axis is the ratio of largest estimated standard error to smallest estimated standard error ($max[\widehat{SE}(\hat{\beta})]/min[\widehat{SE}(\hat{\beta})]$ across all estimators, in effect size units.

Precision is meaningfully worse when we move from finite- to super-population estimators or from person- to site-average estimators. SE estimates from the finite-population *site*-average estimators are 29% larger than their finite-population *person*-average counterparts (averaged across outcomes). Compared with the *finite*-population person-average SEs, the *super*-population person-averaged SEs are 25% larger and the super-population site-averaged SEs are 30% larger (although there was much instability here, see below). Choosing to generalize to a super population, to estimate the effect for the average site, or to do both generally results in less precise estimated effects, compared to finite-population person-weighted.

Next, we examine whether the choice of standard error estimator matters once an estimand has been selected. To do so we calculate the ratio of the largest to smallest estimated standard error among the estimators that target a specific estimand, for each outcome. Results are on Figure 6, below. For example, the top left corner of Figure 6 plots the 34 ratios of the largest to smallest estimated standard error for each of the six estimators that target the finite population person estimand.

Figure 6. Ratio of Largest to Smallest Estimate of $SE(\beta)$ Among all Estimators Targeting a Given Estimand from 12 Studies (34 outcomes)



First, if a finite-population person, finite-population site, or super-population person estimand is chosen, then the choice of $\boldsymbol{SE(\beta)}$ estimator generally does <u>not</u> matter. For these estimands, the average ratio of the maximum to minimum estimated standard error are 102%, 105%, and 104%, respectively. For the finite-population person estimand, these results have several interesting implications. First, the potential precision gains in the fixed effects estimator's bias-precision trade-off are small in practice. In addition, correcting for heteroskedasticity when using a fixed effects estimator is generally inconsequential in these trials. Finally, results confirm that the multilevel model with random intercept and fixed treatment coefficient (RICC) should *not* be considered an estimator that allows for a super population inference. Rather, the estimated standard error is finite population and nearly identical to the classic fixed effect estimator.

The second broad finding from Figure 6 is that if the super-population site estimand is chosen (bottom right quadrant), then the choice of $\boldsymbol{SE(\beta)}$ estimator can matter a lot. Here, the comparison is between the design-based estimator and the multi-level model FIRC. The largest ratio of the maximum estimated standard error to the minimum estimated standard error is 236% (an outlier), with an average ratio of 123%. In 59% of cases, the estimated standard errors differ by more than 10% across the two options. Compared with the design-based super population site estimator, we generally expected FIRC to result in smaller standard error estimates since the multi-level model estimator will give more weight to larger sites in most instances. This held true for the most part; however, 32 percent of the time the design-based estimator had a smaller estimated standard error than FIRC. Estimated standard errors can differ due to difficulties with estimation or fundamental differences in the true uncertainty of the estimators. We examine this next.

Are the wide-ranging standard error estimates due to instability in estimation?

Surprisingly, sometimes the super population estimators give smaller estimated standard errors than the finite population estimators. This appears to be driven by instability in estimating the super population targeted standard errors. Furthermore, FIRC and the design-based site superpopulation standard errors frequently differed substantially, with the FIRC standard errors ranging from 58% smaller to 62% larger than the design-based site superpopulation standard errors. To understand where these instabilities were occurring, we turn to our simulation study, details and results of which are more fully discussed in online Appendix B.

In our simulation we generated multisite data with a variety of structures mimicking our empirical studies under a correctly specified multilevel data generating process. We then compared several measures: (1) the true standard errors of the different estimators, (2) the overall accuracy of the estimated standard errors, and (3) the variability of the estimated standard errors. We generally found that, across our scenarios, the super population estimators have more difficulty estimating standard errors. Superpopulation standard error estimates tend to be around 14-19 percent off the truth, with many scenarios tending towards 30 percent or more. In contrast, the site-averaged, finite-population standard errors tend to be estimated to within 6 percentage points of accuracy, and the person-averaged finite population even better than that. This means that super population standard error estimators can frequently return erroneously small values, smaller even than the finite-sample counterparts. In fact, for many of our scenarios with a cross-site variation (τ) of 0.10σ, $\hat{\beta}_{DB-SP-sites}$ gave estimated SEs smaller than its finite-sample counterpart more than 30 percent of the time, and $\hat{\beta}_{ML-FIRC}$ did so 15-20 percent of the time.

Comparing the two site-average superpopulation estimators, we found the design-based super-population estimator is more variable than FIRC, with 17% larger true SEs, averaged across the scenarios. It also has more difficulty *estimating* those standard errors: the average standard deviation of the SE estimates, relative to the true SEs, was 34% larger on average for the design based. In other words, the design-based estimator is not only generally more variable in truth, it is more unstable in its SE estimates. This estimator appears to be paying a heavy price for unbiasedness and less model dependence, while the adaptive FIRC will tend towards precision-weighted, which gives it greater stability. On the other hand, in those scenarios with substantial bias, the FIRC's lower SEs coupled with the biased point estimate undermines coverage; we explore that below.

Our simulations are conducted under correct model specification; it could be that that the greater instability of some of the estimators is a function of the reliance on fewer assumptions. Under misspecification, models such as FIRC could conceivably break down, resulting in poorly estimated SEs and even greater bias in point estimates. Regardless, the instability is a concern. We leave exploring how misspecification plays a role to future work.

How do the theoretically possible bias-precision tradeoffs play out in real data?

There are two types of biased estimator we consider: the fixed effect estimator for estimating person-weighted impacts, and the multilevel modeling estimators for estimating site average impacts. We investigate whether there are gains in using these estimators above and beyond the costs due to the potential bias.

For the fixed effect estimator, the bias (with respect to the person-average effect estimand) tends to be quite small. We saw little to no difference between the fixed effect (precision weighted) effect estimates and the unbiased effect estimates. The standard errors also tended to be the same, correlating at above 0.99 across the studies. The fixed effect standard errors were about 99% of the design based standard errors, on average. The two outcomes with the greatest estimated standard error reduction showed fixed effects standard errors around 6 and 11% smaller than the design based. Overall, in our considered studies, the fixed effect estimators do not appear to give much improved precision over the design-based estimators. They do not, however, appear to be heavily biased either.

On the other hand, FIRC often differs from the design-based super population site estimator. This can be driven by bias in the multilevel models and/or instability in either of the estimators. Disentangling these two factors requires knowing the truth, which we do not know in the empirical examples. We therefore turn to our simulations. Here we find that with variable site size but no variation in impacts across sites, the multilevel model was, under correct model specification, more stable, resulting in a substantially lower root mean squared error (RMSE) than the site-average design-based estimator. As cross-site impact variation increased, however, if that cross-site variation was correlated with site size, the tendency of FIRC towards the person-weighted estimates cost in an increasingly large bias. That said, less than 2% of our considered scenarios had the FIRC fair worse, in terms of RMSE; FIRC was slightly worse when there were many (80) large sites that varied in size, and site size correlated with site impact. In these scenarios the nominal coverage of FIRC was also worse (occasionally dropping to 85%) due to this bias; see the final section of Supplement B for a detailed comparison of FIRC to DB-SP-Sites.

Remark. All of the estimators need to be used with care, and many have specific implementation details or aspects that can be easily missed, and which can lead to error. For example, the reweighted estimators rely on weighted regression methods taken from survey sampling, *not* precision-weighting from classic generalized least squares; if the incorrect statistical methods are used, these methods can give smaller estimated standard errors than the true variability. This issue, and others, are discussed in greater detail in our supplementary material. In particular, in our technical supplement (Appendix A) we discuss each of the methods in turn and attempt to highlight the intuitions driving these estimators along with the common pitfalls one might come across in their use.

Conclusion

Defining an estimand is critical to the design, analysis, and interpretation of a multisite RCT. Even when one is interested in estimating an ATE, careful consideration must be given to defining the target of inference. For example, the choice of estimand influences what formula is appropriate when conducting power calculations – yet many write-ups of power calculation are ambiguous or silent with regards to the estimand chosen. Relatedly, registering studies and creating analysis plans are becoming the norm when conducting RCTs, yet even newly created registries, such as this society's *Registry of Educational Effectiveness Studies* (REES: https://sreereg.icpsr.umich.edu/pages/checklist.php#rt) do not require an estimand be defined. Consequently, readers of an analysis plan are left to assume an implied estimand based on the power calculation formula used and/or estimator selected. Similarly, many scholarly articles and reports do not state a target of inference, making it difficult to assess whether the chosen estimator is appropriate and obscuring the goal of the research. **We recommend defining the estimand(s) early in the research process and clearly stating them in important products.**

We find that the choice of estimand can matter in field trials. In a series of empirical examples, we find that the decision occasionally influences effect estimates and frequently influences the estimated standard error. Most notably, interest in super population or site averaged effects often comes with substantial cost in terms of precision. Appreciating this may help ensure that studies targeting these estimands are adequately powered.

For three of the estimands considered, once an estimand is selected, the choice of estimator does not matter – estimated effects and standard errors are very similar across estimation approach. Researchers, proposal reviewers, journal article reviewers, technical working groups advisors, etc. ought not fret over whether, for example, a fixed effects estimator did or did not use heteroskedastic robust standard errors. It is interesting to note that, at least in the studies considered, choosing estimators that rely on modeling assumptions (which we might believe to be more precise and more sensitive to model misspecification) do not, in the end, appear to differ substantively from their design-based or robust counterparts. It would be interesting to extend our simulation framework to delineate what circumstances these estimators would diverge.

Where estimator choice may matter is when the estimand is $\beta_{SP-site}$. Here, the multilevel FIRC model can result in different treatment effect estimates than its unbiased design-based counterpart and frequently results in different estimated standard errors as well. This is a result of both the bias-precision trade-off and the instability of all the super population standard error estimators. In our simulation studies we found that (under correct model specification) FIRC was almost always superior in terms of RMSE, compared to DB-SP-Sites. This is true even in those scenarios where FIRC was biased due to tending towards person weighted impacts. Moreover, the latter had, much more than the former, demonstrable and serious instability in most contexts considered. Coverage, on the other hand, showed the design-based being somewhat more reliable in the face of cross-site impact variation. This is especially true when there were many variable-sized sites with site size correlated with site impact. Regardless of estimator, standard error estimation for the super population site-average impact, even under correct model specification, is difficult. To be clear – super population standard errors ought to be appropriately larger than their finite population counterparts. However, the challenge is that we cannot be certain that the *estimated* standard error is an accurate reflection of *true* uncertainty. The differences of FIRC and DB-SP-Sites may be the consequence of the choice between modeling and making fewer assumptions: our simulations show that when data are approximately normal, the design-based approaches can be costly. We leave exploring how the assumptions behind FIRC could undermine good inference under model misspecification to future work.

Given these findings, researchers may naturally ask what they should do. The starting point is selecting an estimand based on the study goal and designing a study with a realistic chance of achieving that goal. Pragmatically, this can mean acknowledging that the ideal goal is unachievable (e.g., the budget does not align with the sample size needs for adequate precision).

The result may be a compromise: conducting a study acknowledging limited precision with respect to a site and/or super population estimand or selecting an estimand that is not the top choice, but still a meaningful quantity to estimate. In some cases, these practical considerations may yield the conclusion that the study that could be achieved is not worth doing.

When it comes to analyses, we generally suggest starting simple. It is difficult to go wrong starting a presentation of findings with the finite-sample person estimand. Even if the ultimate goal is understanding a program's effects in a super population, it is difficult to imagine situations where it is of no interest to know the extent that a program was effective for the study participants at the study sites.

After estimating effects for the individuals and sites in the evaluation, next consider how estimated uncertainty increases as the inference extends to a super population. For ease of exposition, we have occasionally presented a false choice between finite- and super-population estimands. However, it may be prudent to consider both in this way. As an example, consider the Head Start Impact Study and the program's average estimated effect on the *WJ-III LW early reading* outcome for the person finite- and person super population estimands. Across the reasonable estimators of these estimands, the average effect estimate is consistently 0.232σ. However, the estimated standard error for the person *finite* sample estimators are all 0.035, whereas they are 0.041 and 0.042 for the person *super* population estimators. The 17 percent increase in the estimated standard error reflects the increased uncertainty that comes with the generalization implied by the super population inference. In this study, the data provide clear evidence that Head Start had a positive effect on early reading for the children attending the 253 Head Start Centers included in the analyses *and* for children attending a super population of centers from which these 253 Centers could reasonably have been randomly drawn.

In the above instance we can be confident of positive effects on the students in the study and in a larger super population. However, one can conceive of cases where we are confident that an intervention positively affected the people in the study, but we are far less certain whether the intervention would positively affect people on average in the super population. In an efficacy study with a finite population estimand, this situation may be grounds for a large-scale effectiveness trial. In an effectiveness trial with a super population estimand, this may complicate the interpretation of the findings in a way that is important to acknowledge and appreciate.

We have seen for super population goals, estimates of uncertainty are generally larger and estimators are more unstable than their finite population counterparts. In some ways the instability is unsurprising – estimating uncertainty when targeting a super population requires knowing how much treatment effects vary across sites. Recent theoretical work by Bloom and Spybrook (2017) and empirical work by Weiss et al. (2017) demonstrate that under many real-world circumstances this cross-site impact variation is itself estimated with considerable uncertainty. This uncertainty is likely the cause of the instability of the super population standard error estimators. Gaining a deeper understanding of the conditions (e.g., number of sites, number of people per site, etc.) that allow for acceptably reliable standard error estimation for standard errors of this type is an area ripe for future simulation and theoretical research.

While work presented in this paper is broad, with many studies and many estimation approaches, there are several limitations of note. First, our findings apply to multisite trials. How different estimation strategies compare in other contexts, such as cluster randomized trials, is an important area for future work. There are also many other estimators we might consider such as hybrid approaches such as cluster-robust standard errors on top of a multilevel models. We have also not investigated how decisions regarding missing data or handling of attrition could move estimates, and whether the sensitivity to those decisions is on par with, smaller than, or larger than the decisions regarding estimand and estimator selection. It would also be interesting to see if different estimation strategies were more or less robust to such decisions. We also acknowledge that our simulations only show relative estimator performance when the multilevel models are correctly specified; seeing if misspecification changed the story is an important area of future work. And of course any investigation of a collection of studies suffers from the questions of whether the studies are representative. Although the simulations suggest that only in fairly extreme circumstances would the story change, we should seek to verify this with real data. Only with continued comparisons such as our empirical investigation can we accumulate best practices and rule-of-thumb guidelines.

Finally, it is again worth noting that when intervention effects vary, a one-number summary of the overall average effect masks important information. For example, if effects vary across sites, an estimate of the overall average effect from a multisite trial may not accurately predict the effect at any individual site, limiting the value of a one-number summary for local decision-makers (Orr et al., 2019). At a minimum, it is important to attempt to understand the distribution of site ATEs and the sources of this heterogeneity (Raudenbush and Bloom, 2015; Bloom et al., 2017; Weiss, Bloom, and Brock, 2014), as well as the average impact, however it may be defined.

Table 3. Statistics from 12 Multi-site RCTs.

			Site Size				Proportion Treated			Est. SD of
		#	10th		90th	SD over	10th		90th	Site-ave.
Project/Outcome	N	Sites	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Effects*
Early childhood-Elementary school										
Head Start Impact Study (HSIS)										
Externalizing behavior problems	3,404	257	3	13	22	0.67	0.50	0.61	0.73	0.10
PPVT-III receptive vocabulary	3,360	251	3	13	22	0.66	0.50	0.61	0.75	0.07
Self-regulation skills	3,404	254	3	13	22	0.67	0.50	0.61	0.75	0.18
WJ-III AP early numeracy	3,353	253	3	13	22	0.67	0.50	0.61	0.75	0.09
WJ-III LW early reading	3,377	253	3	13	22	0.66	0.50	0.62	0.75	0.29
WJ-III OC oral comprehension	3,288	247	3	13	22	0.66	0.50	0.62	0.75	0.12
After School Reading										
SAT-10 total reading	1,904	25	57	76	94	0.17	0.51	0.58	0.65	0.05
After School Math										
SAT-10 total math	1,984	25	58	79	101	0.26	0.49	0.56	0.65	0.11
High school										
Enhanced Reading Opportunities										
GRADE reading comprehension	4,584	34	92	135	169	0.22	0.56	0.59	0.63	0.00
GRADE reading vocabulary	4,584	34	92	135	169	0.22	0.56	0.59	0.63	0.00
% of required credits earned, yr 1	5,225	34	92	154	192	0.25	0.53	0.57	0.63	0.00
% of required credits earned, yr 2	4,554	34	80	134	167	0.25	0.53	0.58	0.63	0.00
Communities in Schools										
Chronic absenteeism	1,816	28	26	65	130	0.70	0.43	0.52	0.65	0.14
Failed at least one course	2,123	28	30	76	166	0.73	0.45	0.52	0.67	0.00
Early College High Schools										
On track in ninth grade	3,793	19	56	200	466	0.67	0.44	0.59	0.84	0.28
Earned a high school diploma	2,792	19	52	147	353	0.75	0.42	0.60	0.84	0.00
Career Academies										
Earned HS diploma/GED, yr 5	1,533	9	54	170	259	0.44	0.52	0.56	0.59	0.00
Enrolled in postsecondary	1,482	9	50	165	252	0.45	0.52	0.56	0.58	0.08
Avg. annual earnings, yrs 1-4	1,458	9	52	162	255	0.43	0.53	0.55	0.58	0.00
Avg. annual earnings, yrs 5-8	1,405	9	46	156	235	0.44	0.53	0.55	0.58	0.00
Avg. months worked annually, yrs 1-4	1,458	9	52	162	255	0.43	0.53	0.55	0.58	0.00
Avg. months worked annually, yrs 5-8	1,405	9	46	156	235	0.44	0.53	0.55	0.58	0.00
Postsecondary education										
Learning Communities										
Targeted credits earned, 1 sem	6,974	11	198	634	1,089	0.67	0.50	0.60	0.67	0.17
Cumulative targeted credits earned, 3 sem	6,974	11	198	634	1,089	0.67	0.50	0.60	0.67	0.10
Total credits earned, 1 sem	6,974	11	198	634	1,089	0.67	0.50	0.60	0.67	0.05
Cumulative total credits earned, 3 sem	6,974	11	198	634	1,089	0.67	0.50	0.60	0.67	0.00
Performance-Based Scholarships										
Cumulative total credits earned, yr 1	6,935	15	63	462	1,081	0.78	0.50	0.58	0.62	0.00
Cumulative total credits earned, yr 3	6,935	15	63	462	1,081	0.78	0.50	0.58	0.62	0.00
Earned a degree, yr 3	6,968	15	63	465	1,081	0.78	0.50	0.58	0.62	0.00
Encouraging Summer Enrollment 1										
Enrolled in summer	7,118	10	90	712	2,317	1.30	0.49	0.50	0.51	0.00
Credits earned	7,118	10	90	712	2,317	1.30	0.49	0.50	0.51	0.00
Encouraging Summer Enrollment 2										
Enrolled in summer	7,103	10	88	710	2,310	1.30	0.49	0.50	0.50	0.06
Credits earned	7,103	10	88	710	2,310	1.30	0.49	0.50	0.50	0.07
Labor/Workforce										
Welfare-to-Work Program										
Ave. annual earnings, quarter 1-8	69,399	59	322	1,176	3,011	0.95	0.50	0.66	0.84	0.09

*The last colum (SD of Dist. of Site-ave. Effects) is estimated from the FIRC model.

Table 4. Estimates of the ITT effect ($\widehat{\beta}$) and its standard error ($\widehat{SE}(\hat{\beta})$) from 12 Multi-site RCTs.

			ESTIMATORS (BETA)													
			Design-based				Linear Regression								MLM	
		#	FP	FP	SP	SP		FE	FE	FE	FE-IPTW	FE-IPTW	FE-Inter	FE-Inter		
Project/Outcome	N	Sites	Person	Site	Person	Site	FE	HW	CR	Club	Person	Site	Person	Site	RICC	FIRC
Early childhood-Elementary school																
Head Start Impact Study (HSIS)																
Externalizing behavior problems	3404	257	-0.094	-0.063	-0.094	-0.063	-0.094	-0.094	-0.094	-0.094	-0.094	-0.063	-0.094	-0.063	-0.096	-0.094
			(0.034)	(0.04)	(0.035)	(0.039)	(0.034)	(0.034)	(0.036)	(0.035)	(0.034)	(0.038)	(0.035)	(0.041)	(0.034)	(0.035)
PPVT-III receptive vocabulary	3360	251	0.131	0.156	0.131	0.156	0.129	0.129	0.129	0.129	0.131	0.156	0.131	0.156	0.129	0.129
			(0.03)	(0.035)	(0.03)	(0.035)	(0.029)	(0.03)	(0.031)	(0.03)	(0.03)	(0.033)	(0.029)	(0.034)	(0.03)	(0.03)
Self-regulation skills	3404	254	-0.018	-0.021	-0.018	-0.021	-0.019	-0.019	-0.019	-0.019	-0.018	-0.021	-0.018	-0.021	-0.018	-0.020
			(0.034)	(0.04)	(0.036)	(0.048)	(0.032)	(0.033)	(0.037)	(0.035)	(0.034)	(0.04)	(0.032)	(0.037)	(0.033)	(0.036)
WJ-III AP early numeracy	3353	253	0.119	0.132	0.119	0.132	0.115	0.115	0.115	0.115	0.119	0.132	0.119	0.132	0.114	0.114
			(0.032)	(0.037)	(0.034)	(0.039)	(0.031)	(0.032)	(0.034)	(0.033)	(0.032)	(0.036)	(0.031)	(0.036)	(0.032)	(0.033)
WJ-III LW early reading	3377	253	0.232	0.223	0.232	0.223	0.232	0.232	0.232	0.232	0.232	0.223	0.232	0.223	0.236	0.228
			(0.035)	(0.038)	(0.041)	(0.043)	(0.035)	(0.035)	(0.042)	(0.041)	(0.035)	(0.037)	(0.035)	(0.04)	(0.035)	(0.04)
WJ-III OC oral comprehension	3288	247	-0.004	0.019	-0.004	0.019	-0.007	-0.007	-0.007	-0.007	-0.004	0.019	-0.004	0.019	-0.007	-0.006
			(0.029)	(0.032)	(0.03)	(0.038)	(0.029)	(0.029)	(0.031)	(0.03)	(0.029)	(0.033)	(0.029)	(0.034)	(0.029)	(0.031)
After School Reading																
SAT-10 total reading	1904	25	-0.090	-0.093	-0.090	-0.093	-0.092	-0.092	-0.092	-0.092	-0.090	-0.093	-0.090	-0.093	-0.098	-0.092
			(0.042)	(0.043)	(0.044)	(0.042)	(0.042)	(0.042)	(0.045)	(0.044)	(0.042)	(0.043)	(0.042)	(0.043)	(0.042)	(0.044)
After School Math																
SAT-10 total math	1984	25	0.087	0.098	0.087	0.098	0.086	0.086	0.086	0.086	0.087	0.098	0.087	0.098	0.086	0.089
			(0.04)	(0.041)	(0.045)	(0.053)	(0.041)	(0.041)	(0.046)	(0.045)	(0.041)	(0.041)	(0.04)	(0.042)	(0.041)	(0.046)
High school																
Enhanced Reading Opportunities																
GRADE reading comprehension	4584	34	0.060	0.065	0.060	0.065	0.059	0.059	0.059	0.059	0.060	0.065	0.060	0.065	0.061	0.060
			(0.029)	(0.03)	(0.034)	(0.034)	(0.029)	(0.029)	(0.035)	(0.035)	(0.029)	(0.03)	(0.029)	(0.03)	(0.029)	(0.035)
GRADE reading vocabulary	4584	34	-0.011	-0.014	-0.011	-0.014	-0.010	-0.010	-0.010	-0.010	-0.011	-0.014	-0.011	-0.014	-0.009	-0.010
			(0.03)	(0.03)	(0.028)	(0.03)	(0.029)	(0.029)	(0.028)	(0.028)	(0.029)	(0.03)	(0.03)	(0.031)	(0.029)	(0.029)
% of required credits earned, yr 1	5225	34	0.068	0.070	0.068	0.070	0.068	0.068	0.068	0.068	0.068	0.070	0.068	0.070	0.068	0.068
			(0.026)	(0.028)	(0.031)	(0.033)	(0.026)	(0.026)	(0.031)	(0.031)	(0.026)	(0.028)	(0.026)	(0.028)	(0.026)	(0.031)
% of required credits earned, yr 2	4554	34	0.028	0.042	0.028	0.042	0.028	0.028	0.028	0.028	0.028	0.042	0.028	0.042	0.027	0.030
			(0.028)	(0.029)	(0.031)	(0.033)	(0.028)	(0.028)	(0.032)	(0.031)	(0.028)	(0.029)	(0.028)	(0.029)	(0.028)	(0.031)
Communities in Schools																
Chronic absenteeism	1816	28	0.046	-0.025	0.046	-0.025	0.051	0.051	0.051	0.051	0.046	-0.025	0.046	-0.025	0.053	0.037
			(0.046)	(0.046)	(0.051)	(0.08)	(0.047)	(0.047)	(0.051)	(0.05)	(0.047)	(0.047)	(0.047)	(0.058)	(0.047)	(0.056)
Failed at least one course	2123	28	0.034	-0.016	0.034	-0.016	0.036	0.036	0.036	0.036	0.034	-0.016	0.034	-0.016	0.038	0.036
			(0.041)	(0.046)	(0.035)	(0.048)	(0.041)	(0.041)	(0.035)	(0.035)	(0.041)	(0.046)	(0.041)	(0.052)	(0.041)	(0.041)
Early College High Schools																
On track in ninth grade	3793	19	0.246	0.148	0.246	0.148	0.248	0.248	0.248	0.248	0.246	0.148	0.246	0.148	0.242	0.161
			(0.031)	(0.034)	(0.094)	(0.073)	(0.027)	(0.029)	(0.103)	(0.106)	(0.031)	(0.034)	(0.027)	(0.035)	(0.028)	(0.073)
Earned a high school diploma	2792	19	0.119	0.149	0.119	0.149	0.121	0.121	0.121	0.121	0.119	0.149	0.119	0.149	0.119	0.122
			(0.039)	(0.052)	(0.027)	(0.044)	(0.038)	(0.038)	(0.029)	(0.029)	(0.039)	(0.048)	(0.039)	(0.051)	(0.037)	(0.038)
Career Academies																
Earned HS diploma or equivalent	1533	9	0.001	0.012	0.001	0.012	0.001	0.001	0.001	0.001	0.001	0.012	0.001	0.012	-0.002	0.001
			(0.051)	(0.061)	(0.03)	(0.032)	(0.051)	(0.051)	(0.03)	(0.03)	(0.051)	(0.061)	(0.051)	(0.058)	(0.051)	(0.051)
Enrolled in postsecondary	1482	9	-0.009	-0.009	-0.009	-0.009	-0.009	-0.009	-0.009	-0.009	-0.009	-0.009	-0.009	-0.009	-0.013	-0.007
			(0.052)	(0.06)	(0.059)	(0.058)	(0.052)	(0.052)	(0.06)	(0.06)	(0.052)	(0.06)	(0.052)	(0.059)	(0.052)	(0.058)
Avg. annual earnings, yrs 1-4	1458	9	0.157	0.126	0.157	0.126	0.158	0.158	0.158	0.158	0.157	0.126	0.157	0.126	0.155	0.157
			(0.056)	(0.06)	(0.048)	(0.054)	(0.057)	(0.056)	(0.048)	(0.048)	(0.056)	(0.06)	(0.057)	(0.064)	(0.056)	(0.056)
Avg. annual earnings, yrs 5-8	1405	9	0.088	0.093	0.088	0.093	0.087	0.087	0.087	0.087	0.088	0.093	0.088	0.093	0.086	0.087
			(0.048)	(0.053)	(0.032)	(0.03)	(0.048)	(0.048)	(0.032)	(0.032)	(0.048)	(0.052)	(0.048)	(0.055)	(0.049)	(0.049)
Avg. months worked annually, yrs 1-4	1458	9	0.098	0.078	0.098	0.078	0.098	0.098	0.098	0.098	0.098	0.078	0.098	0.078	0.095	0.098
			(0.051)	(0.059)	(0.043)	(0.049)	(0.051)	(0.051)	(0.043)	(0.043)	(0.051)	(0.059)	(0.051)	(0.057)	(0.051)	(0.051)
Avg. months worked annually, yrs 5-8	1405	9	0.062	0.110	0.062	0.110	0.061	0.061	0.061	0.061	0.062	0.110	0.062	0.110	0.061	0.061
			(0.053)	(0.064)	(0.048)	(0.059)	(0.053)	(0.053)	(0.048)	(0.048)	(0.053)	(0.064)	(0.053)	(0.061)	(0.053)	(0.053)
Postsecondary education																
Learning Communities																
Targeted credits earned, sem 1	6974	11	0.166	0.099	0.166	0.099	0.171	0.171	0.171	0.171	0.166	0.099	0.166	0.099	0.171	0.116
			(0.023)	(0.031)	(0.071)	(0.053)	(0.023)	(0.024)	(0.071)	(0.073)	(0.023)	(0.031)	(0.023)	(0.033)	(0.023)	(0.058)
Cumulative targeted credits earned, sem 3	6974	11	0.084	0.033	0.084	0.033	0.087	0.087	0.087	0.087	0.084	0.033	0.084	0.033	0.087	0.054
			(0.023)	(0.031)	(0.045)	(0.036)	(0.023)	(0.023)	(0.046)	(0.047)	(0.023)	(0.031)	(0.023)	(0.033)	(0.023)	(0.041)
Total credits earned, sem 1	6974	11	0.088	0.056	0.088	0.056	0.089	0.089	0.089	0.089	0.088	0.056	0.088	0.056	0.088	0.080
			(0.023)	(0.03)	(0.028)	(0.029)	(0.023)	(0.023)	(0.029)	(0.03)	(0.023)	(0.03)	(0.023)	(0.033)	(0.023)	(0.029)
Cumulative total credits earned, sem 3	6974	11	0.030	-0.001	0.030	-0.001	0.031	0.031	0.031	0.031	0.030	-0.001	0.030	-0.001	0.029	0.031
			(0.023)	(0.028)	(0.017)	(0.024)	(0.022)	(0.023)	(0.017)	(0.018)	(0.023)	(0.028)	(0.023)	(0.032)	(0.023)	(0.023)
Performance-Based Scholarships																
Cumulative total credits earned, yr 1	6935	15	0.120	0.071	0.120	0.071	0.120	0.120	0.120	0.120	0.120	0.071	0.120	0.071	0.119	0.122
			(0.023)	(0.038)	(0.032)	(0.055)	(0.023)	(0.023)	(0.032)	(0.033)	(0.023)	(0.038)	(0.023)	(0.037)	(0.023)	(0.032)
Cumulative total credits earned, yr 3	6935	15	0.066	0.028	0.066	0.028	0.066	0.066	0.066	0.066	0.066	0.028	0.066	0.028	0.065	0.067
			(0.022)	(0.032)	(0.023)	(0.057)	(0.022)	(0.022)	(0.024)	(0.024)	(0.022)	(0.033)	(0.022)	(0.035)	(0.022)	(0.024)
Earned a degree, yr 3	6968	15	0.051	0.043	0.051	0.043	0.051	0.051	0.051	0.051	0.051	0.043	0.051	0.043	0.052	0.054
			(0.024)	(0.04)	(0.028)	(0.038)	(0.024)	(0.024)	(0.028)	(0.029)	(0.024)	(0.04)	(0.025)	(0.039)	(0.024)	(0.027)
Encouraging Summer Enrollment 1																
Enrolled in summer	7118	10	0.118	0.133	0.118	0.133	0.118	0.118	0.118	0.118	0.118	0.133	0.118	0.133	0.118	0.118
			(0.024)	(0.045)	(0.021)	(0.035)	(0.024)	(0.024)	(0.021)	(0.022)	(0.024)	(0.045)	(0.024)	(0.044)	(0.024)	(0.024)
Credits earned	7118	10	0.075	0.065	0.075	0.065	0.075	0.075	0.075	0.075	0.075	0.065	0.075	0.065	0.075	0.075
			(0.024)	(0.046)	(0.021)	(0.035)	(0.024)	(0.024)	(0.021)	(0.023)	(0.024)	(0.046)	(0.024)	(0.045)	(0.024)	(0.024)
Encouraging Summer Enrollment 2																
Enrolled in summer	7103	10	0.277	0.250	0.277	0.250	0.277	0.277	0.277	0.277	0.277	0.250	0.277	0.250	0.277	0.261
			(0.025)	(0.046)	(0.052)	(0.028)	(0.025)	(0.025)	(0.052)	(0.061)	(0.025)	(0.046)	(0.025)	(0.046)	(0.025)	(0.039)
Credits earned	7103	10	0.182	0.134	0.182	0.134	0.182	0.182	0.182	0.182	0.182	0.134	0.182	0.134	0.182	0.161
			(0.025)	(0.046)	(0.045)	(0.037)	(0.025)	(0.025)	(0.045)	(0.053)	(0.025)	(0.046)	(0.025)	(0.046)	(0.025)	(0.038)
Labor/Workforce																
Welfare-to-Work Program																
Avg. annual earnings, quarters 1-8	69399	59	0.098	0.101	0.098	0.101	0.091	0.091	0.091	0.091	0.098	0.101	0.098	0.101	0.094	0.097
			(0.009)	(0.013)	(0.014)	(0.017)	(0.009)	(0.008)	(0.013)	(0.013)	(0.009)	(0.013)	(0.009)	(0.013)	(0.008)	(0.016)

Note: Cells represent estimated effects with standard errors in parentheses below.

References

Abadie, A., S. Athey, G. W. Imbens, & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* Technical report, National Bureau of Economic Research.

Aronow, P. M., Green, D. P., & Lee, D. K. (2014). Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3), 850-871.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, 10(4), 817-842.

Bloom, H. S., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness*, 10(4), 877-902.

Cameron, A. C., J. B. Gelbach, & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics.* 90(3), 414-427.

Clark, M. A., Gleason, P., Tuttle, C. C., & Silverberg, M. K. (2011). *Do Charter Schools Improve Student Achievement? Evidence from a National Randomized Study*. Working Paper. Mathematica Policy Research, Inc.

Cochran, W. G. 1977. Sampling techniques. 3rd edn. New York: John Wiley and Sons.

Edmunds, J. A., Unlu, F., Glennie, E., Bernstein, L., Fesler, L., Furey, J., & Arshavsky, N. (2017). Smoothing the transition to postsecondary education: The impact of the early college model. *Journal of Research on Educational Effectiveness*, 10(2), 297-325.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*: Cambridge university press New York, NY, USA.

Hodges J. S. & Clayton, M. K. (2011). *Random Effects Old and New*.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, New York.

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*(1), 103-127.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*. 7(1), 295-318.

Lohr, S. L. (2019). *Sampling: Design and Analysis: Design and Analysis*. CRC Press.

Mayer, A., Patel, R., Rudd, T., & Ratledge, A. (2015). *Designing scholarships to improve college success: Final report on the performance-based scholarship demonstration.* Retrieved from New York: MDRC.

Miratrix, L. W., Sekhon, J. S. & Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 75(2), 369-396.

Miratrix, L., Sekhon, J. S., Theodoridis, A., & Campos L., (2018). Worth Weighting? How to Think About and Use Sample Weights in Survey Experiments. *Political Analysis*, 26(3), 275-291.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis Management*, 32(1), 107-121.

Orr, L.L., Olsen, R.B., Bell, S.H., Schmid, I., Shivji, A. and Stuart, E.A. (2019), Using the Results from Rigorous Multisite Evaluations to Inform Local Policy Decisions. *J. Pol. Anal. Manage.*, 38: 978-1003. doi:10.1002/pam.22154

Parsons, J., Weiss, C., & Wei, Q. (2016). Impact evaluation of the adolescent behavioral learning experience (ABLE) program. *New York: Vera Institute of Justice. https://www. vera. org/publications/rikersadolescent-behavioral-learning-experience-evaluation*.

Pashley, N. E., & Miratrix, L. W. (2020). Insights on variance estimation for blocked and matched pairs designs. *arXiv preprint arXiv*:.10342.

Pustejovsky, J. E. & Tipton, E. (2018). Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. *Journal of Business and Economic Statistics.* 36(4), 672-683.

Raudenbush, S. W., & Schwartz, D. (2020). Randomized Experiments in Education, with Implications for Multilevel Causal Inference. *Annual Review of Statistics and Its Application*, *7*(1), 177–208. http://doi.org/10.1146/annurev-statistics-031219-041205

Raudenbush, S. W., & Bloom, H. S. (2015). Learning About and From a Distribution of Program Impacts Using Multisite Trials. *American Journal of Evaluation,* 36(4), 475-499. doi:10.1177/1098214015600515

Richburg-Hayes, L., Visher, M., & Bloom, D. (2008). Do learning communities effect academic outcomes? Evidence from an experiment in a community college. *Journal of Research on Educational Effectiveness* 1(1), 33-65.

Rosenbaum, P. R. (2010). *Design of Observational Studies*: Springer New York.

Sarndal, C.-E., Swensson, B., & Wretman, J. (2003). Model assisted survey sampling. Springer.

Schochet, P. Z. (2016). *Statistical theory for the RCT-YES software: Design-based causal inference for RCTs, Second Edition (NCEE 2015–4011)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from http://ies.ed.gov/ncee/edlabs.

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

Tipton, E. & Olsen, R. (2018) A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, 47(8): 516-524.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, 10(4), 843-876.

Weiss, M. J., Bloom, H. S., Brock, T. (2014), A Conceptual Framework for Studying the Sources of Variation in Program Effects. *J. Pol. Anal. Manage.*, 33: 778-808. doi:10.1002/pam.21760

Weiss, M. J., Ratledge, A., Sommo, C., Gupta, H. (2019). Supporting Community College Students from Start to Degree Completion: Long-Term Evidence from a Randomized Trial of CUNY's ASAP. *American Economic Journal: Applied Economics*. 11(3).

Yeager, D. S., Hanselman, P., Paunesku, D., Hulleman, C., Dweck, C., Muller, C., . . . Duckworth, A. (2017). *National Study of Learning Mindsets - One Year Impact Analysis.* Retrieved from https://osf.io/tn6g4/.