Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1537

An Applied Researchers' Guide to Estimating Effects from Blocked Cluster Randomized Trials: Estimands, Estimators, and Estimates

Luke Miratrix
Harvard University

Michael J. Weiss
MDRC

Sophie Litschwartz
MDRC

Colin Hill
MDRC

Kayla Warner
MDRC

Blocked cluster randomized trials (CRTs) are widely used to evaluate educational interventions. In such trials, researchers face critical choices about how to define and estimate the average treatment effect (ATE). These choices—both the estimand (e.g., person- vs. cluster-weighted) and the estimator (e.g., regression models, multilevel models)—can influence study conclusions. This paper provides an applied guide to estimands and estimators for blocked CRTs and empirically examines their performance using 26 large-scale blocked CRTs. We estimate ATEs and standard errors for 50 outcomes using 19 estimators and compare results across and within estimands. Findings show that estimator choice can substantially affect point estimates and precision: ranges in ATE estimates exceed 0.10 SD in 18% of cases, and standard error estimates differ by more than 50% in 16% of cases. Some estimators fail or produce unstable results in common real-world designs. These findings underscore the importance of pre-specifying estimation strategies in a public pre-analysis plan. We provide practical guidance for researchers planning and analyzing blocked CRTs.

VERSION: July 2026

Suggested citation: Miratrix, Luke, Michael J. Weiss, Sophie Litschwartz, Colin Hill, and Kayla Warner. (2026). An Applied Researchers' Guide to Estimating Effects from Blocked Cluster Randomized Trials: Estimands, Estimators, and Estimates. (EdWorkingPaper: 26-1537). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/tb6j-0e41

An Applied Researchers' Guide to Estimating Effects from Blocked Cluster Randomized Trials: Estimands, Estimators, and Estimates

Luke Miratrix*
Harvard Graduate School of Education
13 Appian Way
Cambridge, MA 02138.

Michael J. Weiss* (Corresponding Author)
Michael.Weiss@mdrc.org
MDRC
200 Vesey Street, 23rd Floor
New York, NY 10281.

Sophie Litschwartz
MDRC

Colin Hill
MDRC.

Kayla Warner
MDRC

*Co-lead authors.

Abstract

Blocked cluster randomized trials (CRTs) are widely used to evaluate educational interventions. In such trials, researchers face critical choices about how to define and estimate the average treatment effect (ATE). These choices—both the estimand (e.g., person- vs. cluster-weighted) and the estimator (e.g., regression models, multilevel models)—can influence study conclusions. This paper provides an applied guide to estimands and estimators for blocked CRTs and empirically examines their performance using 26 large-scale blocked CRTs. We estimate ATEs and standard errors for 50 outcomes using 19 estimators and compare results across and within estimands. Findings show that estimator choice can substantially affect point estimates and precision: ranges in ATE estimates exceed 0.10 SD in 18% of cases, and standard error estimates differ by more than 50% in 16% of cases. Some estimators fail or produce unstable results in common real-world designs. These findings underscore the importance of pre-specifying estimation strategies in a public pre-analysis plan. We provide practical guidance for researchers planning and analyzing blocked CRTs.

An Applied Researchers Guide to Estimating Effects from Multisite Cluster Randomized Trials: Estimands, Estimators, and Estimates

Introduction[1]

Success for All (SFA) is a widely known elementary school reform that combines a rigorous reading program, whole-school reform strategies, and a focus on continuous improvement. But does SFA improve student achievement?

To answer this question, researchers conducted a blocked cluster randomized trial (CRT) as part of the U.S. Department of Education's Investing in Innovation (i3) program. Thirty-seven schools ("clusters") across five school districts ("blocks") were randomly assigned to implement SFA or serve as controls, and student outcomes were tracked over time. Blocked CRTs such as this one are a powerful tool for evaluating the effectiveness of educational interventions. Researchers typically characterize such effectiveness with estimates of the average treatment effect (ATE) for a specified population.

Even the simple concept of an ATE involves important nuances in a blocked CRT, similar to multisite (blocked) individually randomized trials, where individuals are randomized to treatment and control *within* each site (Miratrix et al., 2021).[2] For example, should we focus on the effect for the average student (person-level ATE) or the average school (cluster-level ATE)? If SFA's effect varies across schools and correlates with school size, the ATE will differ depending on the target of inference. Such distinctions can influence estimation and interpretation.

[1] Portions of this paper borrow, with permission, from a related paper by Miratrix, Weiss, and Henderson (2021).
[2] There is a tension in language here in the literature in that for multisite trials, individuals are randomized within sites (or within randomization blocks within site). The "site" here often corresponds to the "cluster" in a cluster randomized trial. Both might be a school, for example, if we randomize students within school vs. randomize whole schools. More confusingly, for blocked cluster randomized trials, the blocks themselves can be called "sites."

Another key question is whether we target the ATE for the people, clusters, and blocks in the study (i.e., finite-sample inference) or for a larger population (i.e., super-population inference). In this paper, we treat persons and clusters as sampled from a larger population, while blocks are treated as fixed, representing the population of interest. Others might instead want to view the blocks as a sample from a larger population, we unpack these choices more below.

Complicating matters further, researchers use many different *estimators* to calculate intervention effects and standard errors. Some apply regression on individual-level data with cluster robust variance estimation (e.g., see: Banerjee, Karlan, Trachtman, & Udry, 2021; Duflo, Dupas, & Kremer, 2011; Mbiti et al., 2019; Visher, Butcher, & Cerna, 2010). Others recommend regression with design-based variance estimation (Schochet, 2015). Occasionally, data are aggregated to the cluster level before regression (Andrabi, Das, & Khwaja, 2017; Word et al., 1990). Multilevel models, also called hierarchical linear models or random effects models, are another common approach (e.g., see: Clements, Sarama, Spitler, Lange, & Wolfe, 2011; Morris et al., 2014; Morris, Mattera, & Maier, 2016).

As we show in this work, when cluster and block sizes differ, these estimators can produce different ATEs and standard errors, potentially changing study conclusions. Furthermore, if treatment effects vary across clusters or blocks these differences can be driven by systematic shifts rather than mere noise. These concerns are not just theoretical; empirical work shows instances of substantial cross-site variation in intervention effects (Weiss et al., 2017).

The potential difference between person- and cluster-weighting is not new in the methodological literature. For example, Kahan, Li, Copas, and Harhay (2022) explicitly discuss person- vs. cluster-level estimands in the medical and biomedical arena, including a deep investigation of binary outcomes. Li et al. (2025) unpack informative cluster sizes (where size of

cluster is related to impact) and provide a unifying estimation framework that targets either the person-average or cluster-average regardless of the initial model fit. They do this with a second step of working with the model predictions and residuals via a survey sampling or double-robust approach. As part of this, they design a test for whether cluster size is informative, by comparing the person-weighted and cluster-weighted average effect. Schochet, Pashley, Miratrix, and Kautz (2022) use a design-based analysis of linear regression to show it can be viewed as model agnostic, providing an alternative standard error estimator based on the design-based analysis. Similarly, Su and Ding (2021) design estimators justified by random assignment to target different estimands. Wang, Park, Small, and Li (2024) consider cases where cluster size depends on treatment (e.g., due to differential attrition) while keeping the tension between estimands at the forefront. They also allow for using a machine-learning model for covariate adjustment, which could provide additional precision gains. See Li et al. (2025) for further references going back to Donner and Klar (2000) on these themes.

As is common, we follow the potential outcome framework, where we do not assume a parametric model but instead view each individual has having possible outcomes that depend on treatment. Under this framing, Imai, King, and Nall (2009) investigate pair-matching, showing how blocking can bring important gains in precision if cluster size vary. We include larger blocks, which broadens the pool of possible analytic choices. Middleton and Aronow (2015) use this framing to provide exact-unbiased estimators for person-average effects, something which is not accomplished by most other methods targeting this quantity.

We build on these literatures in two primary ways. First, we focus on *blocked* CRTs and frame the problem in terms of potential outcomes coupled with specific statements about how units are obtained from possible superpopulations. We believe this gives an accessible

framework for which to think about estimator performance and choice in most education contexts. By contrast, several of the prior works leave blocking as an implied extension. For example, Li et al. (2025) builds blocking into the assignment mechanism, with the blocks treated as a covariate in their working model. Second, we investigate how differences in estimator and framing appear in practice by comparing a range of estimators across a collection of datasets. Specifically, to assess how much the choice of estimand and estimator matters in practice, we re-analyze data from 26 large-scale blocked CRTs. For each trial, we estimate the (person weighted or cluster weighted) ATE, β, and its standard error, $SE(\beta)$, using 19 estimators across key outcomes. (See Schochet (2009) for a similar empirical comparison for non-blocked experiments.) We then compare estimates from these estimates to examine how—and under what conditions—these choices affect results. We select estimators from the design-based literature, classic regression (based on what we found by reviewing the original studies for our primary datasets), and multilevel modeling (prevalent in education). Our list is comprehensive, but not complete. We did not include some of the more novel estimators, such as machine learning (e.g., Wang et al., 2024), or generalized estimating equation approaches (see a discussion of these in Li et al., 2025). We do not use the exact-unbiased estimators of Middleton and Aronow (2015) due to found instability in our empirical examples. Our general approach follows Miratrix et al. (2021), who studied similar questions in multisite *individually* randomized trials.

Research Questions

We address three sets of questions:

1. **Impact Estimates**

To what extent does the choice of estimator influence the estimated impact ($\hat{\beta}$)?

- How much do impact estimates vary across all estimators?
- For a given estimand (person- vs. cluster-average), how much do estimates vary across reasonable estimators?

2. **Standard Errors**

To what extent does the choice of estimator for the standard error ($SE(\beta)$) affect the estimated standard error of the impact ($\widehat{SE}(\hat{\beta})$)?

- How much do standard error estimates vary across all estimators?
- For a given estimand (person- vs. cluster-average), how much do standard error estimates vary across reasonable estimators?

3. **Precision and Stability**
    - Are differences in standard error estimates primarily due to estimation error or to genuine differences in estimator precision?
    - Are there contexts in which some ATE or standard error estimators are unstable or otherwise unsuitable?

Contributions

Our work makes several contributions to the literature on blocked CRTs:

1. **Unified overview:** We consolidate commonly used estimands and estimators, concepts often discussed separately.
2. **Empirical evidence:** We apply these estimators to 26 large-scale blocked CRTs to show how much the choice of estimand and estimator can influence results in practice.
3. **Practical guidance and support:** We use these findings to provide guidance on designing studies, analyzing data, and interpreting results. We also provide an R package,

clusterRCT,[3] which implements all estimators discussed and provides additional tools for describing and exploring blocked, clustered data.

This paper is accompanied by two supplemental documents. A *Technical Supplement* provides detailed descriptions of the estimators used in this study, as well as a few others of interest, along with more detail on sampling frameworks and defining terms of interest. The *Technical Supplement* is intended as a reference document. A *Simulation Supplement* reports a series of empirically grounded simulations designed to clarify issues that are difficult to understand without full knowledge of the data-generating process.

Key Findings

How one analyzes the data matters. Both the estimated ATE and its standard error can vary substantially depending on the estimand and estimator selected. Some estimators fail to produce standard error estimates or yield unstable estimates in many real-world trials. Even after excluding these estimators and choosing an estimator aligned with an estimand, estimator choice can still meaningfully affect the estimated impact and its estimated precision. These results underscore the importance of pre-specifying estimation strategies in a public pre-analysis plan.

The Estimands

We focus on the common case where a blocked CRT has a three-level structure: persons (e.g., students) nested in clusters (e.g., classrooms or schools, which are randomized), nested in blocks (e.g., schools, districts, or other randomization blocks). Estimands differ based on how we average impacts across the persons, clusters, and blocks, and whether we treat units at each level as sampled from a larger population.

[3] R package is on GitHub at https://github.com/lmiratrix/clusterRCT

For internal consistency, we assume the same averaging rule applies within and between blocks. Specifically:

- If the target is the effect for the **average person within blocks** (i.e., weighting clusters by their size), then the overall target also weights blocks by the number of persons.
- If the target is the effect for the **average cluster within blocks** (i.e., weighting clusters equally), then the overall target weights blocks by the number of clusters, not persons.

The Finite Sample Case and Potential Outcomes

We begin with the completely finite sample case and use the potential outcomes framework (Imbens & Rubin, 2015; Rosenbaum, 2010) to state a precise description of our estimands. Each person has two potential outcomes:

$$Y_{ijk}(1), Y_{ijk}(0)$$

representing the outcome for person i in cluster j in block k under the treatment and control conditions, respectively.

We assume a variant of the Stable Unit Treatment Value Assumption, or SUTVA (Rubin, 1980): the treatment status of one cluster does not effect the outcomes in other clusters. We work with potential outcomes within each cluster as a single group, to allow for arbitrary within-cluster relationships (Middleton & Aronow, 2015; Schochet, Pashley, Miratrix, & Kautz, 2021). Any spillover within cluster is considered part of the treatment impact. Under these assumptions, the intention-to-treat (ITT) effect for a person is:

$$\beta_{ijk} = Y_{ijk}(1) - Y_{ijk}(0)$$

which exists conceptually but is unobservable. Cluster jk's average effect, β_{jk}, is then defined as the average of β_{ijk} for all people in cluster j in block k:

$$\beta_{jk} = \sum_{i=1}^{N_{jk}} \frac{1}{N_{jk}} \beta_{ijk} = \sum_{i=1}^{N_{jk}} \left[\frac{1}{N_{jk}} Y_{ijk}(1) - \frac{1}{N_{jk}} Y_{ijk}(0) \right], \quad (1)$$

where N_{jk} is the number of persons in the study in cluster j in block k.

Each cluster has its own average impact β_{jk}. The β_{jk} can vary within and across blocks. Unlike multisite individually randomized trials—where site-level impacts can be estimated because treatment and control units exist within each site—blocked CRTs cannot estimate *any* β_{jk} directly because entire clusters are assigned to treatment or control.

Within Block ATE estimands

Within each block, we consider two ATE estimands:

1. **Person-Weighted ATE (within block)**

The effect for the **average person** in block k, where each person counts equally:

$$\beta_{k,person} = \sum_{j=1}^{J_k} \frac{N_{jk}}{N_k} \beta_{jk} = \sum_{j=1}^{J_k} \frac{N_{jk}}{N_k} \bar{Y}_{jk}(1) - \sum_{j=1}^{J_k} \frac{N_{jk}}{N_k} \bar{Y}_{jk}(0) \quad (2)$$

where:

J_k = number of clusters in block k

$N_k = \sum_{j=1}^{J_k} N_{jk}$ = total persons in block k.

2. **Cluster-Weighted ATE (within block)**

The effect for the **average cluster** in block k, where each cluster counts equally:

$$\beta_{k,cluster} = \sum_{j=1}^{J_k} \frac{1}{J_k} \beta_{jk} = \sum_{j=1}^{J_k} \frac{1}{J_k} \bar{Y}_{jk}(1) - \sum_{j=1}^{J_k} \frac{1}{J_k} \bar{Y}_{jk}(0). \quad (3)$$

We present each of these estimands in two ways. First, as a weighted average of the cluster-specific ATEs (the β_{jk}), which is conceptually intuitive. Second, as a weighted average of the cluster-average potential outcome levels in the experimental arms (the $\bar{Y}_{jk}(1)$ and $\bar{Y}_{jk}(0)$), which connects directly to the estimators discussed later (see, for example, Table 2).

Overall Estimands Across Blocks

Using consistent weighting within and between blocks, we can estimate the ATE for the average person or for the average cluster, defined as follows:

1. **Fully person-weighted ATE**

Average impact for all persons—counts each person equally:

$$\beta_{person} = \sum_{k=1}^{K} \frac{N_k}{N} \beta_{k,person} = \sum_{k=1}^{K} \sum_{j=1}^{J_k} \frac{N_{jk}}{N} \beta_{jk} \quad (4)$$

$$= \sum_{k=1}^{K} \frac{N_k}{N} \left(\sum_{j=1}^{J_k} \frac{N_{jk}}{N_k} \bar{Y}_{jk}(1) - \sum_{j=1}^{J_k} \frac{N_{jk}}{N_k} \bar{Y}_{jk}(0) \right),$$

where N is the total number of people in the sample.

2. **Fully cluster-weighted ATE**

Average impact for all clusters—counts each cluster equally.

$$\beta_{cluster} = \sum_{k=1}^{K} \frac{J_k}{J} \beta_{k,cluster} = \sum_{k=1}^{K} \sum_{j=1}^{J_k} \frac{1}{J} \beta_{jk} \quad (5)$$

$$= \sum_{k=1}^{K} \frac{J_k}{J} \left(\sum_{j=1}^{J_k} \frac{1}{J_k} \bar{Y}_{jk}(1) - \sum_{j=1}^{J_k} \frac{1}{J_k} \bar{Y}_{jk}(0) \right),$$

where J is the total number of clusters in the sample.

Finite vs. Superpopulation Extensions

So far, we have assumed a finite sample perspective. These estimands extend naturally when persons, clusters, or blocks are treated as sampled from larger populations.

In this work we treat the blocks as finite because in the real-world CRTs we evaluated the blocks were frequently not policy relevant units nor could they easily be thought of as sampled from a population. Even if the blocks were policy relevant sampleable units, most real-world multisite CRTs involve too few blocks and too few clusters per block to precisely estimate the ATE for a super-population of blocks. Our *Technical Supplement* discusses a few estimators that treat the blocks as sampled, however.

Most of our estimators in practice treat the clusters as sampled from broader populations (with clusters within a block being treated as sampled from a population of clusters in that block). The estimands are then natural extensions of the above equations, and are essentially weighted averages over the full (infinite) population.[4] Other models treat the persons as also sampled from larger populations within each cluster; the estimands can again extend provided we are able to define "cluster sizes" for these infinite-sized clusters.

Importantly, whether we assume finite or superpopulation inference at the cluster or individual level often has little practical consequence. First, because treatment is assigned at the cluster level, we cannot statistically distinguish between contexts where the individuals within each cluster are considered a sample from a population and one where the individuals within each cluster are considered whole populations. In other words, at the individual level we cannot distinguish between finite and superpopulation approaches. This extends to the clusters themselves: if we think of them as the units of interest, because they are fully treated, or not, we again cannot meaningfully distinguish between the blocks being comprised of finite sets of clusters vs. clusters sampled from a larger population. For the clusters, the inability to discern the difference is akin to simple random assignment contexts, where it is sometimes referred to as the "correlation of potential outcomes problem." See, for example, Miratrix, Sekhon, and Yu (2013) or, for the case with blocking, Pashley and Miratrix (2020) and Pashley and Miratrix (2021). We discuss finite vs. superpopulation inference further in our simulations, but the consequence is that any of our standard error estimators can be viewed as conservative, if we view the clusters as

[4] With infinite populations, defining the weighting becomes a bit more subtle. With superpopulations of clusters, for example, the J_k are infinite. We then define the estimand as a function of the relative weight given to the blocks. For an infinite superpopulation of individuals, the cluster average effect is simply the expected value of the individual treatment effects in the cluster.

a finite population, or as appropriate if we consider the clusters as coming from a larger superpopulation (see *Simulation Supplement*).

Estimand Selection

Miratrix et al. (2021) outline reasons why researchers might target different estimands. We highlight a few key considerations.

Targeting the effect for the **average person** aligns with utilitarian values, where each individual counts equally. Person-weighted estimands make sense when cluster and block sizes reflect meaningful real-world structures, rather than arbitrary research artifacts.

Cluster-averaging can also be appropriate. For example, when the clusters are schools, school-level decision makers—such as principals—may care about the effect at the average school, weighting each school equally regardless of size.

In principle, the choice between finite vs. superpopulation inference can matter a great deal. Finite population inference brings focus to the specific experimental units and makes no statistical claims to the generalizability of a result. This approach suites efficacy trials, where the interest lies in assessing whether an intervention *can* work in a specific site or set of sites. Finite population inference may also be appropriate when organizations are evaluating themselves with nearly all their sites included in the study, as their interest is confined to the specific people and sites in the evaluation (Miratrix et al., 2021).

In contrast, super population inferences, which views the evaluation sample as representative of a larger whole, is often preferred for policy formulation. This approach aims to generalize findings beyond the specific study sample, arguing that the results provide insights into broader contexts.

For CRTs the choice of targeting a finite vs. superpopulation estimand is, to some extent, moot: As noted earlier, the estimators, which we describe next, all target a superpopulation of clusters (and sometimes people) and treat blocks as a fixed, finite sample. This framing usually is in tension with reality: most blocked CRTs are not conducted on random samples of clusters from clearly defined larger populations, and therefore superpopulation inference implicitly is generalizing to a vaguely defined set of units from which the study sample could reasonably have been randomly drawn. Furthermore, in many cases, the blocks themselves are not meaningful groupings, but are instead constructed as part of the experimental design, further complicating imagining the clusters as sampled (Pashley & Miratrix, 2020). We therefore take the uncertainty estimates as conservative estimates regarding the finite evaluation sample itself.

As noted earlier, estimators that treat blocks as sampled from a population exist and could be used for broader generalization. We exclude them here because blocks in most studies are artificial and few in number, making population-level inference tenuous. When samples are not representative of a clear population, finite-sample inference combined with subject-matter expertise and an understandings of mechanisms (Shadish, Cook, & Campbell, 2002) may provide a stronger basis for generalization.

Estimators of $\boldsymbol{\beta}$ and $\boldsymbol{SE(\beta)}$

Once an estimand is identified, researchers must decide how to estimate it from observed data. They also must obtain an uncertainty estimate (standard error) for the estimate. We thus have *effect estimators* and *standard error estimators*. In this section we describe major classes of effect estimators and their variants. We identify what estimand each estimator is targeting (person-average or cluster-average). We also outline approaches for variance estimation.

Table 1. Estimator names, weighting, notation, and target estimands.

Estimator	Effect Estimator Notation	Weighted Regression	Step 2 Aggregation Across Blocks	Estimand	Variance Estimation
Linear Regression on Individual Data (LR)					
Block Fixed Effects	LR_FE	No	Implicit	Person	crve
					db
	LRcw_FE	1 / Cluster Size	Implicit	Cluster	crve
					db
Fully Interacted (Block & Block x T Interactions)	LR_FIpw	No	Person-weighted	Person	crve
					db
	LRcw_FIcw	1 / Cluster Size	Cluster-weighted	Cluster	crve
					db
Linear Regression on Cluster Level Data (AR)					
Block Fixed Effects	AR_FE	No	Implicit	Cluster	het
					db
	ARpw_FE	Cluster Size	Implicit	Person	het
					db
Fully Interacted (Block & Block x T Interactions)	AR_FIcw	No	Cluster-weighted	Cluster	het
					db
	ARpw_FIpw	Cluster Size	Person-weighted	Person	het
					db
Multilevel Models (MLM)					
Random cluster intercept, block fixed effects, fixed T indicator	MLM_FE	No	Implicit	Cluster	ML
Random cluster intercept, random block intercepts, fixed T indicator	MLM_RE	No	Implicit	Cluster	ML
Random cluster intercept, fully interacted (Block & Block x T Interactions)	MLM_FIcw	No	Cluster-weighted	Cluster	ML

Notes: LR = Linear Regression on individual data; LRcw = Linear Regression on individual data weighting each observation by 1/cluster size; AR = Linear Regression on aggregated (cluster-level) data. ARpw = Linear Regression on aggregate (cluster-level) data weighting each observation by cluster size. MLM = Multilevel Model; FE = Fixed Effects; FIcw = Fully interacted (i.e., block x treatment interactions) with aggregation across blocks based on the number of clusters in the block; FIpw = Fully interacted (i.e., block x treatment interactions) with aggregation across blocks based on the number of persons in the block. crve = Cluster Robust Variance Estimation; db = Design-Based variance estimation; het = heteroskedastic robust standard errors.

Overview of Estimators

All primary estimators we consider are regression-based. Some work with individual-level data, others use cluster-level aggregates. Some regressions weight the observations; others do not. For each fitted model, we identify different ways one might calculate the standard errors.

We identified a core set of 19 estimators, summarized in Table *1*, which lists: (1) estimator form, (2) notation, (3) regression weights, (4) block-level aggregation (explicit for full-interacted models; implicit otherwise), (5) variance estimation method, and (6) target estimand.[5]

We proceed by first discussing the estimators, possible weighting, and aggregation, and then discuss variance estimation.

Linear Regression on Individual Level Data (LR)

A common estimation strategy for blocked CRTs is **linear regression on individual-level data (LR)**. In blocked CRTs, two main versions are used:

- **Block fixed effects**
- **Fully interacted models** with treatment-by-block interaction terms

The fully interacted specification allows treatment effects to vary by block, while block fixed effects assume a common treatment effect across blocks.

In both cases, researchers can apply weights to target different estimands:

- **Person-weighted**: weights observations by cluster size to target the average person.
- **Cluster-weighted**: weights each cluster equally to target the average cluster.

For uncertainty estimation, two common approaches are:

[5] We omit some estimators from the literature because they focus on cluster randomized trials, rather than blocked cluster-randomized trials, and are without guidance on how to include covariates for precision gains. Others we exclude based on performance. For example, Middleton and Aronow (2015) provide an exact unbiased estimator, but we found it was extremely unstable across our empirical examples. Su and Ding (2021) provide a range of estimators, but it is unclear how to include covariate adjustment while still aggregating across blocks. These approaches, motivated by random assignment and minimal assumptions, may become more useful as they evolve.

- **Cluster-robust variance estimation (CRVE)**
- **Design-based variance estimation (DB)**

LR with block fixed effects (LR_FE): One common estimation approach is to regress the outcome on block fixed effects, the treatment indicator, and individual or cluster-level covariates to improve precision (Bloom, Richburg-Hayes, & Black, 2007). For this approach, the Ordinary Least Squares (OLS) working model for estimating impacts is:

$$Y_{ijk} = \sum_{q=1}^{K} \alpha_q Block_{q,ijk} + \beta Z_{jk} + \gamma' X_{ijk} + \epsilon_{ijk}$$

which can also be written as:

$$Y_{ijk} \equiv \alpha_k + \beta Z_{jk} + \gamma' X_{ijk} + \epsilon_{ijk}, \quad (6)$$

where:

$Y_{ijk} =$ outcome for student i in cluster j in block k

$Block_{q,ijk}$ = indicator for block q ($= 1$ if person ijk is in block q, 0 otherwise)

$Z_{jk} =$ treatment indicator for cluster jk ($= 1$ if treated, 0 otherwise)

$X_{ijk} =$ vector of baseline covariates (individual or cluster level).

The second version of Equation (6) uses α_k instead of the sum to represent the K block fixed effects to simplify notation. This model produces the estimate $\hat{\beta}_{LR\_FE}$. LR_FE is shorthand for Linear Regression on individual data with block Fixed Effects.

To understand the fixed effect estimator, it is helpful to consider how it operates without covariates (X_{ijk}). In that case it is equivalent to a weighted average of the block-specific simple difference (sd) impact estimates, where the block-specific estimates are calculated as the average outcome for treatment group members minus the average outcome for control group members:

$$\hat{\beta}_{k,sd} = \sum_{i=1}^{N_k(1)} \frac{1}{N_k(1)} Z_{ijk} Y_{ijk} - \sum_{i=1}^{N_k(0)} \frac{1}{N_k(0)} Z_{ijk} Y_{ijk},$$

where $N_k(1)$ is the total number of program group persons in block k, $N_k(0)$ the total number in control, and Z_{ijk} is the treatment condition of person i in cluster j of block k. The overall ATE is calculated by taking the weighted average of the block specific ATEs, with weights proportional to $N_k p_k(1-p_k)$, where N_k is the total number of persons in block k and p_k is the proportion of persons treated in block k. Blocks with more people and more balanced treatment assignment receive greater weight. This implicit weighting scheme ideally results in a bias-precision tradeoff, described below.

LR_FE logically aligns with the fully person-weighted ATE, β_{Person}. However, this estimator is biased with respect to β_{Person} if there is a correlation between the block ATEs and $p_k(1-p_k)$. In practice, this bias appears minimal, as it also appeared to be in real-world multisite individually randomized trials (Miratrix et al., 2021).

This bias risk comes with potential precision advantages as this estimator is nominally "precision-weighted." Generally, block-average impacts are estimated more precisely in larger blocks and the ratio of program to control students is closer to 1:1 (under an assumption of homoskedasticity across blocks). This aligns with the fixed effects estimator weighting scheme. Thus, the FE estimators could generally be more precise then, for example, ones that just weight each block by the number of individuals in it.

The FE estimator, however, does not exactly maximize precision. The weighting of each block's ATE is by the number of *people* in the block, not the number of *clusters* in the block. If there was a block with a small number of very large clusters, we may have a large uncertainty in that block along with a large "precision" weight, which would undermine the precision-weighting goal. In other words, these weights are not necessarily capturing the actual uncertainty due to the cluster treatment assignment within a block. Thus, this estimator is a bit strange—it is

neither unbiased for any logical estimand nor necessarily optimally precise. Nonetheless, it is important to consider because (a) it is commonly used in developmental economics CRTs (where clusters are often all of similar size), (b) it is likely nearly precision weighted, despite not being optimal, and (c) its potential for bias is often minimal.

Weighted LR with block fixed effects (LRcw_FE): The fixed effects estimator in equation (6) can be adapted for targeting the cluster-weighted ATE by applying weights in the regression. Each observation is weighted inversely to its cluster size:

$$w_{ijk} = \frac{\frac{N}{J}}{N_{jk}}, \tag{7}$$

where N is the total number of persons in the sample and J is the total number of clusters in the sample. Individuals in larger clusters are down-weighted and individuals in smaller clusters are up-weighted, making the total sample weight per cluster equal. This model produces the estimate $\hat{\beta}_{LRcw\_FE}$. $LRcw\_FE$ is shorthand for Linear Regression (Cluster Weighting) on individual data with block Fixed Effects.

By weighting clusters equally, this estimator targets the average cluster effect ($\beta_{cluster}$. For this estimator, each block will be weighted as $J_k p_k^c (1 - p_k^c)$, where J_k is the number of clusters in block k and p_k^c is the proportion of the clusters in block k assigned to treatment. Blocks with more clusters and balanced treatment assignment receive greater weight, which could improve precision. This would be precision-weighted if we assume the uncertainty in all the cluster means is the same (unlikely, if the clusters are actually different sizes). Thus, this estimator in principle also offers a classic bias-precision tradeoff. Bias can arise if the average treatment effect in the blocks are correlated with the proportion of clusters treated in the blocks. Precision is prioritized because, all else equal, blocks with more clusters and with an equal number of treatment and control clusters are given more weight as their impacts are probably

estimated more precisely. But this is all without consideration of the different uncertainty induced by different sized clusters and any underlying heteroskedasticity.

LR with block by treatment interactions (LR_FI). The block fixed-effects estimator in equation (6) can be extended to allow treatment effects to vary across blocks by adding block-by-treatment interaction terms. The model becomes:

$$Y_{ijk} = \sum_{q=1}^{K} \alpha_q Block_{q,ijk} + \sum_{q=1}^{K} \beta_q Block_{q,ijk} Z_{jk} + \gamma' X_{ijk} + \epsilon_{ijk},$$

also written as:

$$Y_{ijk} = \alpha_k + \beta_k Z_{jk} + \gamma' X_{ijk} + \epsilon_{ijk}. \quad (8)$$

This specification produces K block-specific treatment effect estimates ($\hat{\beta}_q$ or $\hat{\beta}_k$), each equivalent to a difference-in-means estimator within its block when no covariates are included.

To calculate an overall ATE from equation (8), we average the K block-specific ATE estimates.

To target the fully person weighted estimand, the estimated β_ks from Equation (8) are weighted by their block sample sizes:

$$\hat{\beta} = \sum_{k=1}^{K} \frac{N_k}{N} \hat{\beta}_k, \quad (9)$$

where N_k is the number of persons in block k and N is the total sample size. This ATE estimator is notated $\hat{\beta}_{LR\_FIpw}$. LR_FIpw is shorthand for Linear Regression on individual data with Full Interactions (block-by-treatment interactions) and Person Weighted block aggregation.

This ATE estimator targets the effect for the average person (β_{person})—within blocks it weights each person equally and in aggregating across blocks it weights blocks by their sample sizes. This estimator is nearly unbiased with respect to β_{person}, with some nuances discussed in the *Technical Supplement*. By ignoring the proportion assigned to treatment when averaging the $\hat{\beta}_k$s, this estimator could be less precise than the similar fixed effect estimator (equation (6)), as

blocks with higher uncertainty could receive more weight. That said, it could be more precise as well, depending on the cluster sizes, variation within cluster, and so forth.

Weighted LR with block by treatment interactions (LRcw_FI): The fully-interacted model in equation (8) can be adapted to target the cluster weighted ATE by applying the weights described in equation (7) during estimation. After fitting the model, we aggregate the block-specific estimates β_k) using cluster counts:

$$\hat{\beta} = \sum_{k=1}^{K} \frac{J_k}{J} \hat{\beta}_k, \tag{10}$$

where J_k is the number of clusters in block k and J is the total number of clusters in the study. This ATE estimator is denoted β_{LRcw_FIcw}. $LRcw\_FIcw$ is shorthand for Linear Regression (Cluster Weighted) on individual data with Full Interactions (block-by-treatment interactions) and Cluster Weighted block aggregation. β_{LRcw_FIcw} targets the effect for the average cluster since within blocks it weights each cluster equally and in aggregating across blocks it weights blocks by their number of clusters. This estimator is unbiased with respect to $\beta_{cluster}$. By ignoring the proportion assigned to treatment when averaging $\hat{\beta}_k$s, one might expect it to be less precise than the corresponding fixed effect estimator (but it does not have to be).

Standard Error Estimation for Linear Regression on Individual Level Data

So far, we have defined four ATE estimators: unweighted and weighted linear regression on individual level data with block fixed effects and unweighted and weighted linear regression on individual level data with block fixed effects and block by treatment interactions. For each estimator, we consider two approaches to variance estimation: cluster robust variance estimation (CRVE) and design-based variance estimation (DB).

Cluster Robust Variance Estimation (CRVE). CRVE treats clusters within each block as sampled from a larger superpopulation from that block. It accounts for within-cluster

correlation by estimating the covariance matrix of residuals within clusters and aggregating across clusters. See the *Technical Supplement* for further discussion.

For fully interacted models, we first compute cluster-robust standard errors for each block-specific estimate ($\hat{\beta}_k$), then aggregate them to obtain an overall standard error.

Without covariates, we can average using the same weights we used when averaging the block ATEs to calculate the overall ATE estimate. Because randomization happens independently within block, the $\hat{\beta}_k$s are independent, giving the generic standard error formula:

$$\widehat{SE}\left(\hat{\beta}\right) = \sqrt{\sum_{k=1}^{K}(w_k)^2\widehat{SE}_k^2}, \tag{11}$$

where w_k is the weight for block k. For the person-weighted aggregation from Equation (9), $w_k = \frac{N_k}{N}$. For the cluster-weighted aggregation from Equation (10), $w_k = \frac{J_k}{J}$.

With covariates, block-specific estimates may be correlated. In this case the fitted regression provides a variance-covariance matrix capturing the overall joint uncertainty of all the parameter estimates. The standard error for the ATE estimator can be estimated by averaging this matrix, as described in the *Technical Supplement*. In short, we extract the sub-matrix describing how our impact estimates vary and covary and multiply this matrix by the vector of block weights (twice) to get an estimated uncertainty for the overall ATE.

Design-Based Variance Estimation (DB). DB estimators rely on the random assignment mechanisms and the sampling mechanism as the sources of uncertainty (Schochet, 2015). They rely on weaker assumptions than other estimators; e.g., homoskedasticity assumptions or assumptions of a random normal residual are not needed. Schochet et al. (2022) show how the (possibly weighted) linear regressions discussed above can be analyzed from a design-based perspective and how the randomization itself justifies this estimator, with or without inclusion of

covariates. This approach loses the neat algebraic expression of classical design-based estimators as simple differences in means of various quantities and is instead more of a justification of linear regression from a different set of assumptions as are typical. See our *Technical Supplement*, Section 5.5, for the formulae, along with further discussion. Also see the *Technical Supplement*'s Section 6 for an overview of purely design based approaches.

These design-based standard error estimators, presented by Schochet et al. (2022), are akin to cluster-robust standard errors, and are calculated from the residuals of the fit linear model. Schochet et al. (2022) show that random assignment justifies these standard errors as being conservative estimates of finite-sample uncertainty. Interestingly, these standard error estimators also coincide, asymptotically, with the superpopulation-targeting cluster robust standard errors (with fixed blocks) from linear regression; this underscores how the finite-sample targeting estimated standard errors are conservative. In short, Schochet et al. (2022) argues that the design-based approach offers estimators that are an improved form of CRVE estimation tailored to the context of CRTs and motivated by random assignment.

Linear Regression on Aggregate Level Data (AR)

Another approach to ATE estimation is to aggregate the data to the cluster level and then fit a linear regression model to the result. The starting point is to compute each cluster's average outcome ($\bar{Y}_{jk}$), defined in terms of person-level potential outcomes:

$$\bar{Y}_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} Y_{ij} = Z_{ij}\bar{Y}_{jk}(1) + \left(1 - Z_{ij}\right)\bar{Y}_{jk}(0). \tag{12}$$

With aggregation, we either observe the average of the treated potential outcomes ($\bar{Y}_{jk}(1)$) or control ($\bar{Y}_{jk}(0)$). Once aggregated, these average outcomes are regressed on the treatment indicator. This approach emphasizes the random assignment at the cluster level and typically treats clusters as the primary unit of analysis. Other literature (e.g., Su & Ding, 2021) also

aggregates clusters to the *total* outcome, and works with those; as they show, these essentially coincide with equation (12) and we do not consider them further in this document.

Cluster-level covariates can be directly included in the regression, and individual-level covariates may be aggregated and also incorporated. As with regression on individual-level data, we consider two versions of LR on aggregate level data: models with block fixed effects and models with block-by-treatment interactions. As before, weighted regression can be used to shift the target estimand. For standard errors, researchers typically rely on heteroskedasticity-robust or design-based standard errors. We begin by discussing the ATE estimators.

LR with block fixed effects (AR_FE)*:* The simplest estimator is a linear regression with block fixed effects:

$$\bar{Y}_{jk} = \alpha_k + \beta Z_{jk} + \delta' X_{jk} + \epsilon_{jk}. \quad (13)$$

This estimator is denoted β_{AR_FE}, where AR_FE is shorthand for linear Regression on Aggregated data with block fixed effects. β_{AR_FE} implicitly weights the blocks in proportion to $J_k p_k^C (1 - p_k^C)$, where p_k^C is the proportion of clusters treated in block k. It targets the cluster-weighted estimand, $\beta_{cluster}$. Within blocks the clusters are equally weighted and the implicit weighting across blocks includes the number of clusters in the block.

This approach is *nearly* identical to the weighted linear regression on individual level described earlier (equation (6) with weights from equation (7)) (β_{LRcw_FE}). Without individual-level covariates the two estimators coincide. When individual-level covariates are included, however, they differ slightly because the individual-level regression adjusts using person-level covariates directly, whereas the aggregated regression uses cluster-level averages, leading to minor differences in adjustment (see Schochet et al., 2022).

Weighted LR with block fixed effects (ARpw_FE). The fixed effect estimator in equation (13) can also be adapted for targeting the person weighted ATE by applying weights proportional to cluster size:

$$w_{ijk} = \frac{N_{jk}}{\frac{N}{J}}. \tag{14}$$

Larger clusters receive greater weight, ensuring that the weighted contribution of each cluster reflects its size. This estimator is denoted β_{ARpw_FE}, where $ARpw\_FE$ is shorthand for linear Regression (Person Weighted) on Aggregated data with block fixed effects. β_{ARpw_FE} targets the person-weighted estimand.

LR with block by treatment interactions (AR_FI). The block fixed-effects estimator in equation (13) can also be extended to allow treatment effects to vary across blocks by including block-by-treatment interaction terms:

$$\bar{Y}_{jk} = \alpha_k + \beta_k Z_{jk} + \delta' X_{jk} + \epsilon_{jk}. \tag{15}$$

This specification produces K block-specific treatment effect estimates ($\hat{\beta}_k$), each representing the cluster-average impact within its block. When no covariates are included, these estimates are equivalent to difference-in-means calculations on the cluster means.

To obtain an overall ATE, the block-specific estimates must be aggregated. To target the cluster weighted estimand, the $\hat{\beta}_k$ values are weighted in proportion to the number of clusters in the block (J_k) as in Equation (10). This estimator, denoted β_{AR_FIcw}, weights clusters equally within blocks and weights blocks by their cluster counts, thereby targeting the effect for the average cluster. AR_FIcw is shorthand for linear Regression on Aggregate data with Full Interactions (block-by-treatment interactions) and Cluster Weighted block aggregation.

Weighted LR with block by treatment interactions (ARpw_FI): The fully-interacted estimator in equation (15) can be adapted to target the person weighted ATE by applying the

weights defined in equation (14) during estimation. We would then need to weight the $\hat{\beta}_k$ to target the fully person weighted estimand, as in Equation (9). This estimator, denoted β_{ARpw_FIpw}, targets the effect for the average person (β_{person}) because clusters are weighted by their size within blocks and blocks are weighted by their total number of persons in the aggregation step. $ARpw\_FIpw$ is shorthand for linear Regression (Person Weighted) on Aggregate data with Full Interactions (block-by-treatment interactions) and Cluster Weighted block aggregation.

Standard Error Estimation for Linear Regression on Aggregate Level Data

For aggregate-level regressions, we consider four ATE estimators: unweighted and weighted regressions with block fixed effects, and unweighted and weighted regressions with block-by-treatment interactions. For each, standard errors can be computed using either heteroskedasticity-robust variance estimation or design-based variance estimation. We do not consider CRVE for these models because doing so would mean treating the blocks as sampled from a superpopulation, introducing challenges discussed elsewhere in this paper.

We begin with heteroskedastic robust variance estimation.

Heteroskedastic Robust Variance Estimation (het): In the traditional regression framework, covariates are treated as fixed and residuals as random. For fixed-effects models estimated on aggregated data (as in Equation (13)), blocks are considered fixed, and uncertainty arises from the residual variation in cluster-level outcomes. Heteroskedastic robust standard errors—often referred to as Huber–White errors—are designed to accommodate heterogeneity in residual variance, which is particularly relevant when clusters differ in size. Since cluster-level outcomes are averages of varying numbers of individual outcomes, residual variances may differ systematically across clusters.

This flexibility also allows for differences in outcome variance between treatment and control groups, relaxing the assumption of homogeneity across arms. Importantly, heteroskedastic robust standard errors are appropriate for finite-sample inference, where blocks are treated as fixed units (Miratrix et al., 2021).

For fully interacted models, the procedure mirrors that used with individual-level data. After estimating block-specific treatment effects ($\hat{\beta}_k$) using heteroskedastic robust standard errors, the overall standard error for the ATE is computed by aggregating these estimates. Equation (11) still applies.

Like before, with covariate adjustment, the standard errors may no longer be independent and we will need to use the entire variance-covariance matrix capturing the overall joint uncertainty of all the parameter estimates as before.

Design-Based Variance Estimation (DB): Interestingly, design-based variance estimation follows the individual level regression approach. In fact, in both cases the variance estimators work with the average residuals at the cluster level. See the individual level discussion, above, for further detail.

Multilevel Models (MLM)

Multilevel models (MLMs) are among the most widely used estimators for analyzing data from cluster randomized trials in education. These models explicitly account for the hierarchical structure of the data. MLMs tend to rely on specific models, opening the door of model misspecification undermining estimation and inference; that said, Wang et al. (2021) proves, with a few caveats, that such risks are minimal. The randomization itself ensures these models work well even under near arbitrary misspecification.

Block fixed effect (FE): A common MLM specification for blocked CRTs includes random intercepts for clusters, fixed treatment effect, and fixed effects for blocks. The model is structured as follows:

Level One (Persons)

$$Y_{ijk} = \theta_{0,jk} + \gamma' X_{ijk} + \epsilon_{ijk}, \qquad \epsilon_{ijk} \sim N(0, \sigma^2)$$

Level Two (Clusters)

$$\theta_{0,jk} = \alpha_k + \beta Z_{jk} + u_{0,jk}, \qquad u_{0,jk} \sim N(0, \tau_0^2)$$

Level Three (Blocks)

$$\alpha_k = \alpha_k.$$

The reduced form is:

$$Y_{ijk} = \alpha_k + \beta Z_{jk} + \gamma' X_{ijk} + u_{0,jk} + \epsilon_{ijk}. \tag{16}$$

This specification assumes a constant treatment effect across clusters and blocks and is denoted β_{MLM_FE}.[6] MLM_FE is shorthand for MultiLevel Model with block Fixed Effects. It is analogous to the fixed-effects linear regression model (equation (6)), though the error structure is modeled explicitly via random effects rather than left unspecified beyond the clusters being assumed independent. As a result, both the ATE and its standard error may differ from those produced by linear regression.

Unlike fixed-effects regression, which closely targets the average person within blocks, MLMs adaptively weight cluster-level outcomes. The weight assigned to each cluster depends on its size (n_{jk}) and the estimated between-cluster variance ($\hat{\tau}_0^2$), which relates to the intraclass correlation. Specifically, within blocks, MLMs approximately weight the cluster mean outcomes

[6] For consistency in notation with the linear regression models, our MLM includes covariates X_{ijk} at level one. These covariates can be individual- or cluster-level covariates. Traditionally, if they are cluster-level covariates they would be presented under level two.

proportional to $\frac{1}{\tau_0^2+\sigma^2/n_{jk}}$.[7] As τ_0^2 approaches zero, the model weights clusters by their sample size. As τ_0^2 increases, the model weights clusters equally. For a single block study, an "observed data" maximum likelihood estimator is approximately:

$$\hat{\beta}_{MLM} = \sum_{j=1}^{J} \frac{\left(\tau_0^2+\frac{\sigma^2}{n_j}\right)^{-1}}{C(1)} Z_{jk}\bar{Y}_j - \sum_{j=1}^{J} \frac{\left(\tau_0^2+\frac{\sigma^2}{n_j}\right)^{-1}}{C(0)} \left(1 - Z_j\right)\bar{Y}_j \tag{17}$$

with normalizing constant $C(1) = \sum_{j=1}^{J} Z_{jk}\left(\tau_0^2 + \frac{\sigma^2}{n_j}\right)^{-1}$, and C(0) defined analagously.

If cluster size is uncorrelated with cluster-level treatment effects (β_{jk}), this weighting does not introduce bias. However, if such a correlation exists, MLMs may be biased with respect to both person- and cluster-average estimands, depending on the magnitude of τ_0^2. Despite this, MLMs are traditionally interpreted as targeting the average cluster effect within blocks, and we adopt that interpretation here. Simulation results support this view, showing that even modest between-cluster variation tends to shift the model toward estimating cluster-average effects.

MLM with block random effects (RE): An alternative specification replaces the fixed block intercepts with random intercepts, assuming $\alpha_k \sim N(\alpha, \sigma_\alpha^2)$. If the treatment effect is not allowed to vary across blocks, this model—denoted β_{MLM_RE}, yields nearly identical estimates to β_{MLM_FE}. MLM_RE is shorthand for Multilevel Model with block Random Effects.

MLM with treatment by block interactions (FI): The MLM in Equation (16) can be extended to allow treatment effects to vary across blocks by including block-by-treatment interaction terms. This specification estimates a separate ATE for each block, producing block-specific treatment effect parameters (β_k). The model is:

[7] The authors thank and greatly appreciate Stephen Raudenbush for deriving this approximation and Equation (22). The ideas draw from Raudenbush's prior work in Raudenbush and Bryk (2002) which Raudenbush notes draws heavily from Dempster, Laird, and Rubin (1977).

Level One (Persons)

$$Y_{ijk} = \theta_{0,jk} + \gamma' X_{ijk} + \epsilon_{ijk}, \qquad \epsilon_{ijk} \sim N(0, \sigma^2).$$

Level Two (Clusters)

$$\theta_{0,jk} = \alpha_k + \beta_k Z_{jk} + u_{0,jk}, \qquad u_{0,jk} \sim N(0, \tau_0^2)$$

Level Three (Blocks)

$$\alpha_k = \alpha_k$$
$$\beta_k = \beta_k$$

The reduced form is:

$$Y_{ijk} = \alpha_k + \beta_k Z_{jk} + \gamma' X_{ijk} + u_{0,jk} + \epsilon_{ijk}. \tag{18}$$

As in Equation (16), the α_k represent block-specific intercepts, but here the β_k capture block-specific treatment effects. To obtain an overall ATE, these estimates must be aggregated. For the cluster-weighted estimand, the block-specific effects are weighted by the number of clusters in each block:

$$\hat{\beta}_{MLM\_FIcw} = \sum_{k=1}^{K} \frac{J_k}{J} \hat{\beta}_k. \tag{19}$$

This estimator is denoted β_{MLM_FIcw}, where MLM_FIcw is shorthand for Multilevel Model with Full Interactions (block-by-treatment interactions) and Cluster Weighted block aggregation. With no covariates, the standard error for this estimator can be estimated using Equation (11), with block specific standard error estimates obtained directly from the fit MLM and weights equal to J_k/J. One can also use the matrix averaging approach on the covariance matrix for the parameter estimates, as before.

MLMs can also be extended to allow treatment effects to vary randomly across blocks. This approach enables estimation of the variance in true block-level ATEs and implies treating blocks as sampled from a superpopulation. However, this specification is rarely advisable in practice. First, it substantially reduces precision when true ATEs vary across blocks, and, second, the standard error depends on estimating cross-block treatment effect variance accurately—a task

that generally requires precise block-specific estimates (Bloom, Raudenbush, Weiss, & Porter, 2017; Bloom & Spybrook, 2017), which most blocked CRTs cannot provide. Given the limited number of blocks and the difficulty of powering such models adequately, random treatment effects at the block level are generally impractical except in exceptionally large multisite trials where blocks themselves are policy-relevant units.

A Unified Representation of Effect Estimators

Purpose and relevance for applied researchers. Applied researchers often select estimators based on their discipline of training or software defaults, without a clear sense of how those estimators translate observed outcomes into an effect estimate. This lack of transparency is especially common for multilevel models (MLMs), which rely on maximum likelihood estimation, and for block fixed-effects estimators, whose implied block-level weights are rarely made explicit. In both cases, the connection between the estimator and concrete data elements—such as cluster mean outcomes and block-specific treatment balance—can be difficult to see, yet these implicit weighting choices have direct implications for interpretation and identifying potential bias.

The goal of this section is not to derive estimators, but to make their behavior transparent. When covariates are excluded, the estimators discussed in this paper can be expressed (or closely approximated) as closed-form weighted averages of cluster-level mean outcomes. Writing estimators in this unified way serves three practical purposes for applied researchers:

1. **Interpretability.** Closed-form expressions that do not rely on linear algebra or optimization routines clarify how estimators implicitly weight observed outcomes.

2. **Comparability.** The unified representation makes clear which estimators yield identical effect estimates under the no-covariate model[8]—those with the same implied weighting.
3. **Bias awareness**. By making weighting explicit, the representation clarifies when bias can arise—for example, when estimators place greater weight on blocks with particular treatment proportions.

Decomposing Effect Estimators into Weighted Averages. The unified representation expresses an ATE estimator as a weighted average of cluster mean outcomes ($\bar{Y}_{jk}(1)$ and $\bar{Y}_{jk}(0)$). Two distinct sets of weights appear:

- **Cluster-level weights** (w_{jk}) determine how clusters contribute to block-specific ATE estimates
- **Block-level weights** (b_k) determine how block-specific ATE estimates are combined to form the overall ATE.

Viewed this way, even estimators that do not explicitly produce block-level estimates can be interpreted as computing implicit block-specific ATEs and aggregating them using implied block weights.

Estimating the ATE within blocks. For block k, the estimated ATE can be written as:

$$\hat{\beta}_k = \sum_{j=1}^{J_k} w_{jk}\, Z_{jk} \bar{Y}_{jk}(1) - \sum_{j=1}^{J_k} w_{jk}\, (1 - Z_{jk}) \bar{Y}_{jk}(0) \tag{20}$$

$$\text{with } \sum_{j=1}^{J_k} w_{jk} Z_{jk} = 1 \text{ and } \sum_{j=1}^{J_k} w_{jk}(1 - Z_{jk}) = 1.$$

Depending on the estimand, the intuitive choices for w_{jk} are as follows: When targeting the effect for the average person, weight clusters proportional to their size ($w_{jk} \propto N_{jk}$). When

[8] Covariates can influence this because aggregate estimators, which use aggregate covariates, can yield different estimates compared with similar non-aggregate estimators, which use individual-level covariates.

targeting the effect for the average cluster, weight clusters equally ($w_{jk} \propto 1$). The latter is unbiased and the former is nearly unbiased.

MLMs, which target the effect for the average cluster, use adaptive cluster-level weights that fall between these extremes, depending on the estimated between-cluster outcome variance (τ_0^2). As $\tau_0^2 \to 0$, weights approach proportional-to-size weighting; as τ_0^2 increases, weights approach equal weighting of clusters. While often motivated by efficiency, these adaptive weights can induce bias if cluster size is correlated with cluster ATEs and τ_0^2 is small.

Aggregating Blocks into an Overall ATE. The overall ATE is a weighted average of block-specific effect estimates:

$$\hat{\beta} = \sum_{k=1}^{K} b_k \hat{\beta}_k, \qquad \text{with } \sum_{k=1}^{K} b_k = 1 \tag{21}$$

Here, intuitive choices are that b_k can be proportional to block sample size ($b_k \propto N_k$) when targeting the person-average effect across blocks, or proportional to the number of clusters ($b_k \propto J_k$) when targeting the cluster-average effect.

Estimators with block fixed effects use implied block weights that depend on treatment proportions within blocks—involving terms such as $p_k(1-p_k)$—which can improve precision under some conditions but may introduce bias if treatment balance is correlated with block-level effects.

Combining the two weights yields a general expression for the overall ATE estimator:

$$\begin{aligned} \hat{\beta} &= \sum_{k=1}^{K} b_k \left(\sum_{j=1}^{J_k} w_{jk} Z_{jk} \bar{Y}_{jk}(1) - \sum_{j=1}^{J_k} w_{jk} (1 - Z_{jk}) \bar{Y}_{jk}(0) \right) \\ &= \sum_{k=1}^{K} \sum_{j=1}^{J_k} b_k w_{jk} Z_{jk} \bar{Y}_{jk}(1) - \sum_{k=1}^{K} \sum_{j=1}^{J_k} b_k w_{jk} (1 - Z_{jk}) \bar{Y}_{jk}(0). \end{aligned} \tag{22}$$

The product $b_k w_{jk}$ characterizes how each cluster contributes to the overall estimate. Making these combined weights explicit is often informative. For example, when targeting the person-average effect within blocks and across blocks, with $w_{jk} = N_{jk}/N_k(1)$ and $b_k = N_k/N$, treated clusters are approximately weighted N_{jk}/pN. This shows that, across blocks, we are weighting the treated clusters by their size.

Table *2* summarizes the implied cluster- and block-level weights for all estimators discussed in this paper.

Table 2. Summary of implied weights of common effect estimators (no covariates)

Weight Name	**Within each Block k, Weight Applied to $\overline{Y}_{jk}(z)$**	**Block Weights Applied to $\widehat{\beta}_k$**	**Estimators using this Weighting**
Person-weighted	$w_{jk} \propto N_{jk}$	$b_k \propto N_k$	$\hat{\beta}_{LR\_FIpw}$, $\hat{\beta}_{ARpw\_FIpw}$
Precision weighted average of person weighted average		$b_k \propto N_k p_k(1-p_k)$	$\hat{\beta}_{LR\_FE}$, $\hat{\beta}_{ARpw\_FE}$
Precision weighted average of random-effect weighted	$w_{jk} \propto \frac{1}{\tau_0^2 + \frac{\sigma^2}{n_{jk}}}$ (approximately)	$b_k \propto N_k p_k(1-p_k)$	$\hat{\beta}_{MLM\_FE}$, $\hat{\beta}_{MLM\_RE}$
Cluster weighted average of random-effect weighted		$b_k \propto J_k$	$\hat{\beta}_{MLM\_FIcw}$
Precision weighted average of cluster weighted average	$w_{jk} \propto 1$	$b_k \propto J_k p_k^c(1-p_k^c)$	$\hat{\beta}_{LRcw\_FE}$, $\hat{\beta}_{AR\_FE}$
Cluster-weighted		$b_k \propto J_k$	$\hat{\beta}_{LRcw\_FIcw}$, $\hat{\beta}_{AR\_FIcw}$

Note: p_k refers to the proportion of individuals assigned to treatment in block k. p_k^c refers to the proportion of clusters assigned to treatment in block k.

Fully Interacted Models, Degrees of Freedom, and Overfitting

Before turning to our empirical examples, it is important to note that model overfitting can be a serious concern in blocked CRTs, especially for the five fully-interacted ATE estimators. Many of these estimators effectively operate on cluster means, so the effective degrees of freedom are closer to the number of clusters than the number of individuals. An

estimator with K block fixed effects uses K degrees of freedom. Fully-interacted estimators add another K degrees of freedom for block-by-treatment terms. In several of our empirical examples, this leaves 10 or fewer cluster-level degrees of freedom, and even fewer after including covariates for precision. Such conditions can lead to unstable estimates and, in some cases, failure to produce an overall ATE or standard error.

Blocks with "Singletons"

A related issue arises when blocks contain only a single treated cluster or a single control cluster—a common occurrence in our empirical examples. Most fully interacted estimators cannot compute uncertainty in these cases, even when overall degrees of freedom are adequate. These estimators require multiple clusters per treatment arm within each block to account for possible heteroskedasticity across blocks and treatment arms. When this condition is not met, standard errors are undefined. In most of our empirical examples, this condition is not met.

Some of the fully interacted estimators with cluster robust standard errors, as implemented in the R packages we use, technically do give a defined uncertainty estimate, but these estimates are not well identified and can thus be extremely unstable. We drop their estimates for the datasets with singletons in this paper.

The fully interacted multilevel model can give overall standard errors because it assumes homoscedasticity across blocks. Other models could be developed, e.g., ones with different residual variances for treated and control clusters, to allow for a middle ground here; we do not consider these as we have not seen them used in the literature.

In summary, standard errors that are undefined or effectively undefined when degrees of freedom are low or blocks contain singletons. In our empirical examples we summarize over

defined quantities, meaning some results are over fewer estimators than other results. We now turn to the empirical examples to which these estimators will be applied.

Empirical Examples

We analyze a convenience sample of 26 blocked cluster randomized trials (CRTs) evaluating 29 predominantly educational interventions. Several trials were multi-arm, assessing multiple interventions within the same study. All studies randomly assigned clusters—typically classrooms or schools—within blocks. The 10th and 90th percentiles for sample size characteristics are as follows: 850 to 17,500 individuals, 30 to 495 clusters, and 5 to 85 blocks. These figures indicate that our findings are most relevant to large-scale CRTs; results may be different for smaller studies. For each dataset we estimated parameters such as intraclass correlations (ICCs) at the cluster and block levels; these statistics are reported in an online *Excel Supplement*.

To facilitate implementation, we restricted analyses to datasets readily available to our team. The selected studies span a wide range of programs, from early childhood interventions to adult education. Outcomes include academic achievement, progress toward and completion of high school and college, financial stability, and health. Interventions vary substantially, encompassing, for example, a middle school cooperative math program, a financial literacy program for adults, and a college mentoring initiative.

Our core sample includes 11 well-documented blocked CRTs for which the research team already possessed or could obtain access to the data; in most cases these data came from restricted-use datasets. We augmented this set with a publicly available dataset (Gilbert, Himmelsbach, Soland, Joshi, & Domingue, 2025)[9] originally developed to study item-level

[9] These datasets are publicly available in the Item Response Warehouse (IRW, (Domingue et al., 2024)) under the prefix "gilbert_meta". https://redivis.com/datasets/as2e-cv7jb41fd

heterogeneous treatment effects (Gilbert, Kim, & Miratrix, 2023). This additional data comprises 38 trials (15 blocked, 23 unblocked), primarily international, with outcome ranging from language and math achievement to psychological and health measures. These data also allow us to examine the performance of estimators in non-blocked CRTs; results for the 23 unblocked trials are presented in online *Appendix B*.

For each intervention, we estimate the overall ATE and its standard error using the methods described earlier and summarized in Table 1. Most trials include multiple outcomes; we focus on primary outcomes, yielding 50 estimates across 29 interventions.

For comparability and simplicity, we standardized all estimates and report results in standardized effect size units (standardized to the control group standard deviation). We present estimates from estimators using a single "best" covariate (beyond blocks) to minimize the degrees of freedom issues described above. We select the best covariate as the one that explained the largest proportion of the outcome variance (this would generally be a pre-test, if there is one available). We also conducted all analyses without covariates and with a more comprehensive set of covariates (see online *Excel Supplement* those estimates).

Online *Appendix A* summarizes the 26 studies and 29 interventions, including program descriptions, target populations, research design, and the primary outcomes, with citations for further detail. Table 3 (above) provides descriptive statistics for each study. Complete estimates of effects and standard errors for all estimators appear in online *Excel Supplement*.

Table 3. Descriptive Statistics from 26 Blocked CRTs of 29 Interventions.

Intervention Name	# of Outcomes	N Persons[a]	J Clusters	K Blocks
Early Childhood				
Making Pre-K Count	2	1,389	69 Schools	16 Complex
Maternal Literacy & Child Home Activities & Materials Packet	2	14,576	8,375 Villages	12 Complex
Building Blocks - TRIAD	1	1,208	38 Schools	7 Test Scores
Building Blocks - TRIAD Scale-up	1	409	16 Schools	8 Complex
EMERGE - T1: Building Blocks	2	523	54 Class/School[b]	5 Complex
- T2: Building Blocks SSR	2	550	59 Class/School[b]	5 Complex
CARES - T1: Incredible Years	3	1,312	52 Centers	22 Grantees
- T2: Preschool PATHS	3	1,237	52 Centers	22 Grantees
- T3: Tools of Mind	2	1,281	52 Centers	22 Grantees
Elementary School				
Ganitha Kalika Andolana	1	3,207	293 Schools	26 Complex
Continuous & Comprehensive Evaluation & Learning Enhancement Program	1	11,966	400 Schools	94 Complex
Tennessee Class Size Experiment[c]	2	5,830	325 Class	79 School & grade
Model of Reading Engagement 1	2	3,365	39 Schools	7 Complex
Model of Reading Engagement 2	1	2,174	30 Schools	7 Complex
Model of Reading Engagement 3	2	1,360	30 Schools	7 Complex
The Teacher Community Assistant Initiative	3	27,201	498 Schools	40 Complex
My Math Academy	1	926	41 Teachers	16 Complex
Success for ALL	1	2,902	37 Schools	5 Districts
FasterReading	1	2,307	74 Schools	10 Complex
Teaching at the Right Level	2	9,194	400 Schools	94 Complex
Structured Pedagogy Programs	1	3,068	180 Schools	10 Complex
Middle School				
Power Teaching	1	16,902	58 Schools	5 Districts
Math PD	3	2,130	40 Schools	10 Districts
Secondary School				
Content Literacy Continuum	2	13,010	30 Schools	10 Complex
High School Financial Education	1	16,852	845 Schools	431 Complex
Postsecondary				
Beacon Mentor	3	2,114	83 Class	23 Complex
Adulthood				
Bilhar Rural Livelihoods Project	1	8,987	179 Panchayats	85 Complex
Psychiatric Care & Livelihoods Assistance	1	887	485 Villages	9 Complex
COVID-19 Mask Program	1	102,873	509 Villages & Households	225 Complex

(continued)

Minimum:	409	16	5
Median	2,307	59	12
Maximum:	102,873	8,375	431

Notes: For the multi-arm trials, sample sizes N, J, and K include the control group and those treated by the intervention listed in that row. The sample sizes for Math PD and Content Literacy Continuum are rounded to the nearest 10 due to the National Center for Education Statistics' data restrictions. [a]*Averaged across outcomes.* [b]*Random assignment was done separately for schools/centers with only one classroom (Group A) & those with two classrooms (Group B). Group A randomized schools, which were also single classrooms. Group B randomized classrooms within schools/centers.* [c]*Technically, Tennessee STAR was a three-arm trial; however, it has become common to merge the large class and large class with aide into a single control group, and we do so here.*

Results

To What Extent Can the Choice of Estimator of $\boldsymbol{\beta}$ Result in a Different $\widehat{\boldsymbol{\beta}}$?

To assess the greatest degree that estimator (or estimand) can matter, for each outcome we first consider the range of effect estimates across all 19 effect estimators ($max(\hat{\beta}) - min(\hat{\beta})$), noting that only 11 estimators are unique. For example, consider the CARES study of the *Preschool Paths* intervention and the *Competent Responses* outcome from the *Challenging Situations Task*, which measured children's social problem-solving skills. For this outcome, the smallest effect estimate was 0.08σ (from multiple estimators) and the largest effect estimate was 0.16σ (from the MLM-RE estimator), yielding a range of 0.08σ. This difference is non-trivial and illustrates how estimator choice can meaningfully affect conclusions.

Figure 1 presents the range of effect estimates for each of the 50 outcomes, along with an indicator of whether the underlying data structure involved low degrees of freedom (fewer than 10) when using the fully interacted estimators. These ranges may be a consequence of differences in estimand as well as differences in estimator.

To assess whether the ranges in Figure 1 reflect estimators targeting different estimands or variation among estimators targeting the same estimand, we also calculated, for each outcome, the range of effect estimates among estimators targeting the same estimand. Results appear in Figure 2.

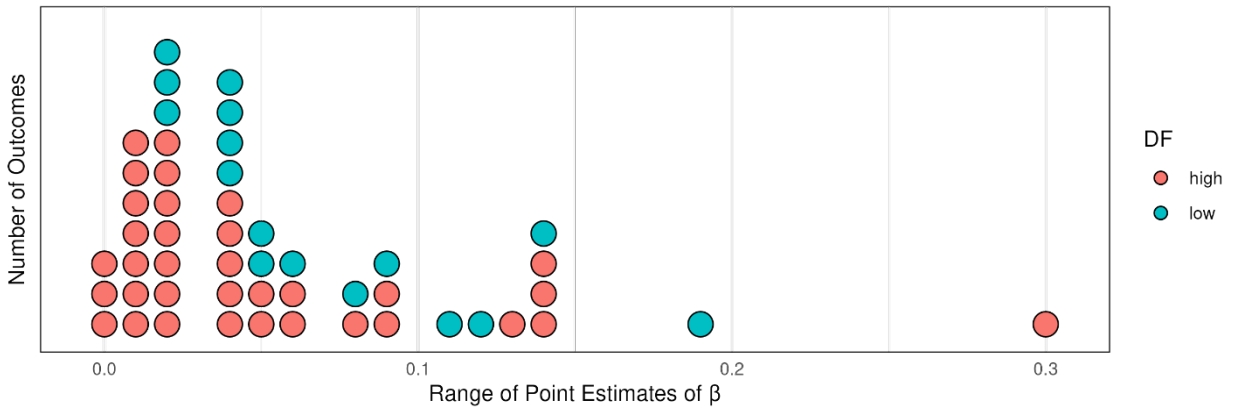


Figure 1. The Range (across 11 unique estimators) of Estimates of β for 50 Outcomes

Note: See online *Excel Supplement* for all estimates. In most cases, only 11 of the 19 estimators produce estimates. "low" df means below 10 under the conservative degree of freedom calculation of J – 2K – g, where g is the number of control covariates.

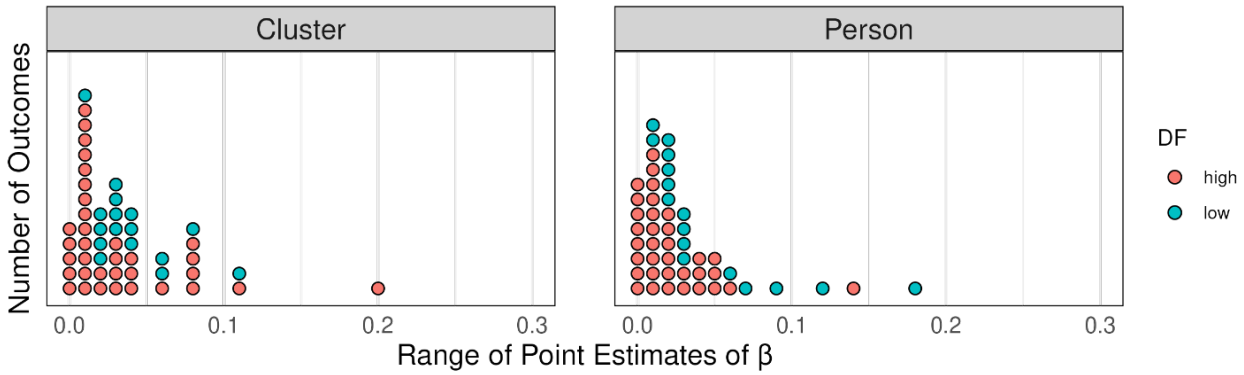


Figure 2. The Range of Estimates of β for all Estimators Targeting a Given Estimand for 50 Outcomes

Estimator choice often matters. Across estimands, in 9 cases (18%), the range of estimates exceeds 0.10σ, which we consider a large difference. In 19 cases (38%), the range exceeds 0.05σ, which we find to be non-trivial. Conditional on selecting an estimand, only 3 ranges (6%) exceed 0.10σ and 16-22% (depending on estimand) exceed 0.05σ.

Whether such differences are practically meaningful depends on the study context. For low-cost interventions where an effect of 0.05σ might be considered worthwhile, small

differences matter. In settings targeting effects of 0.25σ or more, only larger discrepancies may be clinically relevant.

Estimator choice appears to matter more in studies with low degrees of freedom. This is evident in the larger ranges observed for low–degrees-of-freedom studies (blue dots) compared with high–degrees-of-freedom studies (red dots); on average, the ranges are 36% larger in the former. Simulation results suggest that this pattern reflects greater estimator instability in low–degrees-of-freedom settings, rather than increased bias among particular estimators (see *Simulation Supplement*).

More broadly, estimator choice can meaningfully shift estimates across settings. Beyond degrees of freedom, greater variation in cluster size is associated with larger ranges, creating more scope for idiosyncratic divergence between person-weighted and cluster-weighted estimands. When cluster sizes are relatively uniform, estimates from different approaches tend to align more closely. Finally, higher overall estimation uncertainty (i.e., larger standard errors) also amplifies these differences, as estimators that respond differently to features of the data have more room to diverge.

To What Extent Can the Choice of $SE(\beta)$ Estimator result in a different $\widehat{SE}(\hat{\beta})$?

Next, we examine how the estimated standard error of the impact estimator varies by choice of estimator. For each outcome, we compute the ratio of the largest to the smallest estimated standard error across all estimators ($Ratio = max\left[\widehat{SE}(\hat{\beta})\right]/min\left[\widehat{SE}(\hat{\beta})\right]$. This ratio indicates the maximum extent to which estimator choice can influence perceived precision.

We restrict analysis to estimators with defined standard errors. Most (68%) of our studies had at least one singleton, and thus undefined standard errors for some estimators..

Consider the CARES study of *Preschool Paths* and the *Competent Responses* outcome. Among estimators with defined standard errors, the minimum is 0.06σ and the maximum is 0.082σ, a 29% increase. Here, the choice of standard error estimator matters substantially. Figure 3 shows the ratio for each of 50 outcomes.

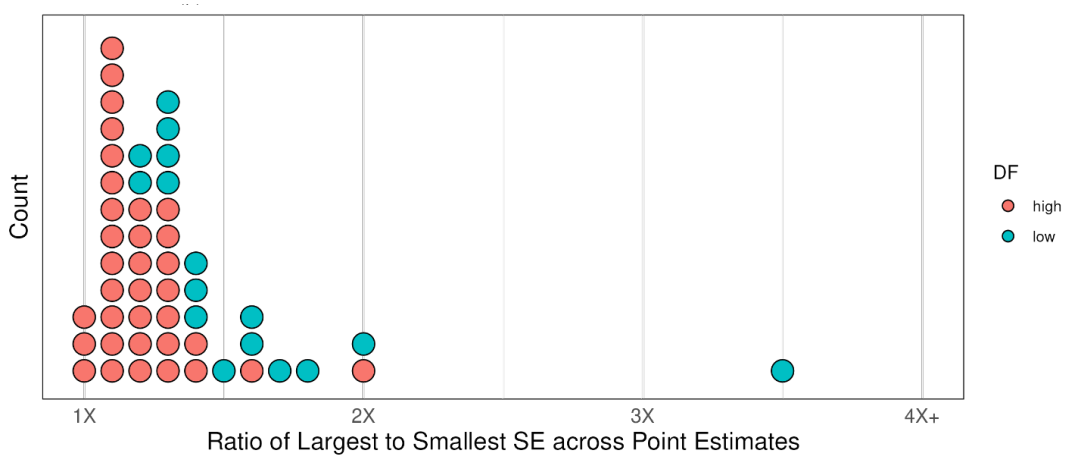


Figure 3. The Ratio (across 19 estimators) of the Largest to Smallest Estimates of $SE(\beta)$ for 50 outcomes.

As shown, the ratio from the Preschool Paths intervention example is not unusual. For eight outcomes (16%) the largest estimated standard error is more than 1.5x the smallest—a substantial difference. For three outcomes (6%), the ratio exceeds two. On average, the largest standard error is about 33% larger than the smallest across studies. Overall, we find estimator choice strongly affects the standard error estimate.

Differences could potentially be because estimating person-average effects might be more or less difficult than estimating cluster-average effect within the same experiment. To explore this, we calculate, for each outcome, the ratio of the largest to smallest estimated standard error among estimators targeting the same estimand. Results are in Figure 4.

Comparing Figure 3 with Figure 4, focusing within estimand yields only modest reductions in the ratio of the largest to smallest estimated standard error. Even after fixing the

estimand, the choice of standard error estimator can still matter a lot. For example, within estimand, eight of the fifty outcomes (16%) had ratios above 1.5, meaning the largest SE estimate was more than 50% greater than the smallest.

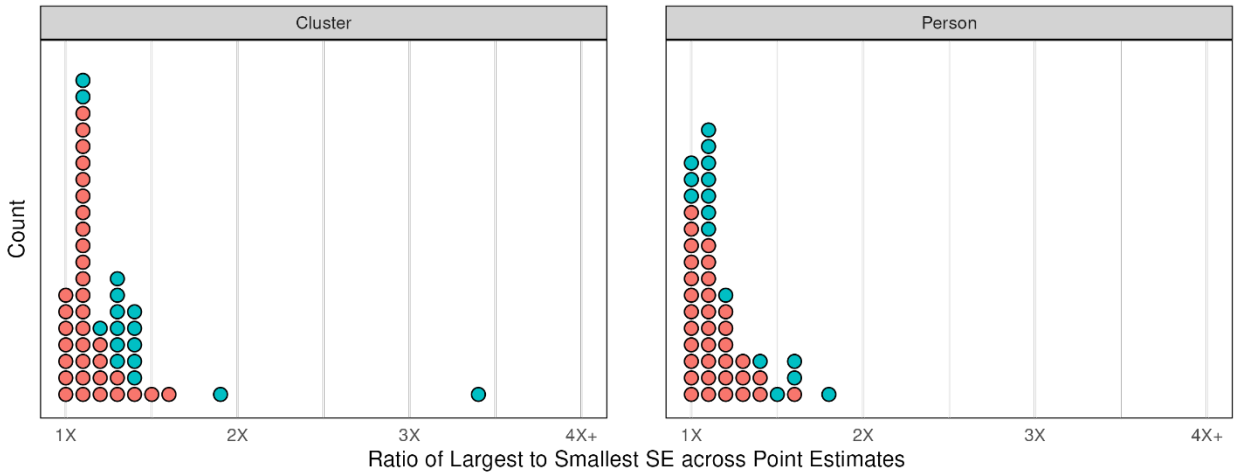


Figure 4. The Ratio (across 19 estimators) of the Largest to Smallest Estimate of $SE(\beta)$ Among Estimators Targeting a Given Estimand for 50 Outcomes

The SE variation may reflect true differences in estimator precision or differences in how estimation error is influenced by design factors. One such factor is the balance between the number of parameters and sample size. We calculate a conservative degrees of freedom for interacted estimators as J – 2K – g, with g being the number of covariates excluding the treatment indicator). In Figures 3 and 4, studies with fewer than 10 degrees of freedom are highlighted. Note that blue dots (low df) tend to appear larger than the red dots (high df): when blocks are small, variation in standard error estimates across methods is greater.

We examined the individual studies where there were notable differences in the SE ratio between the person- and cluster-targeting estimators; one stems from an extremely high uncertainty estimate from a fully interacted multilevel model (where other such models lacked defined SEs). The remaining five show no clear pattern. Overall, we found no systematic explanation for when to expect greater variation in the SE estimates across estimators.

In multisite individually randomized trials, standard error estimator choice also mattered a lot, partly because targeting the effect for the average site rather than the average person decreased precision (Miratrix et al., 2021). The reason is that, in individually randomized trials, targeting the average site down-weights larger sites (with more precise estimates) and up-weights smaller sites (with less precise estimates), reducing overall precision compared to weighting by sample size.

Interestingly, this trend did not occur for large blocked CRTs—targeting the average cluster versus the average person does not consistently affect precision. In CRTs, weighting by cluster size, if there is substantial variation in cluster size, can destabilize an estimate due to an entire cluster being moved from treatment to control. However, larger clusters also yield more precise means, so weighting by cluster size should help. These forces offset each other, and empirically, person-weighted SEs were only about 3% smaller than cluster-weighted SEs, on average. Thus, while SE estimates regularly differ across estimators, targeting the average person versus cluster does not consistently improve precision in blocked CRTs.

A brief aside: We reiterate that design-based and robust standard error estimators, when paired with a fully interacted linear regression estimator, have an important limitation. Design-based SEs allow cluster-level outcome variance to differ across experimental arms within each block—a flexible, model-free approach. However, for fully interacted models, this requires at least two clusters per arm in every block. Only 16 of 50 outcomes in our sample meet this condition; thus, design-based SEs are undefined for 34 outcomes. The same restriction applies to heteroskedastic robust and cluster-robust SEs. Consequently, for most outcomes we compare only 11 estimated SEs rather than 19 (see online *Excel Supplement*).

Additional Simulation Evidence

We supplemented our empirical analysis with two sets of simulations. The first is a large multifactor simulation to explore if and when the different estimators had different performance characteristics. The second is a calibrated simulation where, for each outcome, we ran a simulation by permuting treatment assignment in the original data to assess what differences in estimator performance we might see in practice. We next summarize our overall findings here, but please see the simulation supplement for details.

Systematic Multifactor Simulation. Overall, we found differences across estimators to be fairly minor, with all being generally unbiased and of similar precision, with a few exceptions under what we view as extreme data generating mechanisms that we did not observe in our empirical examples.

First, because MLMs target cluster-weighting, but tend towards person-weighting if the cluster-level ICC is small, they can have up to 30% smaller standard errors than the other estimators when there is *both* a large degree of variation in cluster size *and* the cluster means are not different (i.e., when the cluster-level ICC is zero); in these cases the MLMs largely ignores the clustering, which stabilizes estimates considerably, while the other cluster-average estimators are destabilized by the small clusters. This tendency towards person-weighting can also cause bias: when cluster-level ICC is zero but cluster size strongly correlates with treatment effect (a fairly peculiar circumstance) some MLM estimators can have bias in excess of 0.04.

The only other major source of bias was the fixed effect estimators, whose bias from precision weighting could grow to 0.20 (quite large), but only when there are extremely high correlations between treatment impact and proportion treated along with a very large degree of

treatment effect variation (again, not something we would normally expect in a real study). Fully interacted estimators had virtually no bias across all scenarios.

Overall, all the methods give standard errors that reasonably capture the uncertainty of a superpopulation framework with fixed blocks. When they give SEs at all, the fully interacted estimators (except MLM-FI) produce SEs that are slightly too small relative to the true superpopulation. Fixed effect estimators are well calibrated under typical conditions but can substantially overestimate when there is substantial variation in treatment effects across blocks; this is because the treatment variation gets pushed into the remaining residuals, creating uncertainty akin to a fully superpopulation framework with clusters being sampled from a single population pooled across all the blocks. MLM estimators are generally well calibrated, with some exceptions tied to unbalanced treatment coupled with treatment variation, creating a violation of the homoskedasticity assumption.

Overall, all estimated SEs are better characterized as targeting a superpopulation of clusters rather than finite-population uncertainty. That said, superpopulation and finite population uncertainty only meaningfully diverge when the cluster-level ICC is zero and there is substantial cross-cluster treatment effect variation, which seems like an unlikely thing to see in practice.

Finally, for smaller sample sizes and fewer numbers of clusters, the estimators appear to differ due to estimator instability, not bias. As said above, bias is generally small, and the sources of bias are not primarily driven by sample size concerns.

Calibrated Simulation. We wanted to explore whether the ranges in point estimate and standard error estimate across estimators were evidence for notably different person- vs cluster-average estimands. Our calibrated simulation, a simulation with no treatment impacts for any units, but

where we used the actual empirical data, showed that we would regularly see differences on the scale of our primary results. This suggests that…

Two Illustrative Examples

How do the findings from the above help in practice. To illustrate, we walk through two hypothetical examples.

For our first example, consider a blocked cluster RCT of a whole-school reform, where schools are randomized within blocks defined by district and baseline performance. Under this design, several blocks only have two or three schools. The primary audience is school principals deciding whether to adopt the program. In this setting, the relevant estimand is arguably the effect for the average school: each school represents a single adoption decision, and there is little reason for large schools to count more heavily than small ones. This perspective naturally favors estimators that weight schools more equally and estimate a school-average treatment effect. The small block sizes make fully interacted regression specifications paired with design-based heteroskedastic- or cluster-robust standard error estimators infeasible, as these variance estimators fail to produce standard errors under such block structures. We therefore must align our choice of estimator not only with the estimand of substantive interest, but also with the constraints imposed by the realized randomization design. In this case, a good estimator choice could be an aggregate linear regression with block fixed effects and heteroskedastic robust variance estimator (*AR_FE_het)* or a multi-level model with random intercepts for clusters, fixed treatment effect, and fixed effects for blocks (*MLM_FE*). If we thought treatment variation were minimal, we could also use *MLM-FI*.

For our second example, consider a different blocked cluster RCT of a whole-school reform, using a similar design but aimed at informing state-level policymakers deciding whether

to scale the intervention statewide. Our randomization approach is by district, and we are targeting a set of fairly large districts so our randomization blocks are large. Here, the estimand of interest is more naturally the effect for the *average student*, since the policy objective is to maximize aggregate student outcomes, implying choosing an estimator that weights larger schools more heavily than smaller ones. In this application, most blocks contain multiple treated and control schools, so fully interacted regression estimators are feasible. This example illustrates how reasonable differences in the intended use of evidence—rather than differences in the level of intervention delivery—can justify different estimands and, in turn, different estimators.

Conclusion

Both theory and practice show estimand and estimator choice can influence impact estimates and their precision. Some consistency can be achieved by clearly defining the estimand and selecting estimators that target it. However, even after fixing the estimand, different estimators for that estimand can yield different estimated ATEs and standard errors—potentially leading to different study conclusions. This creates substantial researcher degrees of freedom, with limited guidance for decision-making.

Overall, of the estimators considered, we do not find substantial differences in performance. In our simulations we found all the estimators have similar true precision, and bias concerns are generally minimal. While appealing due to their directly targeting an identified estimand, the fully interacted estimators are difficult to use since a sole singleton cluster in any treatment arm of any block will undercut any standard error estimate. Fortunately, the potential bias of the fixed effect estimators did not seem large. Overall, while different estimators give different estimates for a given dataset, these differences do not appear to be driven by systematic

bias: in our simulations we easily see ranges (Max($\hat{\beta}$) - Min($\hat{\beta}$)) in alignment with our empirical results, even when there is no treatment effect variation whatsoever.

Our main advice is to precisely specify an estimator in a public pre-analysis plan and stick with it. Given the risks of p-hacking and unconscious researcher bias, the best safeguard, given how different estimators can give different estimates simply due to noise, is to minimize researcher degrees of freedom with respect to estimator choice in blocked CRTs. If we were running a blocked CRT today, we would probably go with $\hat{\beta}_{LR\_FE\_CRVE}$ if targeting β_{person} due to the ease of implementation and preference not to use weighted regression without a strong justification. If targeting $\beta_{cluster}$, we would simply aggregate, and use $\hat{\beta}_{AR\_FE\_het}$, another simple and straightforward choice that makes it clear that the clusters are the units of interest.

In a departure from Schochet (2009), we *do not* recommend exploring impact estimates across model specifications as a sensitivity check. Our simulations tend to show that any choice is a valid choice, but that the standard error estimates and, to some extent, the point estimates, can vary due to different idiosyncrasies of a given dataset. Even when models are correctly specified and there is no treatment effect variation, we naturally expect variation in ATE estimates across estimators. Sensitivity analyses in this context risk creating a misleading impression that results are fragile or heavily model-dependent, undermining the rigor of the evaluation.

Acknowledgments

We would like to thank Stephen Raudenbush for deriving Equation (17).

Funding

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through R305D220046 to MDRC. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Achilles, C.M., Bain, H. P., Bellott, F. Boyd-Zaharias, J., Finn, J., Folger J., Johnston, J.; Word, E. (2008), Tennessee's Student Teacher Achievement Ratio (STAR) project, https://doi.org/10.7910/DVN/SIWH9F, Harvard Dataverse, V1, UNF:3:Ji2Q+9HCCZAbw3csOdMNdA== [fileUNF]

Andrabi, T., Das, J., & Khwaja, A. I. (2017). Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets. *American Economic Review, 107*(6), 1535-1563. doi:10.1257/aer.20140774

Banerjee, A., Karlan, D., Trachtman, H., & Udry, C. R. (2021). *Does poverty change labor supply? Evidence from multiple income effects and 115,579 bags* (Working Paper 27314). Retrieved from Cambridge, MA: https://www.nber.org/system/files/working_papers/w27314/w27314.pdf

Blair, C., & Raver, C. C. (2014). *The efficacy of an intervention synthesizing scaffolding designed to promote self-regulation with an early mathematics curriculum: Effects on executive function.* Developmental Psychology, 50(2), 565–580. https://www.researchgate.net/publication/258932213 (researchgate.net in Bing)

Bloom, H. S., Raudenbush, S., 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. doi:https://doi.org/10.1080/19345747.2016.1264518

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using Covariates to Improve Precision for Studies that Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

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.

Clements, D., Sarama, J., Spitler, M., Lange, A., & Wolfe, C. (2011). Mathematics Learned by Young Children in an Intervention Based on Learning Trajectories: A Large-Scale Cluster Randomized Trial. *Journal for Research in Mathematics Education, 42*, 127-166. doi:10.5951/jresematheduc.42.2.0127

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1-38. Retrieved from http://www.jstor.org/stable/2984875

Domingue, B., Braginsky, M., Caffrey-Maffei, L., Gilbert, J., Kanopka, K., Kapoor, R., . . . Zhang, L. (2024). Solving the problem of data in psychometrics: An introduction to the Item Response Warehouse (IRW). In: PsyArXiv.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research.* (Vol. 27). London: Arnold.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review, 101*(5), 1739-1774. doi:10.1257/aer.101.5.1739

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., & Domingue, B. W. (2025). Estimating heterogeneous treatment effects with item-level outcome data: Insights from Item Response Theory. *Journal of Policy Analysis and Management*. doi:10.1002/pam.70025

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023). Modeling Item-Level Heterogeneous Treatment Effects With the Explanatory Item Response Model: Leveraging Large-Scale Online Assessments to Pinpoint the Impact of Educational Interventions. *Journal of Educational and Behavioral Statistics, 48*(6), 889-913. doi:10.3102/10769986231171710

Grossman, J. B. Evaluation of Success for All PowerTeaching in Middle School Grades, United States, 2012-2016. Inter-university Consortium for Political and Social Research [distributor], 2018-06-06. https://doi.org/10.3886/ICPSR37046.v2

Herrera, C., Grossman, J. B., Kauh, T. J., Feldman, A. F., & McMaken, J. (2007). *Making a difference in schools: The Big Brothers Big Sisters school-based mentoring impact study.* MDRC. https://www.mdrc.org/sites/default/files/full_382.pdf

Hofer, K. G., Lipsey, M. W., Dong, N., & Farran, D. C. (2013). *Results of the Early Math Project: Scale-Up cross-site results* (Working paper). Vanderbilt University, Peabody Research Institute. https://cdn.vanderbilt.edu/vu-my/wp-content/uploads/sites/1185/2019/04/14113258/Early-Math-Cross-Site-Technical-Report_Working-Paper.pdf

Imai, K., King, G., & Nall, C. (2009). The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation. *Statistical Science, 24*(1), 29-53, 25. Retrieved from https://doi.org/10.1214/08-STS274

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

Kahan, B. C., Li, F., Copas, A. J., & Harhay, M. O. (2022). Estimands in cluster-randomized trials: choosing analyses that answer the right question. *International Journal of Epidemiology, 52*(1), 107-118. doi:10.1093/ije/dyac131

Li, F., Tong, J., Fang, X., Cheng, C., Kahan, B. C., & Wang, B. (2025). Model-Robust Standardization in Cluster-Randomized Trials. *Statistics in medicine, 44*(20-22), e70270. doi:https://doi.org/10.1002/sim.70270

Mattera, S. K., Jacob, R., & Morris, P. A. (2018). *Strengthening children's math skills with enhanced instruction: The impacts of Making Pre-K Count and High 5s on kindergarten outcomes.* MDRC. https://www.mdrc.org/sites/default/files/MPC-High_5s_Impact_FR_0.pdf

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania*. *The Quarterly Journal of Economics, 134*(3), 1627-1673. doi:10.1093/qje/qjz010

Middleton, J., & Aronow, P. (2015). Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments. *Statistics, Politics and Policy, 6*(1-2).

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. doi:10.1111/j.1467-9868.2012.01048.x

Miratrix, L. W., Weiss, M. J., & Henderson, B. (2021). An Applied Researcher's Guide to Estimating Effects from Multisite Individually Randomized Trials: Estimands, Estimators, and Estimates. *Journal of Research on Educational Effectiveness*, 1-39. doi:10.1080/19345747.2020.1831115

Morris, P. A., Mattera, S. K., Castells, N., Bangser, M., Bierman, K., & Raver, C. (2014). *Impact Findings from the Head Start CARES Demonstration National Evaluation of Three Approaches to Improving Preschoolers' Social and Emotional Competence*. Retrieved

from New York: https://eric.ed.gov/?q=Impact+Findings+from+the+Head+Start+CARES+Demonstration&id=ED546649

Morris, P. A., Mattera, S. K., & Maier, M. F. (2016). *Making Pre-K Count: Improving Math Instruction in New York City*. Retrieved from New York: https://eric.ed.gov/?q=Making+pre-K+count&id=ED569994

Morris, P.. Head Start CARES Demonstration: National Evaluation of Three Approaches to Improving Preschoolers' Social and Emotional Competence, 2009-2015. Inter-university Consortium for Political and Social Research [distributor], 2017-03-06. https://doi.org/10.3886/ICPSR35510.v3

National Center for Education Statistics. (2005). *Middle School Mathematics Professional Development Impact Study.* NCES. https://ies.ed.gov/use-work/evaluations/middle-school-mathematics-professional-development-impact-study

National Center for Education Statistics. (2012). *Improving Adolescent Literacy Across the Curriculum in High Schools (Content Literacy Continuum, CLC)*. NCES. https://nces.ed.gov/use-work/dataset/improving-adolescent-literacy-across-curriculum-high-schools-content-literacy-continuum-clc

Pashley, N., & Miratrix, L. (2020). Insights on Variance Estimation for Blocked and Matched Pairs Designs. *Journal of Educational and Behavioral Statistics, 46*(3), 271-296. doi:10.3102/1076998620946272

Pashley, N., & Miratrix, L. (2021). Block What You Can, Except When You Shouldn't. *Journal of Educational and Behavioral Statistics*.

Quint, J.. Impacts and Implementation of the i3-Funded Scale-Up of Success for All. Inter-university Consortium for Political and Social Research [distributor], 2016-09-12. https://doi.org/10.3886/ICPSR36387.v1

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Newbury Park, CA: Sage Publications.

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

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association, 75*(371), 591-593. doi:10.2307/2287653

Schochet, P. Z. (2009). Technical Methods Report: The Estimation of Average Treatment Effects for Clustered RCTs of Education Interventions (NCEE 2009-0061 rev.). *National Center for Education Evaluation and Regional Assistance*.

Schochet, P. Z. (2015). Statistical Theory for the RCT-YES Software: Design-Based Causal Inference for RCTs. *Mathematica Policy Research, Inc.*

Schochet, P. Z., Pashley, N., Miratrix, L., & Kautz, T. (2021). Design-Based Ratio Estimators and Central Limit Theroems for Clustered, Blocked RCTs. *Journal of the American Statistical Association*. doi:10.1080/01621459.2021.1906685

Schochet, P. Z., Pashley, N. E., Miratrix, L. W., & Kautz, T. (2022). Design-Based Ratio Estimators and Central Limit Theorems for Clustered, Blocked RCTs. *Journal of the American Statistical Association, 117*(540), 2135-2146. doi:10.1080/01621459.2021.1906685

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

Su, F., & Ding, P. (2021). Model-Assisted Analyses of Cluster-Randomized Experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 83*(5), 994-1015. doi:10.1111/rssb.12468

Visher, M., Butcher, K. F., & Cerna, O. S. (2010). *Guiding Developmental Math Students to Campus Services: An Impact Evaluation of the Beacon Program at South Texas College.* Retrieved from New York: http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED517927

Wang, B., Harhay, M. O., Tong, J., Small, D. S., Morris, T. P., & Li, F. (2021). On the mixed-model analysis of covariance in cluster-randomized trials. *arXiv preprint arXiv:2112.00832*.

Wang, B., Park, C., Small, D. S., & Li, F. (2024). Model-robust and efficient covariate adjustment for cluster-randomized experiments. *J Am Stat Assoc, 119*(548), 2959-2971. doi:10.1080/01621459.2023.2289693

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. doi:10.1080/19345747.2017.1300719

Word, E., Johnston, J., Bain, H., Fulton, D., Boyd-Zaharias, J., Lintz, M., . . . Breda, C. (1990). Final Report: Student/teacher achievement ratio (STAR): Tennessee's K-3 class-size study. *Nashville, TN: Tennessee State Department of Education.*

Online Appendix A – Studies

Appendix A. *Studies*

Ganitha Kalika Andolana (GKA)	
Program	A program that promotes mathematics learning in government primary schools.
Target Pop.	4th Grade
Study Design	Schools (i.e., clusters) were randomly assigned to either participate in the program or to a business-as-usual control condition.
Outcomes	Math
Size	Individuals: 3,207 students Clusters: 293 schools Blocks: 26
Report	de Barros, et al. (2024)
Beacon Mentor	
Program	A light touch mentoring program
Target Pop.	College students enrolled in a lower-level math course.
Study Design	Classrooms (i.e., clusters) were randomly assigned within blocks either to receive the mentor program or to a business-as-usual control condition where no mentor was assigned. Blocks were defined by (1) developmental or college-level course, (2) teacher, (3) time of day.
Outcomes	GPA, Credits Earned, Postprogram persistence (%), whether a student registered spring 2008 and fall 2008
Size	Individuals: 2,165 students Clusters: 83 classrooms Blocks: 40
Report	Visher, Butcher, & Cerna (2010)
Building Blocks - TRIAD	
Program	Math intervention based on research-based learning trajectories; includes teacher PD & class materials.
Target Pop.	Pre-K students living in urban, diverse, low-income communities
Study Design	Schools (i.e., clusters) were randomly assigned to one of three groups. Random assignment was done within districts and math test scores (i.e., blocks). The two groups receiving the intervention differed only in that one included a follow-through component implemented when the students moved to kindergarten and first grade.
Outcomes	REMA Rasch Score (Pre-K Post)
Size	Individuals: 1,305 students Clusters: 42 schools Blocks: 9
Report	Clements et al. (2011)
Building Blocks - TRIAD Scale-Up	
Program	Pre-K teacher math training program and children's mathematical learning curriculum based on learning trajectories.
Target Pop.	Pre-K teachers and students living in urban, diverse, low-income communities
Study Design	Schools (i.e., clusters) were randomly assigned to the treatment group or to a buisness as usual control condition within groups (i.e. blocks) defined by school characteristics.
Outcomes	REMA Rasch Score (Pre-K Post)
Size	Individuals: 409 students Clusters: 16 schools Blocks: 8
Report	Hofer et al. (2013)

Bilhar Rural Livelihoods Project	
Program	A program that provides communities with sustainable livelihood opportunities.
Target Pop.	Adults
Study Design	Panchayats (i.e, clusters) were randomly assigned to either pariticipate in the program or to a a buisness- as-usual control condition.
Outcomes	Indebtedness
Size	Individuals: 8,987 adults Clusters: 179 panchayats Blocks: 85
Report	Hoffmann, et al. (2018)
CARES	
Program	Three distinct classroom-based approaches to enhancing children's social-emotional development
Target Pop	Pre-K students living in low-income communities.
Study Design	Centers (i.e., clusters) were randomly assigned to one of the three enhancements (the three program groups) or to a control group that continued with "business as usual." Random assignment was done within groups (i.e., blocks) made up of similar centers within grantees.
Incredible Years Outcomes	Executive function (Pencil Tap task), Behavior problems (Behavior Problems Index), Learning behaviors (Cooper-Farran Behavioral Rating Scales: work-related skills subscale)
Preschool PATHS Outcomes	Emotion knowledge (facial emotions identification task), Social problem-solving skills (Challenging Situations Task: competent response), Social behaviors (Social Skills Rating Scale)
Tools of the Mind Outcomes	Executive function (Pencil Tap task), Learning behaviors (Cooper-Farran Behavioral Rating Scales: work-related skills subscale)
Size	Individuals: 2,670 students Clusters: 104 centers Blocks: 22 grantees
Report	Morris et al. (2014)
Content Literacy Continuum	
Program	A literacy and content learning program, combining whole-school and targeted approaches.
Target Pop.	High school students living in low-income communities.
Study Design	High schools (i.e., clusters) within districts and other groupings (i.e., blocks) were randomly assigned either to implement CLC or to a business-as-usual control condition.
Outcomes	GRADE comprehension composite standard score, Cumulative credits earned as a percentage of credits required for graduation: ELA, Math, Science, History (see documentation for details)
Size	Individuals: 16,640 students Clusters: 30 schools Blocks: 10
Report	Corrin et al. (2012)
Continuous & Comprehensive Evaluation & Learning Enhancement Program	
Program	Programs that offers alternative approaches to teaching and student achievement evaluation.
Target Pop.	Elementary Students
Study Design	Schools (i.e., clusters) were randomly assigned to one of four groups that received CCE alone, LEP alone, CCE and LEP together, or a buisness-as-usual control condition. An additional 100 schools were randomly assigned to receive either CCE alone or no treatment.
Outcomes	Academic Acheivement
Size	Individuals: 11,966 students Clusters: 400 schools Blocks: 94
Report	Duflo, et al. (2015)

COVID-19 Mask Program	
Program	A program that promotes use of COVID-19 masks.
Target Pop.	Households
Study Design	Villages and households were randomized to either recieve the benefits the program or to a buisness- as-usual control condition.
Outcomes	COVID Symptoms
Size	Individuals: 102,873 people Clusters: 509 villages and households Blocks: 255
Report	Abaluck, et al. (2022)
EMERGE	
Program	Two programs, a math curriculum & a math curriculum synthesized with scaffolding of play to promote executive function.
Target Pop.	Pre-K students living in diverse, low-income, and ELL communities.
Study Design	Classrooms (i.e., clusters) in schools or early childhood centers were randomly assigned to one of three conditions (either one of the two treatment groups or a business-as-usual control condition). Random assignment was done separately for schools/centers with only one classroom (Group A) & those with two classrooms (Group B). Group A was randomized in five blocks (defined by half-day vs full-day classes & by district).
Outcomes	TEAM- Scaled Score (Total) (Pre-K Post), ECLS-B Raw Score (Pre-K Post)
Size	Individuals: 1,091 students Clusters: 84 classrooms Blocks: 5
Report	Clements et al. (2020)
FasterReading (FR)	
Program	A literacy program that includes a phonics-based approach to reading instruction.
Target Pop.	Elementary students
Study Design	Eligibile schools (i.e., clusters) were randomly assigned within stratum to either participate in the program or to a buisness- as-usual control condition.
Outcomes	Literacy
Size	Individuals: 2,307 students Clusters: 74 schools Blocks: 10
Report	Sebele, et al. (2023)
High School Financial Education	
Program	A comprehensive high school financial education program.
Target Pop.	11th Grade Students
Study Design	Schools (i.e., clusters) were randomly assigned to either receieve financial education material and teacher training or to a buisness- as-usual control condition.
Outcomes	Financial Literacy
Size	Individuals: 16,852 students Clusters: 845 schools Blocks: 431
Report	Bruhn, et al. (2016)
Making Pre-K Count	
Program	Enhanced math instruction aligned across prekindergarten (pre-K) and kindergarten.
Target Pop.	Pre-K & kindergarten students living in low-income communities in New York City.
Study Design	Pre-K programs (i.e., clusters) were randomly assigned to the treatment group or to a buisness as usual control condition. Programs were randomized within groups of 4 to 5 schools (i.e., blocks). Blocking occurred by borough, venue (public vs. community org), & racial/ethnic composition.
Outcomes	ECLS-B Scale Score (Pre-K Post); WJ W- Applied Problems (Pre-K Post)
Size	Individuals: 1,389 students Clusters: 69 Pre-K programs Blocks: 16
Report	Mattera, Jocob, & Morris (2018)

Maternal Literacy & Child Home Activities & Materials Packet	
Program	Programs that supports adult literacy and parental involvement.
Target Pop.	Children
Study Design	Villages were randomly assigned to one of four groups. In the first group, mothers in the village were offered the Maternal Literacy (ML) program, consisting of daily language and math classes. In the second, mothers were given the Child Home Activities and Materials Packet (CHAMP) program: materials, activities, and training were provided each week to promote enhanced involvement in their children's education at home. In the third, mothers were offered both the literacy and home learning programs (ML-CHAMP). The fourth group served as a control with no intervention.
Outcomes	Language, Math
Size	Individuals: 14,576 children Clusters: 8,375 villages Blocks: 12
Report	Banerji, et al. (2017)
Math PD	
Program	A professional development program for mathematics teachers
Target Pop.	Seventh graders living in low-income communities.
Study Design	Schools (i.e., clusters) within districts and within groups based on similar characteristics within districts (i.e., blocks) were randomly assigned either to receive the PD program or to a business-as-usual control condition.
Outcomes	NWEA Total Scaled Score (Spring 2009), NWEA Fractions & Decimals Scaled Score (Spring 2009), NWEA Ratio & Proportions Scaled Score (Spring 2009)
Size	Individuals: 2,130 students Clusters: 40 schools Blocks: 10 districts
Report	Garet et al. (2011)
Model of Reading Engagement 1	
Program	A content literacy program.
Target Pop.	1st and 2nd Grade Students
Study Design	Elementary schools (i.e., clusters) were randomized to either recieve content literacy lessons in Grades 1 through 3, or a control condition of lessons only in Grade 3. schools were blocked (i.e., stratified) by demographic and achievement characteristics and then randomized to conditions for two consecutive years.
Outcomes	Reading Self Concept, Vocabulary
Size	Individuals: 3,365 students Clusters: 39 schools Blocks: 7
Report	Kim, et al. (2021)
Model of Reading Engagement 2	
Program	A content literacy program.
Target Pop.	2nd Grade Students
Study Design	Elementary schools (i.e., clusters) were randomized to either recieve content literacy lessons in Grades 1 through 3, or a buisness-as-usual control condition. schools were blocked by three levels of school size (i.e., small, medium, large schools based on total student enrollment), two levels of prior reading proficiency (i.e., above and below the median on end-of-grade 3 reading performance), and prior experience with MORE and then randomized to treatment or control conditions for the 12-month study period.
Outcomes	Reading Comprehension
Size	Individuals: 2174 students Clusters: 30 schools Blocks: 7
Report	Kim, et al. (2023)

Model of Reading Engagement 3	
Program	A content literacy program.
Target Pop.	Elementary Students
Study Design	Elementary schools (i.e., clusters) were randomized to either recieve content literacy lessons in Grades 1 through 3, or a control condition of lessons only in Grade 3. Schools were blocked by three levels of school size (i.e., small, medium, and large schools based on total student enrollment), two levels of prior reading proficiency (i.e., above and below the median on end-of-Grade 3 reading performance), and prior experience with MORE and then randomized to a treatment (full MORE spiral curriculum from Grades 1 to 3) or a control condition.
Outcomes	Vocabulary, Reading Comprehension
Size	Individuals: 1,360 students Clusters: 30 schools Blocks: 7
Report	Kim, et al. (2024)
My Math Academy	
Program	A program that provides assessments to support existing curricula.
Target Pop.	Kindergarteners and 1st Grade students
Study Design	Teachers were randomly assigned to one of two conditions: a treatment condition where teachers would implement the program, or a control condition where teachers would implement the math curricula and interventions the district had in place. Schools and grade were blocked.
Outcomes	Math
Size	Individuals: 926 students Clusters: 41 teachers Blocks: 16
Report	Bang, et al. (2023)
PowerTeaching	
Program	A cooperative math learning program
Target Pop.	Middle school students living in urban areas.
Study Design	Schools (i.e., clusters) within districts (i.e., blocks) were randomized to receive PowerTeaching or to a business-as-usual control condition.
Outcomes	Grade 7 standardized state math test score
Size	Individuals: 17,002 students Clusters: 58 schools Blocks: 5 districts
Report	Rappaport et al. (2017)
Psychiatric Care & Livelihoods Assistance	
Program	A program that treats depression.
Target Pop.	Adults
Study Design	Villages (i.e., clusters) were randomly assigened to one of four groups that received PC alone, LA alone, PC and LA together, or a buisness-as-usual control condition.
Outcomes	Depression (PHQ-9)
Size	Individuals: 887 adults Clusters: 485 villages Blocks: 9
Report	Angelucci, et al. (2024)
Structured Pedagogy Programs	
Program	Programs that engages students in cross-linguistic transfer.
Target Pop.	Early Elementary Students
Study Design	Schools (i.e., clusters) were randomly assigned to either participate in the program or to a buisness- as-usual control condition.
Outcomes	Oral Reading Fluency
Size	Individuals: 3,068 students Clusters: 180 schools Blocks: 10
Report	Mohohlwane, et al. (2023)

Success for All	
Program	Whole-school reform including enhanced reading instruction.
Target Pop.	Elementary school students in low-income communities.
Study Design	Schools (i.e., clusters) within districts (i.e., blocks) were randomly assigned to either receive treatment in all grades or continue with business-as-usual.
Outcomes	Woodcock-Johnson Letter-Word Identification
Size	Individuals: 5,015 students Clusters: 38 schools Blocks: 5 districts
Report	Quint, Zhu, Balu, Rappaport, Delaurentis (2015)
The Teacher Community Assistant Initiative	
Program	A student acheivement program.
Target Pop.	1st, 2nd, and 3rd Grade Students
Study Design	One hundred schools (i.e., clusters) were randomly allocated into each of the five treatment designations (four treatment arms and a control arm), stratified by region, above/below median average baseline student test score, and above/below median pupil teacher ratio.
Outcomes	Math, English, Local Language
Size	Individuals: 27,201 students Clusters: 498 schools Blocks: 40
Report	Duflo, et al. (2024)
Teaching at the Right Level	
Program	A program that reorganizes groups of school children based on level.
Target Pop.	1st, 2nd, 3rd, and 4th Graders
Study Design	Schools (i.e., clusters) were randomly assigned to either participate in the program or to a buisness-as-usual control condition.
Outcomes	Hindi, Math
Size	Individuals: 9,194 students Clusters: 400 schools Blocks: 94
Report	Banerjee, et al. (2017)
Tennessee Class Size Experiement	
Program	Small class size.
Target Pop.	Elementary school students in rural, urban, suburban, and inner-city schools.
Study Design	Classrooms (i.e., clusters) were randomly assigned into three experimental arms, within each unique school & grade (i.e., blocks).
Outcomes	Scholastic Aptitude Test (SAT) total math scaled score kindergarten, Scholastic Aptitude Test (SAT) total reading scaled score kindergarten
Size	Individuals: 11,601 students Clusters: 325 classrooms Blocks: 79 schools
Report	Word et al. (1990)

Note: The sample sizes for Content Literacy Continuum and Math PD are rounded to the nearest 10 due to the National Center for Education Statistics' data restrictions.

Online Appendix B – Results for Non-Blocked Experiments

We replicate our core analyses in the main paper with a set of identified CRTs, taken from the collection of datasets assembled by Gilbert, where there was no blocking. For unblocked experiments, our set of estimators is reduced. In particular, the fully interacted estimators are no longer applicable, as there is no blocking variable to interact treatment with. This leaves 9 estimators. For point estimators, we have individual regression, individual regression with weights to target the cluster average, aggregated regression, and aggregated regression with weights to target the person average. For each of these we have design based or robust standard errors. We finally have a multilevel model, with random effects at the cluster level.

In our collection of datasets, we have 38 outcomes across 23 studies. Most of the CRTs had a large number of clusters (an average of around 95 clusters), but a good quarter of the studies had 25 clusters or fewer; for these small studies, we would generally expect high levels of uncertainty, and we would also expect cluster robust estimators to be fairly unstable. Across the middle 80% of the studies, the average cluster size ranges 10 to 47 units, with a median of about 20 units per cluster. One very small study (with psychological distress as an outcome) had an average of less than 2 units per cluster, and another large study on reading had an average of 71.

As with the main paper, we can examine the range of point estimates, finding that most studies have only a modest range across estimators:

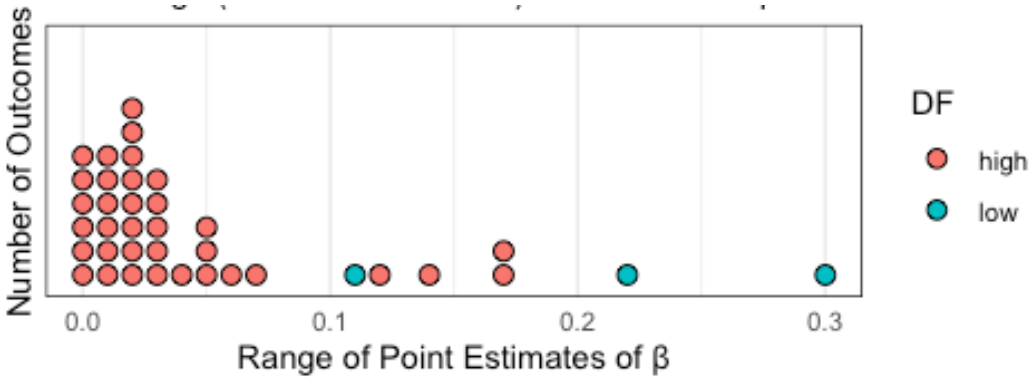


Figure 1. The Range (across 9 estimators) of Estimates of β for 38 Outcomes

Note: "low" df means below 10 under the conservative degree of freedom calculation of J – 2K – g, where g is the number of control covariates.

Across the outcomes, we have 23% of the ranges being 0.05σ or larger, and 18% 0.10σ or larger. As before, if we target a Person vs Cluster average estimand, we find that the range of estimates often shrinks. Across outcomes, the average reduction is around 46%, but 25% of the outcomes had a reduction of 20% or less.

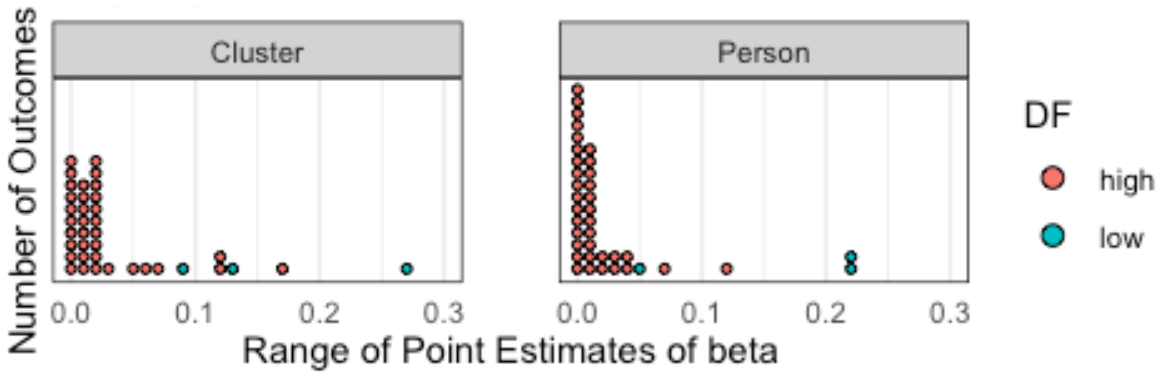


Figure 2. The Range of Estimates of β for all Estimators Targeting a Given Estimand for 38 Outcomes

Note: "low" df means below 10 under the conservative degree of freedom calculation of J – 2K – g, where g is the number of control covariates.

Just as with the blocked experiments, the standard errors range substantially across our estimators:

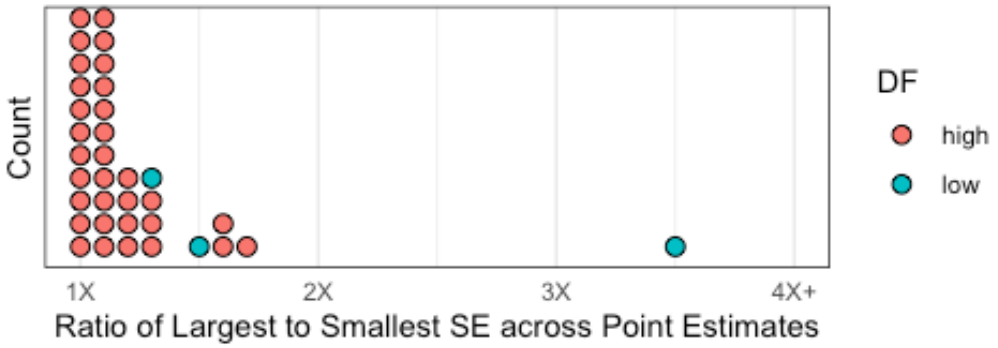


Figure 3. The Ratio (across 9 estimators) of the Largest to Smallest Estimates of SE(B) for 38 outcomes)

Note: "low" df means below 10 under the conservative degree of freedom calculation of J – 2K – g, where g is the number of control covariates.

The single outcome with an extreme ratio of more than three times was a tiny study of 10 clusters with 124 total units. The aggregated, person-weighted linear regression with design-based standard errors gave an outlier standard error estimate of 0.10σ (unusual estimated standard errors is common when using cluster robust or design-based methods with few clusters), while the cluster weighted estimate was 0.36σ. The remaining estimators were all around 0.2σ. Overall, with small cluster randomized experiments, uncertainty estimation can be quite unstable.

And as before, the variation in uncertainty estimates does not seem to be driven by the estimand targeted. When we look within estimand, the ratio of largest to smallest standard error shrinks only modestly (a median reduction of 2%, with 75% of the ratios shrinking by only 8% or less).

Simulation Supplement for "An Applied Researchers Guide to Estimating Effects from Multisite Cluster Randomized Trials: Estimands, Estimators, and Estimates"

Contents

1 Introduction

Blocked cluster randomized controlled trials (CRTs) are a widely used experimental evaluation design in which entire groups (clusters) such as schools or classrooms are randomly assigned to treatment or control conditions within defined strata or blocks. The effectiveness of the program or treatment is then typically quantified using the average treatment effect (ATE). However, within this relatively straightforward quantity, researchers must make a variety of choices. Researchers need to decide whether to target a person-weighted average or

cluster-weighted average; they need to decide whether they are targeting an estimate just for the clusters in the study (finite sample) or an estimate representative of some broader population (superpopulation); and they need to decide from among different methods or estimators that can be used to estimate the ATE.

This manuscript contains simulations that complement research empirically examining the implications of different choices researchers make when performing CRTs (Miratrix et al., 2026). Simulations, unlike empirical analyses, allow us to compare estimated quantities to their true values, which we cannot observe in empirical contexts. This in turn allows us to evaluate how well different statistical methods perform. Additionally, we can systematically vary data conditions and identify which factors impact method performance.

We present two collections of simulations. The first, Simulation I, is a multifactor simulation where we explore how a set of different estimators react to different data generating conditions, such as variation in the size of clusters or systematic differences between the blocks. Simulation I is designed to answer the following research questions:

1. Are any of the estimators ever substantially biased? When?
2. Are any of the estimators generally more precise than others? When?
3. Are any of the standard error estimators inaccurately calibrated? When?
4. Do any of the estimated standard errors target finite versus super-population uncertainty?
5. Why does the range in the estimates produced by the different estimators get larger as the degrees of freedom get smaller? Is the cause estimator bias or estimator instability?

Simulation II, by contrast, is a set of calibrated simulations where, for each of our empirical examples (excepting a few due to logistical considerations), we ascertained the properties of all estimators in the very narrow context represented by those data. These simulations were in reaction to the surprising range in both point estimates and standard errors we found in the empirical data.

In Simulation II, we, for each outcome, permute treatment assignment and examine how the various treatment effect estimators react to different treatments. Because permutation ensures these are not any actual treatment effects, we can explore how the different estimators are similar or different in the context of no actual treatment. This gives insight into whether the results of the empirical study are driven by systematic variation in treatment impact or just idiosyncratic aspects of the contexts explored. These simulations are precisely calibrated to the structure of the data. In particular, by simply permuting treatment, we preserve all features such as covariate relationships and cluster sizes.

Our core question for Simulation II is then:

1. Are the ranges of point estimates in the empirical data consistent with what we would expect under no treatment effect?
2. What about the range of estimated standard errors?
3. Are some of the estimators more precise across permutations than other estimators?

Importantly, Simulation II covers a range of contexts both more broad than Simulation I and also less comprehensive. We simulate over the very heterogeneous set of empirical datasets we have, but we do not systematically vary data generating parameters as in Simulation I. Simulation II has some contexts with many numbers of very small blocks (including blocks of size 2, i.e. matched pairs experiments, which Simulation I does not cover).

2 Overview of Findings

Before diving into the details, at a high level the answers to our core research questions from Simulation I are as follows:

1. Only under extreme conditions do we see notable bias. Fully interacted estimators have virtually no bias. Fixed effect estimators can show large bias (even up to 0.2 in size), but only under the most extreme messy-block conditions with strong correlation between block average effects and proportion treated (this is the precision-weighting bias). Some MLM estimators can exceed 0.04 bias when cluster-level

ICC is zero and cluster size correlates with treatment effect, because the estimator shifts toward person-weighting.

2. Most estimators have comparable relative precision in most contexts. MLMs can be more precise than other cluster-average targeting estimators when cluster-size variation is large and cluster-level ICC is zero; in those cases the MLM largely ignores the clustering, which stabilizes estimates considerably. That said, even in this case, the MLM estimators had standard errors only 30% smaller at most.

3. Overall, all the methods give standard errors that reasonably capture the uncertainty of a superpopulation framework with fixed blocks. In particular, fully interacted estimators (except MLM-FI) produce SEs that are slightly too small relative to the true superpopulation SE; the magnitude, however, is generally negligible. Fixed effect estimators are well calibrated under typical conditions but can substantially overestimate when there is substantial variation in treatment effects across blocks. MLM estimators are generally well calibrated, with exceptions tied to unbalanced treatment and zero cluster ICC.

4. All estimated SEs are better characterized as targeting superpopulation rather than finite-population uncertainty. The two quantities diverge meaningfully only when cluster-level ICC is zero and there is substantial cross-cluster treatment effect variation.

5. The wider range of estimates across methods in low-df settings reflects greater estimator instability (larger true SEs), not greater bias. Bias shows no systematic relationship with degrees of freedom.

We also, on Table 1, show general findings on each estimator's performance across the simulations, identifying under what conditions they fail.

For the calibrated simulation (Simulation II), we found general alignment between the ranges of estimates and ranges in standard errors across estimator that we saw in the simulation and what we saw in the empirical data. Echoing Simulation I, we also found that the true uncertainty (under permutation) of most estimators were about the same across most datasets, but that for some datasets, some estimators were up to 30% more variable than others. These differences are much more modest than the apparent differences when we compare the *estimated* Standard Errors.

Table 1: General properties of the different estimators across the simulations. See text for details.

Method	Treatment Effect	Standard Errors
Cluster Weighted		
Fully Interacted	Virtually no bias across all conditions.	Very small underestimate of the true superpopulation SE under all conditions where the SE is defined. Cannot be used if any of the blocks only have two clusters.
Fixed Effect	Bias can exceed 0.05, and under the most extreme conditions (messy blocks with a high correlation between block-level proportion treated and block-level ATE, combined with unbalanced treatment assignment) can reach -0.2. Bias is negligible under same and reasonable block conditions.	Well estimated under same-block conditions. Modestly elevated for reasonable blocks. Substantially overestimated in messy-block conditions due to model misfit when treatment effects vary across blocks; overestimation is sharpest when cluster-level ICC is zero.
Multi-Level Model	Bias can exceed 0.04 when there is a strong correlation between cluster size and treatment effect and zero cluster-level ICC (causing the estimators to lean person-weighted instead of cluster-weighted). In the non-fully interacted MLM models (MLM-RE and MLM-FE), bias also can exceed 0.04 under extreme conditions akin to the fixed effect estimators discussed above. More precise than other estimators under combination of large variation in cluster size and zero cluster ICCs.	The estimated standard errors can be substantially too large when there is a confluence of unbalanced treatment, zero cluster-level ICC, and cluster-level treatment variation. The FE and RE estimators also have problems in the face of cross-block treatment effect variation, akin to the other FE estimators considered. Coverage can fall below 90% for larger experiments (20+ clusters per block) when cluster size correlates with treatment impact, ICC is zero, cluster sizes vary, and treatment is unbalanced.
Person Weighted		
Fully Interacted	Minimal bias under all conditions.	Generally well estimated with slight downward bias under nearly all conditions. Cannot be used if any of the blocks only have two clusters.
Fixed Effect	Bias can exceed 0.05, and under the most extreme messy-block conditions with unbalanced treatment can reach magnitudes similar to the cluster-weighted FE. Bias is negligible under same and reasonable block conditions.	Well estimated under same-block conditions; person-weighted variants can run slightly low even under typical conditions. Substantially overestimated in messy-block conditions due to the same fixed-effect model misfit as the cluster-weighted variant.

3 The Design of Simulation I

A multifactor simulation is a simulation where we systematically vary a set of factors to understand how these factors influence the performance of the various estimators. In general we are close to full multi-factorial designs, where we vary all factors independently, but we do drop combinations that do not make sense (in particular, we drop scenarios that call for correlations between quantities that do not vary).

3.1 The data generation parameters

The main simulation is a nearly fully factored experiment with the data generating parameters given on Table~2.

Table 2: Simulation Data Generating Parameters

Parameter	**Value(s)**
Parameters fixed across all scenarios	
Number of Blocks	5
Average Treatment Effect	.2
Block-Level ICC	.2
Average Number of Observations Per Cluster	20
Number of Block Level Covariates	0
Block-Level (Varies across scenarios)	
Block Type (blk_type)	Same (Sam), Reasonable (Res), Messy (Mss)
Average Number of Clusters per Block (J.bar)	5 (S), 10 (M), 20 (L)
Proportion of Clusters Treated (p.bar)	.5 (Bal), .7 (UnB)
Cluster-Level (Varies across scenarios)	
Correlation of Impact and Size of Cluster (cor_J_tau)	0 (0), .8 (+)
Cluster Level Treatment Effect Standard Deviation (sd.clu)	0 (0), .3 (+)
Cluster Level ICC (ICC.clu)	0 (0), .2 (+)
Variation in cluster size, coefficient of variation (cv.clu)	0 (0), .33 (S), .6 (L)
Number of Cluster-Level Covariates (num.cov)	0 (0), 1 (S), 5 (L) [R^2 Always .5]

One of our factors is "block type," which is a set of seven parameters that define how similar or different the blocks are from each other (see Table~3). We did not independently vary these factors so we could control the total number of simulation scenarios. We have three levels of block type, which we call "Blocks All The Same," "Blocks All Reasonable," and "Blocks All Messy." The "Same" case is where all blocks are identical in their characteristics. The "Reasonable" case is what we might imagine we would see in actual practice. The "Messy" case pushes our simulation to the limits so we can see how various methods fail in extreme circumstance.

All code is public; others can easily adapt the code to run additional scenarios that are of interest.

Table 3: Block-Type Parameters

Parameter	Blocks All The Same	Blocks All Reasonable	Blocks All Messy
Variation across blocks of number of clusters in the blocks (coefficient of variation)	0 (none)	.33 (average)	.85 (big)
Variation across blocks of the average cluster size (coefficient of variation)	0	.33	.33
Block-Level Average Treatment Effect Standard Deviation	0	.2	.6
Correlation Between Block Average Treatment Effect and Block Average Cluster Size	NA	.3	.8
Correlation Between Block Average Cluster Size and Total Number of Clusters	NA	.3	.8
Correlation Between Control-Side Cluster Average Outcome and Cluster Size	NA	.3	.8
Correlation of Block's Proportion of Clusters Treated and Block Average Treatment Effect	NA	.3	.8
Does the proportion of clusters treated vary across blocks?	No	Yes	Yes

3.2 The data generating process

For each scenario defined by sample size, target correlation, predictiveness of covariates, and so forth, we used the following data generating process (DGP):

1. Randomly generate a single set of blocks based on the simulation scenario parameters. Each generated block has a set of characteristics such as the average treatment effect of the block, average cluster size, amount of cluster variation, and so forth.
2. Then, for each simulation iteration:
    a. Generate a sample of clusters within each block according to the block characteristics.
    b. Generate individual-level data within each cluster.
    c. Calculate the true ATEs (both cluster-weighted and individual-weighted) for the generated dataset. These are our target finite-sample estimands.
    d. Assign a proportion of clusters to treatment according to the target block-level proportion.

At this point we have a complete set of generated data. Next, for each simulation iteration we actually conduct a small number of iterations. For each internal iteration we:

a. Assign treatment to the clusters within blocks.
b. Calculate observed outcomes based on the generated data and the treatment assignment.
c. Estimate the ATE and SE using all our methods.

The nested permutation approach means that for each fixed set of clusters and blocks, we can calculate finite sample performance. For example, the standard deviation of the ATE estimates within a given dataset across the permutations gives us the true finite sample standard error for that dataset. If we look at the standard deviation of the ATE estimates across all our datasets within a scenario, we have the true cluster-level superpopulation standard error. This standard error *does not* take into account uncertainty at the block level, as the block characteristics are all fixed. So, for example, if there is treatment variation across blocks,

our true standard error will not reflect uncertainty due to which blocks were generated, only uncertainty due to which clusters and individuals were generated.

For our target estimands we can use, within each generated dataset, the finite sample estimand. We can also calculate the superpopulation estimand (both person-weighted and cluster-weighted) by averaging the finite sample estimands across all generated datasets within a scenario.

We provide more technical details of how we implement the components of the data generating process at the end of this manuscript.

4 Simulation I Results

We walk through the five research questions, providing plots showing estimator performance with respect to each question.

In this report, we use a set of boxplots customized for our particular data so we can see central spread more clearly. Below is a sample boxplot of the mean bias of an estimator relative to its target estimand across simulation runs for all methods and scenarios. The faint horizontal lines extend to the full range of the data. The bars mark the 2.5th and 97.5th percentiles, capturing 95% of the data. The boxes span the 10th and 90th percentiles, corresponding to the middle 80% of the data.

Not only does this plot show our modification of the box plot, but it provides an important high-level finding: across all the estimators there was negligible bias across the vast majority of simulation conditions considered. A classic 50% boxplot would have no visible "box" here—the 80% box makes variation visible at all.

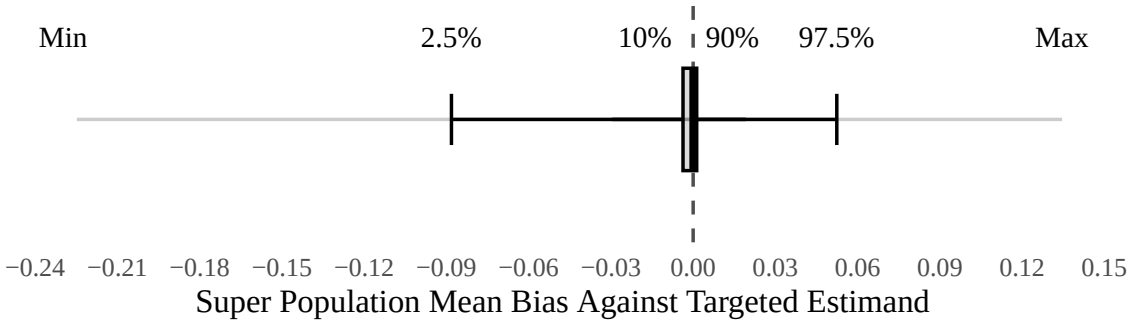


Figure 1: Sample boxplot for paper. This boxplot shows the superpopulation mean bias of each estimator relative to its target estimand across scenarios and methods. The faint horizontal lines extend to the full range, the bars span the 2.5th to 97.5th percentiles, and the box spans the 10th to 90th percentiles.

4.1 Are any of the estimators ever substantially biased? When?

Before diving into the results, we note that many of our estimators give the exact same point estimates (though often different standard errors). In particular, the aggregation regression estimators are identical (in terms of point estimate) to the individual regression estimators, as we did not include level 1 covariates in this simulation. Overall, we have a core set of 7 estimators that produce different point estimates.

Across all estimators the magnitude of bias was less than 0.01 for more than 80% of the simulation scenarios tested (Figure 1). For LRcw-FIcw we see virtually no bias across any of the scenarios considered. For the other estimators, bias is small for the bulk of simulation scenarios considered; see how narrow the 80% boxes generally are on Figure 2. That said, we do see evidence of bias larger than this under some circumstances for several different methods.

To understand when and why bias occurs, we fit regression trees predicting the magnitude of the bias for each estimator using the data generating parameters as predictors. We fit one tree for each estimator (except LRcw-FIcw, which had no bias). The six trees are displayed on Figure 3.

Consider the top-left tree in Figure 3, for LR-FE. Each branch divides all the scenarios into two groups, and each leaf represents all scenarios that had a given combination of factors defined by the splits of the tree. The green terminal nodes at the end give the final predicted bias. So, for our LR-FE example, if blk_type is *not* Same or Reasonable (i.e., it is Messy) AND p.bar is not balanced AND cv.clu is 0 (i.e., there is no variation in cluster sizes) then we have an average bias of 0.08. In this case, the unequal proportion of clusters treated across blocks combined with the messy blocks (which induces correlation between block-level average treatment effect and proportion treated) causes bias because of the fixed-effect "precision-weighting" aspect of LR-FE (see the technical supplement for further discussion of this).

Across Figures 2 and 3 we have several findings: First, the estimators that theory predicts are essentially unbiased indeed appears to have no bias (or possibly very low bias) across the full range of simulations. Second, other than the MLM estimators, our simulations almost never show substantial bias under the data generating conditions we expect to see in practice (i.e. for the same and reasonable block type conditions).

In some of the simulations, the person-weighted fully interacted estimator (LR-FIpw) can have bias as large as .03. The regression tree (top right facet of Figure 3) shows this comes from a correlation between the treatment effect and cluster size combined with a small number of clusters per block, large cluster size variation, and unbalanced treatment assignment. Overall it is a stable performer with little potential for bias.

For the fixed effect estimators (LR-FE, LRcw-FE), we do observe bias as severe as -0.2 in some simulations. This bias is primarily driven by the combination of messy blocks and unbalanced treatment assignment of clusters across blocks. It is the cost of the "precision weighting" we get by having a single common treatment effect in the model. The magnitude here is large, but it is important to note that the messy block condition was deliberately designed to push parameters to the extreme. In particular, the messy block condition has a 0.8 correlation between the proportion of each block assigned to treatment and the block-level treatment effect.

The multilevel fixed effect models (MLM-RE and MLM-FE) are also subject to the fixed effect bias discussed above. The multilevel fully interacted model (MLM-FIcw) performs well even with messy blocks and unbalanced treatment assignment, although it can have positive bias in some scenarios. In particular, when there is a correlation between cluster size and treatment effect coupled with zero cluster-level ICC, the estimator leans more person-weighted (due to the zero ICC), which is considered bias as the estimator is viewed as targeting the cluster-weighted estimand. The MLM-RE and MLM-FE estimators are also subject to this, so they face two possible sources of biases that could stack or partially cancel out.

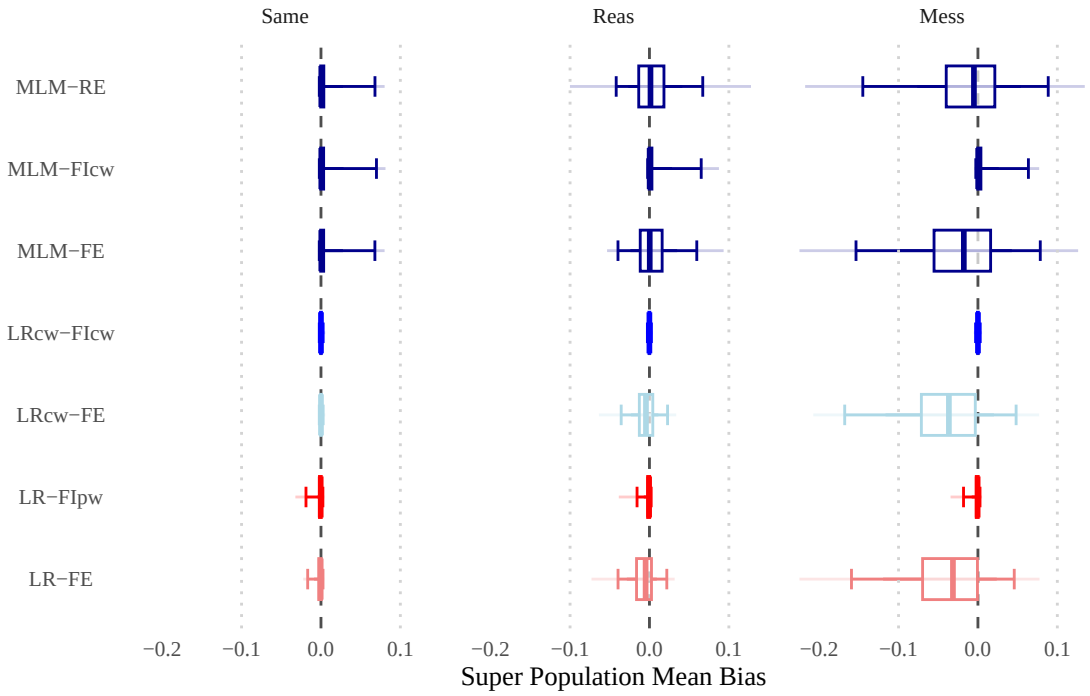


Figure 2: Boxplot of superpopulation mean bias of each estimator relative to its target estimand across scenarios, separated by method. The faint horizontal lines extend to the full range, the bars span the 2.5th to 97.5th percentiles, and the boxes span the 10th to 90th percentiles. Aggregated regression methods coincide with their non-aggregated counterparts when there are no level 1 covariates.

4.2 Are any of the estimators generally more precise than others? When?

All the estimators will be more precise when there is more data, and less precise when there is less. To assess *relative* precision of the estimators, therefore, we compare each estimators true precision (the true SE) to the median precision of the estimators in that scenario. We call this the "relative true standard error":

$$SE^{\text{rel}}_{m,s} = \frac{SE_{m,s}}{\text{median}(SE_{\cdot,s})}$$

where $SE_{m,s}$ is the true standard error for method m in scenario s, and $median(SE_{\cdot,s})$ is the median true standard error across all methods in scenario s.

Unfortunately, the cluster weighted and person weighted estimator can be systematically different if there is variation in cluster size. To assess the magnitude of these differences, we compare the ratio of the true SE of LR_FIpw to the true SE of AR_FIcw (these being the unbiased estimators for each of the two estimands). Results on Figure 4. We see that when there is substantial cluster size variation, then if there is zero cluster-level ICC, the person-weighted estimator can be substantially more precise than the cluster-weighted estimator. By contrast, when there is positive cluster-level ICC, the person-weighted becomes more variable than cluster-weighted.

Given the differences across scenarios in cluster vs. person weighting, we compare each estimator to the median of the estimators *within its group* – so, for example, we compare all of the person weighted estimator to the median of the person weighted estimator for a given scenario. We plot these within-estimand relative SEs for all scenarios on Figure 5. In general, the different estimators have comparable precision, with the boxes all hovering around 1. In some cases, some estimators can have standard errors that are 10% larger than the median.

The multilevel estimators can have a true SE substantially smaller than median the other cluster-average

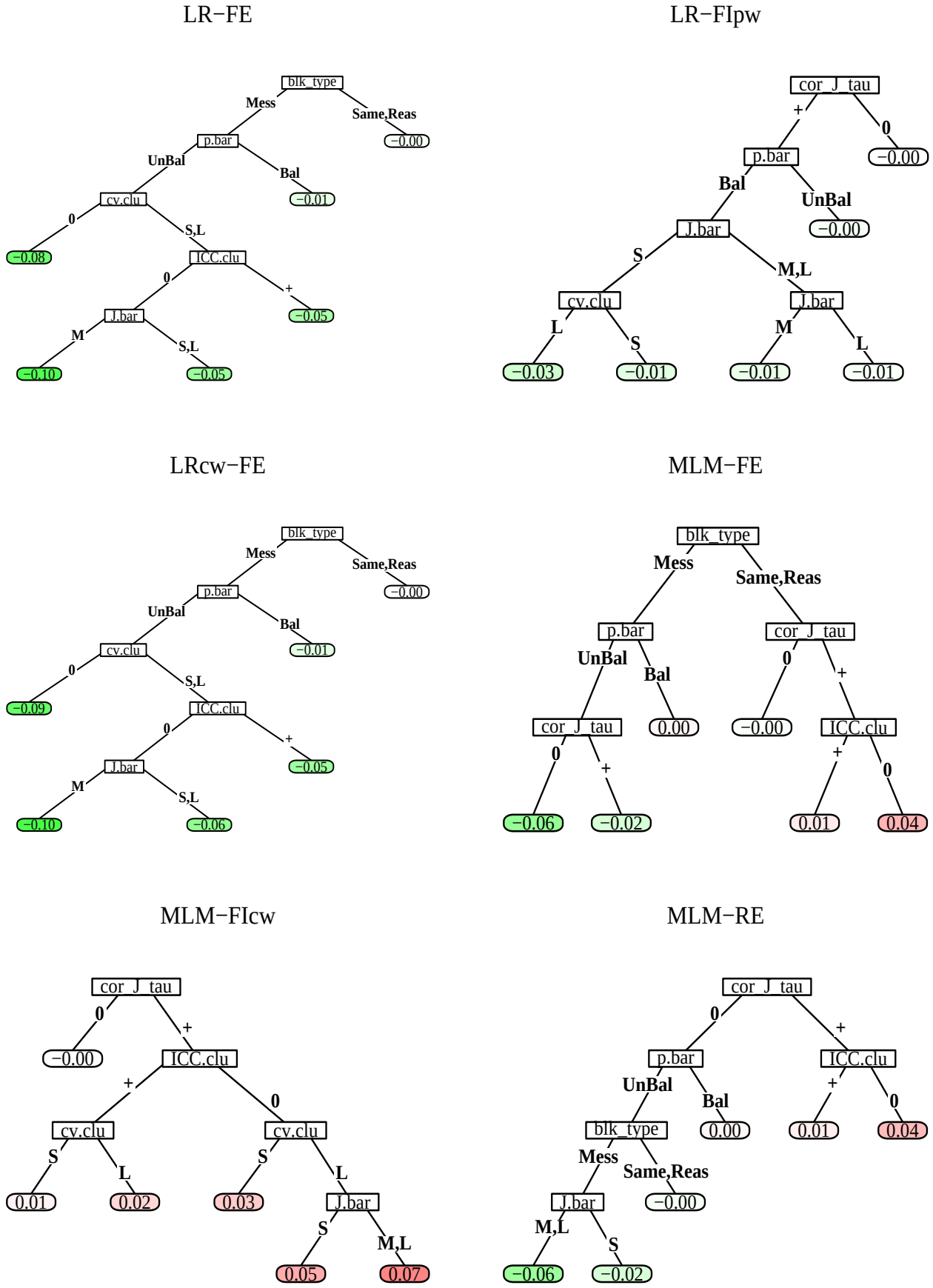


Figure 3: Bias of the estimators as shown by regression trees by method. LRcw-FIcw is not shown as it had negligible bias across all scenarios.

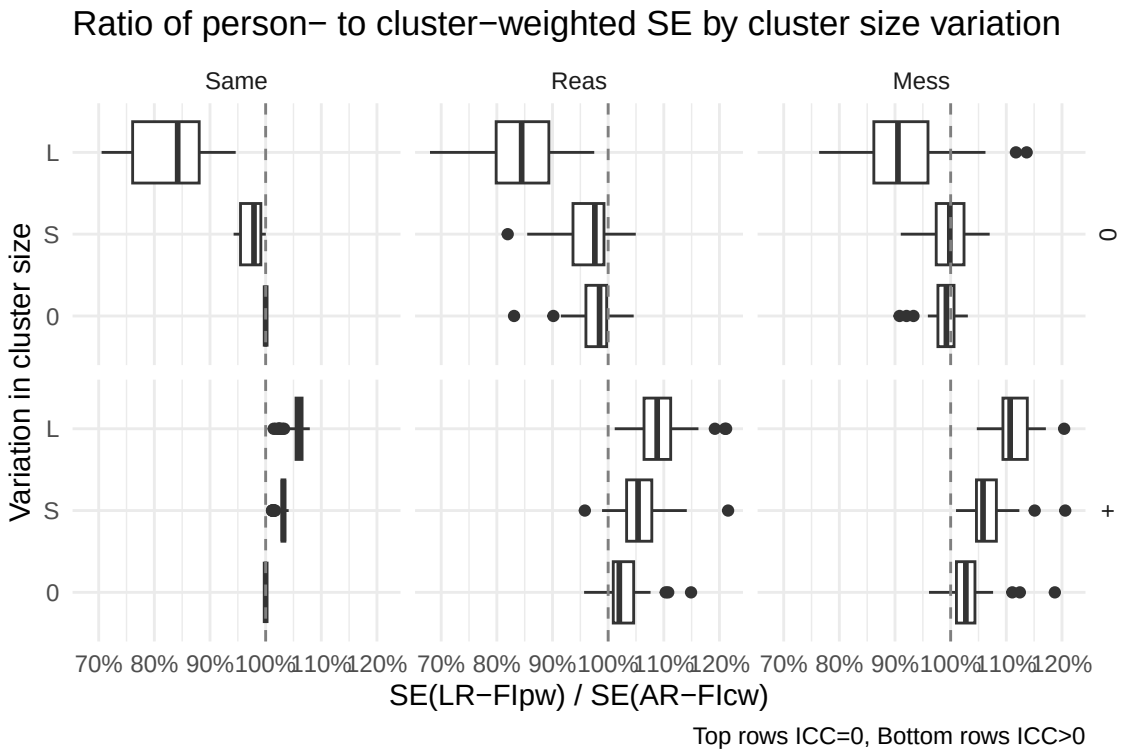


Figure 4: Ratio of the true SE of LR-FIpw to the true SE of AR-FIcw by cluster size variation, cluster-level ICC, and block type. These are normal boxplots, not 80% boxplots.

targeting estimators. The multilevel estimators become more precise than the other cluster-weighted estimators when there is a combination of large variation in cluster size and zero cluster ICC; otherwise, their precision is close to that of the other estimators (see the regression tree on Figure 6). When ICC is low, MLMs essentially ignore the clustering, which stabalizes the estimator considerably (but also makes it target the person-weighting estimand, which can create bias if the cluster size is correlated with impact as described earlier).

4.2.1 Range of Min SE to Max SE across scenarios

We can further explore precision differences among estimators by, for each scenario, calculating the ratio of the maximum to minimum true standard error across all methods for scenarios with no treatment variation. When this ratio is close to 1, all estimators have similar precision; when it is substantially above 1, some estimators are considerably more efficient than others.

For the no impact variation scenarios, the bulk of SE ratios are near 1, with only 5% having a ratio larger than 1.40, consistent with the earlier finding that most estimators have comparable precision. Figure 7 shows a regression tree identifying when these ratios are largest: the primary driver is a zero cluster-level ICC with large variation in cluster-size, exactly the condition identified above in which MLM estimators are substantially more precise than the other cluster weighted estimators. Under conditions typical of most applied studies — non-zero cluster ICC or modest cluster-size variation — the ratio stays close to 1, meaning the choice of estimator has little practical impact on precision.

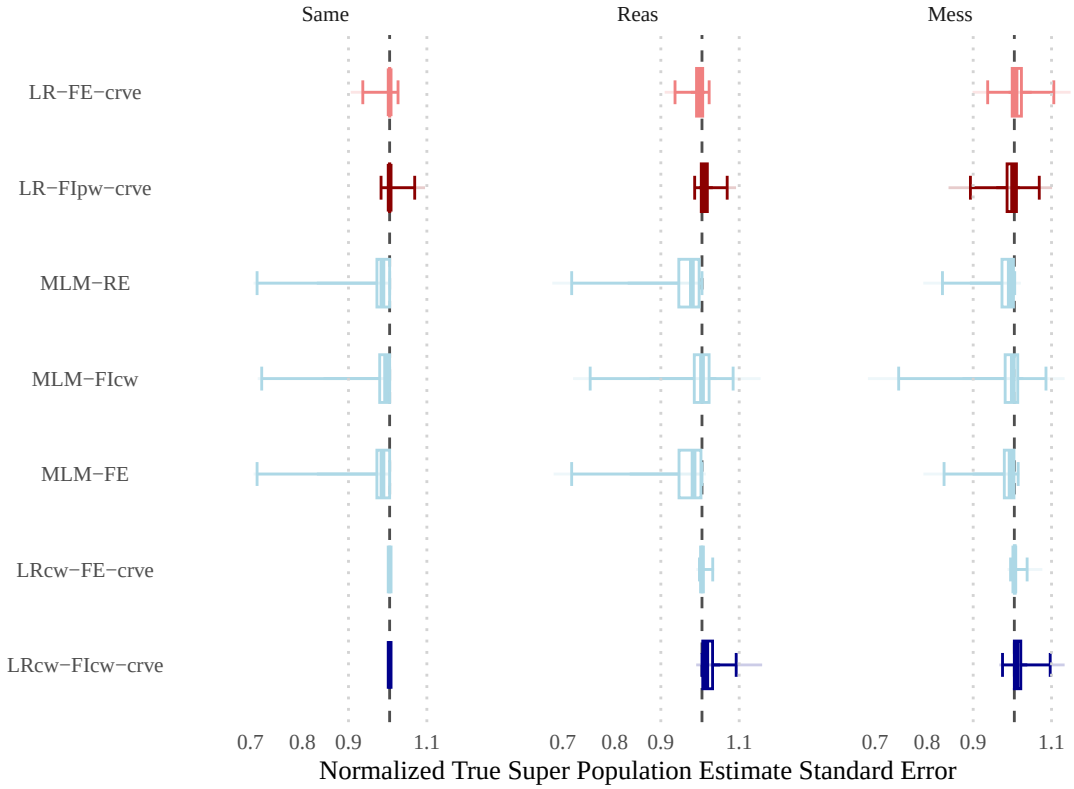


Figure 5: Relative true standard errors across the different scenarios by method. The relative true standard error is the true standard error for a given method divided by the median true standard error across methods within a given simulation scenario. Therefore each data point is adjusted for the core uncertainty of the corresponding DGP. The faint horizontal lines extend to the full range, the bars span the 2.5th to 97.5th percentiles, and the boxes span the 10th to 90th percentiles.

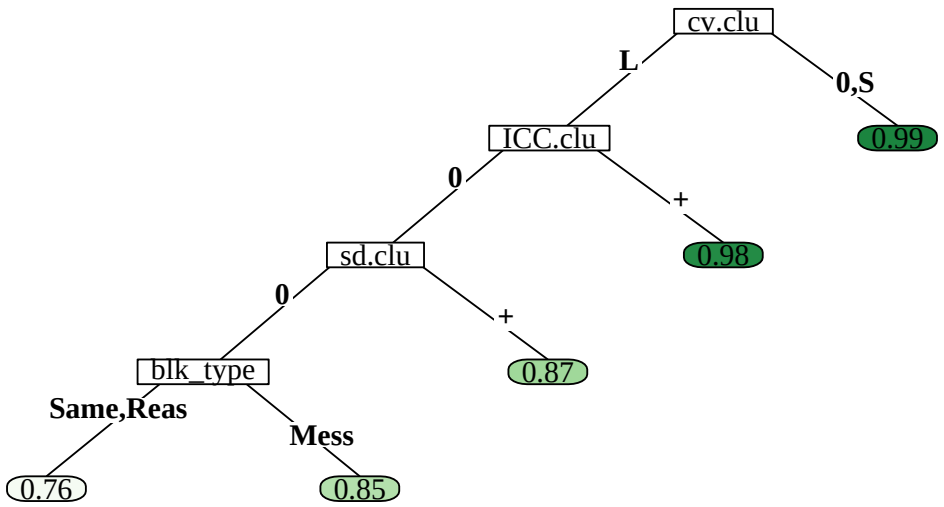


Figure 6: Regression tree predicting the normalized true estimate standard error under different data generating parameters for MLM models. Large cluster variation coupled with 0 ICC leads to smaller SEs.

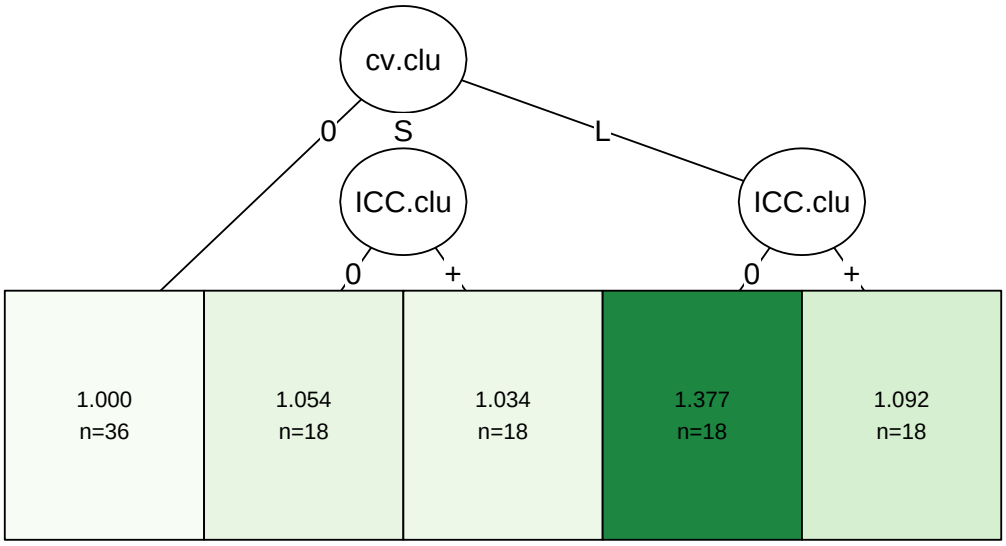


Figure 7: Regression tree predicting the ratio of maximum to minimum true superpopulation SE across estimators within each simulation scenario. A ratio close to 1 means all methods have similar precision; larger values indicate that some estimators are substantially more efficient. The primary driver is a combination of zero cluster-level ICC and high cluster-size variation, the conditions under which MLMs gain a precision advantage.

4.3 Are any of the standard error estimators inaccurately calibrated? When?

A *calibrated* standard error estimator is a standard error estimator that, on average, equals the true standard error. For each estimator and scenario we assess calibration by calculating

$$\text{inflation} = \left(\frac{E[\hat{SE}^2]}{SE^2} \right)^{1/2}$$

where $E[\hat{SE}^2]$ is the average estimated squared standard error across simulation runs within a scenario, and SE is the true superpopulation standard error for that estimator in that scenario. An inflation factor of around 1 means our estimator is calibrated. If it is less than 1 our estimator is systematically too small.

Note that in around 50% of our simulation scenarios, the fully interacted estimators simply did not have defined standard errors due to singleton units in the treatment or control side for at least one block. For these estimators we had to drop 18% of the scenarios with small numbers of clusters, 55% for medium numbers, and 79% for large numbers (for medium and large, we only had to drop scenarios with variable proportion of clusters treated).

Figure 8 shows the distribution of inflation scores across scenarios for all of our estimators. Note we have a full suite of estimators now, not just the core set of point estimators from before. We color them by the estimand they target, and whether they are fixed effect vs. fully interacted.

Across scenarios, the standard error estimates for *fully interacted* estimators, except the fully interacted MLM, are too small relative to the true super population standard errors (Figure 8); the person-weighted ones are slightly more vulnerable to this. However, across almost all simulations, the magnitude of this underestimate is very small, except for the person-weighted variants for some scenarios. We again note that these estimators do not have defined standard errors for around half the scenarios considered.

For the fixed effect estimators, the standard errors are generally well estimated in most of the "same" block condition, although the person-weighted variants can be a bit low under some scenarios. For the "reasonable" blocks, the standard errors are systematically too high and for "messy" they are substantially too high.

The overestimation comes from a model misfit issue, where the fixed effects cannot account for the treatment variation across blocks, and so the residuals get inflated, leading to higher standard errors. The fixed effect estimators assume, in effect, that the average treatment effect across blocks is equal, which is increasingly untrue as we move from "Same" to "Reasonable" to "Messy" blocks. The standard errors are particularly too large when there is no cluster-level ICC (Figure 9). We further investigate this misfit issue below, when we investigate whether our estimators are targeting finite vs. superpopulation standard errors. It is important to remember that the messy blocks condition was designed to test when the estimators break and are not conditions we expect to see in practice.

For the MLM methods, the amount of overestimation is generally small across scenarios, with some exceptions even for the "Same" block condition, as illustrated by the long tails of the boxplots. The inflation is driven by a few different factors. First, the FE versions of MLM (MLM-RE is essentially a variant of fixed effects) suffer from the same fixed effect issue discussed above. Second, for non-Messy blocks we see inflation when the proportion of clusters assigned to treatment varies across blocks, there is zero cluster-level ICC, and there is cross-cluster treatment effect variation (Figure 9).

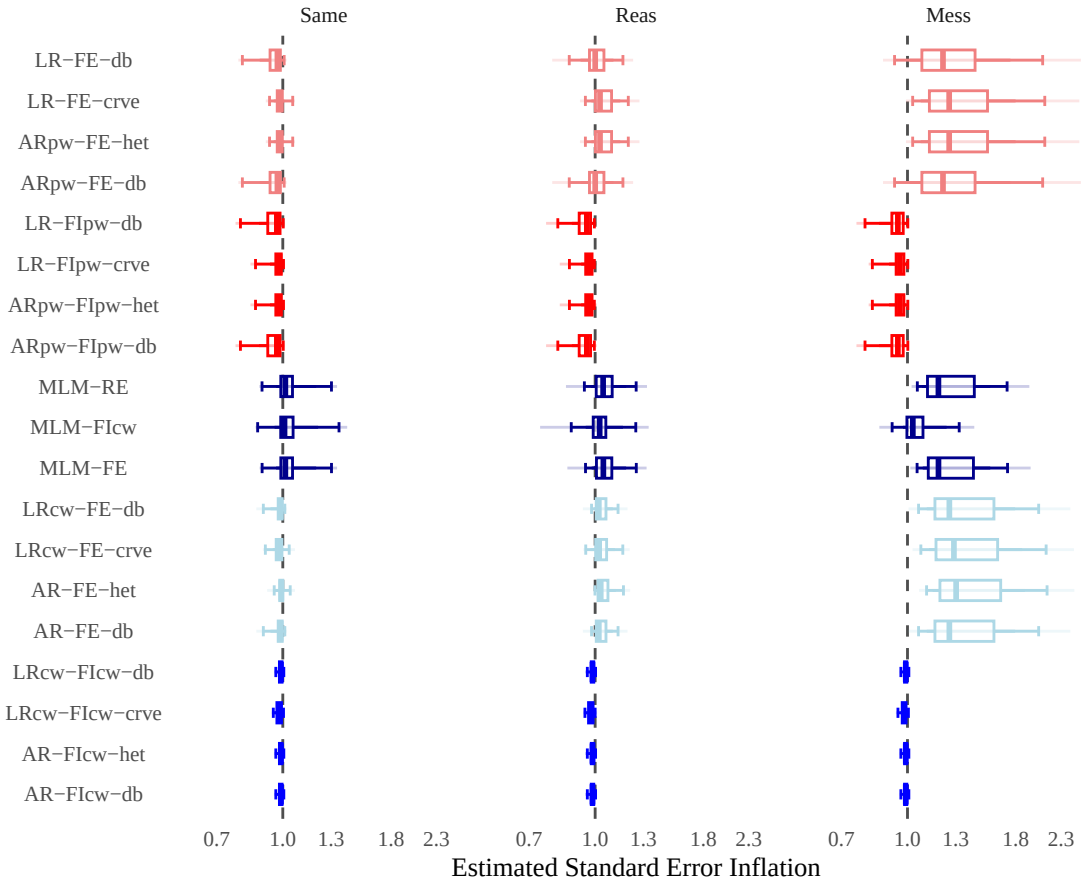
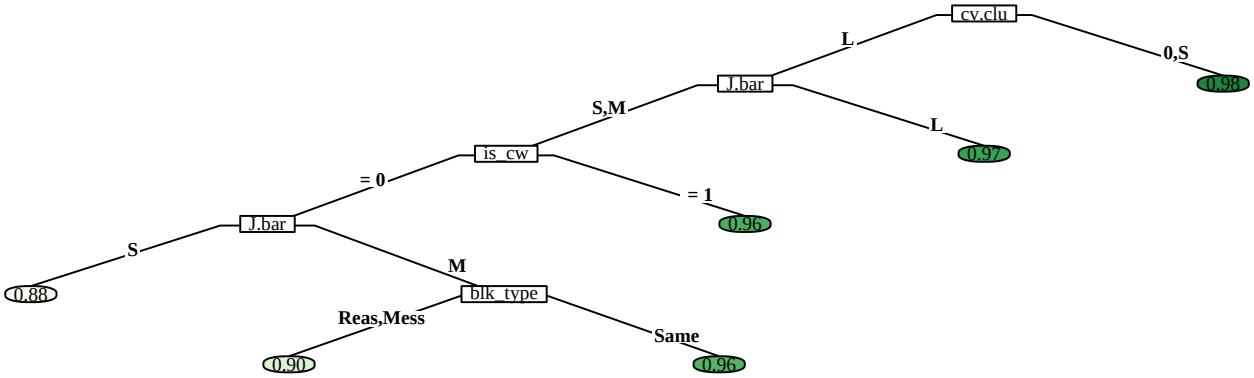
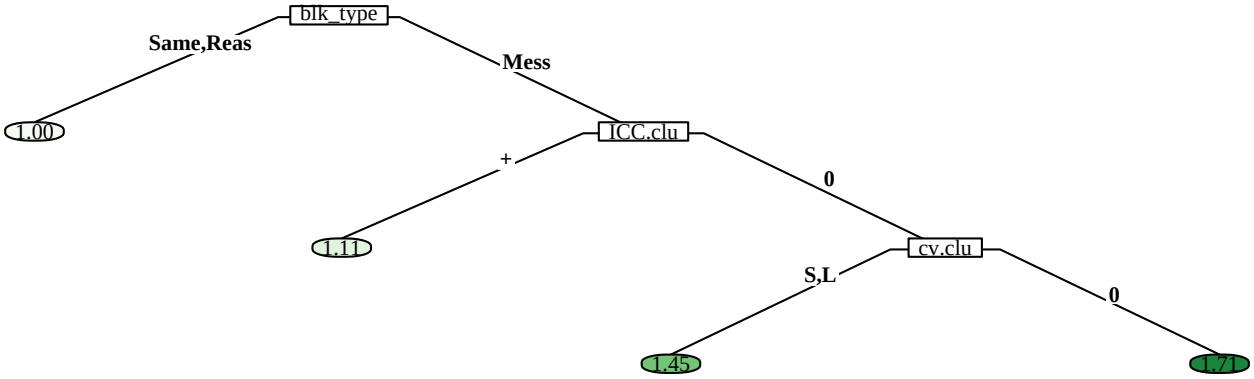


Figure 8: Inflation of estimated standard errors across simulation scenarios by method. Inflation is measured as the ratio of the average estimated standard error over the true super population standard error. Values above 1 indicate systemic overestimation, values below 1 indicate underestimation. The faint horizontal lines extend to the full range, the bars span the 2.5th to 97.5th percentiles, and the boxes span the 10th to 90th percentiles.

FI models



FE models



Multilevel Models

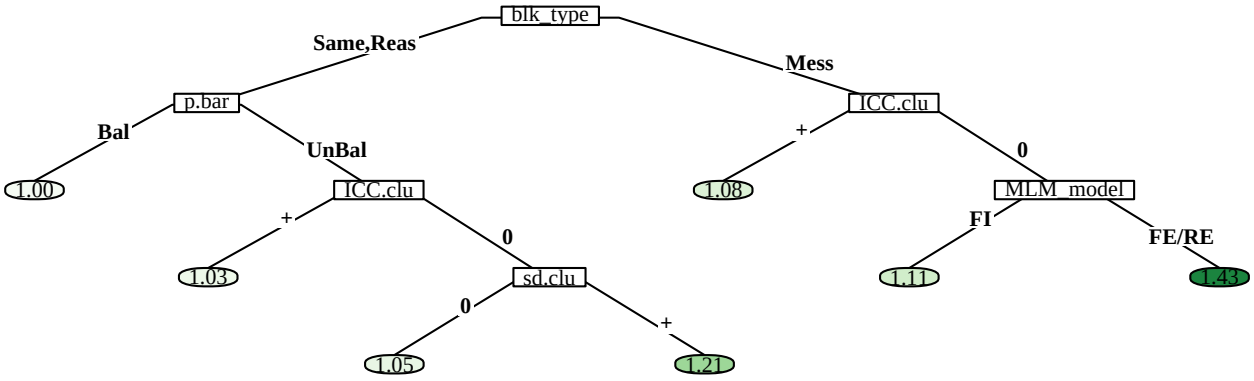


Figure 9: Standard error inflation under different data generating parameters for different estimators.

4.3.1 Coverage

Consistent with the standard errors being generally well estimated across simulation conditions and estimators, the coverage rate is close to 95% across most simulation conditions and estimators (Figure 10), although the person-weighted estimators can be slightly undercovering in a few scenarios.

In particular, the person-weighted design-based standard error estimators can have notably low coverage when there is a lot of variation in cluster size, with few clusters per block (Figure 11, bottom).

For the MLM models, coverage can fall below 90% for the larger experiments when there is correlation between number of clusters and treatment impact, no cluster-level ICC, variation in cluster size, and an unbalanced treatment assignment (Figure 11, top). Under these conditions the weighting for the MLM models shifts from being cluster weighted to person weighted, which causes these estimators to be biased against the cluster weighted estimand, which in turn causes the coverage rate to drop.

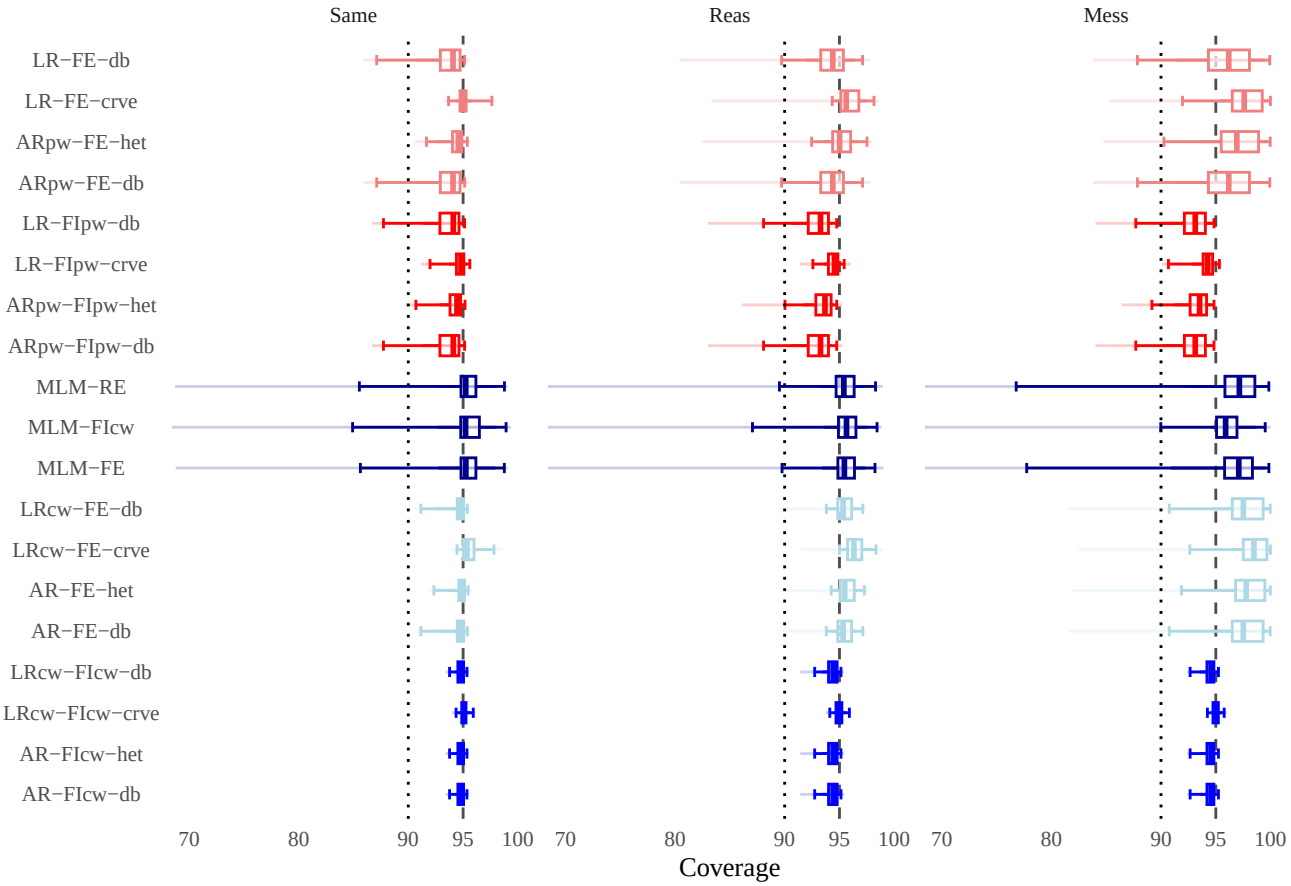


Figure 10: Boxplot of coverage rates across different scenarios by method. The vertical line indicates the target 95% coverage rate. The faint horizontal lines extend to the full range, the bars span the 2.5th to 97.5th percentiles, and the boxes span the 10th to 90th percentiles.

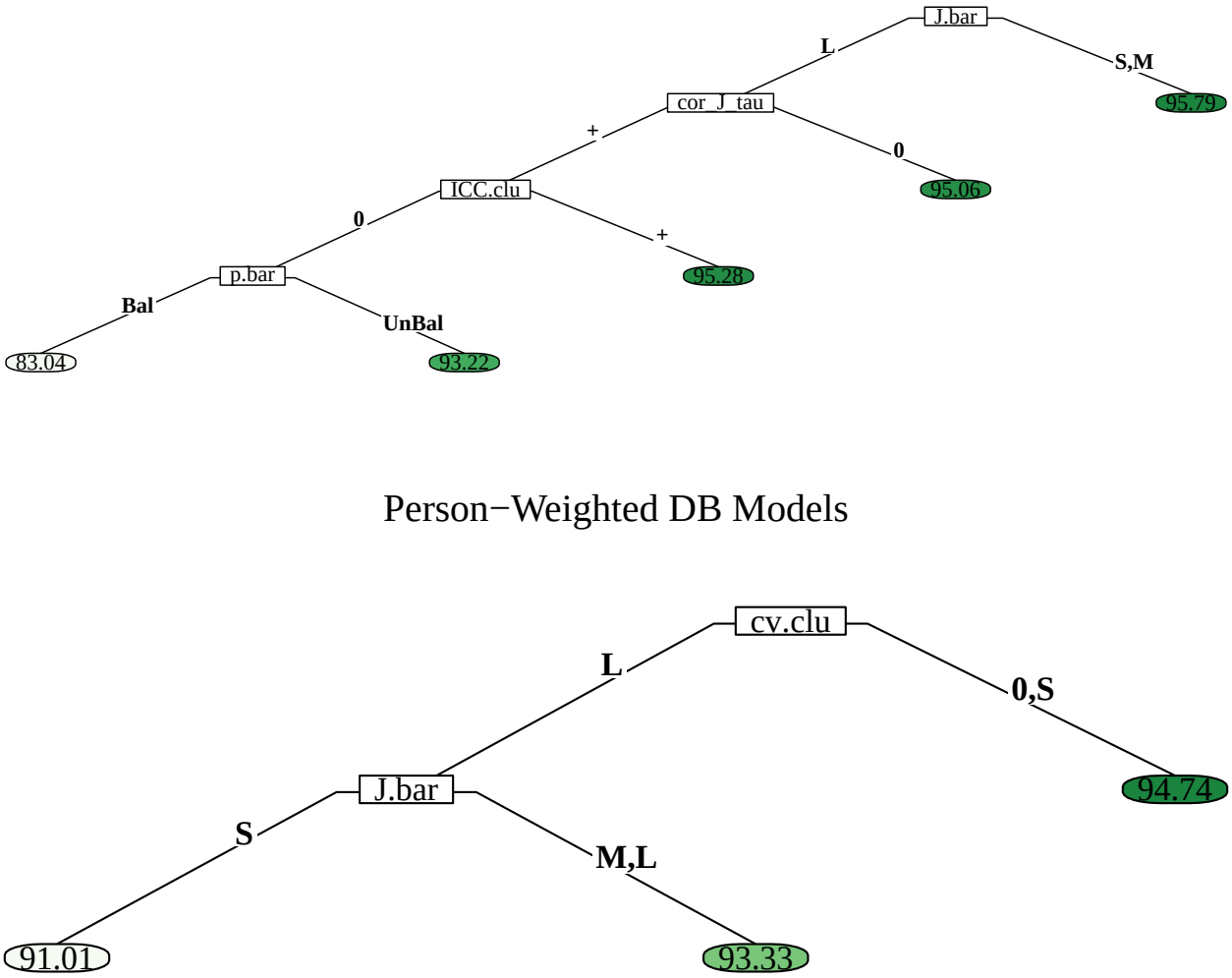


Figure 11: Regression trees for coverage rates under different data generating parameters

4.4 Do any of the estimated standard errors target finite versus super population uncertainty?

Our simulation allows for calculating the true finite population standard errors as well as the true super population standard errors for each scenario and estimator. This allows us to assess what the estimated standard errors are targeting.

For each estimator, we first calculate the ratio of the true finite population standard error over the true super population standard error for each scenario. The true super-population standard errors are only noticeably different from the true finite population standard errors when the cluster-level ICC is zero and there is cross-cluster treatment effect variance (see Figure 12).

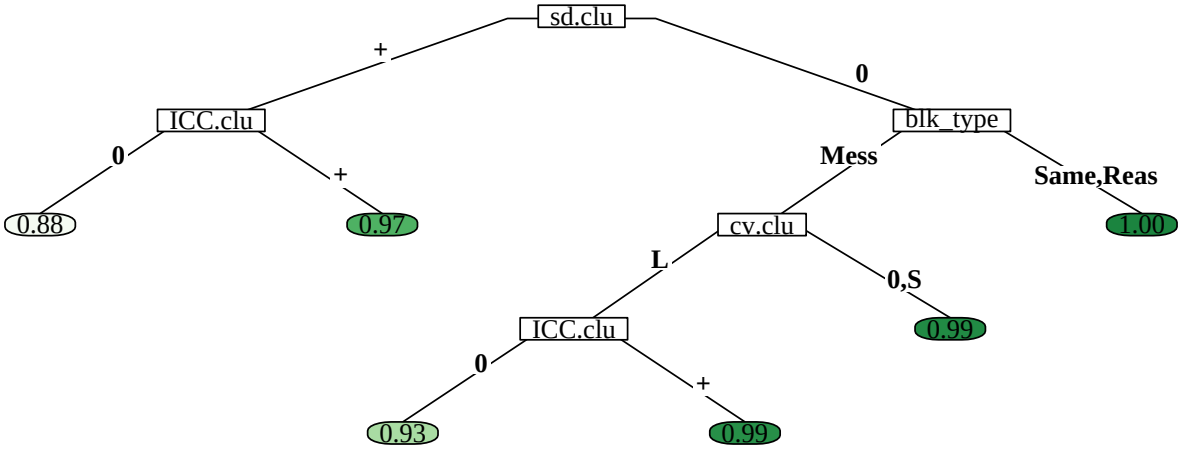


Figure 12: Regression tree which predicts the ratio of true finite population standard error over true super population standard error under different data generating parameters. These two quantities are the same when this ratio is one.

To then see which standard error our estimators are estimating, we subset to those simulations where there is cluster level impact heterogeneity but the cluster ICC is 0, and then calculating the ratios of the average estimated standard error to the true finite population standard error and true superpopulation standard error. Figure 12 compares these, and we see that the superpopulation ratios are generally closer to 1 than finite population (Figure 13). We thus conclude that the standard errors are more accurately characterized as targeting the super-population standard errors (with the blocks being fixed and the clusters being sampled from block specific superpopulations).

That said, there is still systematic inflation of the fixed effect estimators at left of Figure 13, meaning these estimated standard errors are potentially accounting for variation not present in our DGP. As discussed above, we believe this inflation occurs when the blocks have different average treatment impacts. In this case, because the fixed effect estimator estimates a common impact, the additional variation from the treatment variation gets forced into the residuals.

Indeed, when we fit a tree of only the fixed effect estimators on the subset of scenarios from Figure 13, we see that the most inflation is in those scenarios with large block variation and generally homogeneous cluster sizes (see Figure 14).

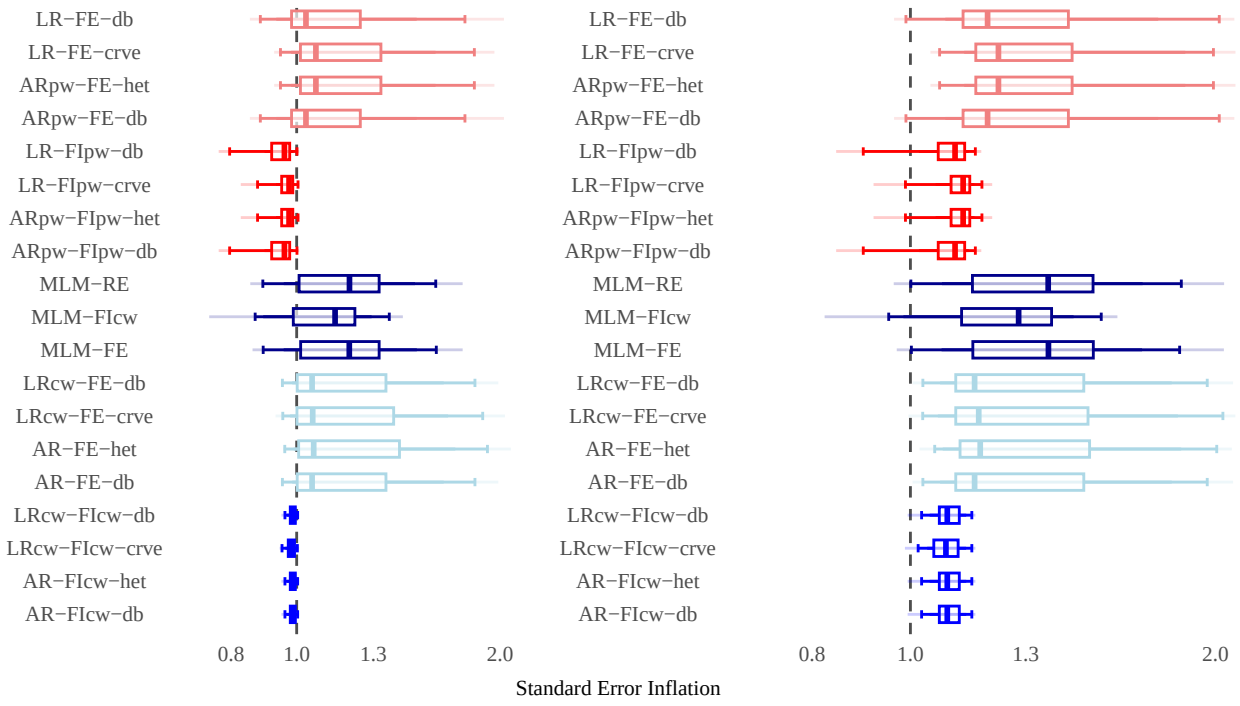


Figure 13: Figure 13: Boxplot of estimated standard error inflation across simulation scenarios with no cluster-level ICC and non-zero treatment effect heterogeneity by method. Standard error inflation is measured as the ratio of the estimated standard error over the true standard error. The faint horizontal lines extend to the full range, the bars span the 2.5th to 97.5th percentiles, and the boxes span the 10th to 90th percentiles.

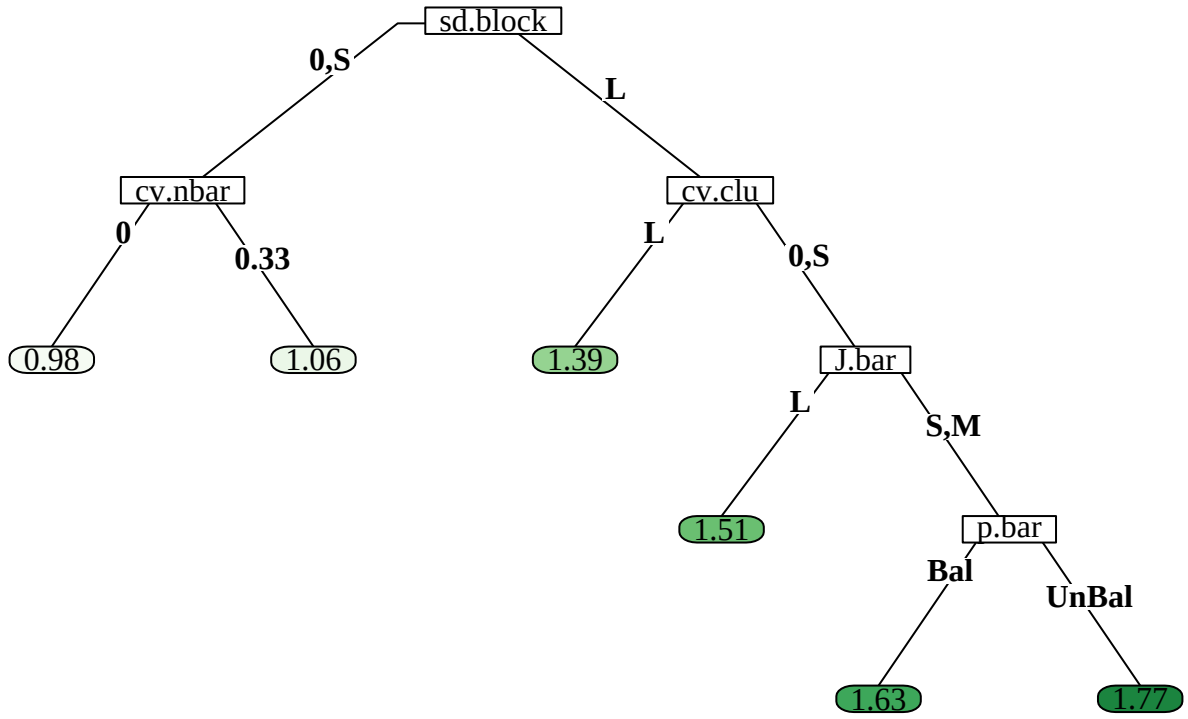


Figure 14: Tree predicting superpopulation standard error inflation for fixed effect estimators on those scenarios with notable finite population versus super population standard error differences.

4.5 Why does the range in the estimates produced by the different estimators get larger as the degrees of freedom get smaller? Is the cause estimator bias or estimator instability?

In the empirical studies the range of estimates produced by different estimators was larger for the low degrees of freedoms studies. We examined this with our simulations. We find that as the degrees of freedom get smaller, the estimate instability gets larger (Figure 15). On the other hand, we see no clear relationship between mean bias and degrees of freedom in a given scenario (Figure 16). Overall we conclude that the range of estimates increases because the estimates are more unstable, and not because of anything to do with estimator bias.

5 The Design of Simulation II

We do not generate data for Simulation II, but instead used the following process for each outcome j in our set of outcomes across each of our empirical datasets (excepting a few studies that were inaccessible):

1) Load the empirical data for outcome j, patch missing data, etc.
2) Repeatedly, for simulation runs $r = 1, \ldots, R$ (with $R = 20$), permute the treatment assignment across clusters and within blocks.
3) For each permutation, estimate the treatment effect and associated empirical standard error using each method of estimation.
4) For each run and outcome, calculate the range of impact estimates and the ratio of the largest to smallest standard error, across all the estimators that were defined for that run. We do this both overall, and for the group of estimators within each estimand (Person or Cluster).

Due to the permutation, there is no actual average treatment effect or cross-block treatment effect variation. All individuals have the same outcome whether assigned to treatment or control. Any actual treatment effect that may have existed in the empirical data is therefore pushed into individual and cluster variation. For

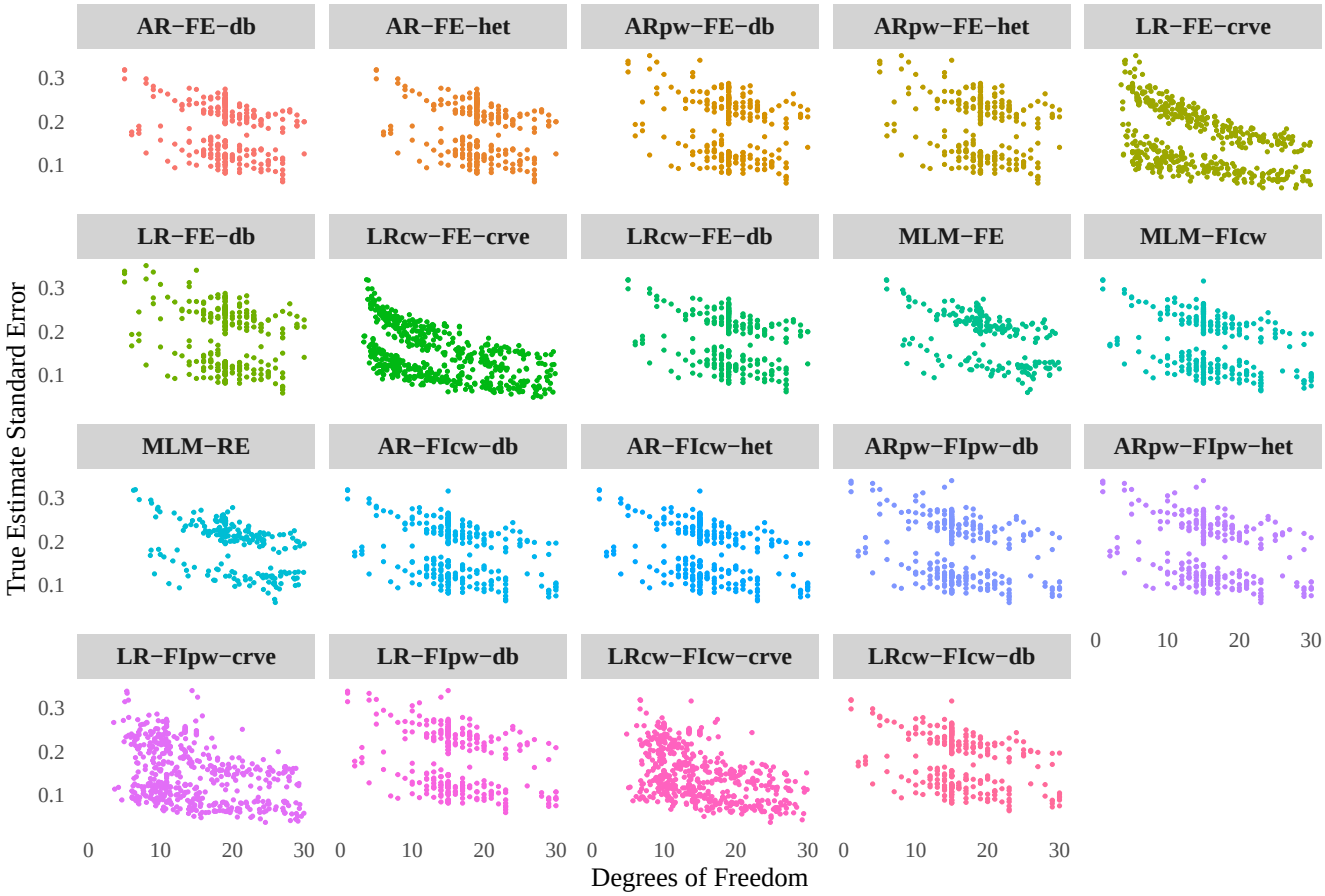


Figure 15: Scatter plot of true estimate standard error versus degrees of freedom by method. Each dot is a given DGP scenario and each panel shows the relationship for a different estimation method. (For 30 df or less.)

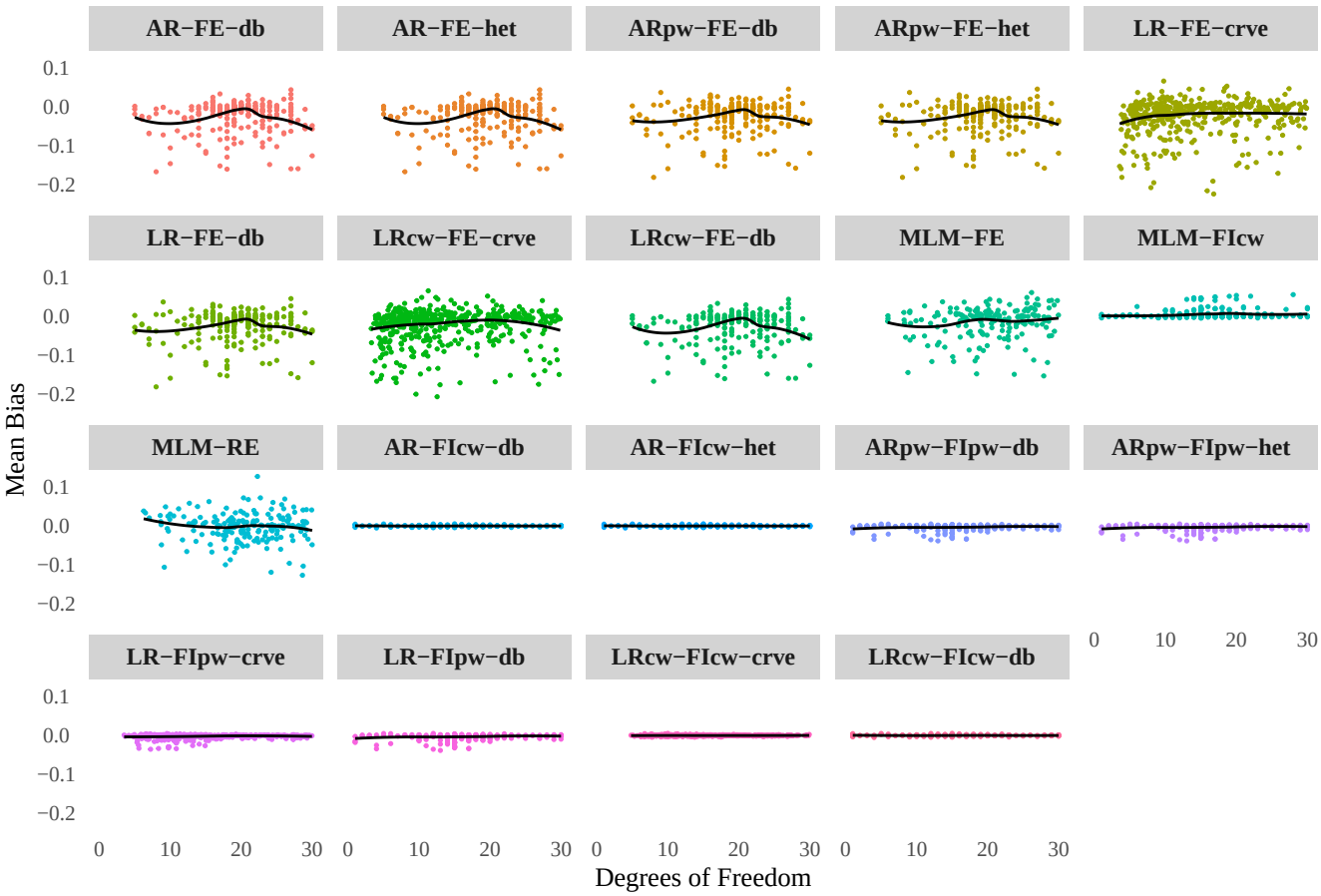


Figure 16: Scatter plot of mean bias versus degrees of freedom by method. Each dot is a given DGP scenario and each panel shows the relationship for a different estimation method. (For 30 df or less.)

example, if a cluster had a positive average effect, the permutation simulation views this cluster as just having a slightly higher intercept than it would have had, had it been a control cluster in reality. The structure of the data, the relationship of covariates to outcome, etc., is all preserved by this approach.

6 Simulation II Results

We next systematically go through our Simulation II results, following our research questions.

6.1 How do the ranges of ATE estimates we saw empirically compare to the ranges we get from the simulation?

For each study, we can look at the actual range of ATEs across estimators from the empirical data (red dot) vs. the ranges we get from the permutations (the boxplot). The boxplot shows what kinds of ranges we would "naturally" see when there is no treatment effect or treatment variation.

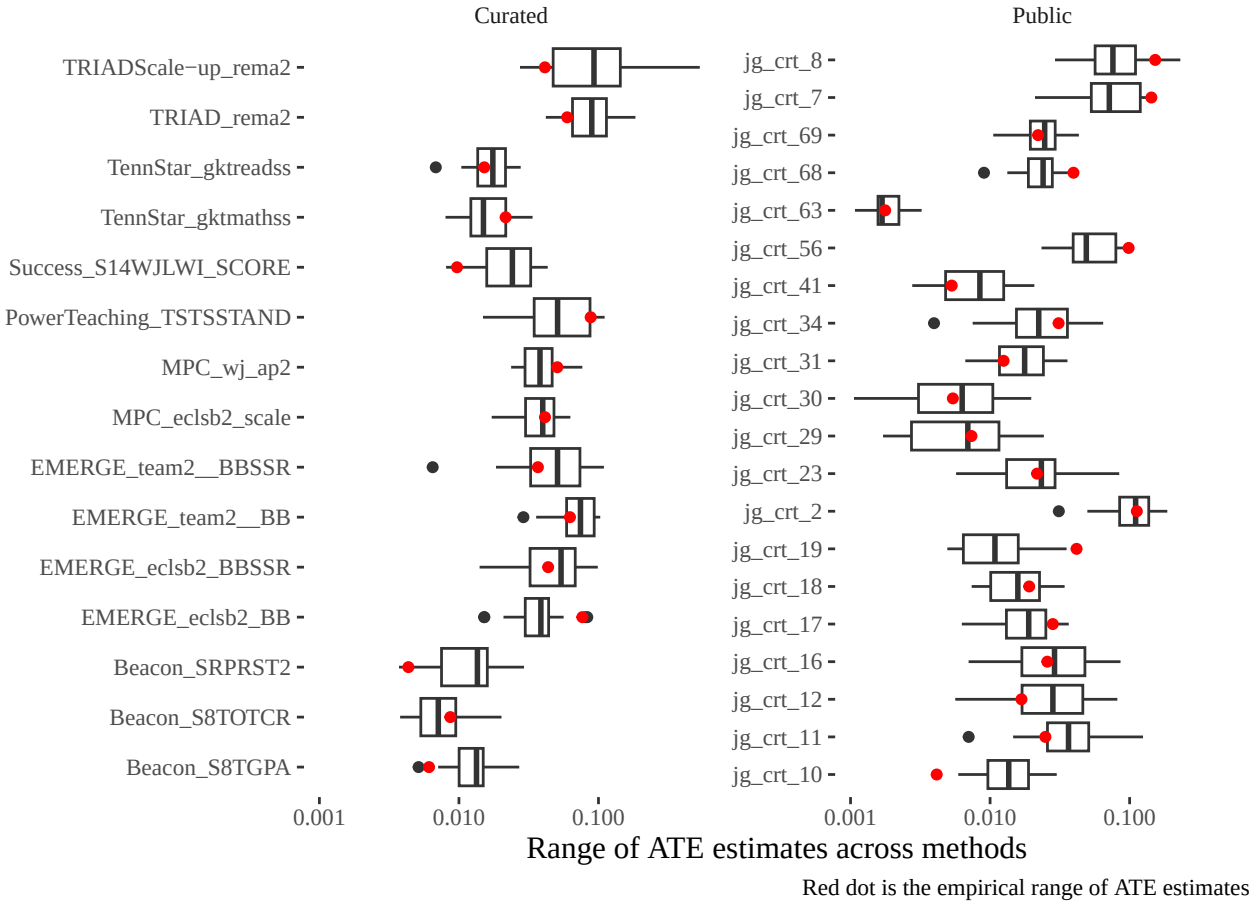


Figure 17: Boxplot of the range of ATE estimates across estimators for each study, with the red dot indicating the empirical range of ATE estimates across estimators for that study. The boxplots show the distribution of ranges we would see if there were no treatment effect variation whatsoever.

The red dots generally lie in the range of the boxplots; this means the range we saw in our empirical data is entirely consistent with the ranges we would see if there were no variation in treatment effects whatsoever. Our observed range, therefore, does not appear to be driven by the estimators targeting different estimands.

To quantify this, we calculated the percent of studies where the empirical range was outside the 90% interval of the simulated ranges. Two studies were notably too low, and three were notably too high (out of 35). We

looked within estimand, and results were broadly the same.

6.2 How does the min-max ratio of estimated SEs compare to the min-max ratio we get from the simulation?

For each simulation iteration we do the following:

1) calculate the max/min ratio of estimated standard errors across all the estimators for that trial.
2) Calculate the inner-quartile range of these ratios.

We then compare the distribution of these IQRs of ratios to the empirical IRQ of the empirical ratios we got from the original data. For the empirical data (for the 35 outcomes used in this simulation), the IQR was 0.25. The average across our simulations was 0.27 with a standard deviation of 0.05. The empirical range was well within the distribution of synthetic ranges, suggesting again that the variation we saw in the empirical data is consistent with what we would see if there were no treatment effect variation whatsoever.

We note that we are intentionally keeping the exploration of estimated SEs fairly minimal. The permutation approach does not change within cluster variation, or block variation, meaning that the estimated standard errors for many of the estimators are remarkably stable across permutations. This means idiosyncrasies of the initial data set cause the estimated standard errors to be systematically, larger or smaller than the true standard error as found with the permutation approach. This has interesting implications regarding understanding fully finite sample uncertainty in the context of a fixed data set with cluster variation.

To attempt to address the stability of the SE estimators, we could extend the idea of calibrated simulation to generate new clusters and blocks with features similar to the estimated features of the empirical data. We leave this for a future work.

6.3 Are some of the estimators more precise across permutations than other estimators?

We compare standard errors of methods by looking at the ratio of the true SE vs. median true SE within study, and then making a boxplot. We look across all estimators as there are no treatment impacts, so all the estimators are estimating the same estimand (an ATE, regardless of weighting, of 0).

All the methods have a SE ratio that is, at the median, the same as the other methods. This suggests that the methods all have the same average overall performance across datasets. That said, some methods have long right tails, meaning there are some studies where their true SEs are up to 30% larger than the median performing method.

We can organize these same values by study to see if some studies naturally lead to bigger ranges of true SE:

The second plot shows that some studies indeed have a wider range of true SEs than other studies. For example, some estimators in `jg_crt_2` have larger SEs and some have smaller SEs. By contrast, almost all estimators in `Beacon_S8TGPA` have the same true SE (note the very narrow boxplot).

Using the true standard errors, we can recreate the dot plot where each dot is the max/min SE ratio for a specific outcome. We can compare the max/min ratio for the true SE and the estimated SE.

These ratios are all inflated by simulation error (we do not fully know the true SEs), but even so, we are not seeing major ratios in the *true* SE, even though we do see them in the *estimated* SE; recall for the estimated SEs we had ratios well above 2. This suggests the large range we see in the empirical data is because the *estimated* SEs vary a lot, not because the estimators are necessarily very different in terms of their precision. In particular, if some estimators have very poorly estimated SEs, the ratios can be driven quite high.

7 Conclusion

Overall, across the two simulations, we find remarkable similarity between estimator performance across a wide range of estimators and approaches. That said, in some circumstances, there are notable differences, but

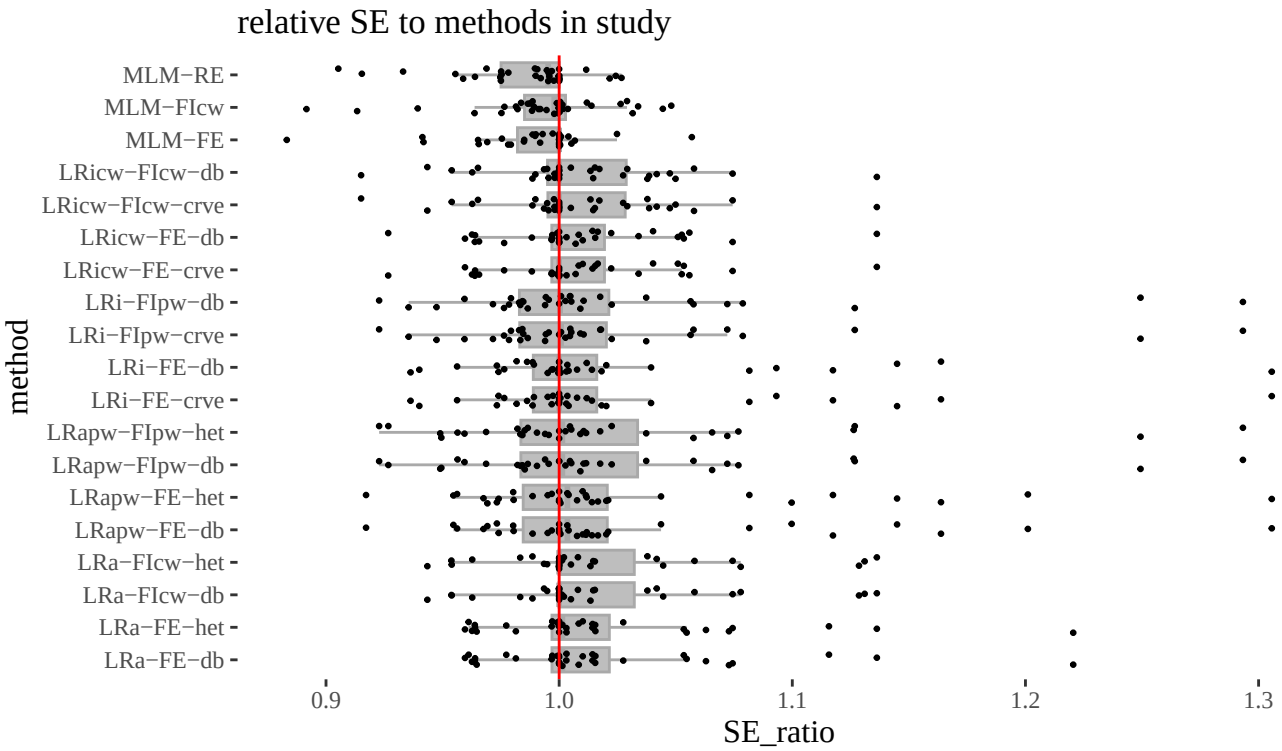


Figure 18: Ratio of true SE to median true SE by method across all studies

those differences tend to arise in rather extreme scenarios.

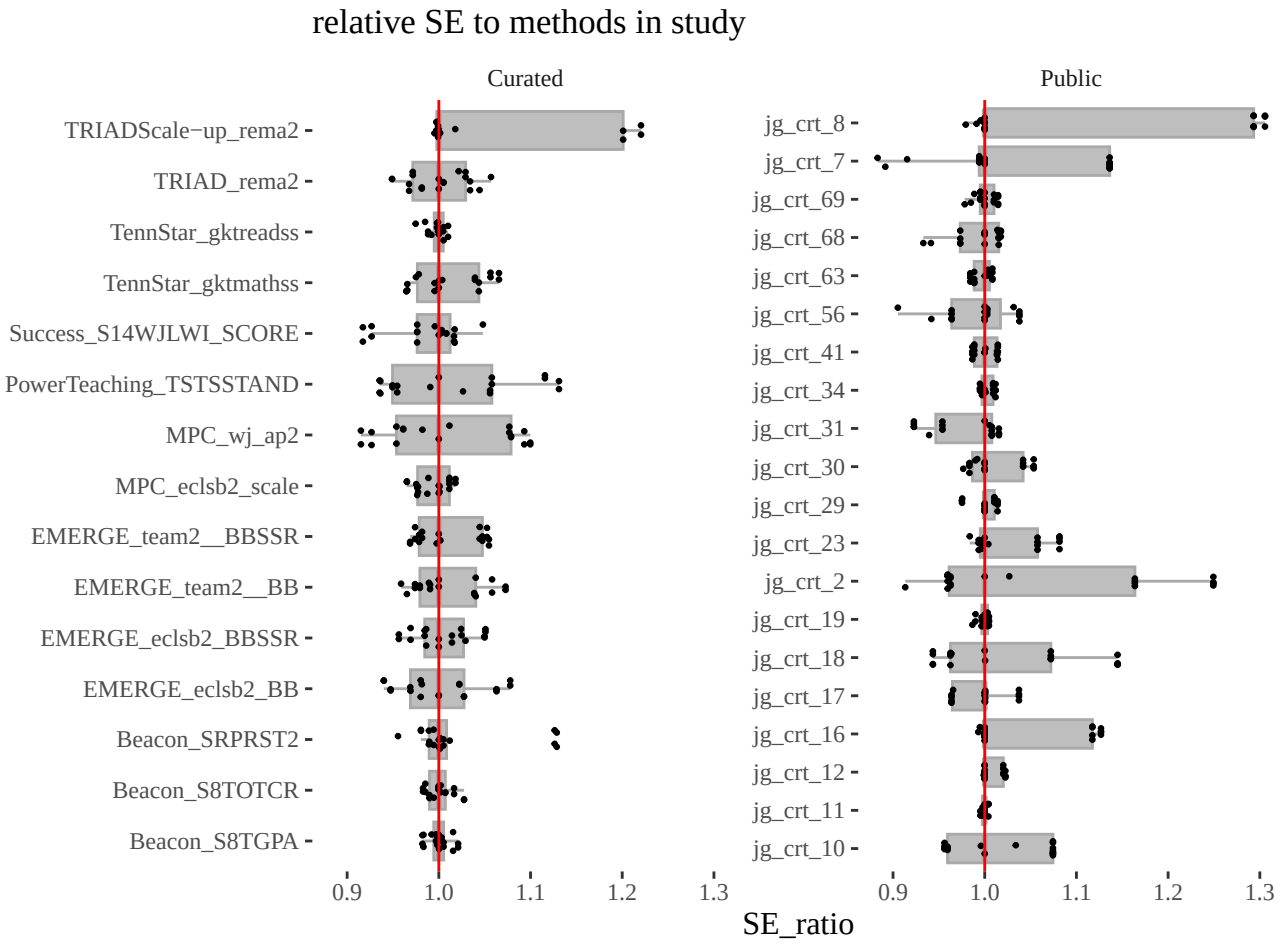


Figure 19: Ratio of true SE to median SE across all estimators grouped by study

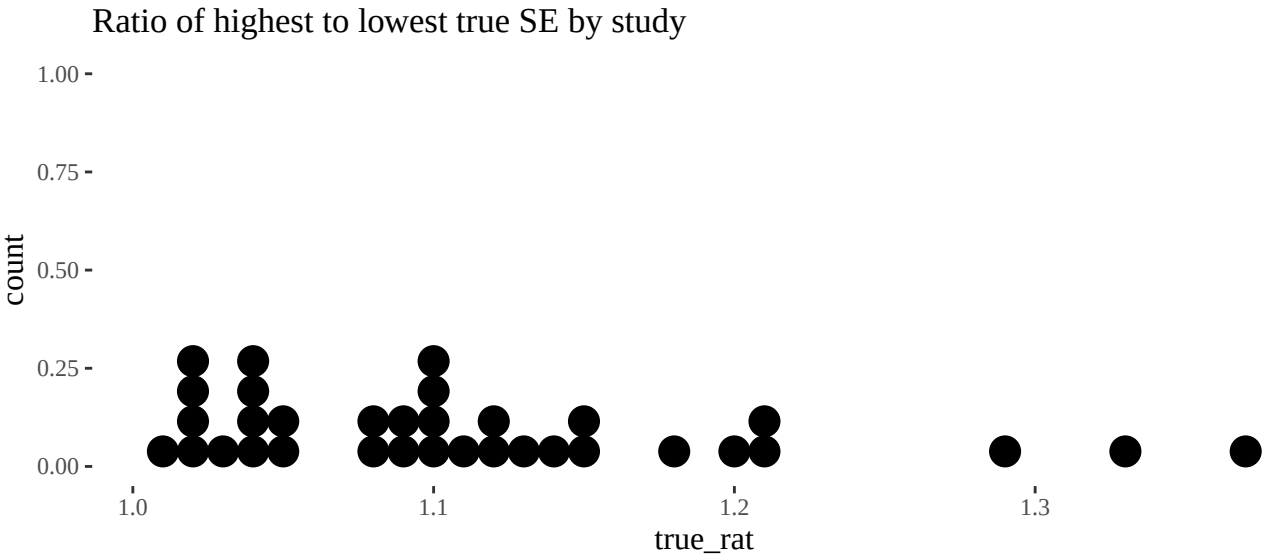


Figure 20: Ratio of max to min true SE by study

8 Appendix: Additional details of Simulation I's DGP

We next give some further technical detail of how we implemented Simulation I.

8.0.1 How the data are generated

For a given scenario, we start with a set of data generating parameters as described above. For example, Table 4 shows one of the more extreme scenarios. Using these parameters, we first generate the "grid" of individual blocks. The individual blocks describe different superpopulations of clusters, one for each block. Once the blocks are created, we then generate the clusters within each block.

Table 4: Simulation Scenario Parameters

variable	value
ATE	0.2
ICC.block	0.2
ICC.cluster	0.2
J.bar	10
K	5
R2.cluster	0.5
cv.block	0.85
cv.cluster	0.33
cv.nbar	0.33
n.bar	20
nbar.J.cor	0.8
num.cov.cluster	5
num.cov.indiv	0
p.bar	0.7
proptx.impact.cor	0.8
sd.block	0.6
sd.cluster	0.3
size.impact.cor	0.8
size.impact.cor.cluster	0.8
size.intercept.cor	0.8
variable.p	TRUE

Generating the individual blocks. To generate the individual blocks, we generate the block-level pairs of average control-side outcome and average treatment effect (in the case of treatment variation) with a multivariate normal distributions. We then generate the other block characteristics (e.g., proportion of clusters treated) according to various specified distributions described below. Finally, for any features of blocks that are to have non-zero correlations, we induce a desired level of correlation by permuting one of the two features across blocks (see below for how we do this).

When generating block mean outcomes and mean effects, we generate our values in a way[1] to exactly match the desired overall average and standard deviation. Doing so mitigates worry that the simulation results might be dependent on *which* blocks we have generated; otherwise we would have to worry that a particular set of blocks had randomly more, or less, variation than desired. That said, we are generating *block level characteristics*, but, if there is cluster variation, the cluster-average effects in a generated sample could deviate from these targets quite substantially, especially if number of clusters and treatment effect are correlated.

Overall, however, the finite nature of the blocks should not be a large concern. In particular, because we have a large number of factors in our experiment, when we look at the main effect of any given factor, we will

[1]Technically, we use the "empirical" setting of the multivariate normal generator to do this.

be averaging across many different scenarios, each with a different specific set of blocks. The averaging would ameliorate any dependencies on which specific set of blocks were generated.

Once generated, we store the block characteristics as a small dataset, which we call a "grid," with each row being a block, and each column a characteristic we need for generating the within-block data (number of clusters, average cluster size, etc.). To illustrate, here is the "grid" of block-level characteristics for the above scenario in our main simulation:

did	tau	alpha	J	p	n.bar	ICC.cluster	k.clu	R2.cluster	cv.cluster	cor.clu	k.indiv
1	0.2	0.5	46	0.7	21	0.2	5	0.5	0.3	0.8	0
2	0.8	0.1	3	0.7	14	0.2	5	0.5	0.3	0.8	0
3	-0.3	-0.7	2	0.5	9	0.2	5	0.5	0.3	0.8	0
4	-0.4	0.0	3	0.7	20	0.2	5	0.5	0.3	0.8	0
5	0.8	0.2	4	0.8	12	0.2	5	0.5	0.3	0.8	0

Given our block statistics, we can describe the data we would get from this fixed set of blocks:

```
## K=5  J.bar=6.20 (3.70, cv=0.60)  n.bar=17.42 (4.21 cv=0.24)  p.bar=0.66 (0.13)
## District stats:
##  block ATE=0.20 (0.60)   block ICC=0.45
##  cor(J,tau)=0.93     cor(ptx,tau)=0.80
##  cor(J,alpha)=0.80       cor(nbar, J)=0.81
## Cluster stats:
##  cluster ICC=0.20 (0.00)
##  avg cluster size = 17.4 (range 5--36, sd = 7.2, cv=0.41)
##  cor(n,tau) = 0.66 (across blocks)
## ATE information:
##  Within-block cluster ATE and variation: 0.20 (0.27)
##  Total cluster ATE: 0.46 (0.59)
##  Total individual ATE: 0.62 (0.59)
##
##   * Some stats estimated on 43156 generated observations across 2480 clusters.
```

The block ATE is the simple average of the block-level cluster-average ATEs. It is as we targeted. The bottom lines of the display show the ATEs for a large dataset generated for this set of blocks: these are our approximated target superpopulation parameters for this scenario.

In this specific example, these are different from each other and from the block average because of the variation in block sizes and other aspects of the DGP. We also see fairly high correlations of all sorts of features of the data. The numbers in parentheses are the standard deviations of a given value.

Generating clusters and individuals within blocks. Given our grid of block characteristics we, for each block in turn, use a multilevel model data generating process (as described more fully in Miratrix et al., 2021) to generate the target number of clusters and individuals.

To keep everything in effect size units, we have a target within-block total variation of $1 - ICC_3$, where ICC_3 is the target intraclass correlation for the blocks. Note that we have $ICC_3 + ICC_2 + \sigma^2_\epsilon = 1$, with ICC_2 being the cluster-level ICC and σ^2_ϵ the individual-level residual variance.

8.0.2 How block-level quantities are correlated

Given two vectors of numbers, x and y, that we wish to correlate with target correlation ρ, we do the following:

1. Generate a bivariate standard normal with correlation ρ.
2. Re-order the y values so the pairs of the ranks of x and the ranks of the reordered y correspond to the pairs of ranks of the bivariate distribution.

If x and y are not normal, then this will not produce correlations that have the target correlation on average, but it will generally give high correlations when ρ is high, and negative correlations when ρ is negative. We are correlating based on the normal transform of the ranks, so this is somewhat akin to a rank correlation.

If we have multiple correlations, then we do these in sequence, holding one of the variables as fixed so we do not undo any of the other correlations. For example, say we need to generate a scenario with both a correlation between the number of clusters in a block and average impact in the block, and a correlation of proportion treated in a block and average impact in the block. To do this, we first generate the three vectors of number of clusters, average impact, and proportion treated. We then rearrange, as described above, the number of clusters so it correlates with average impact, and next rearrange the proportion of clusters treated so it correlates with average impact. This will induce a correlation between number of clusters in a block and proportion treated in a block, because they are both correlated with the block-level ATE.

8.0.3 How cluster and block sizes are generated

We have three kinds of sizes to generate: number of clusters per block, number of individuals per cluster, and average size of cluster for each block.

For each of these we use a similar approach. We first specify the desired mean and coefficient of variation (cv). For example, we might want an average of 8 clusters per block with a cv of 0.75. This means the standard deviation of our values would be $sd = cv * mean = 0.75 * 8 = 6$.

We then have three possibilities: either the variance is less than the mean, in which case we can use a binomial distribution, or the variance is larger than the mean, in which case we can use a negative binomial distribution. If the variance is very close to the mean, we use a poisson distribution.

We use this approach for the average cluster size across blocks, even though the average cluster sizes could be fractional.

Technical Supplement for "An Applied Researchers Guide to Estimating Effects from Multisite Cluster Randomized Trials: Estimands, Estimators, and Estimates"

Contents

1 Introduction

Cluster randomized controlled trials (RCTs) are a powerful and popular way to estimate the average causal effect of education programs, policies, or practices. This document contains technical details for estimators used to estimate the average treatment effect (ATE) in the context of Blocked, Cluster RCTs. By "clustered" we mean entire groups of units are tied together and collectively assigned to treatment or control. By "blocked" we mean that random assignment occurs within different groups of units. That is, we might randomize a pre-determined proportion of clusters (e.g., schools) within each block (e.g., district). We look at outcomes on the students (persons) nested within the schools, nested within the districts.

A researcher has to first decide what statistic they want to estimate, including over what population of units (referred to as the estimand or parameter of interest). For example, we might target the ATE across all people in a study's analytic sample. Once this estimand is identified, a researcher must then decide how to estimate it, given the observed data, and also how to obtain uncertainty estimates (standard errors) for those estimates. We call the ATE estimators "effect estimators" and the uncertainty estimators "standard error estimators."

In the following sections, we systematically go through four general classes of effect estimators, discussing variants of implementing each. The four general classes of estimators are:

- linear regression on individual data,
- linear regression on cluster level aggregate data,
- multilevel models (aka hierarchical linear models or random effects), and

- strictly design-based.

For each class of effect estimator, we note possible standard error estimators, but discuss uncertainty estimation more completely in Section 5.

Because blocking is so prevalent for cluster RCTs, we generally discuss estimators for this more complex context throughout, leaving the simple cluster RCT as a special case (although we sometimes focus on it to motivate overall concepts).

Before we dive into the estimators themselves, we use the remainder of this section to walk through a series of cross-cutting considerations such as how covariates are generally included and the definition of the Intraclass Correlation Coefficent (ICC).

Notation-wise, we have individuals i (e.g., people) in clusters j (e.g., schools), nested in blocks k (e.g., districts). Our outcome is then Y_{ijk}, and treatment assignment is an indicator Z_{jk}. We have K total blocks and J total clusters, with J_k clusters in block k. Similarly, we have N_k people in block k and N_{jk} people in cluster j of block k. Our full set of notation is described on Table 1.

We generally assume we have at least one treated and one control cluster in each block k, although many of our uncertainty estimators require at least two in each (see Section 5.2).

Table 1: Notation Reference

Symbol	**Description**
Indices	
i, j, k	Individuals (e.g., students), Clusters (e.g., classrooms or schools), Blocks (e.g., schools or districts)
Sample sizes	
N_{jk}, N_k, N	Cluster, block, and total sample sizes (people)
N_{zk}	People in block k with treatment status z
$\bar{n}$	Harmonic mean cluster size
J_k, J	Number of clusters in block k; total clusters
J_{zk}	Clusters in block k with treatment status z
K	Total number of blocks
Outcomes and treatment	
Y_{ijk}	Observed outcome (Y_{ij} when no blocks)
$Y_{ijk}(z)$	Potential outcome under condition $z \in \{0, 1\}$
$\bar{Y}_{jk}(z)$	Cluster mean potential outcome under condition z
$\bar{Y}_{jk}$	Cluster mean observed outcome
$\bar{Y}_z$	Overall observed mean for treatment condition z
Y^T_{jk}	Cluster outcome total, $N_{jk}\bar{Y}_{jk}$
Z_{jk}	Treatment assignment indicator for cluster j, block k
Covariates	

(continued on next page)

(continued from previous page)

Symbol	Description
X_{ijk}	Individual-level covariate vector
$\bar{X}_{jk}$	Cluster mean of X_{ijk}
W_{jk}	Cluster-level covariate vector (typically includes $\bar{X}_{jk}$)
$S_{q,ijk}$	Indicator that person ijk is in block q; i.e., $1_{\{q=k\}}$
Proportions and weights	
p_k, p_k^c	Proportion of people / clusters treated in block k; p, p^c for overall
w_k, w_{jk}	Block- and cluster-level weights
Variance components	
ICC_2, ICC_3	Cluster- and block-level intraclass correlations
R_1^2, R_2^2	Proportion of level-1 / level-2 outcome variance explained by covariates
η_{11}	SD of block-average treatment effects across the superpopulation of blocks
τ_0^2, σ^2	Variance of cluster random effects $u_{0,jk}$; individual residuals ϵ_{ijk}
Estimands	
β_{ijk}, β_{jk}	Person- and cluster-level treatment effects; not observable
$\beta_{k,\text{cluster}}$, $\beta_{k,\text{person}}$	Cluster- and person-weighted block-level average effects
β_{cluster}, β_{person}	Overall cluster- and person-weighted average impacts
Estimators	
$\hat{\beta}_k$	Estimated ATE for block k; $\widehat{\beta}_{k,w}$ when targeting weight w
$\widehat{\beta}_{type}$	Estimated overall ATE of a given type
$\widehat{SE}_k$	Estimated standard error for $\hat{\beta}_k$

1.1 How to read this manuscript

We do not recommend reading this manuscript as a traditional paper. The bulk of the manuscript is a reference of the different estimators, capturing some technical details as to what they are. We recommend reading Section 1 on the overall framing of all of the estimators, and then dipping into sections 2 through 4 given specific interest. Our table of contents, above, should help guide this process.

Section 5 provides an overview of uncertainty estimation, and Section 6 an overview of true unbiased estimators. This last section is supplementary; we did not use these estimators in our main paper as they were found to be too unstable in the empirical examples we explored.

Finally, Section 7 gives an high level overview and some remarks regarding our companion R package.

1.2 Estimands

Before we look at estimators, let's consider estimands. An estimand is the quantity we are trying to estimate. We can represent these using the potential outcomes framework (see, e.g., Rosenbaum et al. (2020)). In particular, for each person we have $Y_{ijk}(0)$ and $Y_{ijk}(1)$, which represents their outcome under the control and treatment conditions, respectively. Thus, the person-level treatment effect is defined as $\beta_{ijk} = Y_{ijk}(1) - Y_{ijk}(0)$. Technically, this is the treatment impact on a person from treating a person's entire cluster, which would include spillover effects and so forth within the cluster. For any person, we can never observe both potential outcomes, and thus we will never know β_{ijk}, regardless of how we assign treatment.

Similarly, for each cluster we have the true average treatment effect (ATE) in that cluster as

$$\beta_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} Y_{ijk}(1) - Y_{ijk}(0),$$

where N_{jk} is the size of cluster jk. β_{jk} is the ATE of the people in cluster jk. Similar to the β_{ijk}, we have no ability to estimate a cluster's β_{jk} because we treat or do not treat the entire cluster. This is very different from trials that randomize individual people within clusters (aka multisite individually randomized trials), where, within each cluster, we get to see some of the people treated and some not treated, allowing for (perhaps noisy) estimates of all the cluster ATEs.

Because of this, a cluster-randomized experiment can be thought of as an individually randomized experiment where the individuals are the clusters themselves. For example, we can think of the clusters as the unit of treatment, and the mean outcome of the people in the cluster as the cluster's outcome. This view is prevalent in much of the design based causal inference literature. It is also the intuition behind why we might think of some very large (in terms of N, the total number of individuals) experiments as quite small (in terms of J, the total number of clusters).

Within a block k we have different possible estimands of interest. Similar to Miratrix et al. (2021), we can have the simple average of the β_{jk} (the **cluster-average effect**) of

$$\beta_{k,cluster} = \sum_j \frac{1}{J_k} \beta_{jk} = \sum_j \frac{1}{J_k} \bar{Y}_{jk}(1) - \sum_j \frac{1}{J_k} \bar{Y}_{jk}(0),$$

or we can have the person-weighted average of the β_{jk} (the **person-average effect**) of

$$\beta_{k,person} = \sum_j \frac{N_{jk}}{N_k} \beta_{jk} = \sum_j \frac{N_{jk}}{N_k} \bar{Y}_{jk}(1) - \sum_j \frac{N_{jk}}{N_k} \bar{Y}_{jk}(0),$$

where $N_k = \sum_j N_{jk}$ is the total number of people in block k. One can also write $\beta_{k,person}$ as the average across all individuals in the block, with $\beta_{k,person} = \sum_{i,j} Y_{ijk}(1) - Y_{ijk}(0)$.

Both the above estimands are weighted averages of the cluster ATEs. The former is the average effect of the cluster ATEs, counting each cluster equally, regardless of its size. The latter estimand is the average effect of the individual ATEs.

We write both estimands in two ways: first, as an average of the β_{jk} and second, as a difference in $\overline{Y}_{jk}(z)$. We find the former to be more conceptually intuitive. The latter representation closely ties to the estimators, discussed later, where we have to estimate these targets of inference using cluster average outcomes due to the cluster assignment. Thus, the latter representation is helpful when considering the connections between estimands and estimators.

We can then weight these β_k across the blocks in different ways: a simple average (weighting each block equally), giving the block-average of the block-level estimands, regardless of block size; a cluster-weighted average, weighing each block by the number of clusters in it; or a person-weighted average, weighing each block by the number of people in it.

It is not typically sensible to weight the different levels differently, for example, by calculating a person-weighted average across blocks of cluster-average impacts within blocks. The weighting at level 3 should typically be aligned with the weighting at level 2. This results in two primary estimands: the average impact across all people in the experiment, and the average impact across all clusters in the experiment:

1. **person-weighted**: The average impact for all people:

$$\begin{aligned}\beta_{person} &= \sum_k \frac{N_k}{N}\beta_{k,person} \\ &= \sum_k \frac{N_k}{N}\sum_j \frac{N_{jk}}{N_k}\beta_{jk} = \sum_k\sum_j \frac{N_{jk}}{N}\beta_{jk} \\ &= \sum_k \left(\sum_j \frac{N_{jk}}{N}\bar{Y}_{jk}(1) - \sum_j \frac{N_{jk}}{N}\bar{Y}_{jk}(0)\right)\end{aligned}$$

2. **cluster-weighted**: The average across all clusters of the cluster-average impacts:

$$\begin{aligned}\beta_{cluster} &= \sum_k \frac{J_k}{J}\beta_{k,cluster} \\ &= \sum_k\sum_j \frac{1}{J}\beta_{jk} \\ &= \sum_k \left(\sum_j \frac{1}{J}\bar{Y}_{jk}(1) - \sum_j \frac{1}{J}\bar{Y}_{jk}(0)\right)\end{aligned}$$

We write each estimand terms of the underlying cluster-specific ATEs (the β_{jk}s) as well as the average potential outcomes in both arms (the $\bar{Y}(z)_{jk}$s). As noted previously, this latter representation ties more closely to how we estimate our target quantities of interest; see, e.g., Section 6 on design based estimators.

From a modeling perspective, none of the models we consider explicitly model treatment effect variation within cluster (i.e., letting individual people have different impacts), or even allow impact variation within block (i.e., letting clusters have different impacts), and some do not even explicitly allow for blocks to have different average impacts either. When such variation does exist, the models will tend to average across the people, clusters, and blocks

differently, implicitly targeting the different estimands listed above; in our main paper and this document we identify which estimators are targeting which estimand.

1.2.1 Defining block and cluster sizes

We generally assume that the J_k, N_k, and N_{jk} are immediately observed from the data. In some cases, one might know the *actual* block and cluster sizes, and want to use those as weights. Several of the estimators discussed below allow for the inclusion of regression weights, and thus easily allow for this extension. In this case, we would view the experimental sample as a subsample of some finite, but larger, population. See, e.g., Schochet (2015) and Schochet et al. (2021) for more explicit handling of regression weights; these works also note that weights might be used to handle non-response or other forms of attrition.

More broadly, if your goal is to actually estimate a student-average effect, and you are working with a same-size sample of students from a set of schools of varying sizes, then you should weight the units in the larger schools more heavily to target the proper estimand. In all these cases, the estimand formulas above apply: you would simply use sample sizes taken from administrative records rather than counts from the data.

1.2.2 Block averaging

One might also imagine wanting to summarize at the block level for a block-level decision-maker (e.g., a district superintendent). In this case we might target the simple average across blocks of the block-level estimands (whether person- or cluster-weighted). This gives two additional estimands:

1. **block average of person-weighted**: The simple average of person-weighted block-level impacts:

$$\begin{aligned}\beta_{person-block} &= \sum_k \frac{1}{K}\beta_{k,person} \\ &= \sum_k \frac{1}{K}\sum_j \frac{N_{jk}}{N_k}\beta_{jk} \\ &= \sum_k \frac{1}{K}\left(\sum_j \frac{N_{jk}}{N_k}\bar{Y}_{jk}(1) - \sum_j \frac{N_{jk}}{N_k}\bar{Y}_{jk}(0)\right)\end{aligned}$$

2. **block-average of cluster-weighted**: The simple average of cluster-weighted block-level impacts:

$$\begin{aligned}\beta_{cluster-block} &= \sum_k \frac{1}{K}\beta_{k,cluster} \\ &= \sum_k \frac{1}{K}\sum_j \frac{1}{J_k}\beta_{jk} \\ &= \sum_k \frac{1}{K}\left(\sum_j \frac{1}{J_k}\bar{Y}_{jk}(1) - \sum_j \frac{1}{J_k}\bar{Y}_{jk}(0)\right)\end{aligned}$$

We generally think these estimands are not often policy relevant, and so do not focus on them in this work. Miratrix et al. (2021) provide some logical arguments why one might prefer various estimands.

1.3 Covariate adjustment

Cluster RCTs are notoriously difficult to power. To improve precision researchers frequently turn to person-level and cluster-level covariates to explain some of the outcome variation and thus reduce standard errors of an ATE estimator.

Most of the various estimation strategies, in particular the linear regression models and multilevel models, explicitly allow for covariate adjustment: simply add the baseline covariates into the regression. For example, for a non-blocked experiment we might fit a regression of the form:

$$Y_{ij} = \alpha + \beta Z_j + \gamma' X_{ij} + \lambda' W_j + \epsilon_{ij},$$

where X_{ij} is a vector of individual level covariates and W_j is a vector of cluster-level covariates. These covariates are generally aimed at improving the precision of the ATE estimator.

Individual level covariates can contain, in a sense, "hidden" cluster-level covariates if the group mean of the individual covariates varies substantially across clusters. If the cluster-level relationship to outcome is different than individual-level, one might decompose the covariates, fitting:

$$Y_{ij} = \alpha + \beta Z_j + \gamma' \left(X_{ij} - \bar{X}_j \right) + \lambda' W_j + \epsilon_{ij},$$

Now $\bar{X}_j$ is the cluster average of these individual covariates (so we are group-mean centering the covariates), and W_j would be a vector of cluster-level covariates that includes the $\bar{X}_j$ as elements.

Aggregation estimators, which work with cluster averages, can also allow for covariate adjustment; in these cases the person-level covariates are averaged by the cluster just like the outcomes, and then used in the aggregation estimators as cluster-level covariates.

1.4 Impact variation

Some of the estimators specifically model some aspects of possible treatment impact variation (e.g., those estimators that estimate different average impacts for each block and then average those block specific ATEs to obtain an overall ATE), some allow for it via putting no real structure on the residuals (e.g., those estimators that use heteroskedastic robust standard errors), and some make homoskedasticity or homogeneity assumptions that, in effect, presume there is no impact variation along some dimensions.

In terms of the point estimate, models that assume no impact variation will tend to average across any impact variation present. For example, a simple linear regression with a single parameter for the average impact will average individual impacts, providing something close

to a person average estimate. In this document we identify what estimand each method is targeting, which is equivalent to identifying how this weighting works out.

For uncertainty estimation, if there is impact variation then a model that does not allow for it is technically misspecified, unless one is using some sort of heteroskedastic robust standard error. We believe this misspecification is typically of little consequence for estimated standard errors and associated inference. For example, if the variance of the cluster-average treatment impacts is 0.2^2, when the overall variance in outcomes is 1 (so we are in effect size units), then the overall increase in variance on the treatment side, if the treatment heterogeneity is independent of the control potential outcome, is $1 + 0.2^2 = 1.04$, not much different from 1! If the ICC_2 is 0.2 on the control side, this would further assume the ICC_2 on the treatment-side is $0.2 + 0.2^2 = 0.24$, again not a major change. That being said, if the amount of treatment variation were severe (e.g., 0.7^2), and the proportion treated were unbalanced, then the standard errors under a homoskedasticity assumption could be biased.

1.5 Effect Sizes, Intraclass Correlation Coefficients, and R2 values

We generally assume impacts are measured in effect size units, which is equivalent to assuming our data are rescaled so the control-side overall variation across all blocks, clusters, and individuals is equal to 1.

We also have two Intraclass Correlation Coefficients (ICCs), one for level 2 and one for level 3. If we assume a three-level multilevel model for the control-side data, i.e.,

$$Y_{ijk} = \mu + r_k + u_{jk} + \epsilon_{ijk},$$

then ICC_2 is the variance of the u_{jk} over total variation:

$$ICC_2 = \frac{Var(u_{jk})}{Var(r_k) + Var(u_{jk}) + Var(\epsilon_{ijk})},$$

and ICC_3 is the variance of the r_k over total variation. These ICCs are not conditional ICCs; we represent covariate adjustment via the R^2s.

Our R_1^2, R_2^2 and R_3^2 measures then represent total explanatory power of all covariates for each level of the data hierarchy. For example, R_2^2 represents the total explanatory power of all covariates to explain the variation captured by ICC_2. Here, "all covariates" includes level 1 covariates, as a level 1 covariate with different averages for different clusters can be predictive of level 2 variation (see Section 1.3.

1.6 A universal equation of uncertainty

Assuming equal sized blocks and clusters, all the estimators collapse to the same canonical estimator best described by the design-based construction laid out in Section 6.2. The person-weighted, cluster-weighted, and block-weighted estimands will all also coincide.

For this ur-estimator, the true standard error, assuming the blocks are fixed (finite-sample), and the clusters and individuals within block are sampled from hypothetical superpopulations,

can be written as

$$SE(\hat{\beta}) = \sqrt{\frac{\text{ICC}_2(1-R_2^2)}{p(1-p)JK} + \frac{(1-\text{ICC}_2-\text{ICC}_3)(1-R_1^2)}{p(1-p)JKn}}.$$

where the parameters are defined as above, and p is the proportion treated. For a single block, simply set $K = 1$.

This formula is easily derived under an assumption of equal block and cluster sizes, and when these vary it is approximate. In particular, under variable cluster and block sizes, the standard error formula would no longer have clean JK and nJK terms, but instead would have effective sample size terms that would be functions of the sizes and weights given to the units; we do not include these here.

That said, the formula can still be used as an approximation. If only cluster size varies, the n can usually be taken as the harmonic mean of the cluster sizes. One could similarly plug in the harmonic mean of the J_k for J as well.

This ur-formula allow for easy examination of the sources of uncertainty: we see, for example, that the ICC_2 term is persistent even if cluster size, n, grows. The R_2^2 value is, therefore, likely more important than the R_1^2 value for controlling uncertainty.

If we think of the blocks as also sampled from a superpopulation, our expression gains a term at the third level:

$$SE(\hat{\beta}) = \sqrt{\frac{\eta_{11}^2}{K} + \frac{\text{ICC}_2(1-R_2^2)}{\bar{T}(1-p)JK} + \frac{(1-\text{ICC}_2-\text{ICC}_3)(1-R_1^2)}{p(1-p)JK\bar{n}}}.$$

where η_{11} represents how much the average intervention effects vary across blocks (technically, it is the standard deviation of the block averages of the cluster average effects across the superpopulation of blocks).

If η_{11} is substantial, then even if there are many clusters within block, and even if the clusters are large, there will be substantial uncertainty in the ATE estimate given the η_{11}/K term—this is the consequence of trying to generalize to a (typically hypothetical) superpopulation of blocks. This concern is general: any model that allows treatment effects to vary randomly across blocks will pay a cost in precision.

2 Linear Regression on Individual Data (LR)

One of the most common estimation strategies for cluster RCTs in the social sciences is linear regression. For a two-level context (no blocks), a natural choice is to fit a linear regression of the outcome onto the treatment indicator (and some person- or cluster-level control variables).

In particular, we might fit the simplest model of

$$Y_{ij} = \alpha + \beta Z_j + \gamma' X_{ij} + \lambda' W_j + \epsilon_{ij} \tag{1}$$

to the person-level data to obtain the estimate $\hat{\beta}_{LR-NoFE}$. LR-NoFE is shorthand for Linear Regression on individual data with no block Fixed Effects. Here Y_{ij} and Z_j are defined as before, and X_{ij} is a vector of covariates aimed at improving the precision of the ATE estimator, and W_j is a vector of cluster-level covariates. The $\gamma' A$ notation denotes the dot product of γ and A, i.e., $\gamma' A = \sum_p \gamma_p A_p$.

Here the individual nature of the data will naturally target a person-average estimand, as larger clusters will have larger weight. In the no covariate case, the above boils down to a simple difference in means estimate of $\bar{Y}_1 - \bar{Y}_0$.

Given that this model is estimating the mean of all treated units minus the mean of all control, we might expect it to be unbiased with respect to the person-average effect. Unfortunately, it is not, as discussed in Middleton and Aronow (2015). In fact, none of the main person-average targeting estimators are unbiased for the person-average estimand! We give an overview of the reasoning in Section 6.3.

Uncertainty estimation is usually done with cluster robust standard errors although there are some alternatives; see Section 5 for extended discussion.

2.1 Linear regression with block fixed effects (LR-FE)

The above regression model can naturally be extended to three levels by including fixed effects (i.e., a group of K binary indicator variables) for block. We give some notation for this to connect to other regression estimators discussed below. Define K block indicators $S_{q,ijk}$, $q = 1, \ldots, K$ with $S_{q,ijk} = 1$ if person ijk is in block q; in other words, $S_{q,ijk} = 1_{\{q=k\}}$, where $1_{\{a=b\}}$ is 1 if $a = b$ and 0 otherwise. The simplest OLS working model for estimating an impact would then arguably be

$$Y_{ijk} = \sum_{q=1}^{K} \alpha_q S_{q,ijk} + \beta Z_{jk} + \gamma' X_{ijk} + \lambda' W_{jk} + \epsilon_{ijk}, \tag{2}$$

where the α_q are the K fixed effects for the blocks. The X_{ijk} are individual covariates, and the W_{jk} are cluster-level covariates.

This model will obtain the estimate $\hat{\beta}_{LR-FE}$ (LR-FE is shorthand for Linear Regression on individual data with block Fixed Effects). We exclude an overall intercept to have individual fixed effects for each block and no reference block. This estimator is commonly used in the development economics literature (for examples see Banerjee et al. (2020); Berry et al. (2018); Mbiti et al. (2019); Andrabi et al. (2017)).

One advantage of the fixed effects model, compared to a model with treatment by block interactions (discussed later), is that the single impact parameter reduces the degrees of freedom used by the model, which could improve precision in smaller studies. Given how few clusters exist in many education experiments, this could be worthwhile (See Schochet et al. (2021)). That said, it comes with a cost, as we shift the estimator towards something like a precision weighted average of the block average effects.

For notational simplicity, we can leave the block indicator variables implicit by writing the exact same estimator as:

$$Y_{ijk} = \alpha_k + \beta Z_{jk} + \gamma' X_{ijk} + \lambda' W_{jk} + \epsilon_{ijk}.$$

Interestingly, for point estimation this model is the same as the fixed effect linear regression model used for a multisite individually randomized trial, when the blocks are viewed as sites. We know from work in that area (see, e.g., Miratrix et al., 2021) that this means the estimated $\hat{\beta}$ is nominally "precision-weighted" in that the estimated simple-difference-in-means impact in each block will be weighted proportional to $N_k p_k (1 - p_k)$, in the no covariate case, where p_k is the proportion of people treated in block k. For a multisite trial, these weights are the precisions of estimating the ATE within each site, assuming homoskedasticity. For a blocked cluster RCT, these are no longer exactly proportional to precision as they do not capture the uncertainty due to the cluster treatment assignment within a block. Departures can be especially severe if the clusters are unequally sized, and/or differently so across blocks. Thus the targeted estimand of this estimator does not have a clean interpretation—it is neither unbiased for any of our four estimands, nor optimally precise.

2.1.1 Weighted LR with block fixed effects (LRcw-FE)

The fixed effects estimator above can be adapted for targeting the cluster weighted ATE by adding weights. Each observation is weighted proportional to $\frac{1}{N_{jk}}$, where N_{jk} is the number of persons in cluster jk. These exact weights are calculated as:

$$w_{ijk} = \frac{N}{J} \times \frac{1}{N_{jk}}$$

where N is the total number of persons in the sample, J is the total number of clusters in the sample, and N/J is the average cluster size. The result is that students in larger clusters are down-weighted and students in smaller clusters are up-weighted, forcing the weighted number of persons in every cluster to be equal. This model produces the estimate $\hat{\beta}_{LRcw\_FE}$, where "cw" refers to cluster weighting.

By weighting all clusters equally (regardless of size), this estimator targets the effect for the average cluster, β_{cluster}. Due to the single fixed effect, the overall impact estimate will be a weighted average of the block-level cluster average effects, with weights of $J_k p_k^c (1 - p_k^c)$, where J_k is the number of clusters in block k and p_k^c is the proportion of the clusters in block k assigned to treatment. This estimator would be precision-weighted if we assume the uncertainty in all the cluster means is the same (unlikely, if the clusters are actually different sizes). Thus, this estimator in principle offers a classic bias-precision tradeoff. Bias can arise if the block ATEs are correlated with the proportion of clusters treated in the blocks. Precision is prioritized because, all else equal, blocks with more clusters and with an equal number of treatment and control clusters are given more weight as their impacts are estimated more precisely. But this is all without consideration of the different uncertainty induced by different sized clusters.

2.2 Linear regression with treatment by block interactions (Fully Interacted LR-FI)

Extending the above block fixed-effects model with treatment by block interactions, we can estimate impacts within each block. The model is then

$$Y_{ijk} = \sum_{q=1}^{K} \alpha_q S_{q,ijk} + \sum_{q=1}^{K} \beta_k S_{q,ijk} Z_{jk} + \gamma' X_{ijk} + \lambda' W_{jk} + \epsilon_{ijk} \tag{3}$$

$$= \alpha_k + \beta_k Z_{jk} + \gamma' X_{ij} + \epsilon_{ijk}. \tag{4}$$

The treatment-by-block interaction terms means each $\hat{\beta}_k$ will equal what the simple no-block regression model discussed above (see equation 1) would return if fit to just that block. In particular, each $\hat{\beta}_k$ is targeting the average impact for people in block k (albeit with some potential bias, as noted above, if cluster sizes vary within blocks). We end up with K estimates, $\hat{\beta}_k$, $k = 1, \ldots, K$, in total, that then need to be averaged to obtain an overall ATE.

Including level 2 covariates can eat up critical degrees of freedom. For example, including one in a matched pairs design will induce colinearity in regression and make the ATE estimator undefined.

2.2.1 Obtaining the overall estimate

To obtain an overall ATE we have to average our block-average impact estimates. To do this we need to make explicit choices about how to weight the elements. All estimators will be of the form

$$\hat{\beta}_{LR-FI} = \sum_{k=1}^{K} w_k \hat{\beta}_k,$$

where the weights w_k (with, generally, $\sum_k w_k = 1$) control how we aggregate the block-specific effect estimates.

We identify two natural choices of weighting. Others are possible. First, if we wanted to tend towards the person-average effect we would person weight as

$$\hat{\beta}_{LR-FIpw} = \sum_{k=1}^{K} \frac{N_k}{N} \hat{\beta}_k, \tag{5}$$

where N_k are the number of people in block k and N is the total number of people in the study.

Second, we could weight the $\hat{\beta}_k$ equally, as

$$\hat{\beta}_{LR-FIbw} = \sum_{k=1}^{K} \frac{1}{K} \hat{\beta}_k. \tag{6}$$

This would give us the block-average of the person-weighted impacts.

We cannot directly weight the $\hat{\beta}_k$ to target a cluster-average effect because the $\hat{\beta}_k$ themselves do not target clusters. If we are wanting to weight each cluster equally across the entire experiment, we would first need to ensure the $\hat{\beta}_k$ are estimating the cluster average impacts. See next section for how.

2.2.2 Weighted LR with block by treatment interactions (LRcw-FI)

To target any sort of cluster average effect with our fully interacted approach we first need the individual $\hat{\beta}_k$ to target cluster average effects within block. We can do this by adding weights to the initial regression, following the discussion in Section 2.1.1, or via using the aggregated linear regressions discussed below. In particular, we can target the cluster weighted ATEs by fitting the fully interacted regression with weights of $w_{ijk} = \frac{N}{J}/N_{jk}$.

Once we have cluster-average estimates, we would then weight each block estimate by the number of clusters, rather than the number of people:

$$\hat{\beta}_{LR-FIcw} = \sum_{k=1}^{K} \frac{J_k}{J} \hat{\beta}_k, \tag{7}$$

where J_k is the number of clusters in block k and J is the total number of clusters in the study.

This ATE estimator, notated $\hat{\beta}_{\text{LRcw\_FIcw}}$, logically targets the effect for the average cluster since within blocks it weights each cluster equally and in aggregating across blocks it weights blocks by their number of clusters. This estimator is unbiased with respect to β_{cluster}. By ignoring the proportion assigned to treatment when averaging $\hat{\beta}_k$s, one might expect it to be less precise than the corresponding fixed effect estimator (but it does not have to be).

3 Linear Regression on Cluster Level Aggregate Data (AR)

Another regression based approach is to first aggregate the data within each cluster and then fit a linear model to the aggregated data. Begin with the average cluster outcome ($\bar{Y}_{jk}$), defined in terms of the person-level potential outcomes as

$$\bar{Y}_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} Y_{ij} = Z_{jk}\bar{Y}_{jk}(1) + (1 - Z_{jk})\bar{Y}_{jk}(0).$$

Here Z_{jk} is an 0/1 treatment indicator for cluster j, and we observe either the average of the cluster's treatment or control potential outcomes.

We then regress the aggregated outcomes onto their treatment indicators. At this point, we have reduced our blocked cluster-randomized trial to a multisite trial where the cluster average outcomes are our new outcome and the blocks are the sites. Everything from Miratrix et al. (2021) then applies in terms of different estimators we could in principle use, including design-based estimators and multilevel modeling. We have collapsed our three level (in the

case of blocks) regression to a two-level regression. If we did not have blocks, we are simply running an ordinary linear regression after aggregation.

When including covariates, we simply average any individual level covariates within cluster and include them as a cluster-level covariate. E.g., we would include

$$\bar{X}_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} X_{ijk}$$

as a covariate in our linear regression.

Interestingly, the point estimates of the aggregated regressions will be *identical* to the corresponding individual linear regression approach, unless we have aggregated individual-level covariates. In this case, due to aggregation there is no "within cluster" adjustment, and our point estimates can differ. Adjusting for level 2 covariates will not cause these estimators to diverge, however. It is not necessarily the case that one is more precise than the other, however; see Schochet et al. (2021) for further discussion. (In general differences are inconsequential.)

An advantage of aggregation is the data can be used and shared as aggregated data, which can provide privacy at the individual-level.

3.1 AR with block fixed effects (AR-FE)

For example, we could fit a regression with block-level fixed effects, analogous to the person-level regression case, to the aggregates such as

$$\bar{Y}_{jk} = \alpha_k + \beta Z_{jk} + \lambda' W_{jk} + \epsilon_{jk}. \tag{8}$$

This estimator is notated $\hat{\beta}_{\text{AR\_FE}}$. Precision weighting is used to average across blocks, where each block is weighted in proportion to $J_k p_k^C (1 - p_k^C)$, where p_k^C is the proportion of clusters treated in block k. This estimator targets the cluster-weighted estimand, β_{cluster}. Within blocks, the clusters are equally weighted.

3.1.1 Weighted AR with block fixed effects (ARpw-FE)

The fixed effect estimator above ($\hat{\beta}_{\text{AR\_FE}}$) can be adapted for targeting β_{person}, the person weighted ATE. To do so, run a weighted regression where each observation (i.e., cluster) is weighted proportional to its size:

$$w_{ijk} = J \times \frac{N_{jk}}{N}$$

The weights sum to J to represent the J cluster-level observations. This ATE estimator is notated as $\hat{\beta}_{\text{ARpw\_FE}}$.

We could further weight within block so we could, for example, target a simple average of the block average effects. In this case the block weights would be multiplied by the cluster weights to get the overall weights for the regression. See the discussion of weights in Miratrix et al. (2021) for further discussion of customizing weights like this. That said, such complex weighting schemes seem to undermine the transparency of this sort of aggregation approach.

3.2 AR with block by treatment interactions (AR-FI)

As with individual data, one could include block by treatment interactions,

$$\bar{Y}_{jk} = \alpha_k + \beta_k Z_{jk} + \lambda' W_{jk} + \epsilon_{jk}, \tag{9}$$

allowing different blocks to have different average ATEs, and then these block ATEs could be aggregated to estimate the overall ATE, following the person-level regression with treatment by block interaction case (Section 2.2). In particular, we could weight the block-specific impact estimates equally, by number of clusters in the block, or by total people in the block; see discussion following Equation 2.2.1. We would want this weighting to correspond to the weights placed on the individual clusters in the regression itself.

3.2.1 Weighted AR with block by treatment interactions (ARpw-FI)

The aggregated data fully-interacted estimator described in the previous section can be adapted to target the person weighted ATE by using weighted linear regression, with weights $w_{ijk} = \frac{N_{jk}}{N/J}$. We would then need to average the $\hat{\beta}_k$, to target the fully person weighted estimand ($\hat{\beta} = \sum_{k=1}^{K} \frac{N_k}{N} \hat{\beta}_k$) . This estimator is notated $\hat{\beta}_{\text{ARpw\_FIpw}}$. This impact estimator targets the effect for the average person (β_{person}) since within blocks it weights each cluster by its sample size and in aggregating across blocks it weights blocks by their number of persons. Other options for weighing the blocks follow the same pattern as elsewhere.

4 Multilevel Models (MLMs), aka Hierarchical Linear Models (HLMs), random effects models, or ANCOVA

Multilevel modeling explicitly models the hierarchical nature of the data and allows for estimation of the degree of variation in outcomes at each level (and variation in impacts in the three-level case) as well as the overall ATE. Multilevel modeling is probably the most common tool used to analyze data from cluster RCTs in education contexts. As a lens on typical practice, we use the models identified in the technical documentation for PUMP (Hunter et al., 2021), a power calculator which gives a systematic taxonomy of models and designs. This package is based on the popular PowerUp! power calculator (Dong and Maynard, 2013), which also lists these models. Both tools explicitly note different couplings of design and modeling choices, and provide specific uncertainty formulas for several contexts.

MLMs generally assume a specific model, and one might be concerned that model misspecification could ruin inference. This apparently is not as large a concern as one might expect: Wang et al. (2021) prove that these estimators can be consistent under arbitrary misspecification of the working model, and that if the randomization is balanced ($p = 0.5$) then the variance estimators are also consistent. An important caveat is their proofs rely on an independence assumption between cluster size and average impact; they use this assumption to avoid the estimand concerns discussed throughout this document. That all said, the take-away of this paper is these estimators are surprisingly well behaved, estimating what we might expect even under misspecification such as nonnormality and so forth.

The two-level model for non-blocked experiments has random effects to account for the clustering, and is relatively straightforward. The modeling options multiply once we allow for three levels, as we then can potentially model cross-block means and impact variation using either fixed or random effects at level 3. PowerUp! identifies two estimators for this context, which they call fixed and random effects. The random effect model allows the treatment coefficient to vary across blocks (implying a superpopulation of blocks). We discuss the two-level model and then extend to these two in the following.

4.1 Two-level MLMs

For the non-blocked (two-level) case, we use a random intercept model, with a single coefficient for ATE:

Level-One (people)

$$Y_{ij} = \theta_{0,j} + \gamma' X_{ij} + \epsilon_{ij}$$
$$\epsilon_{ij} \sim N\left(0, \sigma^2\right)$$

Level-Two (clusters)

$$\theta_{0,j} = \alpha + \beta Z_j + \lambda' W_j + u_{0,j}$$
$$u_{0,j} \sim N\left(0, \tau_0^2\right)$$

This model assumes a constant treatment effect for all clusters. The reduced form is

$$Y_{ij} = \alpha + \beta Z_j + \gamma' X_{ij} + \lambda' W_j + u_{0,j} + \epsilon_{ij} \tag{10}$$

In R, fitting this multilevel model would be a `lmer()` formula along the lines of:

```
Yobs ~ 1 + Z + X + W + (1 | S.id)
```

While conceptually the MLM targets cluster-average effects, the reality is not quite so simple. The overall estimand targeted by this estimator (which we label $\beta_{MLM-NoFE}$) will land between a person-average and cluster-average estimate. If the ICC is high, the multilevel models will tend towards a simple average of the level two units when estimating overall means. If the ICC is low, by contrast, multilevel modeling will tend towards something like a person-weighted average. This means that, similar to the adaptive nature of FIRC (Miratrix et al., 2021), the larger clusters drive estimation of the treatment and control group means, leading to more of a person-weighted estimand when ICC is low. When ICC is high, the clusters are weighted more evenly, and thus the estimand is more of a cluster average effect. Overall, the implied estimand $\beta_{MLM-NoFE}$ is neither person- nor cluster-weighted.

What is interesting here is the driver of the adaption is variation of the cluster mean outcome levels, *not* variation in the cluster average impacts. Overall, while MLM conceptually targets the cluster average effects, some conditions that include having a strong relationship between cluster effect and cluster size can bias this estimator. While a general workhorse in education in particular, this estimator thus has some interpretive concerns.

Impact heterogeneity technically would be model misspecification for the above model. Impact heterogeneity (i.e., if the variance of β_j, the cluster-level ATEs, were positive) could show up as different variances for the u_{0j} for the treated and untreated groups. High levels of heterogeneity would inflate the overall estimated variance τ_0^2, meaning we would have more of a cluster-weighted estimand again. But also see Section 1.4 for further discussion regarding why we can likely ignore this concern in most contexts.

To handle such heterogeneity explicitly, one could extend the above model for heteroskedasticity by, e.g., allowing for different variances for treated and control cluster random effects, or by using robust standard errors. To implement the modified error structure in R, one could use the older `nlme` package, or use a Bayesian version of multilevel modeling and specify the model via the `brm` package. For robust standard errors, one could use the `clubSandwich` package. Another option would be the "robust" option in STATA.

4.2 MLM with block fixed effects and no treatment effect variation (MLM-FE)

The simplest extension to the two-level MLM from above is to add block fixed effects. This assumes a constant treatment effect across all blocks and all clusters within a block. We extend the above to

Level-One (people)

$$\begin{aligned} Y_{ijk} &= \theta_{0,jk} + \gamma' X_{ijk} + \epsilon_{ijk} \\ \epsilon_{ijk} &\sim N\left(0, \sigma^2\right). \end{aligned}$$

Level-Two (clusters)

$$\begin{aligned} \theta_{0,jk} &= \alpha_k + \beta Z_{jk} + \lambda' W_{jk} + u_{0,jk} \\ u_{0,jk} &\sim N\left(0, \tau_0^2\right) \end{aligned}$$

where the α_k are the third level intercepts modeled as fixed effects.

The reduced form is

$$Y_{ijk} = \alpha_k + \beta Z_{jk} + \gamma' X_{ijk} + \lambda' W_{jk} + u_{0,jk} + \epsilon_{ijk} \tag{11}$$

This model is analogous to the fixed effect regression (Equation (**??**)) with cluster robust standard errors. The key difference is we are modeling the error structure with a cluster random effect, rather than using robust standard errors. We label this estimator β_{MLM-FE}.

This can result in different ATE estimates and different standard errors. Whereas linear regression models closely target the effect for the average person within block, the multilevel model is adaptive—the weight applied to each cluster mean outcome level depends on the cluster's size (n_{jk}) and the amount of between cluster variation in outcomes (τ_0^2) , which relates to the intraclass correlation.

Specifically, within blocks, MLMs weight the cluster mean outcomes proportional to $1/\left(\tau_0^2 + \frac{\sigma^2}{n_{jk}}\right)$. As τ_0^2 approaches zero, MLM weights clusters by their sample size. As τ_0^2

increases, MLM weights clusters equally. Formally, for a single block study our estimator would approximately be

$$\hat{\beta}_{MLM} = \sum_{j=1}^{J} \frac{\left(\tau_0^2 + \frac{\sigma^2}{n_j(1)}\right)^{-1}}{C(1)} Z_{jk}\bar{Y}_j - \sum_{j=1}^{J} \frac{\left(\tau_0^2 + \frac{\sigma^2}{n_j(0)}\right)^{-1}}{C(0)} Z_{jk}\bar{Y}_j \tag{12}$$

with normalizing constant $C(Z) = \sum_{j=1}^{J} \left(\tau_0^2 + \frac{\sigma^2}{n_j(z)}\right)^{-1}$.

If there is no correlation between cluster size (n_{jk}) and cluster-average effects (β_{jk}), then this implied weighting has no bias implications. However, if there is a correlation between n_{jk} and β_{jk}, then the MLM can be biased with respect to the person-average estimands, the cluster-average estimands, or both, depending on τ_0^2.

Overall, multilevel modeling is traditionally considered to be targeting the effect for the average cluster within block.

If there is treatment effect variation across clusters within block, or across blocks, then the model is misspecified as discussed for the prior MLM. This could cause concern if the heterogeneity is very high; see discussion under Section 1.4.

4.3 MLM with block random intercepts with no treatment effect variation (MLM-RE)

One can also use random intercepts at the block-level instead of fixed intercepts by putting a distribution of $\alpha_k \sim N(\alpha, \sigma_\alpha^2)$, yielding estimator β_{MLM-RE}. If there is no treatment by covariate interaction, the random intercept version will be virtually the same as the fixed effect version. For further discussion of this, see the discussion in Miratrix et al. (2021) comparing the Random Intercept, Constant Coefficient (RICC) model to a classic fixed effect model.

4.4 MLM with block fixed effects interacted with treatment (MLM-FI)

One can also fit a three-level model by interacting the block fixed effects with treatment, allowing for cross-block impact heterogeneity by explicitly estimating each block's ATE. This model is analogous to the interacted fixed effect linear model in Section 2.2. The individual block impact estimates would then be averaged together as a second step, with weights selected by the user, just as with the interacted fixed effect model.

The interacted fixed block effect multilevel model, with random effects for cluster, is given by:

Level-One (people)

$$\begin{aligned} Y_{ijk} &= \theta_{0,jk} + \gamma' X_{ijk} + \epsilon_{ijk} \\ \epsilon_{ijk} &\sim N\left(0, \sigma^2\right) \end{aligned}$$

Level-Two (clusters)

$$\theta_{0,jk} = \alpha_k + \beta_k Z_{jk} + \lambda' W_{jk} + u_{0,jk}$$
$$u_{0,jk} \sim N\left(0, \tau_0^2\right)$$

Level-Three (blocks)

$$\alpha_k = \alpha_k$$
$$\beta_k = \beta_k$$

The α_k are estimated as coefficients to K indicator variables for block membership (i.e., block fixed effects). The β_k are the fixed treatment effects in each of the K blocks.

The reduced form is

$$Y_{ijk} = \alpha_k + \beta_k Z_{jk} + \gamma' X_{ijk} + \lambda' W_{jk} + u_{0,jk} + \epsilon_{ijk}. \tag{13}$$

If there is impact variation across clusters within block, this model would again technically be misspecified, as it assumes a common impact β_k for all clusters in block k. See Section 1.4 for further discussion.

The R model is a `lmer` call with formula

```
Yobs ~ 0 + Z * D.id - Z + X + (1 | S.id)
```

The overall ATE of $\hat{\beta}_{MLM\_FI}$ is then the average of the treatment by block interaction terms. All the questions about weighting and averaging from Section 2.2 apply, both for the overall ATE and also for the overall estimated standard error. That said, what exactly the estimates $\widehat{\beta}_k$ themselves are estimating is less clear due to the random cluster (i.e., cluster) intercepts (see the two-level discussion above), although generally we would take them as estimates of the cluster-average effect.

4.5 Full random effects model (MLM-RIRC and MLM-FIRC)

In the fully random effects model we allow the intercept to vary randomly across clusters and blocks, and the average treatment effect is also allowed to vary randomly across blocks. It is still assumed that there is a constant treatment effect across clusters within a block.

These models are distinct in that they are assuming a *superpopulation* of blocks, meaning we are estimating the average impact across blocks in that superpopulation rather than for the blocks in our sample. This means we would expect greater levels of uncertainty; these are analogous to the superpopulation targeting estimators of Miratrix et al. (2021).

The model for estimating impacts is given by:

Level-One (people)

$$Y_{ijk} = \theta_{0,jk} + \gamma' X_{ijk} + r_{ijk}$$
$$r_{ijk} \sim N\left(0, \sigma^2\right).$$

Level-Two (clusters)

$$\theta_{0,jk} = \alpha_k + \beta_k Z_{jk} + \lambda' W_{jk} + u_{0,jk}$$
$$u_{0,jk} \sim N\left(0, \tau_0^2\right)$$

Level-Three (blocks)

$$\alpha_k = \alpha + w_{0,k}$$
$$\beta_k = \beta + w_{1,k}$$

$$\begin{pmatrix} w_{0,k} \\ w_{1,k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \eta_{00}^2 & \eta_{01} \\ & \eta_{11}^2 \end{pmatrix}\right)$$

with reduced form:

$$Y_{ijk} = \alpha + (\beta + w_{1,k}) Z_{jk} + \gamma' X_{ijk} + \lambda' W_{jk} + w_{0,k} + u_{0,jk} + r_{ijk} \tag{14}$$

In Section 1.6, we noted the superpopulation of blocks means the standard error formula has an additional η_{11}/K term, where η_{11} describes how much the intervention effects vary across blocks. If η_{11} is substantial, then even if there are many clusters within block, and even if the clusters are large, there will be substantial uncertainty in the ATE estimate given the η_{11}/K term—this is the consequence of trying to generalize to a (typically hypothetical) superpopulation of blocks. This concern is not unique to this model: any model that allows treatment effects to vary randomly across blocks will pay a cost in precision.

Furthermore, we expect that in nearly all cluster randomized trials η_{11} will be estimated with substantial uncertainty because even if we had precise estimates of the ATE in each block, we would need a reasonable number of blocks to estimate their standard deviation. This means the estimates of the standard errors themselves will likely be more unstable than counterparts focusing on the finite-sample ATE. This is a much more extreme version of a similar challenge in multi-site individually randomized trials that target the superpopulation of sites, as described in Bloom et al. (2017) and Miratrix et al. (2021).

The R model formula for the above is

```
Yobs ~ 1 + Z + X + W + (1 | S.id) + (1 + Z | D.id)
```

and we label this estimator $\beta_{MLM-RIRC}$.

There are other modeling options within this framing, such as an extension of the Fixed Intercept Random treatment Coefficient (FIRC) model (Bloom et al., 2017) with only a distribution on $w_{1,k}$, but leaving the $w_{0,k}$ as fixed effects. For the multisite comparison to this choice, see Miratrix et al. (2021). We label this estimator $\beta_{MLM-FIRC}$.

This model views the experiment as a sample of blocks. By contrast, the interacted fixed effect models (e.g., Section 4.4) put no distribution on the $w_{0,k}$ and $w_{1,k}$, which implies a finite sample of blocks.

One could use the above model to shrink and regularize the individual block impact estimates, akin to a small area estimation problem. One would then average the Empirical Bayes block

estimates proportional to their size (either person or cluster weighted) to get an overall estimate.

5 Uncertainty Estimation

In the prior sections we walked through a series of point estimators for the ATE in cluster RCTs. We next discuss how to get estimated standard errors for these point estimators. Many of the methods for obtaining standard errors apply to many of the different point estimators. We therefore discuss different classes of uncertainty estimation, and then note for which point estimators they apply.

Before we dive in, we discuss two concerns with uncertainty estimation. First, in Subsection 5.1, we talk about what model (finite or superpopulation) we are operating in. Second, many of the uncertainty estimators can struggle if there are few clusters per treatment arm. In particular, some estimators (especially the fully interacted estimators) are not defined if there are any singleton clusters in a treatment or control arm within a block. Furthermore, if adjusting for covariates it is easy to burn enough degrees of freedom that even if this is not the case, the standard errors are no longer defined. We discuss such concerns in Subsection 5.2.

5.1 The sampling model

Uncertainty estimation, implicitly or explicitly, will depend on the population model used. For example, we could focus on estimating the average effect of the intervention *for the sample of people in the sample of clusters in the sample of blocks in the study only* (i.e., finite or "fixed"), with no attempt to make a statistically-based generalization to any broader population. In this situation, uncertainty comes solely from not knowing each person's unobserved counterfactual outcome (that is, for treatment group members, we do not know what their outcome would have been had they been in the control group, and vice versa). This fully finite-sample view is the most well-known for design-based inference; for this general context, see Middleton and Aronow (2015) or Schochet (2015).

Alternatively, we could model the blocks in the study as sampled from some large or infinite population of blocks. Within the blocks, we could again model the clusters as sampled from some block-specific large or infinite populations of clusters. Or we could take the blocks as fixed, but imagine sampling clusters within each block. Or, conversely, we could imagine sampling blocks, with the entire collection of clusters within a block being fixed, once we sample the block. Finally, for each given cluster, the people inside could be viewed as fixed or sampled. Any shift from a fully finite population to a broader population has the potential to increase uncertainty due to the desire to generalize findings beyond the sample people, clusters, and/or blocks in the study.

Instead of infinite superpopulations, finite superpopulations can also be considered; this is the framing taken up by Abadie et al. (2020). Under this framing, we may need to employ a finite population correction: if most of the blocks in the superpopulation are in the experimental sample, then the uncertainty based on the representativeness of our blocks of the larger population must be low.

Each framework points towards different kinds of uncertainty we need to account for. The largest factor is whether blocks are considered a random sample from a larger population, or not. If they are, then we need to account for the uncertainty of whether the blocks we have in our sample are representative, in terms of their true ATEs, of their parent population. If the blocks are not considered a sample from a population, but are instead the population themselves, then we do not need to account for this uncertainty. Alternatively, if we are willing to assume there is no treatment effect variation across blocks, then there is no generalization uncertainty to account for.

Generally, modeling the clusters as randomly sampled, or the people as randomly sampled, will have less impact on uncertainty estimation compared with modeling blocks as randomly sampled. This is partly due to the so-called "correlation of potential outcomes problem," which is that we cannot directly identify treatment heterogeneity at the person-level or cluster-level, given cluster-level treatment assignment. Consequently, we cannot fully take advantage of the clusters or people being considered fixed instead of sampled, as the usual standard error estimators are forced to implicitly assume, in effect, a "worst case" uncertainty that corresponds to a constant treatment effect. This worst case uncertainty turns out to correspond to assuming the clusters and individuals are sampled from infinite superpopulations (e.g., see Ding et al. (2017) for the simple random assignment case). This means that all of the primary methods for estimating standard errors that we considered all targeted the same values; this was further confirmed by our simulations.

Put differently, finite population standard errors "are likely to be conservative, however, because they ignore precision gains from difficult-to-estimate variance terms that represent the extent to which treatment effects vary and co-vary across students in the same cluster. Thus, in theory, the FP estimators could yield more precise ATE estimates than the SP estimators, but it is difficult to realize these precision gains in practice." (Schochet, 2009, p.g. 29)

So what are our estimators we use actually targeting? As demonstrated by our simulation explorations, we find that the fully interacted model standard errors target a context of a set of fixed blocks, with the clusters and individuals within cluster being sampled from infinite populations per block. This is also equivalent to the model where the blocks are still fixed, the clusters are still sampled from infinite populations of clusters, but the individuals in the cluster are fixed (i.e., we are sampling from an infinite population of finite clusters).

The cluster-robust standard errors for the "fixed effect" estimators are also supposing a superpopulation of clusters. In this case, the clusters are considered fixed, and they come with a covariate of what block they are from. This means these models assume the number of clusters within each block is not fixed (as we are sampling from a single superpopulation). The standard errors thus can be inflated when there is cross-block treatment variation, as compared to the true uncertainty we would get when fixing the blocks and sampling clusters within blocks that, for example, we follow in our simulations. The design-based standard errors for the fixed effect models appear to share this feature given the simulation results.

5.2 Singleton clusters, degrees of freedom, and other woes

Generally, for blocked experiments we will see two classes of estimator: those that estimate a single overall average treatment effect, and those that estimate treatment effects within each block, and then average across blocks (these are "fully interacted" estimators). Some of these latter estimators require specific structure on the data in order to function properly. First, the fully interacted estimators need at least one treated and one control cluster in each block so they can estimate a treatment effect block by block.

Furthermore, most of the standard error estimators for the fully interacted estimators require at least two clusters in each treatment arm within each block to estimate a standard error, since the overall uncertainty is an average that depends on estimating variation within each of these subgroups.

A specific and common design that can cause concerns is the matched-pairs design. In this design, each block has two clusters, one of which is treated and the other not. These designs can be quite powerful if the pairs are fairly homogeneous and cluster size varies a lot across the pairs but little within them (Imai et al., 2009). Regardless, in this case, several of the estimators simply cannot be applied, as there are not enough units within each block to estimate uncertainty.

Annoyingly, this concern extends to designs with even a single block that has a single treated or single control unit. What happens is if we are not assuming some kind of homoskedasticity across blocks and treatment arms, we technically have no idea of what kinds of residual uncertainty is in that singleton group. Because overall uncertainty is an average of the individual within block and treatment arm uncertainty, the entire enterprise is ruined. This concern directly appears with the design based estimators which explicitly are built on averages of within-group estimated variances, but it also appears with the robust variance estimators for the same reason. In short, without putting constraints on the residual variation (like the multilevel model does), all the interacted estimators do not give valid uncertainty estimates if there is a singleton cluster in either treatment or control in at least one block.

Additionally, with cluster RCTs it is very easy to run into degrees of freedom issues, especially when using covariate adjustment. The fully interacted estimators, in particular, can run into degrees of freedom issues when adjusting for additional covariates, especially for the case of very few clusters (e.g., two in the matched pairs case) per block.

Our estimators divide roughly into two classes: those that estimate a single impact parameter, and those that estimate treatment within each block. For the former, we can get a rough benchmark degrees of freedom as

$$\text{df} = J - K - g - 1,$$

where g is the number of level 2 covariates beyond the treatment indicator. This is often a conservative estimate, especially if the ICC values tend to be small and we are using multilevel modeling.

For the fully interacted approach, we have a benchmark conservative value of

$$\text{df} = J - 2K - g.$$

To illustrate the degrees of freedom issue, consider a matched pairs design with no covariates and $K = 9$ blocks (and thus $J = 18$ clusters). Under the single overall average approach, we would have 8 degrees of freedom without counting covariates; even adjusting for a few covariates can induce a great deal of instability in uncertainty estimation. For the fully interacted approach, we have a conservative degrees of freedom of 0. This means that without further structure, we will not get standard errors for many of our estimators. If we adjust even with a single covariate, we will have colinearity and not get a point estimate either.

Our accompanying R package returns NAs for those estimators when they cannot be fit.

5.3 Cluster Robust Variance Estimation

The first general tool is cluster robust variance estimation (CRVE). From a linear regression standpoint, handling the clustering could be considered tricky. First, we cannot include fixed effects, often thought useful for handling within-group dependence, at the cluster level, as these would be completely confounded with the treatment assignment variable. Under a regression framework, we thus rely on cluster-robust standard errors, which view the clusters as being sampled from a larger super-population: we are thus doing super-population inference at the cluster level. Due to cluster-robust standard errors, there is no concern of treatment effect variation inducing heteroskedasticity in the residuals; inference is protected.

For CRVE, we do not, in general, cluster at the block level, even when we have multiple blocks. If we did, we would be treating the entire blocks as sampled from a superpopulation of blocks. Especially if we did not have many blocks, our variance estimation would be quite unstable, for the same reasons discussed for the full random effects model discussed in Section 4.5. As blocks are usually artifacts of the researcher's design, and not of the population being studied, targeting a superpopulation of blocks is generally less preferred. Furthermore, under this clustering, if the treatment effects did vary across blocks, the standard errors could be substantially larger.

In a blocked randomization case, we typically account for block differences with fixed effects at the block level. This is not what protects us from concerns about accounting for lack of independence of units within block, however. It is the randomization of clusters within blocks that protects our inference; see, for example, Schochet (2013).

If we are estimating treatment effects within each block (the fully interacted models), then the sampling model is that the blocks are considered fixed, and we are sampling clusters from within each block from a set of hypothetical infinite superpopulations of clusters, one superpopulation of clusters per block. We then combine our block-specific uncertainties as described below to get an overall uncertainty estimate for the overall average ATE.

We can use CRVE for all our individual level regression approaches. For our aggregated regressions, we no longer have clusters, and thus would use heteroskedastic robust standard errors, discussed below, instead.

There are a variety of flavors of CRVE that differently account for possible overfitting of the model to the data. In particular, with few clusters, the sizes of the residuals will tend to be too small, causing the estimated standard errors to also be too small. Various corrections

exist, that try to account for this. These different variants respond differently to singleton units in the case of the fully interacted estimators.

In particular, one can get cluster robust standard errors even for matched pairs experiments, or experiments with singleton units, assuming there are sufficient degrees of freedom and the ATE estimator is defined. That said, for the interacted models, for some CRVE variants these standard errors can be enormously unstable, and come from a pseudo-inverse of an ill-defined matrix. We therefore do not consider such standard errors to be valid, and state that for the fully interacted models, if there is any singleton treated or control unit in any block, then the CRVEs will not be reasonably estimable. In our R package we force these standard errors to NA values (when we do not, we observe extreme instability in both the empirical and simulation data).

There are various methods for calculating the degrees of freedom. One simple conservative technique is to simply use the number of clusters minus number of blocks and covariates (in the fixed effect case). Alternatively, the `clubSandwich` package provides a Satterwhite approximate degree of freedom calculation as well, which could be important when there are few degrees of freedom.

5.4 Heteroskedastic robust standard errors

If we have aggregated our data so we no longer have clusters, then we cannot use CRVE (unless we were doing it at the block level). That said, differing cluster sizes will likely induce heteroskedasticity in an aggregated linear regression model: the average outcome of the N_{jk} people in cluster jk will generally be more uncertain in smaller clusters than larger clusters due to differing sample sizes. We could therefore use heteroskedastic robust standard errors to account for this along with any other heteroskedasticity from the original data.

For the fully interacted estimators, if there is any singleton treated or control unit in any block, then the robust SEs will not be defined. You need at least two treated and two control clusters in each block.

Degree of freedom concerns exist here also. The `clubSandwich` package can again provide some benefits in the small sample cases.

5.5 Design-based inference

Design-based inference directly focuses on the random assignment and sampling mechanisms as the sources of uncertainty, rather than specifying (usually linear) models with random residual components. They are generally considered to rely on weaker assumptions than other approaches; e.g., there is typically no need for homoskedasticity assumptions, or assumptions of a random normal residual. Some good overview papers on design-based inference for clustered randomized trials are Schochet (2013); Schochet et al. (2021) and Middleton and Aronow (2015).

Traditionally, design-based approaches motivate *building* an estimator with a view to principled inference tied to a specifically articulated estimand (e.g., person or cluster weighted). We

illustrate this in Section 6.2, below. Furthermore, as discussed in the introduction, uncertainty estimation depends on whether the target population is finite or super-population. Designed based inference generally makes which is the target quite clear by the construction of the standard error formula.

In addition to constructing estimators from scratch, the principles from design-based inference can also justify the use of linear regression models. For example, starting with the potential outcomes framework and an assumed treatment assignment mechanism, one can prove that the linear regression estimator reliably estimates the ATE (see, e.g., Lin (2013) and Schochet et al. (2021) again). This theory also provides recipes for standard errors, usually based on a plug-in estimate using a derived variance formula. The analysis of regression also allows for the inclusion of control variables more readily.

The connections between design-based inference and linear regression means the point estimates of a "design based estimator" can be identical to these other estimators such as linear regression; it is typically only the methods for calculating the standard error (and thus the standard error estimates themselves) that will differ.

For uncertainty estimation, the fully finite context is the most well-known for design-based inference; see Middleton and Aronow (2015) or Schochet (2015). By contrast, if we view the clusters as sampled from a fixed set of superpopulations, one for each block, with people within the clusters considered fixed (so we are sampling fixed *sets* of individuals), we can use Schochet et al. (2021) for uncertainty estimation. Other models are possible; see Pashley and Miratrix (2021) and Pashley and Miratrix (2022) for further discussion in the context of individual randomization with blocking.

In general, if we consider the blocks and clusters as fixed, we can use design-based uncertainty estimators within each block (with weights to align with the ATE estimator). Without auxiliary covariate adjustment, Design 3 of Schochet (2015) gives, using weights of $w_{jk} = N_{jk}$ (person weighting) or $w_{jk} = 1$ (equal weighting of clusters), the following finite population standard error estimator:

$$\widehat{SE}(\hat{\beta}_k) = \left[\frac{s_{1k}^2}{\bar{w}_{1k}^2 J_{1k}} + \frac{s_{0k}^2}{\bar{w}_{0k}^2 J_{0k}} - \frac{1}{J_k} \left(\frac{s_{1k}}{\bar{w}_{1k}} - \frac{s_{0k}}{\bar{w}_{0k}} \right)^2 \right]^{1/2},$$

with

$$s_{zk}^2 = \frac{1}{N_{zk} - 1} \sum_{j:Z_{jk}=z} w_{jk}^2 (\bar{Y}_{jk} - \bar{\bar{Y}}_{zk})^2$$

and

$$\bar{\bar{Y}}_{zk} = \frac{1}{W_{zk}} \sum_{j:Z_{jk}=z} w_{jk} \bar{Y}_{jk}$$

where

$$W_{zk} = \sum_{j:Z_{jk}=z} w_{jk}.$$

Note that many people simply drop the third term beginning with the $1/J_k$; this term attempts to account for the correlation of potential outcomes problem. We also drop this

term in the implementation in our R package. To estimate the s^2_{zk} terms, we need at least two treatment and two control clusters in every block. As we saw in the main paper, in many real world trials this can present a problem.

Once we have within-block effect estimates and uncertainty estimates, we can aggregate, using weights depending on our target estimand and on how we calculated our ATE. In particular, the standard errors in this case would simply be the aggregation of the block-level standard errors (similar to the interacted linear model case):

The above formula is for the no-covariate case. If adjusting using auxiliary covariates, Schochet et al. (2021) shows that weighted linear regression can be viewed as a covariate-adjusted design-based approach. We follow that here. In Section 4.2, Schochet et al. (2021) gives a formula (Equation 9) for a standard error that turns out to be very similar to CRVE. The formula works with aggregated residuals, where the residuals of the individuals within each cluster are averaged, and the final standard error is a function of how much these averages vary across clusters within blocks.

In particular, let $\hat{e}_{jk}$ denote the cluster-level regression residual (the residual of the cluster mean outcome $\bar{Y}_{jk}$ from the fitted regression model). Then, for the fully interacted model, the block-specific ATE estimate $\hat{\beta}_k$ has variance estimated by

$$\widehat{\mathrm{Var}}(\hat{\beta}_k) = \frac{s^2_k(1)}{J_{1k}} + \frac{s^2_k(0)}{J_{0k}}, \tag{15}$$

where the arm-specific sample variances are

$$s^2_k(1) = \frac{1}{J_{1k} - v^* p^c_k q_k - 1} \sum_{j:\, Z_{jk}=1} \frac{w^2_{jk}}{(\bar{w}^1_k)^2}\, \hat{e}^2_{jk}, \tag{16}$$

$$s^2_k(0) = \frac{1}{J_{0k} - v^* (1 - p^c_k) q_k - 1} \sum_{j:\, Z_{jk}=0} \frac{w^2_{jk}}{(\bar{w}^0_k)^2}\, \hat{e}^2_{jk}. \tag{17}$$

Here $q_k = W_k/W$ is the weighted share of all clusters belonging to block k, and v^* is a degrees-of-freedom correction for the covariates. For the overall standard error, Schochet et al. shows that we can asymptotically assume the within-block estimates are independent, allowing the block estimates to be averaged as discussed below in Section 5.7.

Unfortunately, if there are any blocks with a singleton cluster in treatment or control, then the standard error for that block is not defined, and therefore the overall standard error is no longer defined as well.

Hypothesis testing then uses a t-distribution with

$$\mathrm{df} = J - 2K - v^* \tag{18}$$

degrees of freedom.

For the fixed effect models, the variance estimator (Schochet et al., 2021, equation 15) is

$$\widehat{\text{Var}}(\hat{\beta}_{FE}) = \frac{J}{J - K - v^* - 1} \cdot \frac{\sum_{k=1}^{K} \sum_{j=1}^{J_k} w_{jk}^2 \left(Z_{jk} - p_k^c\right)^2 \hat{e}_{jk}^2}{\left(\sum_{k=1}^{K} J_k \, p_k^c (1 - p_k^c) \, \bar{w}_k\right)^2}, \tag{19}$$

where $\bar{w}_k = W_k / J_k$ is the mean weight of clusters in block k and $\hat{e}_{jk}$ are residuals from the block-fixed-effects WLS regression. Degrees of freedom are

$$\text{df} = J - K - v^* - 1. \tag{20}$$

For both the above sets of models, the cluster-level weights w_{jk} link the estimand to the variance formulas. For the person-weighted estimand (β_{person}), set $w_{jk} = N_{jk}$ (cluster size). For the cluster-weighted estimand (β_{cluster}), set $w_{jk} = 1$.

The design-based SE estimator splits the data into a treatment and control side; by contrast the CRVE and heteroskedastic robust approaches view them as being sampled together with treatment assignment as an additional covariate. This split calculation leads to mild differences in the final estimated standard errors, but the above are asymptotically equivalent to CRVE. For further details see Schochet et al. (2021). Our R package, `clusterRCT`, implements these formula.

5.6 Multilevel models and inference

Multilevel models obtain standard errors via (possibly restricted) maximum likelihood estimation. The standard errors will be asymptotically valid under the assumptions of the model, which typically include homoskedasticity unless alternate error variances are specified.

We can also apply heteroskedastic robust standard errors to multilevel modeling, as a further protection against heteroskedasticity induced by, for example, differing levels of individual variation within clusters, or additional variation at the cluster level due to treatment variation. This is what STATA's "robust" option does.

5.7 Uncertainty for fully interacted models

For our fully interacted models, we do not have a single estimate of an overall treatment effect. We instead average our block-specific $\hat{\beta}_k$ to get our overall $\hat{\beta}$. The standard error strategies listed above will give, for this vector of K impact estimates, a variance-covariance matrix describing their associated uncertainty estimates. The diagonal of this matrix is the vector of squared estimated standard errors.

For CRVE, for example, the standard errors would be taken from the diagonal of the variance-covariance matrix coming out of the cluster-robust standard error calculation (the full matrix would include other parameters such as the fixed effects or other covariates, and we would

take the submatrix corresponding to the $\hat{\beta}_k$). Packages such as `estimatr` (Blair et al., 2022) can give the matrix that we would then work with. Alternatively, the `avg_comparisons()` method of the `marginaleffects` package allows for collapsing these estimates. Our provided code associated with this project also provides methods to extract and average these matrices.

To get an overall standard error for the overall ATE, we use this variance-covariance matrix to, in effect, average the estimated standard errors corresponding to the $\hat{\beta}_k$. Regardless of weighting, we calculate the standard error for the overall $\hat{\beta}$ using the weights used for the initial averaging. Let $w = (w_1, \ldots, w_K)$ be the weights given to the β_k. Then the standard error for the overall average would be

$$SE(\hat{\beta}) = \left(w' \widehat{\Sigma}_\beta w\right)^{1/2},$$

with $\widehat{\Sigma}_\beta$ the subset the relevant rows and columns of $\widehat{\Sigma}$, the estimated variance-covariance matrix from the original full fit regression, that correspond to the $\hat{\beta}_k$.

Because the randomization happens independently within each block, if we are not using covariate adjustment, the individual block average treatment effect estimates are also independent. We can then simplify the above as

$$\widehat{\Sigma}_\beta = diag(\widehat{SE}_1^2, \ldots, \widehat{SE}_K^2),$$

giving

$$SE(\hat{\beta}) = \sqrt{\sum w_k^2 \widehat{SE}_k^2}.$$

While this expression is simpler and more intuitive, the full block covariance matrix would be preferred if there are common covariate adjustments in the regression, as they can make the point estimates no longer independent. Again, the w_k are the weights used in calculating an ATE of $\beta = \sum_k w_k \beta_k$, with $\sum w_k = 1$. Without covariate adjustment, this formula works for all the variants of the fixed block effects interacted with treatment models discussed under linear, multilevel modeling, and aggregation (see Section 2.2).

6 "Pure" Design-based estimators

In our uncertainty estimation section (Section 5) we discussed design-based standard error estimators in terms of obtaining a standard error for a fit regression that is justified by the randomization itself rather than an overall model. We can also use the principles of design-based inference to *construct* our estimators from scratch. We next discuss how to do this, and then introduce some pure unbiased design-based estimators along with a discussion of how *all* the person-weighted estimators discussed so far can have bias when cluster sizes vary. The advantage of building a design-based estimator by hand is that it makes it easier to directly and explicitly tie the estimator to an estimand of interest (Schochet, 2015), as discussed in Miratrix et al. (2021).

6.1 Single-block estimators

We first focus on a single block k, and then extend to multiple blocks. For a single block k, the simplest cluster-average estimator for the average impact would be

$$\hat{\beta}_{k,DB\_cluster} = \sum_{j=1}^{J_k} \frac{1}{J_{1k}} Z_{jk} \bar{Y}_{jk} - \sum_{j=1}^{J_k} \frac{1}{J_{0k}} (1 - Z_{jk}) \bar{Y}_{jk}, \tag{21}$$

where J_{1k} and J_{0k} are the number of clusters assigned to treatment or control status, respectively. The $\bar{Y}_{jk}$ are the observed mean outcomes of the clusters. Our (unadjusted) design based estimate is the aggregation approach (Section 3) without covariates. This estimator targets the average of the cluster average effects, in block k.

The person-average estimator would weight the clusters by size, giving

$$\hat{\beta}_{k,DB\_person} = \sum_{j=1}^{J_k} \frac{N_{jk}}{N_{1k}} Z_{jk} \bar{Y}_{jk} - \sum_{j=1}^{J_k} \frac{N_{jk}}{N_{0k}} (1 - Z_{jk}) \bar{Y}_{jk}, \tag{22}$$

where $N_{1k} = \sum Z_{jk} N_{jk}$ are the number of people assigned to treatment, and N_{0k} is defined similarly. Surprisingly, as discussed below in Section 6.3, this estimator is not actually unbiased when cluster sizes vary (although the bias term will generally be small).

6.2 Design based estimators for three-level data

The classic design-based estimator for a blocked experiment estimates impacts within each block k, and then averages those estimates across blocks, just as the interacted linear model estimators do. For example, a block-size-weighted average of the person-average $\hat{\beta}_{k,DB\_person}$ would give the estimated effect for the average person across the sample. This estimator would be

$$\hat{\beta}_{DB-FI-person} = \sum_{k=1}^{K} \frac{N_k}{N} \hat{\beta}_{k,DB\_person}, \tag{23}$$

where N_k is the total size (in people) of block k and N is the total number of people.

For the cluster-average of cluster averages, we weight as

$$\hat{\beta}_{DB-FI-cluster} = \sum_{k=1}^{K} \frac{J_k}{J} \hat{\beta}_{k,DB\_cluster}, \tag{24}$$

where J_k is the total size (in clusters) of block k and J is the total number of clusters.

For the block-average of person averages, we would weight as

$$\hat{\beta}_{DB-FI-person-block} = \sum_{k=1}^{K} \frac{1}{K} \hat{\beta}_{k,DB\_person}. \tag{25}$$

This would be the simple average of the person-weighted block average impacts.

Finally, for the block-average of the cluster averages, we would weight as

$$\hat{\beta}_{DB-FI-cluster-block} = \sum_{k=1}^{K} \frac{1}{K} \hat{\beta}_{k,DB\_cluster}. \tag{26}$$

We might use block-weighting if we were interested in how much a random block would shift its average cluster, on average, if given treatment. Other weightings options are also possible; the above four align with the estimands introduced in Section 1.2. The overall point is the design approach begins with what one wants to estimate, and then directly constructs an estimator to achieve that end.

The above estimators are equivalent to the interacted linear regressions of Section 2.2 and Section 3.2, which estimate an ATE for each block. Schochet et al. (2021) shows this connection, which allows for covariate adjustment, but also shows how we can also motivate the fixed effect regression approach (see Section 2.1) as also being solely justified by the randomization of the clusters.

6.3 Truly unbiased design based estimators

Middleton and Aronow note that the design-based estimator given above, Equation 22, for the person-average effect is actually biased. (This means, by extension, all person-average targeting estimators discussed up to now are biased.) They then propose a class of estimators that are unbiased for targeting the person-average treatment effect for cluster randomized trials. With multiple blocks, we can apply their approach within each block and then average, pooling the standard error as discussed above (assuming we are holding the blocks as fixed).

To see the bias concern, consider the expected value of the simple difference estimator of Equation 22, where we are weighting each cluster by its size. For this estimator, we divide the weighted sum of the treated clusters by the total number of treated people N_{1k}, and N_{1k} is random if cluster sizes vary. The same holds for control as well. The expected value of the estimator is then the following:

$$\begin{aligned}
E[\hat{\beta}_{k,DB_person}] &= E\left[\bar{Y}_k(1) - \bar{Y}_k(0)\right] \\
&= E\left[\frac{1}{N_{1k}} \sum_{j=1}^{J_k} Z_{jk} N_{jk} \bar{Y}_{jk} - \frac{1}{N_{0k}} \sum_{j=1}^{J_k} (1 - Z_{jk}) N_{jk} \bar{Y}_{jk}\right] \\
&= \sum_{j=1}^{J_k} E\left[\frac{1}{N_{1k}} Z_{jk}\right] N_{jk} \bar{Y}_{jk}(1) - \sum_{j=1}^{J_k} E\left[\frac{1}{N_{0k}} (1 - Z_{jk})\right] N_{jk} \bar{Y}_{jk}(0).
\end{aligned}$$

We can compare the above to the person-weighted estimand of

$$\beta_{k,person} = \frac{1}{N_k} \sum_{j=1}^{J_k} N_{jk} \bar{Y}_{jk}(1) - \frac{1}{N_k} \sum_{j=1}^{J_k} N_{jk} \bar{Y}_{jk}(0).$$

To be unbiased, the expected value of $E\left[Z_{jk}/N_{1k}\right]$ would need to equal $1/N_k$ for each j and k. It does not, if cluster sizes vary, as in this case

$$E[\frac{1}{N_{1k}}Z_{jk}] \neq \frac{1}{N_k}$$

because the N_{1k} term is random: if we end up with a large cluster in treatment, the total is larger, and if we end up with a small one, it is smaller.

For an alternate explanation of where the bias comes from, consider that the mean of the treatment-side in the above is the sum of all outcomes in the treatment group divided by total number of people in the treatment group. Then, the expected value of this across randomizations is

$$E\left[\frac{1}{N_{1k}}\sum_{j=1}^{J_k} Z_{jk}N_{jk}\bar{Y}_{jk}\right] = E\left[\frac{\sum_j N_{jk}Z_{jk}\bar{Y}_{jk}(1)}{\sum_j N_{jk}Z_{jk}}\right] \neq \frac{E\left[\sum_j N_{jk}Z_{jk}\bar{Y}_{jk}(1)\right]}{E\left[\sum_j N_{jk}Z_{jk}\right]} = \bar{Y}_k(1).$$

In other words, because the number of people in the treatment group is random, the expectation does not go through, and we do not end up getting an unbiased estimate of the average of all the people' treatment-side potential outcomes as would be needed for an overall unbiased estimate. The more the cluster sizes vary, and the more lopsided the treatment proportion, the larger the bias can be. That being said, the bias of this estimator is often not large, and the estimator is clearly targeting the average treatment impact of the people in the sample, not the average of the cluster average impacts. Also see Schochet (2015), page 93-94 for another example and discussion of this bias.

One possible fix is to make an estimator that has the *expected* number of treated or control people as the denominators, i.e., replace N_{1k} with $E[N_{1k}] = p_k N_k$, with p_k the fixed proportion of clusters in block k assigned to treatment:

$$\hat{\beta}_{k,DB\_HT} = \frac{1}{p_k N_k}\sum_{j=1}^{J_k} Z_{jk}N_{jk}\bar{Y}_{jk} - \frac{1}{(1-p_k)N_k}\sum_{j=1}^{J_k}(1-Z_{jk})N_{jk}\bar{Y}_{jk}, \tag{27}$$

Now the expectation goes through, and we have

$$E\left[\hat{\beta}_{k,HT}\right] = \beta_{k,person}.$$

$\hat{\beta}_{k,HT}$ is a type of Horvitz-Thompson estimator, taking the cluster totals $N_{jk}\bar{Y}_{jk}$ as the potential outcomes. These estimators are often very unstable.

Middleton and Aronow propose an improvement to this estimator by using a Raj difference estimator, where we adjust all the outcomes to try and stabilize the consequence of different sized clusters when using the Horvitz-Thompson trick. First, restate, by rearranging terms, the Horvitz-Thompson estimator above as

$$\hat{\beta}_{k,DB\_HT} = \frac{J_k}{N_k}\left[\frac{1}{J_{1k}}\sum_{j=1}^{J_k} Z_{jk}Y_{jk}^T(1) - \frac{1}{J_{0k}}\sum_{j=1}^{J_k}(1-Z_{jk})Y_{jk}^T(0)\right],$$

where $Y_{jk}^T = N_{jk}\bar{Y}_{jk}$ is the sum of the outcomes for cluster j in block k. The above equation shows that the Horvitz-Thompson estimator is estimating the average impact on the cluster totals, and then rescaling by J_k to translate this impact estimate to the average impact on the grand total sum of the outcomes across all units, and finally dividing by N_k to get the average impact on the mean of the outcomes, which is our usual ATE.

Next define modified potential outcomes as

$$U_{jk} = Y_{jk}^T - \alpha\left(N_{jk} - \frac{N_k}{J}\right),$$

where α is the regression coefficient from regressing Y_{jk}^T on N_{jk}. Technically α needs to be known in advance; more on this regression coefficient later.

The term $N_{jk} - \frac{N_k}{J}$ is the difference between the size of cluster j and the average cluster size in block k. The U_{jk} are adjusted totals that take into account different sizes of cluster–if the relationship between size and total is linear, these will be more homogeneous than the initial Y_{jk}^T. In other words, the U_{jk} is essentially an adjusted outcome total, adjusted to what it would be if the cluster were average in size.

We then estimate impact as

$$\hat{\beta}_{k,DB\_Raj} = \frac{J_k}{N_k}\left[\frac{1}{J_{1k}}\sum_j Z_{jk}U_{jk} - \frac{1}{J_{0k}}\sum_j (1 - Z_{jk})U_{jk}\right]. \tag{28}$$

In expectation across random assignments, the adjustments average out to 0,

$$\begin{aligned} E\left[\frac{1}{J_{1k}}\sum_j Z_{jk}\alpha\left(N_{jk} - \frac{N_k}{J}\right)\right] &= \frac{1}{J_{1k}}\sum_j E[Z_{jk}]\alpha\left(N_{jk} - \frac{N_k}{J}\right) \\ &= \frac{1}{J_{1k}}p_k\alpha\sum_j\left(N_{jk} - \frac{N_k}{J}\right) = 0, \end{aligned}$$

meaning the overall estimator has the same expectation as $\hat{\beta}_{k,HT}$, i.e., is unbiased. (In the above, note that the only random element is the Z_{jk}, allowing for the simplification of the sum.)

To estimate α we propose to simply regress the cluster totals onto cluster sizes across all the blocks in a single step. For any given block, the value of α will be primarily driven by the other blocks and thus we can mostly avoid the reuse issues discussed in (Middleton and Aronow, 2015). Alternatively, out-of-sample estimation can be used, where α is estimated via the other blocks for each block k; this would be more computationally intensive. See Middleton and Aronow (2015) for further discussion and alternate approaches here.

To get estimated standard errors we have, again taking from Middleton and Aronow (2015),

$$\widehat{SE}(\hat{\beta}_{k,DB\_Raj})^2 = \frac{J_k^2}{N_k^2}\left[\frac{S_{1k}^2}{J_{1k}} + \frac{S_{0k}^2}{J_{0k}}\right],$$

with sample variance of the adjusted clusters totals for treatment group z of

$$S_{zk}^2 = \frac{1}{J_{zk} - 1} \sum_{j:Z_j=z} \left(U_{jk} - \overline{U}_{\cdot kz} \right)^2,$$

where $\overline{U}_{\cdot kz}$ is the average of the U_{jk} in treatment group z. This is simply a Neyman standard error scaled by $(J_k/N_k)^2$.

To get the overall estimates and standard errors, simply average the block-level impacts as discussed in Section 2.2 and elsewhere.

7 Technical notes for the R package

The `clusterRCT` R package[1] implements the methods discussed in this document. These can be applied to cluster RCTs to assess how method choice can impact a final estimated ATE. In our package, we also provide a suite of functions to describe the structure of a cluster RCT, and also methods for generating synthetic data calibrated to provided cluster RCTs, or from whole cloth.

7.1 Describing Data

The `describe_clusterRCT()` method takes an empirical dataset and tallies and estimates a variety of parameters for that dataset. It primarily reports number of blocks, clusters, individuals, and also gives average sizes of these quantities. It further describes how much variation there is in the block and cluster sizes. It has a few other statistics of that nature.

It also estimates ICC values and R^2 values. To estimate ICC, the package fits a simple multilevel model with cluster and block random effects. It does not control for any covariates, and thus reports *total* variation at each level.

To calculate the R^2 values, the package fits a sequence of models and estimates R^2 values for proportion of variance explained at each level by looking at how much the estimated variances are reduced when including covariates.

In particular, to calculate R_1^2 we fit multilevel models of the form

$$Y_{ijk} = \alpha_k + \gamma_X' X_{ijk} + \gamma_W' W_{ijk} + u_{jk} + \epsilon_{ijk},$$

with $\sigma^2 = var(\epsilon_{ijk})$ and $\tau^2 = var(u_{jk})$.

We fit three such models, with varying covariates at level one and two:

$$\begin{aligned} M_{no1} &: Y \sim W + Z + ID_{block} + (1|cluster) \\ M_{no2} &: Y \sim X + Z + ID_{block} + (1|cluster) \\ M_{full} &: Y \sim X + W + Z + ID_{block} + (1|cluster) \end{aligned}$$

[1] On GitHub, link to be added

where X are level 1 covariates, W are level 2 covariates, and Z is the treatment indicator. We have a random effect for cluster and fixed effects for block. We group-mean centered all level 1 covariates at the cluster level (the X), and add group means as part of the level 2 covariates (the W).

We then report

$$R_1^2 = \left(\sigma_{no1}^2 - \sigma_{full}^2\right)/\sigma_{full}^2$$
$$R_1^2 = \left(\tau_{no2}^2 - \tau_{full}^2\right)/\tau_{full}^2$$

The final R^2 values target variation at the given level, so they do not represent total variance explained, but rather variance explained *at the targeted level*. Level 1 covariates can explain variation at both level 1 and level 2; this is why we group mean center and add in the group means as a second level 2 covariate. This allows isolation of contribution of all variables to all relevant levels.

7.2 Comparing Methods

The `compare_methods()` function takes a dataset and returns estimated ATEs and associated standard errors for a variety of methods. It takes a formula of the form $Y \sim Z|C|B$, where Z is the treatment indicator, C are cluster IDs, and B are block IDs. One can pass an additional control formula listing all additional variables to control for.

The main method calls a series of helper functions for the different families of estimator. They can all be called individually if desired.

The package will automatically drop rows of data missing outcome, treatment assignment, or group ID, but impute missing values for all covariates (imputing baseline covariates does not compromise the integrity of the overall ATE estimation, assuming the covariates are not missing as a function of treatment assignment; see, e.g., Zhao et al. (2024)). We refer to this imputation as "patching" the data. This option is customizable.

See the package documentation for further details of use.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296.

Andrabi, T., Das, J., and Khwaja, A. I. (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review*, 107(6).

Banerjee, A., Karlan, D., Trachtman, H., and Udry, C. R. (2020). Does poverty change labor supply? evidence from multiple income effects and 115,579 bags. NBER Working Paper 27314, National Bureau of Economic Research.

Berry, J., Karlan, D., and Pradhan, M. (2018). The impact of financial education for youth in ghana. *World Development*, 102:71–89.

Blair, G., Cooper, J., Coppock, A., Humphreys, M., and Sonnet, L. (2022). estimatr: Fast estimators for design-based inference. R package version 1.0.0.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., and 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.

Ding, P., Li, X., and Miratrix, L. W. (2017). Bridging finite and super population causal inference. *Journal of Causal Inference*, 5(2).

Dong, N. and Maynard, R. (2013). Powerup!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, 6(1):24–67.

Hunter, K., Miratrix, L., and Porter, K. (2021). Power under multiplicity project (pump): Estimating power, minimum detectable effect size, and sample size when adjusting for multiple outcomes. *arXiv preprint arXiv:2112.15273*.

Imai, K., King, G., and Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science*, 24(1):29–53.

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

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., and Rajani, R. (2019). Inputs, incentives, and complementarities in education: Experimental evidence from tanzania. *The Quarterly Journal of Economics*, 134(3):1627–1673.

Middleton, J. A. and Aronow, P. M. (2015). Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments. 6(1-2):312.

Miratrix, L. W., Weiss, M. J., and Henderson, B. (2021). An applied researcher's guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *Journal of Research on Educational Effectiveness*, 14(1):270–308.

Pashley, N. E. and Miratrix, L. W. (2021). Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*, 46(3):271–296.

Pashley, N. E. and Miratrix, L. W. (2022). Block what you can, except when you shouldn't. *Journal of Educational and Behavioral Statistics*, 47(1):69–100.

Rosenbaum, P. R., Rosenbaum, and Briskman (2020). *Design of observational studies*, volume 2. Springer.

Schochet, P. Z. (2009). Technical methods report: The estimation of average treatment effects for clustered rcts of education interventions. Technical report, Mathematica Policy Research.

Schochet, P. Z. (2013). Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference. *Journal of Educational and Behavioral Statistics*, 38(3):219 – 238. Good illustration of using the Neyman model for understanding other procedures.

Schochet, P. Z. (2015). Statistical theory for the rct-yes software: Design-based causal inference for rcts. *National Center for Education Evaluation and Regional Assistance*.

Schochet, P. Z., Pashley, N. E., Miratrix, L. W., and Kautz, T. (2021). Design-based ratio estimators and central limit theorems for clustered, blocked rcts. *Journal of the American Statistical Association*, pages 1–12.

Wang, B., Harhay, M. O., Tong, J., Small, D. S., Morris, T. P., and Li, F. (2021). On the mixed-model analysis of covariance in cluster-randomized trials. *arXiv preprint arXiv:2112.00832*.

Zhao, A., Ding, P., and Li, F. (2024). Covariate adjustment in randomized experiments with missing outcomes and covariates. *Biometrika*, 111(4):1413–1420.