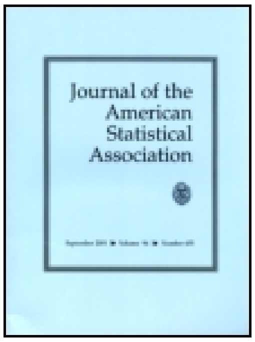

Journal of the
American
Statistical
Association

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Design-Based Ratio Estimators and Central Limit Theorems for Clustered, Blocked RCTs

Peter Z. Schochet, Nicole E. Pashley, Luke W. Miratrix & Tim Kautz

To cite this article: Peter Z. Schochet, Nicole E. Pashley, Luke W. Miratrix & Tim Kautz (2021): Design-Based Ratio Estimators and Central Limit Theorems for Clustered, Blocked RCTs, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1906685

To link to this article: https://doi.org/10.1080/01621459.2021.1906685

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2021.1906685

Taylor & Francis
Taylor & Francis Group

Design-Based Ratio Estimators and Central Limit Theorems for Clustered, Blocked RCTs

Peter Z. Schochet[a], Nicole E. Pashley[b], Luke W. Miratrix[c], and Tim Kautz[a]

[a]Mathematica, Princeton, NJ; [b]Department of Statistics, Rutgers University, Piscataway, NJ; [c]Graduate School of Education, Harvard University, Cambridge, MA

ABSTRACT
This article develops design-based ratio estimators for clustered, blocked randomized controlled trials (RCTs), with an application to a federally funded, school-based RCT testing the effects of behavioral health interventions. We consider finite population weighted least-square estimators for average treatment effects (ATEs), allowing for general weighting schemes and covariates. We consider models with block-by-treatment status interactions as well as restricted models with block indicators only. We prove new finite population central limit theorems for each block specification. We also discuss simple variance estimators that share features with commonly used cluster-robust standard error estimators. Simulations show that the design-based ATE estimator yields nominal rejection rates with standard errors near true ones, even with few clusters.

ARTICLE HISTORY
Received May 2020
Revised February 2021
Accepted March 2021

KEYWORDS
Blocked designs; Clustered designs; Design-based estimators; Finite population central limit theorems; Randomized controlled trials

1. Introduction

There is a growing literature on design-based methods for analyzing randomized controlled trials (RCTs) (e.g., Yang and Tsiatis 2001; Freedman 2008; Schochet 2010, 2016; Lin 2013; Miratrix et al. 2013; Imbens and Rubin 2015; Middleton and Aronow 2015; Li and Ding 2017). These nonparametric methods are built on the potential outcomes framework, introduced by Neyman (1923) and later developed in seminal works by Rubin (1974, 1977) and Holland (1986). They leverage a fundamental component of experimental designs—the known treatment assignment mechanism—to achieve results that rely on minimal assumptions.

The design-based literature has largely focused on non-clustered designs in which individuals are randomly assigned to research conditions. A much smaller literature has considered design-based methods for clustered RCTs where groups (such as schools, hospitals, or communities) rather than individuals are randomized. Clustered designs are common in evaluations that test interventions targeted to a group and are sometimes preferred to non-clustered designs as they can help minimize bias due to the potential spillover of intervention effects from treatment to control subjects. Clustered designs are becoming increasingly prevalent in social policy research (Schochet 2008) and have grown exponentially in medical trials (Bland 2004).

For example, the evaluation of the Social and Character Development (SACD) Research Program was a major federal initiative, co-funded by the Institute of Education Sciences at the U.S. Department of Education and the Centers for Disease Control and Prevention, to test interventions promoting positive social and character development among elementary school children, with the goal of ultimately improving their academic performance (SACD Research Consortium 2010). The study was conducted in seven large school districts (blocks), where half the schools (clusters) within each district were randomly assigned to a treatment group and half to a control group, yielding a final sample of 84 schools (42 treatment and 42 control). Intervention features included materials and lessons on social skills, behavior management, social and emotional learning, self-control, anger management, and violence prevention.

Several key aspects of the SACD study motivate the theory underlying this article. First, neither the sample of seven SACD school districts nor the 10 to 14 study schools per district were randomly sampled from broader populations. Rather, the participating districts and schools were volunteers, yielding a convenience sample, as is often the case in RCTs across disciplines. This suggests a finite population framework for estimating average treatment effects (ATEs), where the sample and their potential outcomes are considered fixed, with treatment assignments being the only source of randomness, and where study results are assumed to pertain to the study sample only (Neyman 1923). This framework differs from typical model-based, super-population approaches where potential outcomes are assumed to be randomly sampled from a broader (often infinite) population—even if vaguely defined—and where study results are assumed to generalize to this population.

A second aspect of the SACD study that motivates our theory is the need for flexible weighting schemes to accommodate decisions on how clusters and blocks are to be weighted to estimate pooled effects and to help adjust for data nonresponse. Third, the theory should address common approaches for including and incorporating block (fixed) effects in the models. Finally, the estimation strategy should allow for the inclusion of model baseline covariates to improve precision;

CONTACT Peter Z. Schochet pschochet@mathematica-mpr.com Mathematica, P.O. Box 2393, Princeton, NJ 08543-2393

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

this is especially important for clustered designs where power is often a concern due to design effects from clustering and the typical high cost of adding clusters to the study.

We achieve these objectives in this work by developing covariate-adjusted design-based methods for obtaining point estimates and associated inference for clustered RCTs. Our results rely on new finite population central limit theorems (CLTs) for design-based ATE ratio estimators that apply to the general case where randomization of clusters is conducted within blocks (strata). We consider ratio estimators for clustered RCTs obtained using weighted least square (WLS) methods, which have intuitive appeal because they parallel differences-in-means and regression-adjusted ATE estimators for non-clustered designs. We allow for general weighting schemes and covariates. We also consider models with block-by-treatment status interactions as well as models with block fixed effects only.

We provide consistent variance estimators and compare them, both analytically and through simulations, to widely used ordinary least-square estimators with cluster-robust standard errors (CRSE) (Liang and Zeger 1986; Cameron and Miller 2015). Our simulations suggest that the design-based ratio estimators yield Type I error rates near nominal levels, even with relatively few clusters. We also conduct an empirical analysis using data from the SACD study to compare different specifications of our estimators to each other and to the standard CRSE estimator.

The rest of this article is structured as follows. Section 2 discusses the literature this work is built on and Section 3 provides our theoretical framework. Sections 4 and 5 present our finite population CLTs and variance estimators. Section 6 presents simulation results and Section 7 presents empirical results using our motivating SACD example. Section 8 concludes.

2. Related Work

Our finite population CLTs build on Li and Ding (2017), who consider CLTs for unbiased estimators for clustered RCTs using the Horvitz-Thompson estimator developed by Middleton and Aronow (2015) with cluster-level covariates, but do not consider ratio estimators or blocked designs with general weighting schemes. Our theory also builds on results in Scott and Wu (1981) who consider CLTs for ratio estimators for finite population totals, but not for clustered designs or RCTs. We extend the design-based results in Imai et al. (2009) who examine clustered RCTs with pairwise matching but not general blocked designs, models with covariates, or CLTs. We also extend the design-based results in Schochet (2013) who examines clustered designs without blocking, and Pashley and Miratrix (2020) and Liu and Yang (2020) who consider blocked designs without clustering. In particular, Liu and Yang (2020) also derive finite population CLTs for blocked designs with covariates for RCTs, but only for designs without clustering.

Other literature in this area has a different focus. Abadie et al. (2017) discuss reasons for adjusting for clustering and investigate differences between the true asymptotic finite population variance and the CRSE variance estimator, but do not consider impact estimation. Hansen and Bowers (2009) propose model-assisted estimators combined with randomization inference for regression models in a specific context without deriving design-based estimators. Samii and Aronow (2012) compare design-based and robust estimators for non-clustered designs, but not for clustered designs or models with covariates. While there is a large statistical literature on related design-based methods for analyzing survey data with complex sample designs (e.g., Fuller 1975, 2009; Cochran 1977; Bickel and Freedman 1984; Rao and Shao 1999; Wolter 2007; Lohr 2009), these works do not focus on RCT settings.

3. Framework and Definitions

We assume that a clustered RCT of m total clusters is conducted across h blocks, with block b having m_b clusters ($b = 1, \ldots, h$). Randomization of clusters is conducted separately by block, with $m_b^1 = m_b p_b$ assigned to the treatment group and $m_b^0 = m_b(1 - p_b)$ assigned to the control group ($0 < p_b < 1$). We assume a sample of n_{jb} individuals in cluster j in block b, with n_b individuals in the block and n individuals in total. For each cluster, either all individuals are treated or not. We index individuals by ijb for individual i in cluster j in block b. Let $Y_{ijb}(1)$ be a person's outcome if assigned to a treated cluster and $Y_{ijb}(0)$ be the outcome in a control cluster. These potential outcomes can be continuous, binary, or discrete. We assume a finite population model, where potential outcomes are assumed to be fixed for the study. Let T_{jb} equal 1 if cluster jb is randomly assigned to the treatment condition and 0 otherwise. Let $S_{ijb,s}$ and $S_{jb,s}$ denote indicator variables of block membership for individuals and clusters (that is, $S_{ijb,s} = 1$ or $S_{jb,s} = 1$ if the specified person or cluster belongs to block s).

We also allow for weights, with individual weights of $w_{ijb} > 0$, cluster weights of $w_{jb} = \sum_{i=1}^{n_{jb}} w_{ijb}$, and block weights of $w_b = \sum_{j=1}^{m_b} w_{jb}$. Depending on the research questions of interest, the weights can be set, for example, so that intervention effects pertain to the average individual in the block ($w_{ijb} = 1$ and $w_{jb} = n_{jb}$) or the average cluster in the block ($w_{ijb} = 1/n_{jb}$ and $w_{jb} = 1$). They can also be further modified to handle various forms of data nonresponse (we do not consider estimation error in the nonresponse weights in our variance formulas).

We assume two conditions that generalize those in Imbens and Rubin (2015) for the nonclustered RCT design to our context. The first is the stable unit treatment value assumption (SUTVA) (Rubin 1986):

(C1): *SUTVA*: Let $Y_{ijb}(\mathbf{T}_{\text{clus}})$ denote the potential outcome for an individual given the random vector of all cluster treatment assignments, $\mathbf{T}_{\text{clus}}$. Then, if $T_{jb} = T'_{jb}$ for cluster j, we have that $Y_{ijb}\left(\mathbf{T}_{\text{clus}}\right) = Y_{ijb}\left(\mathbf{T}'_{\text{clus}}\right)$.

SUTVA allows us to express $Y_{ijb}\left(\mathbf{T}_{\text{clus}}\right)$ as $Y_{ijb}\left(T_{jb}\right)$, so that the vector of individual potential outcomes in cluster jb depends only on the cluster's treatment assignment and not on the treatment assignments of other clusters in the sample. SUTVA could be more plausible for clustered designs than non-clustered designs because there are likely to be fewer meaningful interactions between sample members across clusters than within clusters. SUTVA also assumes a particular treatment unit cannot receive different forms of the treatment.

Under SUTVA, the block b ATE parameter of interest for the finite population model is

$$\beta_{1,b} = \frac{\sum_{j=1}^{m_b} w_{jb}(\bar{Y}_{jb}(1) - \bar{Y}_{jb}(0))}{\sum_{j=1}^{m_b} w_{jb}} = \bar{\bar{Y}}_b(1) - \bar{\bar{Y}}_b(0), \tag{1}$$

where, for $t \in \{1, 0\}$, $\bar{Y}_{jb}(t) = \frac{1}{w_{jb}} \sum_{i=1}^{n_{jb}} w_{ijb} Y_{ijb}(t)$ is the weighted mean potential outcome in the treatment or control condition. The ATE parameter across all blocks is then a weighted average of the block-specific ATE parameters:

$$\beta_1 = \frac{\sum_{b=1}^{h} w_b \beta_{1,b}}{\sum_{b=1}^{h} w_b}. \tag{2}$$

Our second condition is the randomization itself:

(C2): Complete randomization of clusters within blocks: Let $\mathbf{T}_{\text{clus},b}$ be the random vector of cluster treatment assignments in block b. Let $(m_1^1, \ldots, m_h^1)$ be a prespecified vector denoting the number of clusters to assign to treatment within each block. Then, for any vector, $\mathbf{t}_{\text{clus},b} = (t_{1b}, \ldots, t_{m_b b})$ of randomization realizations such that $\sum_{j=1}^{m_b} t_{jb} = m_b^1$, we have that $\text{prob}(\mathbf{T}_{\text{clus},b} = \mathbf{t}_{\text{clus},b}) = \begin{pmatrix} m_b \\ m_b^1 \end{pmatrix}^{-1}$. This holds for all $b \in \{1, \ldots, h\}$, and we further assume that the cluster assignments, $(\mathbf{T}_{\text{clus},1}, \ldots, \mathbf{T}_{\text{clus},h})$, are mutually independent.

4. ATE Estimators for the Finite Population Model

Under the potential outcomes framework and SUTVA, the data-generating process for the observed outcome measure, y_{ijb}, is a consequence of the assignment mechanism:

$$y_{ijb} = T_{jb} Y_{ijb}(1) + (1 - T_{jb}) Y_{ijb}(0). \tag{3}$$

This relation states that we can observe $y_{ijb} = Y_{ijb}(1)$ for those in the treatment group and $y_{ijb} = Y_{ijb}(0)$ for those in the control group, but not both.

Rearranging (3) generates the following nominal regression model for any given block:

$$y_{ijb} = \beta_{0,b} + \beta_{1,b}\left(T_{jb} - p_b^*\right) + u_{ijb}, \tag{4}$$

where $\beta_{1,b} = \bar{\bar{Y}}_b(1) - \bar{\bar{Y}}_b(0)$ is the block-specific ATE parameter, $p_b^* = \frac{1}{w_b} \sum_{j=1}^{m_b} T_{jb} w_{jb}$ is the weighted treatment group assignment probability, $\beta_{0,b} = p_b^* \bar{\bar{Y}}_b(1) + (1 - p_b^*) \bar{\bar{Y}}_b(0)$ is the mean potential outcome in the block, and the "error" term, u_{ijb}, can be expressed as

$$u_{ijb} = T_{jb}\left(Y_{ijb}(1) - \bar{\bar{Y}}_b(1)\right) + (1 - T_{jb})\left(Y_{ijb}(0) - \bar{\bar{Y}}_b(0)\right).$$

We center the treatment indicator in (4) to facilitate the theory without changing the estimator.

In contrast to usual formulations of the regression model, our residual, u_{ijb}, is random solely because of T_{jb} (that is, due to random assignment) (see also Freedman 2008; Lin 2013; Middleton 2018). This framework allows treatment effects to differ across individuals and clusters and is nonparametric because it makes no assumptions about the distribution of potential outcomes. Note that our model does not satisfy key assumptions of the usual regression model for correlated data: over the randomization distribution, $E(u_{ijb})$ is not zero, u_{ijb} is heteroscedastic, $\text{cov}(u_{ijb}, u_{i'jb})$ is not constant for individuals in the same cluster, $\text{cov}(u_{ijb}, u_{i'j'b})$ is nonzero for individuals in different clusters, and u_{ijb} is correlated with the regressor $\left(T_{jb} - p_b^*\right)$ (see Schochet 2016). Under this framework, correlations arise because individuals in the same cluster share the same treatment assignment, and because T_{jb} and $T_{j'b}$ are correlated due to the complete randomization of clusters within the finite population. This differs from the typical model-based framework where correlations arise from shared cluster-specific random effects (e.g., due to common environmental factors).

The model in (4) can also be expressed using block indicator variables as follows:

$$y_{ijb} = \sum_{s=1}^{h} \beta_{1,s} S_{ijb,s} \tilde{T}_{js} + \sum_{s=1}^{h} \beta_{0,s} S_{ijb,s} + u_{ijb}, \tag{5}$$

where $\tilde{T}_{jb} = (T_{jb} - p_b^*)$ is the centered treatment status indicator. Due to blocked random assignment, the errors are independent across blocks. We include terms for all h blocks in the model and exclude a grand intercept term.

For estimation, we use the following working (hypothesized) model that provides covariate-adjusted ATE estimates by including in (5) a $1xv$ vector of fixed, block-mean-centered baseline covariates, $\tilde{\mathbf{x}}_{ijb}$, with associated parameter vector $\boldsymbol{\gamma}$:

$$y_{ijb} = \sum_{s=1}^{h} \beta_{1,k} S_{ijb,s} \tilde{T}_{js} + \sum_{s=1}^{h} \beta_{0,s} S_{ijb,s} + \tilde{\mathbf{x}}_{ijb} \boldsymbol{\gamma} + e_{ijb},$$

where $\tilde{\mathbf{x}}_{ijb} = (\mathbf{x}_{ijb} - \bar{\bar{\mathbf{x}}}_b)$, $\bar{\bar{\mathbf{x}}}_b = \frac{1}{w_b} \sum_{j=1}^{m_b} \sum_{i=1}^{n_{jb}} w_{ijb} \mathbf{x}_{ijb}$, and e_{ijb} is the error term. These covariates, unaffected by the treatment, can be at the individual or cluster level. We assume sufficient degrees of freedom for variance estimation (see Section 4.2). While the covariates do not enter the true RCT model in (4) and the ATE estimands do not change, the covariates will increase precision to the extent they are correlated with the potential outcomes. We do not need to assume that the true conditional distribution of y_{ijb} given $\mathbf{x}_{ijb}$ is linear in $\mathbf{x}_{ijb}$.

We do not consider working models that interact $\tilde{\mathbf{x}}_{ijb}$ and $\tilde{T}_{jb}$ due to associated degrees of freedom losses that can seriously reduce the power of clustered RCTs that, in practice, often contain relatively few clusters for cost reasons. Similarly, we pool $\boldsymbol{\gamma}$ across blocks. As discussed in Section 4.1, $\boldsymbol{\gamma}$ is well defined: it is the finite population regression coefficient that would be obtained if we could run the weighted regression on the full schedule of potential outcomes.

Using individual-level data, we can fit our working model using weighted least squares (WLS) with weights w_{ijb}. This yields the following closed-form WLS estimator for $\beta_{1,b}$ (see

Section A.3 in the supplementary materials for the derivation):

$$\begin{aligned}\hat{\beta}_{1,b} &= \frac{1}{w_b^1}\sum_{j:T_{jb}=1}^{m_b} w_{jb}\bar{y}_{jb} - \frac{1}{w_b^0}\sum_{j:T_{jb}=0}^{m_b} w_{jb}\bar{y}_{jb} \\ &\quad - \left(\frac{1}{w_b^1}\sum_{j:T_{jb}=1}^{m_b} w_{jb}\bar{\mathbf{x}}_{jb} - \frac{1}{w_b^0}\sum_{j:T_{jb}=0}^{m_b} w_{jb}\bar{\mathbf{x}}_{jb}\right)\hat{\boldsymbol{\gamma}} \\ &= \bar{\bar{y}}_b(1) - \bar{\bar{y}}_b(0) - (\bar{\bar{\mathbf{x}}}_b^1 - \bar{\bar{\mathbf{x}}}_b^0)\hat{\boldsymbol{\gamma}}, \qquad (6)\end{aligned}$$

where, for $t \in \{1,0\}$, $\bar{\bar{y}}_b(t)$ is the weighted average of the observed outcome across subjects in the treatment or control group, $w_b^t = \sum_{j:T_{jb}=t}^{m_b} w_{jb}$ is the sum of the weights, and $\hat{\boldsymbol{\gamma}}$ is the WLS estimator for $\boldsymbol{\gamma}$. $\hat{\beta}_{1,b}$ is the WLS estimator that would be obtained using standard statistical packages.

4.1. Theoretical Results

To examine the asymptotic properties of $\hat{\beta}_{1,b}$, we consider a hypothetical increasing sequence of finite populations where $m_b \to \infty$ in each block, so that $m = \sum_{b=1}^{h} m_b \to \infty$. The number of blocks, h, however, remains fixed. In principle, parameters should be subscripted by m, but we omit this notation for simplicity. We further assume that the proportion of all clusters in a block converges to a constant, that is, $m_b/m \to q_b$ as $m \to \infty$. We finally assume that $p_b = m_b^1/m_b$ is (approximately) constant as $m \to \infty$, so that the number of treated and control clusters in each block increases with m (that is, $m_b^1 \to \infty$ and $m_b^0 \to \infty$).

Given this framework, we present a CLT for the WLS estimator that provides design-based standard errors and associated inference. Before presenting our theorem, we first need to define several quantities pertaining to finite population variances and covariances. First, for $t \in \{1,0\}$, we define $D_b(t) = \frac{w_{jb}}{\bar{w}_b}(\bar{Y}_{jb}(t) - \bar{\bar{Y}}_b(t) - (\bar{\mathbf{x}}_{jb} - \bar{\bar{\mathbf{x}}}_b)\boldsymbol{\gamma})$ as the residualized potential outcomes at the cluster level in the treatment and control conditions, where $\bar{w}_b = \frac{1}{m_b}\sum_{j=1}^{m_b} w_{jb}$. Second, we define $S^2_{D_b}(t)$ as the variance of these residuals,

$$S^2_{D_b}(t) = \frac{1}{m_b - 1}\sum_{j=1}^{m_b}\frac{w_{jb}^2}{\bar{w}_b^2}(\bar{Y}_{jb}(t) - \bar{\bar{Y}}_b(t) - (\bar{\mathbf{x}}_{jb} - \bar{\bar{\mathbf{x}}}_b)\boldsymbol{\gamma})^2,$$

and $S^2_{D_b}(1,0)$ as the associated treatment-control covariance,

$$\begin{aligned}S^2_{D_b}(1,0) &= \frac{1}{m_b - 1}\sum_{j=1}^{m_b}\frac{w_{jb}^2}{\bar{w}_b^2}(\bar{Y}_{jb}(1) - \bar{\bar{Y}}_b(1) \\ &\quad - (\bar{\mathbf{x}}_{jb} - \bar{\bar{\mathbf{x}}}_b)\boldsymbol{\gamma})(\bar{Y}_{jb}(0) - \bar{\bar{Y}}_b(0) - (\bar{\mathbf{x}}_{jb} - \bar{\bar{\mathbf{x}}}_b)\boldsymbol{\gamma}).\end{aligned}$$

Third, we define $\mathrm{var}(\hat{D}_b)$ as the variance of the mean difference in residuals between the observed (randomized) treatment and control group samples,

$$\mathrm{var}\left(\hat{D}_b\right) = \frac{S^2_{D_b}(1)}{m_b^1} + \frac{S^2_{D_b}(0)}{m_b^0} - \frac{S^2(D_b)}{m_b}, \qquad (8)$$

where $S^2(D_b)$ is the variance (heterogeneity) of the ATEs across clusters in block b,

$$S^2(D_b) = \frac{1}{m_b - 1}\sum_{j=1}^{m_b}\frac{w_{jb}^2}{\bar{w}_b^2}(\bar{Y}_{jb}(1) - \bar{Y}_{jb}(0) - (\bar{\bar{Y}}_b(1) - \bar{\bar{Y}}_b(0)))^2.$$

Fourth, we define the variance of the weights as $S^2(w_b) = \frac{1}{m_b - 1}\sum_{j=1}^{m_b}(w_{jb} - \bar{w}_b)^2$. Fifth, we need the weighted variances, $S^2_{x_b,k}$, of each covariate k,

$$S^2_{x_b,k} = \frac{1}{m_b - 1}\sum_{j=1}^{m_b}\frac{w_{jb}^2}{\bar{w}_b^2}([\bar{\mathbf{x}}_{jb} - \bar{\bar{\mathbf{x}}}_b]_k)^2,$$

and the weighted variance-covariance matrix of the covariates with themselves, $\mathbf{S}^2_{\mathbf{x},b}$, which is analogous to the classic $\mathbf{X}'\mathbf{W}\mathbf{X}$ matrix in WLS,

$$\mathbf{S}^2_{\mathbf{x},b} = \frac{1}{m_b}\sum_{j=1}^{m_b}\sum_{i=1}^{n_{jb}} w_{ijb}\left(\mathbf{x}_{ijb} - \bar{\bar{\mathbf{x}}}_b\right)'\left(\mathbf{x}_{ijb} - \bar{\bar{\mathbf{x}}}_b\right).$$

Finally, we need two matrices that are analogous to the classic $\mathbf{X}'\mathbf{W}\mathbf{Y}$ matrix in WLS,

$$\mathbf{S}^2_{\mathbf{x},Y,b}(t) = \frac{1}{m_b}\sum_{j=1}^{m_b}\sum_{i=1}^{n_{jb}} w_{ijb}\mathbf{x}'_{ijb}Y_{ijb}(t) - \bar{\bar{\mathbf{x}}}'_b\overline{wY(t)}_b$$

and

$$\mathbf{S}^2_{\mathbf{x}Y,b}(t) = \frac{1}{m_b}\sum_{j=1}^{m_b}\sum_{i=1}^{n_{jb}}\left(w_{ijb}\mathbf{x}'_{ijb}Y_{ijb}(t) - \overline{wxY(t)}_b\right)^2,$$

where $\overline{wY(t)}_b = \frac{1}{m_b}\sum_{j=1}^{m_b}\sum_{i=1}^{n_{jb}} w_{ijb}Y_{ijb}(t)$ and $\overline{wxY(t)}_b = \frac{1}{m_b}\sum_{j=1}^{m_b}\sum_{i=1}^{n_{jb}} w_{ijb}\mathbf{x}'_{ijb}Y_{ijb}(t)$.

We now present our CLT theorem, proved in Section A.3 in the supplementary materials, which adapts finite population CLT results in Li and Ding (2017) and Scott and Wu (1981) to our setting.

Theorem 1. Assume $(C1)$, $(C2)$, and the following conditions for $t \in \{1,0\}$ and $b \in \{1,\ldots,h\}$:

$(C3)$ Letting $g_b(t) = \max_{1\le j\le m_b}\left(\frac{w_{jb}}{\bar{w}_b}\left(\bar{Y}_{jb}(t) - \bar{\bar{Y}}_b(t) - (\bar{\mathbf{x}}_{jb} - \bar{\bar{\mathbf{x}}}_b)\boldsymbol{\gamma}\right)\right)^2$, as $m \to \infty$,

$$\frac{1}{(m_b^t)^2}\frac{g_b(t)}{\mathrm{var}(\hat{D}_b)} \to 0.$$

$(C4)$ $f_b^t = m_b^t/m_b$ has a limiting value in $(0,1)$, and $S^2_{D_b}(t)$ and $S^2_{D_b}(1,0)$ also have finite limiting values.

$(C5)$ As $m \to \infty$,

$$\left(1 - f_b^t\right)\frac{S^2(w_b)}{m_b^t\bar{w}_b^2} \to 0.$$

$(C6)$ Letting $h_{b,k}(t) = \max_{1\le j\le m_b}\left\{\frac{w_{jb}}{\bar{w}_b}([\bar{x}_{jb} - \bar{\bar{x}}_b]_k)\right\}^2$ for all k, as $m \to \infty$,

$$\frac{1}{\min\left(m_b^1, m_b^0\right)}\frac{h_{b,k}(t)}{S^2_{x_b,k}} \to 0.$$

($C7$) $S^2_{x_b,k}$, $\mathbf{S}^2_{\mathbf{x},b}$, $\mathbf{S}^2_{\mathbf{x},Y,b}(t)$, and $\mathbf{S}^2_{\mathbf{x}Y,b}(t)$ have finite (positive definite) limiting values.

Then, as $m \to \infty$, $\hat{\beta}_{1,b}$ is a consistent estimator for $\beta_{1,b}$ and

$$\frac{\hat{\beta}_{1,b} - \left(\bar{\bar{Y}}_b(1) - \bar{\bar{Y}}_b(0)\right)}{\sqrt{\operatorname{var}\left(\hat{D}_b\right)}} \xrightarrow{d} N(0,1),$$

where $\operatorname{var}(\hat{D}_b)$ is defined as in (8).

Remark 1. Condition ($C3$) is a Lindeberg-type condition (controlling the tails) that allows us to invoke the CLT in Theorem 4 of Li and Ding (2017) that underlies our finite population CLT. ($C4$) ensures that the treatment and control group samples in each block both grow sufficiently fast, and also ensures limiting values of asymptotic variances and covariances of the residualized potential outcomes. ($C5$) provides a weak law of large numbers for the weights so that $\bar{w}_b^t/\bar{w}_b \xrightarrow{p} 1$, where $\bar{w}_b^t = \frac{1}{m_b^t}\sum_{j:T_j=t}^{m_b} w_{jb}$. ($C7$) specifies limiting values of the covariate variances and outcome-covariate covariances, which in turn, provide regularity conditions on $\hat{\boldsymbol{\gamma}}$.

These conditions imply that within a block, we cannot have one cluster (or a few clusters) asymptotically dominating all other clusters in terms of their weights, so that the weighted covariance matrices for the outcomes and covariates are well defined. In general, this is unlikely to be restrictive: due to normalizing by the mean block weight, even if weighted cluster sizes steadily grow, our results hold as long as the relative block sizes do not get too disparate.

Remark 2. The above theorem is proved as a two-step process. We first assume $\boldsymbol{\gamma} = \left(\sum_{b=1}^h q_b \mathbf{S}^2_{\mathbf{x},b}\right)^{-1} [\sum_{b=1}^h p_b q_b \mathbf{S}^2_{\mathbf{x},Y,b}(1) + \sum_{b=1}^h \left(1 - p_b\right) q_b \mathbf{S}^2_{\mathbf{x},Y,b}(0)]$ is known, where $q_b = m_b/m$ and $p_b = m_b^1/m_b$, and obtain a CLT with this known parameter. This parameter is the (unobserved) WLS coefficient vector that would be obtained using the full set of potential outcomes. We then show that $\hat{\boldsymbol{\gamma}}$ converges to the same asymptotic value as $\boldsymbol{\gamma}$ and use ($C6$) to ensure that $(\bar{\bar{\mathbf{x}}}_b^1 - \bar{\bar{\mathbf{x}}}_b^0)$ is asymptotically normal with zero mean, so that the ATE estimator still converges to a standard normal. Note that we focus on the asymptotic regime where the number of clusters increases. Liu and Yang (2020) instead allow the number of blocks to increase, focusing on settings with a small number of clusters per block and a large number of blocks.

Remark 3. The first two terms in (8) pertain to separate variances for the treatment and control groups because we allow for heterogeneous treatment effects. These variances are based on model residuals averaged to the cluster level and are similar in form to the variance formulas in Li and Ding (2017) for nonblocked designs. The third term pertains to the covariance of cluster-level average potential outcomes in the treatment and control conditions, $S^2_{D_b}(1,0)$, that we express in terms of the heterogeneity of treatment effects across clusters, $S^2(D_b)$. We hereafter label this term the "finite population heterogeneity" term. It cannot be identified from the data but can be bounded (as discussed in Section 4.2).

Remark 4. Under ($C1$)–($C5$), Theorem 1 also applies to models without covariates by setting $\boldsymbol{\gamma} = \mathbf{0}$, yielding a simple differences-in-means ATE ratio estimator, $\hat{\beta}_{1,b} = \bar{\bar{y}}_b(1) - \bar{\bar{y}}_b(0)$. We discuss the finite sample bias of this estimator in Section A.2.2 in the supplementary materials.

Corollary 1. Under the conditions of Theorem 1 with model covariates, the asymptotic variance in (8) is minimized when $\boldsymbol{\gamma} = \boldsymbol{\gamma}_{\mathbf{B}}$, where $\boldsymbol{\gamma}_{\mathbf{B}}$ is the between-cluster regression parameter using data aggregated to the cluster level. The parameter, $\boldsymbol{\gamma}_{\mathbf{B}}$, is defined analogously to $\boldsymbol{\gamma}$ in Remark 2 by replacing $\mathbf{S}^2_{\mathbf{x},b}$ with $\mathbf{S}^{\mathbf{2}}_{\mathbf{x},b,\mathbf{B}} = \frac{1}{m_b}\sum_{j=1}^{m_b} w_{jb}\left(\bar{\mathbf{x}}_{jb} - \bar{\bar{\mathbf{x}}}_b\right)'\left(\bar{\mathbf{x}}_{jb} - \bar{\bar{\mathbf{x}}}_b\right)$, and using parallel cluster-level versions of $\mathbf{S}^{\mathbf{2}}_{\mathbf{x},Y,b}(1)$ and $\mathbf{S}^{\mathbf{2}}_{\mathbf{x},Y,b}(0)$.

This result follows because $S^2_{D_b}(1)$ and $S^2_{D_b}(0)$ in (8) are based on cluster-level residuals. Section A.5 of the supplementary materials derives the asymptotic efficiency loss using the individual data when the covariates vary both within and between clusters. Intuitively, the $\boldsymbol{\gamma}$ parameter is a weighted average of between- and within-cluster population regression parameters. But the within-cluster covariates, $(\mathbf{x}_{ijb} - \bar{\mathbf{x}}_{jb})$, have no effect on $S^2_{D_b}(t)$, and hence, on precision, so they bias $\boldsymbol{\gamma}$. However, as discussed in Section 4.2, these efficiency losses using the individual data can be offset by other precision factors in finite samples.

Corollary 2. Under the conditions of Theorem 1 and assuming $\bar{w}_b = \frac{w_b}{m_b}$ has a finite limit for all b, the pooled ATE estimator across blocks, $\hat{\beta}_1 = \frac{1}{h\bar{w}}\sum_{b=1}^h w_b\hat{\beta}_{1,b}$, is consistent for β_1 in (2) and $\frac{1}{\sqrt{\operatorname{var}(\hat{D})}}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N(0,1)$, where $\operatorname{var}(\hat{D}) = \frac{1}{(h\bar{w})^2}\sum_{b=1}^h w_b^2 \operatorname{var}(\hat{D}_b)$ and $\bar{w} = \frac{1}{h}\sum_{b=1}^h w_b$.

This result follows because the $\hat{\beta}_{1,b}$ estimators are asymptotically independent.

4.2. Variance Estimation

We can estimate the block-specific variance in (8) with a consistent (upper bound) plug-in variance estimator based on the regression residuals averaged to the cluster level as follows:

$$\widehat{\operatorname{var}}\left(\hat{D}_b\right) = \frac{s^2_{D_b}(1)}{m_b^1} + \frac{s^2_{D_b}(0)}{m_b^0}, \tag{9}$$

where

$$s^2_{D_b}(1) = \frac{1}{\left(m_b^1 - v^* p_b^* q_b^* - 1\right)} \sum_{j:T_{jb}=1}^{m_b^1} \frac{w_{jb}^2}{(\bar{w}_b^1)^2}(\bar{y}_{jb} - \hat{\beta}_{0,b} - (1 - p_b^*)\hat{\beta}_{1,b} - \tilde{\bar{\bar{\mathbf{x}}}}_{jb}\hat{\boldsymbol{\gamma}})^{\mathbf{2}},$$

$$s^2_{D_b}(0) = \frac{1}{\left(m_b^0 - v^*(1 - p_b^*) q_b^* - 1\right)} \sum_{j:T_{jb}=0}^{m_b^0} \frac{w_{jb}^2}{(\bar{w}_b^0)^2}(\bar{y}_{jb} - \hat{\beta}_{0,b} + p_b^*\hat{\beta}_{1,b} - \tilde{\bar{\bar{\mathbf{x}}}}_{jb}\hat{\boldsymbol{\gamma}})^{\mathbf{2}},$$

$q_b^* = \frac{w_b}{\sum_{b=1}^{h} w_b}$ is the weighted share of all clusters in block b, and v^* is a degrees of freedom adjustment for the covariates. As discussed in Donald and Lang (2007), plausible values for v^* are $v^* = v$ (the number of covariates), which applies when using cluster-level covariates, or $v^* = 0$, which applies when using individual-level covariates that vary only within clusters and not between clusters. Other approaches have been proposed, such as adjusting v for design effects due to clustering (Hedges 2007) and minimum distance methods (Wooldridge 2006).

In our simulations in Section 6, we also use two variants of (9). First, we multiply (9) by $(1 - R^2_{TXb})^{-1}$, where R^2_{TXb} is the R-squared value from a regression of $S_{ijb,s}\tilde{T}_{jb}$ on $\tilde{\mathbf{x}}_{ijb}$ and the other block-by-treatment status interactions in (5) (with no intercept). This term captures the finite sample collinearity between $\tilde{T}_{jb}$ and $\tilde{\mathbf{x}}_{ijb}$ (which inflates the variances). This estimator performs well in our simulations. The second variant subtracts $\frac{1}{m_b}\left(\sqrt{s^2_{D_b}(1)} - \sqrt{s^2_{D_b}(0)}\right)^2$, a lower bound on the finite population heterogeneity term based on the Cauchy–Schwarz inequality. Aronow et al. (2014) discuss sharper bounds on this heterogeneity term by approximating the marginal distributions of potential outcomes.

The regression model can also be estimated using data averaged to the cluster level, so that setting $v^* = v$ yields well-defined degrees of freedom. For models that use cluster-level covariates only (or no covariates), estimators using the individual and aggregate data will coincide. However, for models with individual-level covariates that vary both within and between clusters, the estimators can differ. In this case, as discussed in Section 4.1, data aggregation yields asymptotically efficient estimators. However, as proved and quantified in Section A.6 (the supplementary materials), in finite samples, precision losses from using individual data could be offset by precision gains due to the reduced collinearity between the covariates and treatment indicators (TX collinearity). In our simulations, we examine precision levels using the individual and aggregate data to assess these counteracting factors.

Finally, we can obtain pooled ATE estimators across all blocks by inserting $\hat{\beta}_{1,b}$ into (2) and using $\frac{1}{(h\bar{w})^2}\sum_{b=1}^{h} w_b^2 \widehat{\text{var}}(\hat{D}_b)$ for variance estimation. Hypothesis testing can be conducted using z-tests. Alternatively, results in Bell and McCaffrey (2002), Hansen (2007), and Cameron and Miller (2015) for the CRSE estimator suggest that t-tests with $(m - 2h - v^*)$ degrees of freedom perform better in small samples and is what we use hereafter.

4.3. *Comparing Design-Based and CRSE Estimators*

The CRSE variance estimator is an extension of robust standard errors (Huber 1967; White 1980) to clustered designs (Liang and Zeger 1986). The CRSE approach, which assumes iid sampling of clusters from some (infinite) super-population, allows for errors to be correlated within clusters but not across clusters. As with the design-based estimators, the CRSE estimator is based on weighted least squares using the pooled data across blocks and is asymptotically normal. Therefore, both approaches yield the same ATE estimator when the covariates and weights are the same, but the variance estimators differ in several ways.

To illustrate the key variance differences, consider first the model without covariates. In this case, as shown in Section A.7 in the supplementary materials, the CRSE variance estimator for a single block-by-treatment ATE estimate, $\hat{\beta}_{1,b}$, is

$$\widehat{\text{var}}_{\text{CRSE}}\left(\hat{\beta}_{1,b}\right) = g\frac{s^{2*}_{D_b}(1)}{m_b^1} + g\frac{s^{2*}_{D_b}(0)}{m_b^0}, \tag{10}$$

where $s^{2*}_{D_b}(1) = \frac{(m_b^1-1)}{m_b^1}s^2_{D_b}(1)$, $s^{2*}_{D_b}(0) = \frac{(m_b^0-1)}{m_b^0}s^2_{D_b}(0)$, and g is a small sample correction term. Here, we use $g = \left(\frac{m}{m-1}\right)\left(\frac{n-1}{n-2h-v}\right)$, a common value in statistical software packages such as Stata (Cameron and Miller 2015), although other approaches have been proposed, such as bias-corrected CRSE estimators (Mackinnon and White 1985; Angrist and Lavy 2002; Bell and McCaffrey 2002; Pustejovsky and Tipton 2018) and bootstrap methods (Cameron et al. 2008; Webb 2013) to adjust for the known Type 1 error inflation of the CRSE estimator in small samples.

Compare (10) to the following parallel expression for the design-based estimator in (9):

$$\widehat{\text{var}}\left(\hat{D}_b\right) = \frac{m_b^1}{(m_b^1 - 1)}\frac{s^{2*}_{D_b}(1)}{m_b^1} + \frac{m_b^0}{(m_b^0 - 1)}\frac{s^{2*}_{D_b}(0)}{m_b^0}. \tag{11}$$

Examining (10) and (11) establishes that the design-based and CRSE variance estimators are asymptotically equivalent, because both correction terms converge to 1 as $m_b^1 \to \infty$ and $m_b^0 \to \infty$. In finite samples, however, there are two key differences between the estimators that pertain to the degrees of freedom adjustments. First, the adjustments for the design-based variance estimator are applied separately for treatments and controls based on m_b^1 and m_b^0, whereas the standard CRSE estimator applies a single adjustment, g, based on total sample sizes (m and n). Second, the design-based estimator uses $(m-2)$ degrees of freedom for the t-tests, reflecting separation of the two research groups, whereas the CRSE estimator commonly uses $(m - 1)$ (Cameron and Miller 2015). These two differences will typically lead to larger design-based variances and lower rejection rates. Note also that the finite population heterogeneity term does not apply to the CRSE estimator as it assumes a super-population sampling framework. Similar results apply to the model with covariates (see Section A.7 in the supplementary materials).

5. Restricted ATE Estimators with Fixed Block Effects Only

A commonly used estimation strategy for blocked designs is to include block indicator variables in the regression model but to exclude block-by-treatment status interaction terms:

$$y_{ijb} = \beta_{1,R}\tilde{T}_{jb} + \sum_{s=1}^{h}\delta_{0,s}S_{ijb,s} + \varepsilon_{ijb}, \tag{12}$$

where ε_{ijb} is the error term. Because this framework imposes restrictions on the assumed data structure, it typically produces asymptotically biased estimates of the true ATE parameter in (2). Nevertheless, it has practical appeal due to its parsimony and additional degrees of freedom.

Consider WLS estimation of (12) where the model includes the $\tilde{\mathbf{x}}_{ijb}$ covariates with parameter vector $\boldsymbol{\gamma}$. As shown in Section A.4 (supplementary materials), the WLS estimator for $\beta_{1,R}$ is a weighted average of block-level ATE estimates with weights, $\tilde{w}_{b,R} = \frac{1}{mw_b} w_b^0 w_b^1$:

$$\hat{\beta}_{1,R} = \sum_{b=1}^{h} \frac{\tilde{w}_{b,R}}{\sum_{a=1}^{h} \tilde{w}_{a,R}} \left(\bar{\bar{y}}_b(1) - \bar{\bar{y}}_b(0) - \left(\bar{\bar{\mathbf{x}}}_b^1 - \bar{\bar{\mathbf{x}}}_b^0 \right) \hat{\boldsymbol{\gamma}} \right). \quad (13)$$

The weights can also be expressed in the limit as $\tilde{w}_{b,R} = q_b p_b \left(1 - p_b\right) \bar{w}_b$, where $q_b = m_b/m$. Thus, this approach uses a form of precision weighting to weight the block-specific treatment effects and is analogous to a fixed effects regression model using nonclustered data. The weights for the restricted model will differ from those for the unrestricted model (causing asymptotic bias), except if the ATEs or $p_b\left(1 - p_b\right)$ values are homogenous across blocks.

5.1. Theoretical Results

We now present a CLT for $\hat{\beta}_{1,R}$ that is proved in Section A.4 in the supplementary materials. Let

$$\beta_{1,R} = \sum_{b=1}^{h} \frac{q_b p_b (1 - p_b) \bar{w}_b}{\sum_{b=1}^{h} q_b p_b (1 - p_b) \bar{w}_b} \left(\bar{\bar{Y}}_b(1) - \bar{\bar{Y}}_b(0) \right)$$

denote the treatment effect parameter for the restricted model. Also define the vector of block-level estimators as a series of triples:

$$\mathbf{t} = (\bar{w}_1^1 \left(\bar{\bar{y}}_1(1) - \bar{\bar{\mathbf{x}}}_1^1 \boldsymbol{\gamma} \right), \bar{w}_1^0 \left(\bar{\bar{y}}_1(0) - \bar{\bar{\mathbf{x}}}_1^0 \boldsymbol{\gamma} \right), \bar{w}_1^1, \ldots, \\ \bar{w}_h^1 \left(\bar{\bar{y}}_h(1) - \bar{\bar{\mathbf{x}}}_h^1 \boldsymbol{\gamma} \right), \bar{w}_h^0 \left(\bar{\bar{y}}_h(0) - \bar{\bar{\mathbf{x}}}_h^0 \boldsymbol{\gamma} \right), \bar{w}_h^1).$$

Theorem 2. Assume $(C1)$, $(C2)$, $(C4)$ for f_b^t, $(C5)$, $(C6)$, $(C7)$ and the following conditions for $t \in \{1, 0\}$:

$(C8)$ As $m \to \infty$,

$$\max_{1 \le b \le h} \frac{a_{Y,b}(t)}{p_b(1 - p_b) m_b v_{Y,b}(t)} \to 0, \text{ where}$$

$$a_{Y,b}(t) = \max_{1 \le j \le m_b} \left(w_{jb} \left(\bar{Y}_{jb}(t) - \bar{\mathbf{x}}_{jb} \boldsymbol{\gamma} \right) - \bar{w}_b \left(\bar{\bar{Y}}_b(t) - \bar{\bar{\mathbf{x}}}_b \boldsymbol{\gamma} \right) \right)^2,$$

$$v_{Y,b}(t) = \frac{1}{m_b - 1} \sum_{j=1}^{m_b} \left(w_{jb} \left(\bar{Y}_{jb}(t) - \bar{\mathbf{x}}_{jb} \boldsymbol{\gamma} \right) - \bar{w}_b \left(\bar{\bar{Y}}_b(t) - \bar{\bar{\mathbf{x}}}_b \boldsymbol{\gamma} \right) \right)^2.$$

$(C9)$ As $m \to \infty$,

$$\max_{1 \le b \le h} \frac{a_{w,b}}{p_b(1 - p_b) m_b v_{w,b}} \to 0,$$

where $a_{w,b} = \max_{1 \le j \le m_b} \left(w_{jb} - \bar{w}_b \right)^2$ and $v_{w,b} = \frac{1}{m_b - 1} \sum_{j=1}^{m_b} \left(w_{jb} - \bar{w}_b \right)^2$.

$(C10)$ The correlation matrix of $\mathbf{t}$ has a finite limiting value $\boldsymbol{\Sigma}$.

$(C11)$ The variance expressions, $v_{w,b}$ and $v_{Y,b}(t)$, have finite limiting values for $b \in \{1, \ldots, h\}$.

$(C12)$ $\bar{w}_b \left(\bar{\bar{Y}}_b(1) - \bar{\bar{\mathbf{x}}}_b \boldsymbol{\gamma} \right) \neq 0$ or $\bar{w}_b \left(\bar{\bar{Y}}_b(0) - \bar{\bar{\mathbf{x}}}_b \boldsymbol{\gamma} \right) \neq 0$ for some b.

Then, as $m \to \infty$, $\hat{\beta}_{1,R}$ is a consistent estimator for $\beta_{1,R}$ and

$$\frac{\hat{\beta}_{1,R} - \beta_{1,R}}{\sqrt{\operatorname{var}\left(\tilde{\beta}_{1,R} \right)}} \xrightarrow{d} N(0, 1), \text{ where}$$

$$\begin{aligned} &\operatorname{var}\left(\tilde{\beta}_{1,R} \right) \quad (14) \\ &= \sum_{b=1}^{h} \frac{1}{m_b(m_b - 1)} \frac{(q_b p_b (1 - p_b) \bar{w}_b)^2}{(\sum_{a=1}^{h} q_a p_a (1 - p_a) \bar{w}_a)^2} \\ &\quad \sum_{j=1}^{m_b} \left(\sqrt{\frac{1 - p_b}{p_b}} \left(\frac{w_{jb} \left(\bar{Y}_{jb}(1) - \bar{\bar{Y}}_b(1) - \left(\bar{\mathbf{x}}_{jb} - \bar{\bar{\mathbf{x}}}_b \right) \boldsymbol{\gamma} \right)}{\bar{w}_b} \right) \right. \\ &\quad + \sqrt{\frac{p_b}{1 - p_b}} \left(\frac{w_{jb} \left(\bar{Y}_{jb}(0) - \bar{\bar{Y}}_b(0) - \left(\bar{\mathbf{x}}_{jb} - \bar{\bar{\mathbf{x}}}_b \right) \boldsymbol{\gamma} \right)}{\bar{w}_b} \right) \\ &\quad \left. + \frac{(1 - 2p_b)}{\sqrt{p_b(1 - p_b)} \bar{w}_b} (\beta_{1,b} - \beta_{1,R})(w_{jb} - \bar{w}_b) \right)^2. \end{aligned}$$

Remark. The first two terms inside the brackets in (14) pertain to block-specific variances for the treatment and control groups that are analogous to the corresponding variance terms in the unrestricted model in (8). The third term represents the covariance between the block treatment effects and block weights that is induced by the restricted model. This term differentiates the variances for the restricted and unrestricted models, along with the block weights used for pooling. This third term is 0 if $p_b = 0.5$ or $\operatorname{cov}\left(\beta_{1,b}, w_{jb} \right) = 0$ for all b, but otherwise can be positive or negative.

5.2. Variance Estimation

A consistent variance estimator for (14) can be obtained by multiplying out the squared term and using plug-in estimators for each of the resulting six terms. However, it is simpler to use the following consistent estimator based on cluster-level model residuals:

$$\widehat{\operatorname{var}}\left(\hat{\beta}_{1,R} \right) = \frac{m}{(m - h - v^* - 1)} \\ \frac{\sum_{b=1}^{h} \sum_{j=1}^{m_b} w_{jb}^2 \tilde{T}_{jb}^2 (\bar{y}_{jb} - \hat{\beta}_{1,R} \tilde{T}_{jb} - \hat{\delta}_{0,b} - \bar{\tilde{\mathbf{x}}}_{jb} \hat{\boldsymbol{\gamma}})^2}{(\sum_{b=1}^{h} m_b p_b^* (1 - p_b^*) \bar{w}_b)^2}, \quad (15)$$

recalling that $\tilde{T}_{jb} = (T_{jb} - p_b^*)$. Following Schochet (2016), the expression in (15) can be justified using the following standard asymptotic expansion for the WLS estimator:

$$\sqrt{m} \left(\hat{\beta}_{1,R} - \beta_{1,R} \right) \\ = \frac{\sum_{b=1}^{h} \sum_{j=1}^{m_b} w_{jb} \tilde{T}_{jb} (\bar{y}_{jb} - \beta_{1,R} \tilde{T}_{jb} - \delta_{0,b} - \bar{\tilde{\mathbf{x}}}_{jb} \boldsymbol{\gamma})}{\sqrt{m} \sum_{b=1}^{h} q_b p_b (1 - p_b) \bar{w}_b} + o_p(1), \quad (16)$$

where $o_p(1)$ signifies a term that converges in probability to zero. Suppose we insert into (16), $\bar{y}_{jb} = T_{jb} \bar{Y}_{jb}(1) + (1 - T_{jb}) \bar{Y}_{jb}(0)$

and $\delta_{0,b} = p_b \bar{\bar{Y}}_b(1) + (1 - p_b) \bar{\bar{Y}}_b(0)$ (see Section A.4 in the supplement), and then add and subtract $(\beta_{1,b} - \beta_{1,R}) \tilde{T}_{jb}$. If we then calculate $\text{var}\left(\hat{\beta}_{1,R}\right)$ over the randomization distribution, we obtain (15) after some algebra. Hypothesis testing can be conducted using t-tests with $(\sum_{b=1}^{h} \left(m_b^1 + m_b^0\right) - \nu^* - h - 1)$ degrees of freedom. Note that with $\nu^* = 0$, estimation only requires at least 1 treatment and 1 control cluster per block rather than two as for the fully-interacted model. A $(1 - R_{TX}^2)^{-1}$ correction can also be applied to (15).

6. Simulation Results

To examine the statistical properties of the design-based estimator, we conducted simulations for a clustered, non-blocked design ($h = 1$). We applied the variance estimator in (9) with and without (1) model covariates, (2) the R_{TX}^2 adjustment, and (3) the finite population heterogeneity term based on the Cauchy-Schwarz inequality. We also included the standard CRSE estimator to identify sources of differences between the two approaches (our goal is not to compare various proposed CRSE estimators to each other). Note that if $h > 1$, differences in block-level estimates between the two approaches will be greater than the simulation results presented here (holding m fixed), because the degrees of freedom adjustments will differ more (see (10) and (11)).

To generate potential outcomes for our primary simulations, we used the following model (see Section B in the supplementary materials for simulation details):

$$
\begin{aligned}
Y_{ij}(0) &= x_{ij1} + x_{ij2} + u_j + e_{ij} \\
Y_{ij}(1) &= Y_{ij}(0) + \theta_j,
\end{aligned}
\tag{17}
$$

where u_j, θ_j (which captures treatment effect heterogeneity), and e_{ij} are each iid mean zero random errors and x_{ij1} and x_{ij2} are independent covariates. For each draw of potential outcomes, we conducted 1000 replications where we randomly assigned clusters to either the treatment or control group. We ran separate simulations for $m = 8$ to 50 clusters.

We examined a range of simulation scenarios for the covariates and model distributions. We generated data with (1) no covariates (excluding x_{ij1} and x_{ij2} from (17)); (2) two individual-level covariates (applying an intraclass correlation coefficient of $\rho_X = 0$); (3) two cluster-level covariates (applying $\rho_X = 1$); and (4) one individual-level and one cluster-level covariate. For the models with individual-level covariates, we calculated the degrees of freedom in three ways, setting ν^* equal to 0, the total number of covariates (ν), or the number of cluster-level covariates (if applicable). We generated data assuming normal, bimodal, and chi-squared distributions for the errors and covariates in (17).

Finally, to compare the true variances of models estimated using individual and aggregate data (see Sections 4.1 and 4.2), rather than directly generating data for x_{ijk}, we instead generated data separately for the between- and within-cluster components, $\bar{x}_{jk}$ and $\left(x_{ijk} - \bar{x}_{jk}\right)$. We then included these components as covariates in (17) to generate the outcomes, allowing for different parameter values, $\boldsymbol{\gamma}_{Bk}$ and $\boldsymbol{\gamma}_{Wk}$. The individual-level models, however, were estimated using x_{ijk} only (not the components) with parameter $\boldsymbol{\gamma}_k$ (see the supplementary materials for details).

6.1. Results Without Covariates

For models without covariates, the simulation results indicate that the design-based estimator without the correction for the finite population heterogeneity term yields Type I errors near the 5% nominal level and standard errors near true values, even with relatively few clusters (see Figure 1 and Table B.1, supplementary materials for the full results). In contrast, the standard CRSE estimator in (10) yields inflated Type I errors similar to those found in the literature using a super-population simulation framework (see, e.g., Cameron et al. 2008; Green and Vavrek 2008; Angrist and Pischke 2009) (Figure 1; Table B.1, supplementary materials). These differences arise because the CRSE estimator applies a *single* degrees of freedom variance adjustment based on the total sample size, whereas the design-based estimator applies a *separate* adjustment for the treatment and control groups, which inflates the variances. A more minor reason is that the CRSE approach uses $(m - 1)$ degrees of freedom for the t-tests rather than $(m - 2)$. The design-based variance estimator that includes a correction for the finite population heterogeneity term also overrejects (so we do not focus on this estimator in what follows), but less so than the CRSE estimator (Figure 1; Table B.4, supplementary materials). We find similar simulation results using different model distributions (Table B.1, supplementary materials) and individual sample sizes (Table B.4, supplementary materials).

Biases of the ATE estimators are negligible (Table B.1, supplementary materials). Further, mean squared errors of the estimated standard errors are nearly identical for the two approaches, suggesting similar stability in estimating uncertainty (Table B.1, supplementary materials). This result occurs because the larger variance of the estimated standard errors for the design-based estimator is offset by the smaller bias in its estimated standard errors (relative to the "true" values as measured by the standard deviation of the 1000 ATE estimates across replications). Finally, for small m, power levels are lower

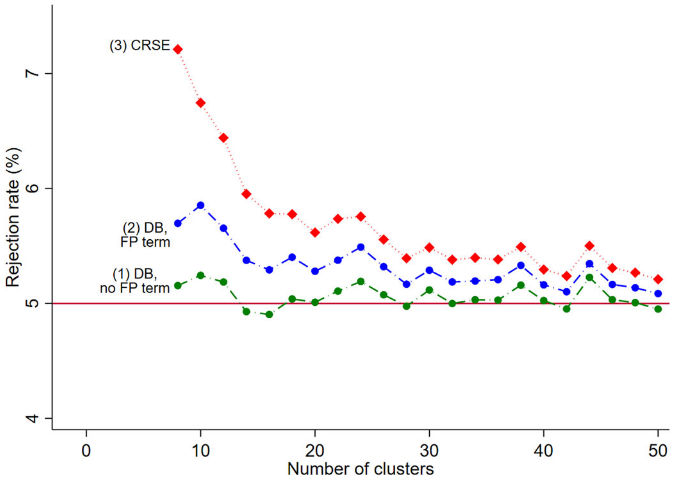

Figure 1. Type I error rates for models without covariates.
Abbreviations. DB = Design-based variance estimator; FP term = Finite population heterogeneity term included based on the Cauchy-Schwarz inequality; CRSE = Standard cluster-robust standard error estimator.

for the design-based than CRSE estimator (due to lower Type I errors), but the design-based estimator more closely matches power levels calculated using the true standard errors (Table B.2, supplementary materials).

6.2. *Results with Covariates*

A similar pattern of results arises when covariates are included in the model (and standard errors decrease) (Figure 2; Tables B.3 and B.4, supplementary materials). For the model with individual-level covariates only ($\rho_X = 0$), we find that the design-based estimator with $v^* = 0$ yields Type I errors near the nominal level (Model (1) in Figure 2). In this case, the R^2_{TX} adjustment has little effect on the results (Table B.3, supplementary materials). If we instead apply $v^* = 2$, the design-based approach becomes conservative (Model (2) in Figure 2). As before, the standard CRSE estimator yields inflated Type 1 errors (Model (3) in Figure 2).

For the model with cluster-level covariates only ($\rho_X = 1; v^* = 2$), which is identical to aggregating the data to the cluster level, the design-based estimator yields Type 1 errors at the nominal level if the R^2_{TX} adjustment is applied, even with $m = 8$, but overrejects without this adjustment (Models (4) and (5) in Figure 2). For this specification, the CRSE estimator produces Type 1 errors that are more inflated than with individual covariates (Model (6) in Figure 2).

For models containing one individual-level covariate ($v_1 = 1$) and one cluster-level covariate ($v_2 = 1$), we find that for the design-based estimator, setting $v^* = 0$ is liberal; setting $v^* = v_2 = 1$ yields Type I errors close to the nominal rate; and setting $v^* = v_1 + v_2 = 2$ is conservative (Table B.4, supplementary materials). As before, the CRSE tends to overreject with few clusters.

Finally, the simulation results comparing the true variances based on the individual and aggregate data—where we generated data using $\bar{x}_{jk}$ and $\left(x_{ijk} - \bar{x}_{jk}\right)$ but estimated the models using x_{ijk} only—support the theory presented in Sections 4.1 and 4.2 (Table B.5, supplementary materials). When $\boldsymbol{\gamma}_{Bk} = \boldsymbol{\gamma}_{Wk}$, the true variances are always smaller using the individual data because $\boldsymbol{\gamma}_k$ is asymptotically efficient in this case and the TX collinearities are always smaller. The differences decrease, however, as m increases and the TX collinearities become negligible. In contrast, when $\boldsymbol{\gamma}_{Bk}/\boldsymbol{\gamma}_{Wk} = 2$, with a single covariate, the aggregate data produce more precise estimates, even when $m = 8$. However, with five covariates, the TX collinearities become more problematic, so even when $\boldsymbol{\gamma}_{Bk}/\boldsymbol{\gamma}_{Wk} = 2$, the individual data yield efficiency gains unless $m \geq 50$. With more covariates, the TX collinearities using the aggregate data are severe, favoring the use of the individual data.

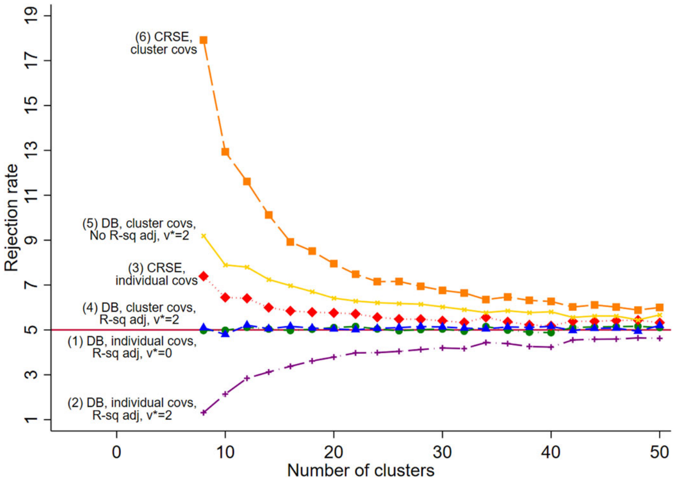

Figure 2. Type I error rates for models with covariates.
Abbreviations. DB = Design-based variance estimator in (9) (without the finite population heterogeneity term); "cluster covs" = Two cluster-level covariates included in the model; "Individual covs" = Two individual-level covariates included in the model; "R-sq-adj" = R^2_{TX} adjustment applied to the DB estimator; $v^* =$ Degrees of freedom adjustment for the covariates for the DB estimator; CRSE = Standard cluster-robust standard error estimator.

6.3. *Super-Population Simulation Results*

To compare our results to those in the literature for the CRSE estimator, we also conducted select simulations using a super-population framework by generating 50,000 separate datasets. These results show a similar pattern to the results above, but with somewhat larger true standard errors (Table B.6, supplementary materials).

6.4. *Discussion*

Overall, the simulation results suggest that the design-based estimator has beneficial statistical properties with few clusters for models with or without covariates. For models with covariates and relatively large n (typical for clustered RCTs in practice), the results suggest that setting $v^* = v_2$ (the number of cluster-level covariates) could be a good general strategy, and that setting $v^* = v$ (the total number of covariates) is conservative. The simulations further indicate that even with individual-level covariates, the design-based estimator using data averaged to the cluster-level with the R^2_{TX} adjustment and $v^* = v$ yields nominal rejection rates. While aggregation can result in losses in statistical power (due to increased TX collinearities), it could be a good strategy with small numbers of covariates if m is moderate.

7. Empirical Application Using the Motivating SACD Example

To demonstrate the considered estimators, we use outcome and baseline data from the SACD evaluation on 4018 4th graders (2147 treatments and 1871 controls) in 84 schools in 7 large school districts. The data were obtained from student reports administered in the classroom, primary caregiver telephone interviews, and teacher reports on students (SACD Consortium 2010). We analyze six primary study outcome scales (Table 1) and adjust for baseline covariates selected in an initial step from the 46 available, along with their two-way interactions, using least absolute shrinkage and selection operator (LASSO) methods with 5-fold cross-validation (Tibshirani 1996). We use the LASSO-WLS hybrid procedure for clustered RCTs developed in Schochet (2020b) based on the design-based estimators presented above (see also Bloniarz et al. 2016). In the first stage, LASSO estimation is conducted using cluster-level averages, and in the second stage, design-based WLS estimation is conducted using the individual data and first stage LASSO covariates. Our

Table 1. Outcome variables for the empirical analysis using SACD RCT data.

Outcome	Data source (Spring 2007)	Description of variable
Problem behavior	Child report	Scale ranges from 0 to 3 and contains 6 items from the Frequency of Delinquent Behavior Scale and 6 items from the Aggression Scale; Reliability = 0.86.
Normative beliefs about aggression	Child report	Scale ranges from 1 to 4 and contains 12 items from the Normative Beliefs About Aggression Scale; Reliability = 0.83.
Student afraid at school	Child report	Scale ranges from 1 to 4 and contains 4 items from the Feelings of Safety at School scale; Reliability = 0.79.
Altruistic behavior	Primary caregiver report	Scale ranges from 1 to 4 and contains 8 items from the Altruism Scale, Primary Caregiver Version; Reliability = 0.88
Positive social behavior	Teacher report	Scale ranges from 1 to 4 and contains 6 items from the Responsibility Scale and 19 items from the Social Competence Scale and 8 items from the Altruism Scale, Teacher Version; Reliability = 0.97.
Problem behavior	Teacher report	Scale ranges from 1 to 4 and contains 14 items from the BASC Aggression Subscale, Teacher Version, 7 items from the BASC Conduct Problems Subscale, Teacher Version and 2 items from the Responsibility Scale; Reliability = 0.95.

Note: See SACD Research Consortium (2010) for a complete description of the construction of these scales.

Table 2. Estimated ATEs and standard errors for the SACD study, by model specification.

	Model with site-by-treatment interaction terms					
	Individuals weighted equally		Schools and sites weighted equally		Model with site fixed effects only	
Outcome variable and covariate specification	Design-based	Standard CRSE	Design-based	Standard CRSE	Design-based	Standard CRSE
Model without covariates						
Problem behavior (CR)	0.006	0.006	0.011	0.011	0.006	0.006
	(0.037)	(0.034)	(0.041)	(0.038)	(0.036)	(0.035)
Normative beliefs about aggression (CR)	0.003	0.003	0.000	0.000	0.003	0.003
	(0.031)	(0.029)	(0.038)	(0.035)	(0.031)	(0.030)
Student afraid at school (CR)	−0.064	−0.064	−0.041	−0.041	−0.064	−0.064
	(0.052)	(0.048)	(0.061)	(0.056)	(0.050)	(0.048)
Altruistic behavior (PCR)	−0.006	−0.006	−0.011	−0.011	−0.006	−0.006
	(0.035)	(0.032)	(0.041)	(0.037)	(0.034)	(0.033)
Positive social behavior (TR)	−0.046	−0.046	−0.036	−0.036	−0.045	−0.045
	(0.061)	(0.056)	(0.065)	(0.060)	(0.060)	(0.056)
Problem behavior (TR)	0.019	0.019	0.006	0.006	0.019	0.019
	(0.040)	(0.036)	(0.044)	(0.040)	(0.039)	(0.038)
Model with covariates						
Problem behavior (CR)	−0.006	−0.006	−0.002	−0.002	−0.006	−0.006
	(0.027)	(0.025)	(0.031)	(0.028)	(0.027)	(0.025)
Normative beliefs about aggression (CR)	−0.005	−0.005	−0.009	−0.009	−0.005	−0.005
	(0.026)	(0.024)	(0.033)	(0.030)	(0.026)	(0.025)
Student afraid at school (CR)	−0.067*	−0.067*	−0.047	−0.047	−0.067*	−0.067*
	(0.040)	(0.035)	(0.045)	(0.040)	(0.037)	(0.036)
Altruistic behavior (PCR)	−0.016	−0.016	−0.013	−0.013	−0.017	−0.017
	(0.028)	(0.024)	(0.031)	(0.027)	(0.027)	(0.026)
Positive social behavior (TR)	−0.011	−0.011	−0.015	−0.015	−0.010	−0.010
	(0.045)	(0.039)	(0.045)	(0.040)	(0.043)	(0.040)
Problem behavior (TR)	−0.009	−0.009	−0.011	−0.011	−0.009	−0.009
	(0.025)	(0.023)	(0.026)	(0.023)	(0.025)	(0.023)

CR = child report; PCR = primary caregiver report; TR = teacher report; CRSE = Cluster-robust standard error estimator.
*Statistically significant at the 10% level, two-tailed test.

goal is not to replicate study results but to illustrate the ATE estimators.

Table 2 presents the estimation results for various model specifications: (1) with and without baseline covariates, (2) with and without block-by-treatment status interaction terms, and (3) with equal weighting of individuals (to estimate ATEs for the average student in the sample) versus equal weighting of sites and clusters (to estimate ATEs for the average school in the average district). Our methods can easily accommodate weights to adjust for data item nonresponse. We present results for both the design-based and standard CRSE estimators.

The results indicate that for all specifications, the behavioral health interventions had no statistically significant effect on any outcome scale, although the negative estimate on the scale measuring fear in school is marginally statistically significant at the 10% level for most models with covariates (Table 2). Across the six outcomes, standard errors are about 16 to 35 percent smaller when covariates are included in the models. Further, we find very similar results for the fully-interacted and restricted models for two reasons: (1) the estimated treatment effects vary little across sites (an average standard deviation of 0.07 across the outcomes) and (2) the two sets of site weights are highly correlated (greater than 0.95) because p_b is about 0.5 in all sites. Further, because sample sizes do not vary substantially across sites (they range from 425 to 650 students in 10 to 14 schools), findings do not materially differ when individuals versus blocks and clusters are weighted equally, although in the latter case, standard errors increase due to design effects from weighting, and the marginally significant impact on the scale measuring fear at school disappears. Finally, consistent with the theory and simulations, standard errors are somewhat larger using the design-based estimators than the parallel CRSE estimators.

8. Conclusions

This article considered design-based ratio estimators for clustered, blocked RCTs using the Neyman-Rubin-Holland model and weighted least-square methods. We developed finite population CLTs for the ATE estimators, allowing for baseline covariates to improve precision, general weighting schemes, and several common approaches for handling blocks in the models. We showed that the design-based ratio estimators are attractive in that they yield consistent and asymptotically normal ATE estimators with simple variance estimators based on cluster-level model residuals; apply to continuous, binary, and discrete outcomes; and yield Type I errors at nominal levels for models with and without covariates, even in small samples. Our theory applies to analyses conducted using either individual or aggregate (cluster-level) data, where our results suggest that in practice, the use of individual data will tend to yield more precise estimates for models with covariates (that vary both within and between clusters), unless the number of covariates is very small.

An unexpected finding is that the "conservative" variance estimator that excludes a correction for the finite population heterogeneity term based on the Cauchy-Schwarz inequality improves statistical performance. Further, for models with covariates, an R^2_{TX} adjustment for the collinearity between the covariates and treatment indicator improves results in designs with few clusters and subjects.

Our findings justify the CRSE estimator from a finite population perspective (even though it estimates a super-population ATE parameter); this contribution follows similar literature for the individual randomized case (see, e.g., Freedman 2008; Lin 2013). However, while the structure of the design-based and standard CRSE variance estimators are similar, differences in their degrees of freedom adjustments do affect their statistical performance in small samples (the standard CRSE estimator overrejects in this case). The key difference is simple: the randomization mechanism leads to separate degrees of freedom adjustments for the treatment and control groups based on their respective numbers of clusters (in each block), whereas the CRSE approach often used in practice applies a single adjustment based on the total number of clusters. These differences tend to increase with more blocks.

As discussed in the article, other corrections for the CRSE estimator have been proposed that can improve the Type 1 error inflation rate in general settings. In the RCT setting, however, the advantage of the design-based variance estimator is that it is tailored to experiments, as it is derived directly from principles underlying them. Further, it is simple to apply and parallels design-based estimators for non-clustered RCTs. The free *RCT-YES* software (www.rct-yes.com), funded by the U.S. Department of Education, estimates ATEs for full sample and baseline subgroup analyses using the design-based methods discussed in this article using either R or Stata, and also allows for multi-armed trials with multiple treatment conditions.

Acknowledgments

The authors thank Charles Tilley, Joel Middleton, and session participants at the Society for Research on Education Effectiveness conference for useful feedback on the manuscript. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Funding

Nicole Pashley, PhD, was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303 while working on this paper.

Supplementary Materials

The supplementary materials provide proofs of the asymptotic results in Theorems 1 and 2. They also provide results on the efficiency of regression-adjusted ATE estimators using the individual and aggregate data in large and finite samples, as well as details comparing the design-based and CRSE estimators. Finally, the supplement provides details on the simulation methods and presents the full set of simulation results.

References

Abadie A., Athey, S., Imbens, G., and Wooldridge, J. (2017), "When Should You Adjust Standard Errors for Clustering?" arxiv: 1710.02926[Math.ST] [2]

Angrist, J. D., and Lavy, V. (2002), "The Effect of High School Matriculation Awards: Evidence from Randomized Trials," *American Economic Review*, 99, 1384–1414. [6]

Angrist, J., and Pischke, S. (2009), *Mostly Harmless Econometrics*, Princeton NJ: Princeton University Press. [8]

Aronow, P. M., Green, D. P., and Lee, D. K. K. (2014), "Sharp Bounds on the Variance in Randomized Experiments," *Annals of Statistics*, 42, 850–871. [6]

Aronow, P. M., and Middleton, J. A. (2013), "A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments," *Journal of Causal Inference*, 1, 135–154.

Bell, R., and McCaffrey, D. (2002), "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology*, 28, 169–181. [6]

Bickel, B. J., and Freedman, D. A. (1984), "Asymptotic Normality and the Bootstrap in Stratified Sampling," *The Annals of Statistics*, 12(2), 470–482. [2]

Bland, J. M. (2004), "Cluster Randomised Trials in the Medical Literature: Two Bibliometric Surveys," *BMC Medical Research Methodology*, 4, 21. [1]

Bloniarz A., Liu H., Zhang C, Sekhon J. S., and Yu B. (2016), "Lasso Adjustments of Treatment Effect Estimates in Randomized Experiments," *Proceedings of the National Academy of Sciences,* 113, 7383–7390. [9]

Cameron, A. C., Gelbach, J. G., and Miller, D. L. (2008), "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, 90, 414–427. [6,8]

Cameron, A. C., and Miller, D. L. (2015), "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 50, 317–372. [6]

Cochran, W. (1977), *Sampling Techniques*, New York: Wiley. [2]

Ding, P., Feller, A., & Miratrix, L. (2018), "Decomposing Treatment Effect Variation," *Journal of the American Statistical Association*, 114, 304–317.

Donald, S. G., and Lang, K. L. (2007), "Inference with Difference-in-Differences and Other Panel Data," *Review of Economics and Statistics*, 89, 221–233. [6]

Freedman, D. (2008), "On Regression Adjustments to Experimental Data," *Advances in Applied Mathematics* 40, 180–193. [1,3,11]

Fuller, W. A. (1975), "Regression Analysis for Sample Survey," *Sankya*, 37, Series C, 117–132. [2]

——— (2009), *Sampling Statistics*, Hoboken, NJ: Wiley. [2]

Green, D.P., and Vavrek, L. (2008), "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches," *Political Analysis*, 16, 138–152. [8]

Hansen, C. B. (2007), "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T Is Large," *Journal of Econometrics*, 141, 597–620. [6]

Hansen, B. B., and Bowers, J. (2009), "Attributing Effects to a Cluster-Randomized Get-Out-the-Vote Campaign," *Journal of the American Statistical Association*, 104, 873–885. [2]

Hedges, L. (2007), "Correcting a Significance Test for Clustering," *Journal of Educational and Behavioral Statistics*, 32, 151–179. [6]

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960. [1]

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Procedures of the Fifth Berkeley Symposium on Math and Statistical Probability*, 1, 221–233. [6]

Imai, K., King, G., and Nall, C. (2009), "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with an Application to the Mexican Universal Health Insurance Evaluation," *Statistical Science*, 24, 29–53. [2]

Imbens G. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4–29.

Imbens, G., and Rubin, D. (2015), Causal Inference for Statistics, *Social, and Biomedical Sciences: An Introduction*, Cambridge, UK: Cambridge University Press. [1,2]

Kish, L. (1995), *Survey Sampling*, New York: Wiley.

Li, X., and Ding, P. (2017), "General Forms of Finite Population Central Limit Theorems With Applications to Causal Inference," *Journal of the American Statistical Association*, 112, 1759–1769. [1,2,4,5]

Liang, K., and Zeger, S. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22. [2,6]

Lin, W. (2013), "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique," *Annals of Applied Statistics*, 7, 295–318. [1,3,11]

Liu, H., and Yang, Y. (2020), "Regression-Adjusted Average Treatment Effect Estimates in Stratified Randomized Experiments, *Biometrika*, asaa038, *https://doi.org/10.1093/biomet/asaa038*. [2,5]

Lohr, S. L. (2009), *Sampling: Design and Analysis*, 2nd ed. Pacific Grove, CA: Duxbury Press. [2]

Mackinnon, J. G., and White, H. (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics* 29, 305–225. [6]

Middleton, J. A., and Aronow, P. M. (2015), "Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments," *Statistics, Politics and Policy*, 6, 39–75. [1,2]

Middleton, J. A. (2018), "A Unified Theory of Regression Adjustment for Design-based Inference," available at *https://arxiv.org/abs/1803.06011*. [3]

Miratrix, L.W., Sekhon, J. B., and Yu, B. (2013), "Adjusting Treatment Effect Estimates in Randomized Experiments," *Journal of the Royal Statistical Society*, Series B, 75, 369–396. [1]

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9," *Statistical Science*, 5, 465–472. [1]

Pashley, N. E., and Miratrix, L. W. (2020), "Insights on Variance Estimation for Blocked and Matched Pairs Designs," *Journal of Educational and Behavioral Statistics*, Online First. [2]

Pustejovsky, J. E., and Tipton, E. (2018), "Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models," *Journal of Business and Economic Statistics*, 36, 672–683. [6]

Rao, J. N. K., and Shao, J. (1999), "Modified Balanced Repeated Replication for Complex Survey Data," *Biometrika*, 86, 403–415. [2]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Education Psychology*, 66, 688–701. [1]

Rubin, D. B. (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Education Statistics*, 2, 1–26. [1]

Rubin, D. B. (1986), "Which Ifs Have Causal Answers?" Discussion of Holland's "Statistics and Causal inference," *Journal of the American Statistical Association*, 81, 961-962. [2]

SACD Research Consortium (2010). *Efficacy of Schoolwide Programs to Promote Social and Character Development and Reduce Problem Behavior in Elementary School Children*, Washington, DC: Final Report: Institute for Education Sciences, U.S. Department of Education. *https://ies.ed.gov/ncer/pubs/20112001/* [10]

Samii, C. and Aronow, P. M. (2012), "On Equivalencies Between Design-Based and Regression-Based Variance Estimators for Randomized Experiments," *Statistics & Probability Letters*, 82, 365–370. [2]

Schochet P. Z. (2020a), "Analyzing Grouped Administrative Data for RCTs Using Design-Based Methods," *Journal of Educational and Behavioral Statistics*, 45: 32–57.

——— (2008), "Statistical Power for Random Assignment Evaluations of Education Programs," *Journal of Educational and Behavioral Statistics*, 33, 62–87. [1]

——— (2010), "Is Regression Adjustment Supported by the Neyman Model for Causal Inference?" *Journal of Statistical Planning and inference*, 140, 246–259. [1]

——— (2013), "Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference," *Journal of Educational and Behavioral Statistics*, 38, 219–238. [2]

——— (2020b), "A Lasso-OLS Hybrid Approach to Covariate Selection and Average Treatment Effect Estimation for Clustered RCTs Using Design-Based Methods." Available at *https://arxiv.org/abs/2005.02502*, Under journal review. [9]

——— (2016), *Statistical Theory for the RCT-YES Software: Design-Based Causal inference for RCTs* (NCEE 2015–4011), 2nd ed. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. *https://ies.id.gov/ncee/pubs/20154011/pdf/20154011.pdf*. [1,3,7]

Scott, A., and Wu, C. F. (1981), "On the Asymptotic Distribution of Ratio and Regression Estimators," *Journal of the American Statistical Association*, 112, 1759-1769. [2,4]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [9]

Webb, M. D. (2013), "Reworking Wild Bootstrap Based Inference for Clustered Errors," Queens Economics Department Working Paper 1315. [6]

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838. [6]

Wolter, K.M. (2007), *Introduction to Variance Estimation*, 2nd edition, Springer Verlag. [2]

Wooldridge, J. M. (2006), *Cluster Sample Methods in Applied Econometrics*, Working Paper. Michigan State University. [6]

Yang, L. and Tsiatis, A. (2001), "Efficiency Study of Estimators for a Treatment Effect in a Pretest–Posttest Trial," *The American Statistician*, 55, 314–321. [1]