Inference for Cluster Randomized Experiments with Nonignorable Cluster Sizes

Federico Bugni

Northwestern University

Ivan A. Canay

Northwestern University

Azeem M. Shaikh

University of Chicago

Max Tabord-Meehan

University of Chicago

We consider the problem of inference in randomized experiments where treatment is assigned at the level of a cluster and cluster sizes are nonignorable, in that cluster-level average treatment effects may depend on the cluster sizes. In a novel superpopulation framework in which cluster sizes are modeled as random and allowing for only a

We thank the coeditor and three anonymous referees for comments and suggestions that have improved the manuscript. We also would like to thank Eric Auerbach, David MacKinnon, Joe Romano, and conference participants at Cowles, the Toulouse School of Economics, and the 2023 annual meeting of the Allied Social Science Associations for helpful comments on this paper. Xun Huang and Juri Trifonov provided excellent research assistance. The third author acknowledges support from the National Science Foundation through grant SES-2419008. The fourth author acknowledges support from the National Science Foundation through grant SES-2149408.

Electronically published April 28, 2025

Journal of Political Economy Microeconomics, volume 3, number 2, May 2025.

https://doi.org/10.1086/732836

subset of the units within each cluster to be sampled, we distinguish between two parameters of interest that differ in how they average the treatment effect across units. For each parameter, we provide methods for inference when treatment is assigned using a covariate-adaptive stratified randomization procedure.

I. Introduction

Cluster randomized experiments, in which treatment is assigned at the level of the cluster rather than at the level of the unit within a cluster, are widely used throughout economics and the social sciences more generally for the purpose of evaluating treatments or programs. Duflo, Glennerster, and Kremer (2007) survey various examples from development economics, in which clusters are villages and units within a cluster are households or individuals. Numerous other examples can be found in, for instance, research on the effectiveness of educational interventions (see, e.g., Raudenbush 1997; Schochet 2013; Raudenbush and Schwartz 2020; Schochet et al. 2021) and research on the effectiveness of public health interventions (see, e.g., Donner and Klar 2000; Turner et al. 2017). In this paper, we consider the problem of inference about the effect of a binary treatment on an outcome of interest in such experiments in a superpopulation framework in which cluster sizes are permitted to be random and nonignorable. By nonignorable cluster sizes, we refer to the possibility that the treatment effects may depend on the cluster size.

Before proceeding, we illustrate this possibility with an example inspired by our empirical application in section V. Suppose clusters represent public primary care facilities, or clinics, in a state. The size of the clinic may be related to the quality of care through a variety of mechanisms. For instance, it is plausible in our empirical application that patients may experience longer waiting times in smaller clinics because these may be comparatively understaffed relative to their patient populations. For this reason, patients may be less sensitive to incentives to go to the clinic more frequently. Similar considerations may, of course, apply in most other examples; we therefore view cluster sizes being nonignorable as the rule rather than the exception.

To model this phenomenon, we adopt, in the spirit of the survey sampling literature (see, e.g., Lohr 2021), a two-stage sampling framework, in which a set of clusters is first sampled from the population of clusters and then a set of units is sampled from the population of units within each cluster. Importantly, in the first stage of the sampling process, each cluster may differ in terms of observed characteristics, including its size, and these characteristics may be used subsequently in the second stage of

the sampling process to determine the number of units to sample from the cluster, including the possibility that all units in the cluster are sampled. We further emphasize that the sampling framework imposes no restrictions on the dependence across units within clusters. Our two-stage superpopulation sampling framework departs from earlier analyses of cluster randomized experiments in which cluster sizes were treated as nonrandom, and subsequently allows us to cohesively consider a large class of stratified treatment assignment mechanisms that can incorporate information on the cluster sizes along with other baseline covariates.

In the context of this framework, we revisit two different parameters of interest previously considered in the literature on cluster randomized experiments (see, e.g., Athey and Imbens 2017; Su and Ding 2021; Wang et al. 2022; Kahan et al. 2023) that differ in the way they aggregate, or average, the treatment effect across units. They differ, in particular, according to whether the units of interest are the clusters themselves or the individuals within the cluster. The first of these parameters takes the clusters themselves as the units of interest and identifies an *equally weighted* cluster-level average treatment effect. The second of these parameters takes the individuals within the clusters as the units of interest and identifies a *size-weighted* cluster-level average treatment effect. When individual-level average treatment effects vary with cluster size (i.e., cluster size is nonignorable) and cluster sizes are heterogeneous, these two parameters are generally different, though, as discussed in remark 3, they coincide in some instances. Importantly, we show that the estimand associated with the standard difference-in-means estimator is a *sample-weighted* cluster-level average treatment effect, which cannot generally be interpreted as an average treatment effect for either the clusters themselves or the individuals within the clusters. We show, however, in section II.B that this estimand can equal the size-weighted or the equally weighted cluster-level average treatment effect for some very specific sampling designs. We argue that a clear description of whether the clusters themselves or the individuals within the clusters are of interest should therefore be at the forefront of empirical practice, yet we find that such a description is often absent. Indeed, we surveyed all articles involving a cluster randomized experiment published in *American Economic Journal: Applied Economics* from 2018 to 2022. We document our findings in appendix section A.3 (appendix is available online). From this survey, we find that most papers do not explicitly discuss their parameter of interest, and that as many as a third of the experiments conduct analyses that, when paired with their corresponding sampling design, do not necessarily recover either of the parameters that we consider in this paper.

For each of the two parameters of interest that we consider, we propose an estimator and develop the requisite distributional approximations to permit its use for inference about the parameter of interest when treatment

is assigned using a covariate-adaptive stratified randomization procedure. In the case of the equally weighted cluster-level average treatment effect, the estimator we propose takes the form of a difference-in-means of cluster averages. This estimator may equivalently be described as the ordinary least squares estimator of the coefficient on treatment in a regression of the average outcome (within clusters) on a constant and treatment. In the case of the size-weighted cluster-level average treatment effect, the estimator we propose takes the form of a weighted difference-in-means of cluster averages, where the weights are proportional to cluster size. This estimator may equivalently be described as the weighted least squares estimator of the coefficient on treatment in a regression of the individual-level outcomes on a constant and treatment with weights proportional to cluster size.[1]

Although both estimators that we propose have previously been studied in the context of completely randomized experiments (see Athey and Imbens 2017; Su and Ding 2021), to our knowledge we are the first to establish results for these estimators when treatment assignment is performed using a stratified covariate-adaptive randomization procedure. As in Bugni, Canay, and Shaikh (2018, 2019), this refers to randomization schemes that first stratify according to baseline covariates and then assign treatment status so as to achieve "balance" within each stratum (see Rosenberger and Lachin 2016 for a textbook treatment focused on clinical trials and Duflo, Glennerster, and Kremer 2007 and Bruhn and McKenzie 2008 for reviews focused on development economics). Our results show that typical hypothesis tests constructed using a cluster-robust variance estimator are generally conservative in such cases, and as a result, we provide a simple adjustment to the standard errors that delivers asymptotically exact tests. In this sense, our inference results generalize those of Bugni, Canay, and Shaikh (2018) for individual-level randomized experiments to settings with cluster-level randomization.

By virtue of its sampling framework, our paper is distinct from a closely related and complementary literature that has analyzed cluster randomized experiments from a finite-population perspective. Important contributions to this literature include Middleton and Aronow (2015), Athey and Imbens (2017), Hayes and Moulton (2017), de Chaisemartin and Ramirez-Cuellar (2020), Schochet et al. (2021), Su and Ding (2021), and Abadie et al. (2023). The primary source of uncertainty in this literature is "design-based" uncertainty stemming from the randomness in treatment assignment, though parts of the literature additionally permit

[1] In app. sec. A.4, we also briefly consider versions of both estimators that allow for linear regression adjustment using additional baseline covariates.

up to two additional sources of uncertainty: the randomness from sampling clusters from a finite population of clusters and the randomness from sampling only a subset of the finite number of units in each cluster. In the context of such a sampling framework, the literature has defined finite-population counterparts to both our equally weighted and size-weighted cluster-level average treatment effects. See, for instance, Athey and Imbens (2017, chap. 8) and Su and Ding (2021, sec. 4). In particular, Su and Ding (2021) provide estimators and methods for inference about each quantity when treatment is assigned completely at random (excluding, as a result, stratified covariate-adaptive treatment assignments). Our contribution may thus be viewed as leveraging our novel superpopulation sampling framework to develop results for general stratified covariate-adaptive randomization procedures; we discuss further comparisons between these approaches in remark 7. With this in mind, our results may be especially relevant for the analysis of data that are sampled from a well-defined larger population; Muralidharan and Niehaus (2017) argue that at least 30% of the experiments they examined in economics satisfied this criterion.

Our paper is also related to a large literature on the analysis of clustered data (not necessarily from experiments) in econometrics and statistics. Prominent contributions to this literature include Liang and Zeger (1986), Hansen (2007), Djogbenou, MacKinnon, and Nielsen (2019), and Hansen and Lee (2019. Additional references can be found in the surveys Cameron and Miller (2015) and MacKinnon, Nielsen, and Webb (2023). These papers are designed as methods for inference for parameters defined via linear models or estimating equations rather than parameters such as our equally weighted or size-weighed cluster-level average treatment effects that are defined explicitly in terms of potential outcomes. Importantly, in almost all of these papers, the sampling framework treats cluster sizes as nonrandom, though we note that in some cases the results are rich enough to permit the distribution of the data to vary across clusters; further discussion is provided in remark 2. Finally, none of these papers seem to explicitly consider the additional complications stemming from sampling only a subset of the units within each cluster.

The remainder of our paper is organized as follows. Section II describes our setup and notation, including a formal description of our sampling framework and two parameters of interest. We then propose in section III estimators for each of these two quantities and develop the requisite distributional approximations to use them for inference about each quantity. In section IV, we demonstrate the finite-sample behavior of our proposed estimators in a small simulation study. Finally, in section V we conduct an empirical exercise to demonstrate the practical relevance of our findings. Proofs of all results are included in the appendix.

II. Setup and Notation

A. *Notation and Sampling Framework*

Let $Y_{i,g}$ denote the (observed) outcome of the ith unit in the gth cluster, A_g denote an indicator for whether the gth cluster is treated or not, Z_g denote observed baseline covariates for the gth cluster, and N_g the size of the gth cluster. Further denote by $Y_{i,g}(1)$ the potential outcome of the ith unit in the gth cluster if treated and by $Y_{i,g}(0)$ the potential outcome of the ith unit in the gth cluster if not treated. As usual, the (observed) outcome and potential outcomes are related to treatment assignment by the relationship

$$Y_{i,g} = Y_{i,g}(1)A_g + Y_{i,g}(0)(1 - A_g). \tag{1}$$

We model the distribution of the data described above in two parts: a superpopulation sampling framework for the clusters and an assignment mechanism that assigns the clusters to treatments. The sampling framework itself can be described in two stages. In the first stage, an i.i.d. sample of G clusters is drawn from a distribution of clusters. In the second stage, a subset of the individual units within each cluster is sampled. A key feature of this framework is that the cluster size N_g is modeled as a random variable in the same way as other cluster characteristics Z_g. While the clusters are (ex ante) identically distributed, we note that they may exhibit heterogeneity in terms of their (ex post) realizations of N_g and Z_g. The second sampling stage allows for settings in which the analyst does not observe all of the units within a cluster. Define $\mathcal{M}_g$ to be the subset of $\{1, \ldots, N_g\}$ corresponding to the observations within the gth cluster that are sampled by the researcher. We emphasize that a realization of $\mathcal{M}_g$ is a set whose cardinality we denote by $|\mathcal{M}_g|$, whereas a realization of N_g is a positive integer. For example, in the event that all observations in a cluster are sampled, $\mathcal{M}_g = \{1, \ldots, N_g\}$ and $|\mathcal{M}_g| = N_g$. Once the sample of clusters is realized, the experiment assigns treatments $A^{(G)} := (A_g : 1 \leq g \leq G)$ using an assignment rule that stratifies according to baseline covariates Z_g and cluster sizes N_g. Formally, denote by P_G the distribution of the observed data

$$(((Y_{i,g} : i \in \mathcal{M}_g), A_g, Z_g, N_g) : 1 \leq g \leq G)$$

that arises from sampling and treatment assignment, and by Q_G the distribution of

$$W^{(G)} := (((Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g), \mathcal{M}_g, Z_g, N_g) : 1 \leq g \leq G).$$

Note that the observed distribution P_G is determined jointly by (1) together with the distribution of $A^{(G)}$ and Q_G, so we will state our assumptions below in terms of these two quantities.

We begin by describing our assumptions on the distribution of $A^{(G)}$. Strata are constructed from the observed, baseline covariates Z_g and cluster sizes N_g using a function $S : \text{supp}((Z_g, N_g)) \to \mathcal{S}$, where $\mathcal{S}$ is a finite set. For $1 \leq g \leq G$, let $S_g = S(Z_g, N_g)$ and denote by $S^{(G)}$ the vector of strata $(S_1, S_2, \ldots, S_G)$. In what follows, we rule out trivial strata by assuming that $p(s) := P\{S_g = s\} > 0$ for all $s \in \mathcal{S}$. For $s \in \mathcal{S}$, let

$$D_G(s) := \sum_{1 \leq g \leq G} (I\{A_g = 1\} - \pi) I\{S_g = s\}, \tag{2}$$

where $\pi \in (0, 1)$ is the "target" proportion of clusters to assign to treatment in each stratum. Note that $D_G(s)$ measures the amount of imbalance in stratum s relative to the target proportion π. Our requirements on the treatment assignment mechanism are then summarized as follows:

ASSUMPTION 1. The treatment assignment mechanism is such that

a. $W^{(G)} \perp\!\!\!\perp A^{(G)} | S^{(G)}$;
b. $\{\{D_G(s)/\sqrt{G}\}_{s \in \mathcal{S}} | S^{(G)}\} \xrightarrow{d} N(0, \Sigma_D)$ almost surely, where

$$\Sigma_D = \text{diag}\{p(s)\tau(s) : s \in \mathcal{S}\}$$

with $0 \leq \tau(s) \leq \pi(1 - \pi)$ for all $s \in \mathcal{S}$.

Assumption 1 mirrors the assumption on assignment mechanisms considered in Bugni, Canay, and Shaikh (2018) for individual-level randomized experiments. Assumption 1a requires that the assignment mechanism be a function only of the strata and an exogenous randomization device. Assumption 1b requires that the randomization mechanism assign treatments within each stratum so that the fraction of units being treated has a well-behaved limiting distribution centered around the target proportion π. For each stratum $s \in \mathcal{S}$, the parameter $\tau(s) \in [0, 1]$ determines the amount of dispersion that the treatment assignment mechanism allows on the fraction of units assigned to the treatment in that stratum. A lower value of $\tau(s)$ implies that the treatment assignment mechanism imposes a higher degree of "balance" or "control" of the treatment assignment proportion relative to its desired target value. Bugni, Canay, and Shaikh (2018) provide several important examples of assignment mechanisms satisfying this assumption that are used routinely in economics. In particular, assumption 1 is satisfied by stratified block randomization (see, e.g., Angelucci, Karlan, and Zinman 2015; Attanasio et al. 2015; Duflo, Dupas, and Kremer 2015), which assigns exactly a fraction π of units within each stratum to treatment, at random. When $\tau(s) = 0$ (as is the case for stratified block randomization) for all $s \in \mathcal{S}$, we say that the assignment mechanism achieves "strong balance." Note further that the assumption also applies in settings without stratification, in which case $|\mathcal{S}| = 1$. Despite its broad applicability, assumption 1 nevertheless precludes treatment

assignment mechanisms with many "small" strata, by the virtue of assuming that $\mathcal{S}$ is a fixed finite set. This precludes, for instance, "matched pairs" designs (see, e.g., Banerjee et al. 2015; Crépon et al. 2015), the analysis of which can be found in the companion paper Bai et al. (2022).

We now describe our assumptions on Q_G. In order to do so, it is useful to introduce some further notation. To this end, define $R_G(\mathcal{M}^{(G)}, Z^{(G)}, N^{(G)})$ to be the distribution of

$$((Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g) : 1 \leq g \leq G) | \mathcal{M}^{(G)}, Z^{(G)}, N^{(G)},$$

where $\mathcal{M}^{(G)} \coloneqq (\mathcal{M}_g : 1 \leq g \leq G)$, $Z^{(G)} \coloneqq (Z_g : 1 \leq g \leq G)$, and $N^{(G)} \coloneqq (N_g : 1 \leq g \leq G)$. Note that Q_G is completely determined by $R_G(\mathcal{M}^{(G)}, Z^{(G)}, N^{(G)})$ and the distribution of $(\mathcal{M}^{(G)}, Z^{(G)}, N^{(G)})$. While N_g obviously determines the length of $(Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g)$, we emphasize that it may affect the location and shape of its distribution as well. In this paper, we say cluster sizes are *ignorable* whenever the individual-level average treatment effect does not depend on the cluster size, in the sense that

$$\begin{aligned} P\{E[Y_{i,g}(1) - Y_{i,g}(0) | N_g] = E[Y_{i,g}(1) - Y_{i,g}(0)] \text{ for all } 1 \leq i \leq N_g\} \\ = 1 \text{ for all } 1 \leq g \leq G. \end{aligned} \tag{3}$$

Consequently, we say that cluster sizes are *nonignorable* whenever (3) fails. Example 1 provides a simple illustration of nonignorability, and remark 3 provides some related discussion of the consequences of assuming that clusters are, in fact, ignorable. Finally, for $a \in \{0, 1\}$, define

$$\bar{Y}_g(a) \coloneqq \frac{1}{|\mathcal{M}_g|} \sum_{i \in \mathcal{M}_g} Y_{i,g}(a).$$

The following assumption states our requirements on Q_G using this notation.

ASSUMPTION 2. The distribution Q_G is such that

a. $\{(\mathcal{M}_g, Z_g, N_g), 1 \leq g \leq G\}$ is an i.i.d. sequence of random variables.
b. The distribution of $((Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g) : 1 \leq g \leq G) | \mathcal{M}^{(G)}, Z^{(G)}, N^{(G)}$ can be factored as

$$R_G(\mathcal{M}^{(G)}, Z^{(G)}, N^{(G)}) = \prod_{1 \leq g \leq G} R(\mathcal{M}_g, Z_g, N_g),$$

where $R(\mathcal{M}_g, Z_g, N_g)$ denotes the distribution of $(Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g)$ conditional on $(\mathcal{M}_g, Z_g, N_g)$.
c. $\mathcal{M}_g \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g) | Z_g, N_g$ for all $1 \leq g \leq G$, that is, $R(\mathcal{M}_g, Z_g, N_g) = R(Z_g, N_g)$.

d. For $a \in \{0, 1\}$ and $1 \leq g \leq G$,

$$E[\bar{Y}_g(a)|N_g] = E\left[\frac{1}{N_g}\sum_{1\leq i\leq N_g} Y_{i,g}(a)|N_g\right] \text{ almost surely.}$$

e. $P\{|\mathcal{M}_g| \geq 1\} = 1$ and $E[N_g^2] < \infty$.
f. For some $C < \infty$, $P\{E[Y_{i,g}^2(a)|N_g, Z_g] \leq C \text{ for all } 1 \leq i \leq N_g\} = 1$ for all $a \in \{0, 1\}$ and $1 \leq g \leq G$.

Assumptions 2a and 2b formalize the idea that our data consist of an i.i.d. sample of clusters, where the cluster sizes are themselves random and possibly related to potential outcomes. An important implication of these two assumptions for our purposes is that

$$\left\{(\bar{Y}_g(1), \bar{Y}_g(0), |\mathcal{M}_g|, Z_g, N_g), 1 \leq g \leq G\right\} \tag{4}$$

is an i.i.d. sequence of random variables, as established by lemma A.1 in the appendix.

Assumptions 2c and 2d impose high-level restrictions on the second stage of the sampling framework. Assumption 2c allows the subset of observations sampled by the experimenter to depend on Z_g and N_g, but rules out dependence on the potential outcomes within the cluster itself. Assumption 2d is a high-level assumption that guarantees that we can extrapolate from the observations that are sampled to the observations that are not sampled. Note that assumptions 2c and 2d are trivially satisfied whenever $\mathcal{M}_g = \{1, \ldots, N_g\}$ for all $1 \leq g \leq G$ with probability 1, that is, whenever all observations within each cluster are always sampled. Assumption 2d is also satisfied whenever assumption 2c holds and there is sufficient homogeneity across the observations within each cluster in the sense that $P\{E[Y_{i,g}(a)|N_g, Z_g] = E[Y_{j,g}(a)|N_g, Z_g] \text{ for all } 1 \leq i, j \leq N_g\} = 1$ for $a \in \{0, 1\}$. Finally, we show in lemma 1 below that if $\mathcal{M}_g$ is drawn as a random sample without replacement from $\{1, 2, \ldots, N_g\}$ in an appropriate sense, then assumptions 2c and 2d are also satisfied.

Assumptions 2e and 2f impose some mild regularity on the (conditional) moments of the distribution of cluster sizes and potential outcomes, in order to permit the application of relevant laws of large numbers and central limit theorems. Note that assumption 2e does not rule out the possibility of observing arbitrarily large clusters but does place restrictions on the frequency of extremely large realizations. For instance, two consequences of assumptions 2a and 2e are that[2]

[2] The first is an immediate consequence of the law of large numbers and the continuous mapping theorem. The second follows from lemma S.1.1 in Bai, Romano, and Shaikh (2021).

$$\frac{\sum_{1\le g\le G} N_g^2}{\sum_{1\le g\le G} N_g} = O_P(1)$$

and

$$\frac{\max_{1\le g\le G} N_g^2}{\sum_{1\le g\le G} N_g} \xrightarrow{P} 0,$$

which mirror heterogeneity restrictions imposed in earlier work on the analysis of clustered data when cluster sizes are modeled as nonrandom (see, e.g., assumption 2 in Hansen and Lee 2019), but are modestly stronger than the heterogeneity restrictions considered in recent work studying cluster randomized experiments from a finite-population perspective (see, e.g., theorem 1 in Su and Ding 2021). We use assumption 2e extensively when establishing asymptotic normality in theorem 5; recent work by Sasaki and Wang (2022) and Chiang, Sasaki, and Wang (2023), however, suggests that one may be able to sometimes obtain asymptotic normality even when $E[N_g^2] = \infty$, provided that certain delicate conditions about the tail behavior of N_g are satisfied.

LEMMA 1. Suppose that $\mathcal{M}_g \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \le i \le N_g) | Z_g, N_g$ for all $1 \le g \le G$, and that, conditionally on $(Z_g, N_g, |\mathcal{M}_g|)$, $\mathcal{M}_g$ is drawn uniformly at random from all possible subsets of size $|\mathcal{M}_g|$ from $\{1, 2, \ldots, N_g\}$. Then, assumptions 2c and 2d are satisfied.

REMARK 1. We could in principle modify our framework so that the distribution of cluster sizes is allowed to depend on the number of clusters G. By doing so, we would be able to weaken assumption 2e at the cost of strengthening assumption 2f to require, for example, uniformly bounded $2 + \delta$ moments for some $\delta > 0$. Such a modification, however, would complicate the exposition and the resulting procedures would ultimately be the same. We therefore see no apparent benefit and do not pursue it further in this paper.

REMARK 2. An attractive feature of our framework is that, by virtue of modeling cluster sizes as random, it is straightforward to permit dependence between the cluster size and other features of the cluster, such as the distribution of potential outcomes within the cluster. In this way, our setting departs from other frameworks in the literature on clustered data in which the cluster sizes are treated as deterministic; see, for example, Hansen and Lee (2019). We note, however, that the results in this literature permit the distribution of the data across clusters to be nonidentically distributed, and in fact the literature has noted that the method described in Liang and Zeger (1986) may fail when cluster sizes are nonignorable; see, in particular, Benhin, Rao, and Scott (2005). Although it is possible that these related papers could be applied to our framework by first conditioning on the cluster sizes, results obtained in this way would

necessarily hold conditionally on the cluster sizes, whereas our results are unconditional.

B. *Parameters of Interest*

In settings with cluster data, there are multiple ways to aggregate, or average, the heterogeneous treatment effect $Y_{i,g}(1) - Y_{i,g}(0)$. In particular, an important consideration is whether the units of interest are the clusters themselves or the individuals within the cluster. This distinction is precisely what motivates the two parameters of interest that we study in this paper. Before introducing these parameters, we present a simple example that will help us illustrate the differences.

EXAMPLE 1. We revisit the example described in the introduction, where clusters represent primary care facilities, or clinics, within a state, as in the empirical application in section V. Let us suppose there are two types of clinics potentially exposed to a treatment: "big" clinics with $N_g = 40$ regular patients, and "small" clinics with $N_g = 10$ regular patients. Suppose further that cluster size is nonignorable in that $Y_{i,g}(1) - Y_{i,g}(0) = 1$ for all patients in a "big" clinic, $Y_{i,g}(1) - Y_{i,g}(0) = -2$ for all patients in a "small" clinic, and that both types of clinics are equally likely; that is,

$$P\{N_g = 40\} = P\{N_g = 10\} = 1/2.$$

Policymakers evaluating the adoption of the treatment may arrive at different conclusions depending on their objective. For instance, if clinics are considered as the unit of analysis, a policymaker may perceive the treatment as harmful, as half of the clinics experience a positive treatment effect of 1, and the other half experience a negative treatment effect of -2, resulting in an "average" treatment effect of $-1/2$. Conversely, if individuals (patients) within the state are considered as the unit of analysis, the treatment may appear beneficial. In this case, $4/5$ of the patients experience a positive treatment effect of 1, while only $1/5$ experience a negative treatment effect of -2, resulting in an "average" treatment effect of $2/5$. The distinction between these two different ways of averaging motivates the formal definitions of the parameters that we consider below.

Example 1 illustrates that characterizing an "average" treatment effect in settings with cluster data depends on whether clusters or individuals are the focus of the analysis. This distinction, as illustrated in the example, is relevant whenever clusters feature cluster size heterogeneity and average treatment effect heterogeneity with respect to cluster size (even if the treatment effect is assumed to be homogeneous within clusters as in example 1). When there is average treatment effect heterogeneity

within and across clusters, the question becomes whether the aggregated information in the clusters, that is, $(1/N_g)\Sigma_{1\le i\le N_g} Y_{i,g}(a)$, should be weighted by the size of the cluster or not.

Motivated by example 1, we consider two different parameters of interest: one that considers the clusters as the units of interest, and one that considers the individuals as the units of interest. Both of these parameters can be written in the form

$$E\left[\omega_g\left(\frac{1}{N_g}\sum_{1\le i\le N_g}\left[Y_{i,g}(1) - Y_{i,g}(0)\right]\right)\right] \tag{5}$$

for different choices of (possibly random) weights ω_g, $1 \le g \le G$, satisfying $E[\omega_g] = 1$. The first parameter of interest corresponds to the choice of $\omega_g = 1$, thus weighting the average effect of the treatment across clusters equally:

$$\theta_1(Q_G) := E\left[\frac{1}{N_g}\sum_{1\le i\le N_g}\left[Y_{i,g}(1) - Y_{i,g}(0)\right]\right]. \tag{6}$$

We refer to this quantity as the equally weighted cluster-level average treatment effect. Here $\theta_1(Q_G)$ can be thought of as the average treatment effect in which the clusters themselves are the units of interest. The second parameter of interest corresponds to the choice of $\omega_g = N_g/E[N_g]$, thus weighting the average effect of the treatment across clusters in proportion to their size:

$$\theta_2(Q_G) := E\left[\frac{1}{E[N_g]}\sum_{1\le i\le N_g}\left[Y_{i,g}(1) - Y_{i,g}(0)\right]\right]. \tag{7}$$

We refer to this quantity as the size-weighted cluster-level average treatment effect. The term $\theta_2(Q_G)$ can be thought of as the average treatment effect in which individuals are the units of interest. Note that assumptions 2a and 2b imply that we may express both $\theta_1(Q_G)$ and $\theta_2(Q_G)$ as a function of R and the common distribution of $(\mathcal{M}_g, Z_g, N_g)$. In particular, neither quantity depends on g or G and so in what follows we simply denote $\theta_1 = \theta_1(Q_G)$, $\theta_2 = \theta_2(Q_G)$.

If treatment effects are heterogeneous with respect to cluster size and cluster sizes are themselves heterogeneous, then θ_1 and θ_2 are indeed distinct parameters. We illustrate this in the context of example 1.

Example 1 (continued). Recall the setting of example 1. The equally weighted cluster-level average treatment effect simply equals

$$\theta_1 = P\{N_g = 10\}E\left[\frac{1}{N_g}\sum_{1\le i\le N_g}\left[Y_{i,g}(1) - Y_{i,g}(0)\right]|N_g = 10\right]$$
$$+ P\{N_g = 40\}E\left[\frac{1}{N_g}\sum_{1\le i\le N_g}\left[Y_{i,g}(1) - Y_{i,g}(0)\right]|N_g = 40\right]$$
$$= \frac{1}{2}\times -2 + \frac{1}{2}\times 1 = -\frac{1}{2}.$$

The parameter θ_1 captures an average treatment effect in which the clusters are the units of interest since both treatment effects 1 and -2 receive the same weight (both types of clinics are equally likely). The size-weighted cluster-level average treatment effect, in turn, equals

$$\theta_2 = \frac{P\{N_g = 10\}}{E[N_g]}E\left[\sum_{1\le i\le N_g}\left[Y_{i,g}(1) - Y_{i,g}(0)\right]|N_g = 10\right]$$
$$+ \frac{P\{N_g = 40\}}{E[N_g]}E\left[\sum_{1\le i\le N_g}\left[Y_{i,g}(1) - Y_{i,g}(0)\right]|N_g = 40\right]$$
$$= \frac{1/2}{25}\times -20 + \frac{1/2}{25}\times 40 = \frac{2}{5}.$$

The parameter θ_2 captures an average treatment effect in which the individuals are the units of interest since both treatment effects, 1 and -2, are weighted by the proportion of the patients in the state that attend each type of clinic.

REMARK 3. While we generally expect θ_1 and θ_2 to be distinct, they are equivalent in some special cases. For example, if all clusters are of the same fixed size k, that is, $P\{N_g = k\} = 1$, then it follows immediately that $\theta_1 = \theta_2$. Alternatively, if treatment effects are constant, so that $P\{Y_{i,g}(1) - Y_{i,g}(0) = \tau \text{ for all } 1 \le i \le N_g\} = 1$, then $\theta_1 = \theta_2$. Beyond these two cases, we have $\theta_1 = \theta_2$ whenever cluster sizes are ignorable in the sense of (3) and the average treatment effects are homogeneous in the sense that $P\{E[Y_{i,g}(1) - Y_{i,g}(0)] = E[Y_{j,g}(1) - Y_{j,g}(0)] \text{ for all } 1 \le i, j \le N_g\} = 1$. Note that this last statement is not generally true if one replaces (3) with

$$P\left\{\mathrm{E}[\mathrm{Y_{i,g}}(1) - \mathrm{Y_{i,g}}(0) \mid \mathrm{N_g}, \mathrm{X_g}] = \mathrm{E}[\mathrm{Y_{i,g}}(1) - \mathrm{Y_{i,g}}(0)\big|\mathrm{X_g}] \text{ for all } 1 \le \mathrm{i} \le \mathrm{N_g}\right\}$$
$$= 1 \text{ for all } 1 \le \mathrm{g} \le \mathrm{G},$$

where X_g represents a collection of cluster-level characteristics (either observed or unobserved). In this sense, even if N_g is simply a proxy for other characteristics X_g, it still plays an important role in our analysis through the distinction between θ_1 and θ_2.

Not all estimands that arise from commonly used empirical strategies take the form in (5). For example, as we show in theorem 1 in the next section, the usual difference-in-means estimator consistently estimates the following population parameter,

$$\vartheta := E\left[\frac{1}{E[|\mathcal{M}_g|]}\sum_{i\in\mathcal{M}_g}\left[Y_{i,g}(1) - Y_{i,g}(0)\right]\right]. \tag{8}$$

This parameter corresponds to a sample-weighted cluster-level average treatment effect and, without assumptions on the sampling process, does not generally identify either an average treatment effect in which the clusters are the units of interest or an average treatment effect in which the individuals are the units of interest. In other words, when treatment effects are heterogeneous and cluster sizes are nonignorable, ϑ need not equal either θ_1 defined in (6) or θ_2 defined in (7). Two specific sampling designs for which ϑ equals either θ_1 or θ_2 are worth highlighting. First, if $P\{|\mathcal{M}_g| = k\} = 1$ for all $1 \le g \le G$, then ϑ is equal to θ_1. This is intuitive, as in this case, ϑ gives equal weights to each cluster and thus behaves as if the units of interest are the clusters themselves. Second, if $|\mathcal{M}_g|$ is a constant fraction of N_g for all $1 \le g \le G$, that is, $P\{|\mathcal{M}_g| = \gamma N_g\} = 1$ for some $0 < \gamma \le 1$, then ϑ is equal to θ_2. This is also intuitive, as in this case, the relative weights of individuals in the parameter ϑ coincide with the weights they would have obtained if the units of interest were the individuals. In general, as illustrated in the following example, ϑ may not equal either θ_1 or θ_2.

EXAMPLE 1 (continued). Recall the setting of example 1. Suppose further that the experimenter samples $|\mathcal{M}_g| = 5$ patients at random without replacement from each "small" clinic, and $|\mathcal{M}_g| = 10$ patients at random without replacement from each "big" clinic. Importantly, note that this sampling scheme does not maintain the relative proportions of these clinics that exist at the population level. It is now straightforward to show that

$$\begin{aligned}\vartheta &= P\{N_g = 10\}\frac{E[|\mathcal{M}_g| \mid N_g = 10]}{E[|\mathcal{M}_g|]}E\left[\frac{1}{N_g}\sum_{1\le i\le N_g}\left[Y_{i,g}(1) - Y_{i,g}(0)\right]|N_g = 10\right]\\ &\quad + P\{N_g = 40\}\frac{E[|\mathcal{M}_g| \mid N_g = 40]}{E[|\mathcal{M}_g|]}E\left[\frac{1}{N_g}\sum_{1\le i\le N_g}\left[Y_{i,g}(1) - Y_{i,g}(0)\right]|N_g = 40\right]\\ &= \frac{1}{2}\times\frac{5}{15/2}\times -2 + \frac{1}{2}\times\frac{10}{15/2}\times 1 = 0,\end{aligned}$$

where the first equality exploits assumption 2c, so that $|\mathcal{M}_g|$ and $Y_{i,g}(1) - Y_{i,g}(0)$ are independent conditional on N_g, and assumption 2d, so that $(1/|\mathcal{M}_g|)\Sigma_{i\in\mathcal{M}_g}$ can be replaced by $(1/N_g)\Sigma_{1\le i\le N_g}$. We conclude that ϑ is

not equal to either θ_1 or θ_2. Indeed, it is straightforward to show that as we vary the distribution of $|\mathcal{M}_g|$ conditional on N_g, ϑ could take any value between -1.72 and 0.93 in this example, so that it could be smaller, in between, or larger than both $\theta_1 = -1/2$ and $\theta_2 = 2/5$.

III. Main Results

A. *Asymptotic Behavior of the Difference-in-Means Estimator*

Given its central role in the analysis of randomized experiments, we begin this section by studying the asymptotic behavior of the difference-in-means estimator

$$\hat{\theta}_G^{\text{alt}} := \frac{\sum_{1\le g\le G}\sum_{i\in\mathcal{M}_g} Y_{i,g}A_g}{\sum_{1\le g\le G}|\mathcal{M}_g|A_g} - \frac{\sum_{1\le g\le G}\sum_{i\in\mathcal{M}_g} Y_{i,g}(1-A_g)}{\sum_{1\le g\le G}|\mathcal{M}_g|(1-A_g)}. \tag{9}$$

Note that $\hat{\theta}_G^{\text{alt}}$ may be obtained as the estimator of the coefficient on A_g in the following ordinary least squares regression:

$$\text{regress } Y_{i,g} \text{ on constant} + A_g.$$

The following theorem derives the probability limit of this estimator:

THEOREM 1. Under assumptions 1 and 2,

$$\hat{\theta}_G^{\text{alt}} \xrightarrow{p} E\left[\frac{1}{E[|\mathcal{M}_g|]}\sum_{i\in\mathcal{M}_g}\left[Y_{i,g}(1) - Y_{i,g}(0)\right]\right] = \vartheta$$

as $G \to \infty$.

As we discussed in the previous section, the quantity ϑ corresponds to a sample-weighted cluster-level average treatment effect. In general, this parameter will not equal either θ_1 or θ_2 as illustrated in example 1. As a result, unless the experimenter is interested in a distinct weighting of the cluster-level treatment effects that differs from those that arise when the clusters themselves are the units of interest or when the individuals within the clusters are the units of interest, care must be taken when interpreting $\hat{\theta}_G^{\text{alt}}$. While it is true, as we discussed in section II.B, that ϑ is in fact equal to either θ_1 or θ_2 for some specific sampling designs, in what follows we consider alternative estimators that are generally consistent for θ_1 and θ_2 without imposing additional restrictions on the sampling procedure.

B. *Equally Weighted Cluster-Level Average Treatment Effect*

In this section, we consider the estimation of θ_1 defined in (6). To this end, consider the following difference-in-"average of averages" estimator:

$$\hat{\theta}_{1,G} \coloneqq \frac{\sum_{1\le g\le G}\bar{Y}_g A_g}{\sum_{1\le g\le G}A_g} - \frac{\sum_{1\le g\le G}\bar{Y}_g(1-A_g)}{\sum_{1\le g\le G}(1-A_g)}, \tag{10}$$

where

$$\bar{Y}_g = \frac{1}{|\mathcal{M}_g|}\sum_{i\in\mathcal{M}_g} Y_{i,g}. \tag{11}$$

For what follows, it will be useful to introduce some notation to denote various types of averages. Given a sequence of random variables $C_g : 1 \le g \le G$, consider the following definitions:

$$\hat{\mu}^C_{G,a}(s) \coloneqq \frac{1}{\sum_{1\le g\le G} I\{A_g = a, S_g = s\}} \sum_{1\le g\le G} C_g I\{A_g = a, S_g = s\},$$

$$\hat{\mu}^C_{G}(s) \coloneqq \frac{1}{\sum_{1\le g\le G} I\{S_g = s\}} \sum_{1\le g\le G} C_g I\{S_g = s\},$$

$$\hat{\mu}^C_{G,a} \coloneqq \frac{1}{\sum_{1\le g\le G} I\{A_g = a\}} \sum_{1\le g\le G} C_g I\{A_g = a\}.$$

Given this notation, $\hat{\theta}_{1,G}$ could alternatively be written as

$$\hat{\theta}_{1,G} = \hat{\mu}^{\bar{Y}}_{G,1} - \hat{\mu}^{\bar{Y}}_{G,0}.$$

Note that $\hat{\theta}_{1,G}$ may be obtained as the estimator of the coefficient on A_g in the following ordinary least squares regression:

$$\text{regress } \bar{Y}_g \text{ on constant} + A_g.$$

As such, we can view $\hat{\theta}_{1,G}$ as estimating the treatment effect for an individual-level randomized experiment in which the clusters are themselves the units of interest. The following theorem derives the asymptotic behavior of this estimator.

THEOREM 2. Under assumptions 1 and 2,

$$\sqrt{G}(\hat{\theta}_{1,G} - \theta_1) \xrightarrow{d} N(0, \sigma_1^2)$$

as $G \to \infty$, where

$$\sigma_1^2 \coloneqq \frac{1}{\pi}\text{Var}[\bar{Y}_g^{\dagger}(1)] + \frac{1}{1-\pi}\text{Var}[\bar{Y}_g^{\dagger}(0)] + E[(\bar{m}_1(S_g) - \bar{m}_0(S_g))^2] + E\left[\tau(S_g)\left(\frac{1}{\pi}\bar{m}_1(S_g) + \frac{1}{1-\pi}\bar{m}_0(S_g)\right)^2\right],$$

with

$$\begin{aligned}\bar{Y}_g^{\dagger}(a) &:= \bar{Y}_g(a) - E[\bar{Y}_g(a)|S_g],\\ \bar{m}_a(S_g) &:= E[\bar{Y}_g(a)|S_g] - E[\bar{Y}_g(a)],\end{aligned} \tag{12}$$

and π, $\tau(\cdot)$ defined as in assumption 1.

From this result it is immediate that $\hat{\theta}_{1,G}$ is most efficient when paired with an assignment mechanism that features $\tau(s) = 0$ for every $s \in \mathcal{S}$ (i.e., strong balance) and least efficient when $\tau(s) = \pi(1-\pi)$ for every $s \in \mathcal{S}$. The next result shows that, as a consequence, the probability limit of the standard heteroskedasticity-robust variance estimator is generally too large relative to σ_1^2.

THEOREM 3. Let $\tilde{\sigma}_{1,G}^2$ denote the heteroskedasticity-robust estimator of the variance of the coefficient of A_g in an ordinary least squares regression of $\bar{Y}_g$ on a constant and A_g. Note that this estimator can be written as

$$\tilde{\sigma}_{1,G}^2 := \frac{1}{(1/G)\sum_{1\le g\le G} A_g}\widehat{\mathrm{Var}}[\bar{Y}_g(1)] + \frac{1}{(1/G)\sum_{1\le g\le G} 1 - A_g}\widehat{\mathrm{Var}}[\bar{Y}_g(0)],$$

where

$$\widehat{\mathrm{Var}}\big[\bar{Y}_g(a)\big] := \hat{\mu}_a^{\bar{Y}2} - \big(\hat{\mu}_a^{\bar{Y}}\big)^2.$$

Then under assumptions 1 and 2,

$$\tilde{\sigma}_{1,G}^2 \xrightarrow{p} \frac{1}{\pi}\mathrm{Var}\big[\bar{Y}_g(1)\big] + \frac{1}{1-\pi}\mathrm{Var}\big[\bar{Y}_g(0)\big] \ge \sigma_1^2, \tag{13}$$

with equality if and only if

$$[\pi(1-\pi) - \tau(s)]\left[\frac{1}{\pi}\bar{m}_1(s) + \frac{1}{1-\pi}\bar{m}_0(s)\right]^2 = 0,$$

for every $s \in \mathcal{S}$.

Note that it can be shown that in the case of Bernoulli random assignment, where $A^{(G)}$ is an i.i.d. sequence with $P(A_g = 1) = \pi$, assumption 1 is satisfied with $\tau(s) = \pi(1-\pi)$ for every $s \in \mathcal{S}$. As such, we obtain from theorem 3 that in this case $\tilde{\sigma}_{1,G}^2$ is a consistent estimator of σ_1^2. Note that $\tilde{\sigma}_{1,G}^2$ is also consistent when there is no stratification (i.e., $|\mathcal{S}| = 1$), since in this case $\bar{m}_a(s) = 0$ for $a \in \{0, 1\}$.

To facilitate the use of theorem 2 for inference about θ_1, we now provide an estimator of σ_1^2 that is consistent. For any $s \in \mathcal{S}$, let

$$G(s) := \sum_{g\in G} 1\{S_g = s\}. \tag{14}$$

Then, define the following estimators:

$$\hat{\xi}^2_{\bar{Y}}(\pi) := \frac{1}{\pi}\left(\hat{\mu}^{\bar{Y}^2}_{G,1} - \sum_{s\in\mathcal{S}}\frac{G(s)}{G}\hat{\mu}^{\bar{Y}}_{G,1}(s)^2\right) + \frac{1}{1-\pi}\left(\hat{\mu}^{\bar{Y}^2}_{G,0} - \sum_{s\in\mathcal{S}}\frac{G(s)}{G}\hat{\mu}^{\bar{Y}}_{G,0}(s)^2\right), \quad (15)$$

$$\hat{\xi}^2_H := \sum_{s\in\mathcal{S}}\frac{G(s)}{G}\left[\left(\hat{\mu}^{\bar{Y}}_{G,1}(s) - \hat{\mu}^{\bar{Y}}_{G,1}\right) - \left(\hat{\mu}^{\bar{Y}}_{G,0}(s) - \hat{\mu}^{\bar{Y}}_{G,0}\right)\right]^2, \quad (16)$$

$$\hat{\xi}^2_A(\pi) := \sum_{s\in\mathcal{S}}\tau(s)\frac{G(s)}{G}\left[\frac{1}{\pi}\left(\hat{\mu}^{\bar{Y}}_{G,1}(s) - \hat{\mu}^{\bar{Y}}_1\right) + \frac{1}{1-\pi}\left(\hat{\mu}^{\bar{Y}}_{G,0}(s) - \hat{\mu}^{\bar{Y}}_{G,0}\right)\right]^2, \quad (17)$$

and define $\hat{\sigma}^2_{1,G} := \hat{\xi}^2_{\bar{Y}}(\pi) + \hat{\xi}^2_H + \hat{\xi}^2_A(\pi)$.

The following theorem establishes the consistency of $\hat{\sigma}^2_{1,G}$ for σ^2_1. In the statement of the theorem, we make use of the following additional notation: for scalars a and b, we define $[a \pm b] := [a - b, a + b]$, and denote by $\Phi(\cdot)$ the standard normal cumulative distribution function.

THEOREM 4. Under assumptions 1 and 2,

$$\hat{\sigma}^2_{1,G} \xrightarrow{p} \sigma^2_1$$

as $G \to \infty$. Thus, for $\sigma^2_1 > 0$ and for any $\alpha \in (0, 1)$,

$$P\left\{\theta_1 \in \left[\hat{\theta}_{1,G} \pm \frac{\hat{\sigma}_{1,G}}{\sqrt{G}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]\right\} \to 1 - \alpha$$

as $G \to \infty$.

REMARK 4. A sufficient condition under which $\sigma^2_1 > 0$ holds is that $\mathrm{Var}[\bar{Y}_g(a) - E[\bar{Y}_g(a)|S_g]] > 0$ for some $a \in \{0, 1\}$. More generally, we expect $\sigma^2_1 > 0$ except in pathological cases such as when the distribution of outcomes is degenerate or in cases with perfect negative within-cluster correlation.

REMARK 5. As mentioned earlier, $\hat{\theta}_{1,G}$ can equivalently be obtained as the estimator of the coefficient on A_g in an ordinary least squares regression of $\bar{Y}_g$ on a constant and A_g. A natural next step would be to include additional baseline covariates in this linear regression. Doing so carefully can lead to gains in efficiency; see Negi and Wooldridge (2021) and references therein for related results in the context of individual-level randomized experiments. In appendix section A.4, we describe one such adjustment strategy.

C. *Size-Weighted Cluster-Level Average Treatment Effect*

In this section, we consider the estimation of θ_2 defined in (7). To this end, consider the following difference-in-"weighted average of averages" estimator:

$$\hat{\theta}_{2,G} \coloneqq \frac{\sum_{1\le g\le G} \bar{Y}_g N_g A_g}{\sum_{1\le g\le G} N_g A_g} - \frac{\sum_{1\le g\le G} \bar{Y}_g N_g (1 - A_g)}{\sum_{1\le g\le G} N_g (1 - A_g)}, \tag{18}$$

where $\bar{Y}_g$ is defined as in (11). Note that $\hat{\theta}_{2,G}$ may be obtained as the estimator of the coefficient on A_g in the following weighted least squares regression:

$$\text{regress } Y_{i,g} \text{ on constant} + A_g \text{ using weights } N_g/\left|\mathcal{M}_g\right|.$$

In the special case in which $\mathcal{M}_g = \{1, 2, \ldots, N_g\}$ for all $1 \le g \le G$ with probability 1, we have $\hat{\theta}_{2,G} = \hat{\theta}_G^{\text{alt}}$ (i.e., the weights collapse to 1). The following theorem derives the asymptotic behavior of this estimator.

THEOREM 5. Under assumptions 1 and 2,

$$\sqrt{G}(\hat{\theta}_{2,G} - \theta_2) \xrightarrow{d} N(0, \sigma_2^2)$$

as $G \to \infty$, where

$$\begin{aligned} \sigma_2^2 \coloneqq \frac{1}{\pi}\text{Var}\left[\tilde{Y}_g^{\dagger}(1)\right] + \frac{1}{1-\pi}\text{Var}\left[\tilde{Y}_g^{\dagger}(0)\right] + E\left[\left(\tilde{m}_1(S_g) - \tilde{m}_0(S_g)\right)^2\right] \\ + E\left[\tau(S_g)\left(\frac{1}{\pi}\tilde{m}_1(S_g) + \frac{1}{1-\pi}\tilde{m}_0(S_g)\right)^2\right], \end{aligned} \tag{19}$$

with

$$\begin{aligned} \tilde{Y}_g(a) &\coloneqq \frac{N_g}{E[N_g]}\left(\bar{Y}_g(a) - \frac{E[\bar{Y}_g(a)N_g]}{E[N_g]}\right), \\ \tilde{Y}_g^{\dagger}(a) &\coloneqq \tilde{Y}_g(a) - E[\tilde{Y}_g(a)|S_g], \\ \tilde{m}_a(S_g) &\coloneqq E[\tilde{Y}_g(a)|S_g] - E[\tilde{Y}_g(a)], \end{aligned}$$

and π, $\tau(\cdot)$ defined as in assumption 1.

REMARK 6. Note that, unlike $\hat{\theta}_G^{\text{alt}}$ and $\hat{\theta}_{1,G}$, the estimator $\hat{\theta}_{2,G}$ cannot be computed without explicit knowledge of $N^{(G)} \coloneqq (N_g : 1 \le g \le G)$. As explained in section II.B, however, θ_2 is in some instances equal to ϑ, which may be consistently estimated using $\hat{\theta}_G^{\text{alt}}$.

REMARK 7. As mentioned in the introduction, our analysis is distinct from the finite-population analyses undertaken in, for instance, Su and Ding (2021). Here we compare our σ_2^2 to the variance of the difference-in-means estimator from such an analysis. Specifically, in the special case in which $\mathcal{M}_g = \{1, 2, \ldots, N_g\}$ and $|\mathcal{S}| = 1$, σ_2^2 could alternatively be written as

$$\sigma_2^2 \coloneqq \frac{1}{E[N_g]^2}\left(\frac{E\left[\left(\sum_{1\le i\le N_g}\varepsilon_{i,g}(1)\right)^2\right]}{\pi} + \frac{E\left[\left(\sum_{1\le i\le N_g}\varepsilon_{i,g}(0)\right)^2\right]}{1-\pi}\right),$$

with

$$\varepsilon_{i,g}(a) = Y_{i,g}(a) - \frac{E[N_g \bar{Y}_g(a)]}{E[N_g]}.$$

It follows from theorem 1 of Su and Ding (2021) that the finite-population design-based variance is given by

$$\sigma^2_{2,G,\text{finpop}} := \left(\frac{G}{N}\right)^2 \left(\frac{1}{G} \sum_{1 \le g \le G} \left[\frac{\left(\sum_{1 \le i \le N_g} \tilde{\varepsilon}_{i,g}(1)\right)^2}{\pi} + \frac{\left(\sum_{1 \le i \le N_g} \tilde{\varepsilon}_{i,g}(0)\right)^2}{1-\pi} \right] - \frac{1}{G} \sum_{1 \le g \le G} \left[\sum_{1 \le i \le N_g} \left[\tilde{\varepsilon}_{i,g}(1) - \tilde{\varepsilon}_{i,g}(0) \right] \right]^2 \right),$$

where

$$N := \sum_{1 \le g \le G} N_g,$$

$$\tilde{\varepsilon}_{i,g}(a) := Y_{i,g}(a) - \frac{1}{N} \sum_{1 \le g \le G} \sum_{1 \le i \le N_g} Y_{i,g}(a).$$

We emphasize that in the finite-population framework adopted by Su and Ding (2021), all of the above quantities are nonrandom, and are derived under complete randomization with $|\mathcal{S}| = 1$. From this, we see that the comparison between $\sigma^2_{2,G,\text{finpop}}$ and σ^2_2 exactly mimics the comparison between the superpopulation and finite-population variance expressions for the difference-in-means estimator in the nonclustered setting (see, e.g., Ding, Li, and Miratrix 2017). In particular, $\sigma^2_{2,G,\text{finpop}}$ is made up of two terms: the first term corresponds to a finite-population analog of σ^2_2, whereas the second term, which enters negatively, can be interpreted as the gain in precision that results from observing the entire population.

Remark 8. As discussed in remark 3, $\theta_1 = \theta_2$ whenever $N_g = k$ for all $1 \le g \le G$. Furthermore, in this case we have $\hat{\theta}_{1,G} = \hat{\theta}_{2,G}$ and thus $\sigma^2_1 = \sigma^2_2$ as well.

In parallel with our development in the preceding section, we note that $\hat{\theta}_{2,G}$ is most efficient when paired with an assignment mechanism that features $\tau(s) = 0$ for all $s \in \mathcal{S}$, and we show that the probability limit of the cluster-robust variance estimator is generally too large relative to σ^2_2.

Theorem 6. Let $\tilde{\sigma}^2_{2,G}$ denote the cluster-robust estimator of the variance of the coefficient of A_g in a weighted least squares regression of Y_{ig} on a constant and A_g, with weights equal to $N_g/|\mathcal{M}_g|$. This estimator can be written as

$$\tilde{\sigma}^2_{2,G} = \tilde{\sigma}^2_{2,G}(1) + \tilde{\sigma}^2_{2,G}(0), \tag{20}$$

where, for $a \in \{0, 1\}$, we define

$$\tilde{\sigma}^2_{2,G}(a) := \frac{1}{\left[(1/G)\sum_{1\le g\le G} N_g I\{A_g = a\}\right]^2} \frac{1}{G} \sum_{1\le g\le G} \left[\left(\frac{N_g}{|\mathcal{M}_g|} \right)^2 I\{A_g = a\} \left(\sum_{i\in\mathcal{M}_g} \hat{\varepsilon}_{i,g}(a) \right)^2 \right], \quad (21)$$

with

$$\hat{\varepsilon}_{i,g}(a) := Y_{i,g} - \frac{1}{\sum_{1\le g\le G} N_g I\{A_g = a\}} \sum_{1\le g\le G} N_g \bar{Y}_g I\{A_g = a\}.$$

Then, under assumptions 1 and 2,

$$\tilde{\sigma}^2_{2,G} \xrightarrow{p} \frac{1}{\pi} \mathrm{Var}\left[\tilde{Y}_g(1)\right] + \frac{1}{1-\pi} \mathrm{Var}\left[\tilde{Y}_g(0)\right] \ge \sigma_2^2 \quad (22)$$

with equality if and only if

$$[\pi(1-\pi) - \tau(s)] \left(\frac{1}{\pi} \tilde{m}_1(s) + \frac{1}{1-\pi} \tilde{m}_0(s) \right)^2 = 0,$$

for every $s \in \mathcal{S}$.

To facilitate the use of theorem 5 for inference about θ_2, we now provide an estimator for σ_2^2 that is consistent. Similarly to Bai et al. (2022) and Liu (2023), we construct our estimator using a feasible analog of $\tilde{Y}_g(a)$ given by

$$\hat{Y}_g := \frac{N_g}{(1/G)\sum_{1\le j\le G} N_j} \left(\bar{Y}_g - \frac{(1/G)\sum_{1\le j\le G} \bar{Y}_j I\{A_j = A_g\} N_j}{(1/G)\sum_{1\le j\le G} I\{A_j = A_g\} N_j} \right).$$

We then define the following estimators,

$$\hat{\xi}^2_{\hat{Y}}(\pi) := \frac{1}{\pi} \left(\hat{\mu}^{\hat{Y}^2}_{G,1} - \sum_{s\in\mathcal{S}} \frac{G(s)}{G} \hat{\mu}^{\hat{Y}}_{G,1}(s)^2 \right) + \frac{1}{1-\pi} \left(\hat{\mu}^{\hat{Y}^2}_{G,0} - \sum_{s\in\mathcal{S}} \frac{G(s)}{G} \hat{\mu}^{\hat{Y}}_{G,0}(s)^2 \right), \quad (23)$$

$$\hat{\xi}^2_H := \sum_{s\in\mathcal{S}} \frac{G(s)}{G} \left[\left(\hat{\mu}^{\hat{Y}}_{G,1}(s) - \hat{\mu}^{\hat{Y}}_{G,1} \right) - \left(\hat{\mu}^{\hat{Y}}_{G,0}(s) - \hat{\mu}^{\hat{Y}}_{G,0} \right) \right]^2, \quad (24)$$

$$\hat{\xi}^2_A(\pi) := \sum_{s\in\mathcal{S}} \tau(s) \frac{G(s)}{G} \left[\frac{1}{\pi} \left(\hat{\mu}^{\hat{Y}}_{G,1}(s) - \hat{\mu}^{\hat{Y}}_{G,1} \right) + \frac{1}{1-\pi} \left(\hat{\mu}^{\hat{Y}}_{G,0}(s) - \hat{\mu}^{\hat{Y}}_{G,0} \right) \right]^2, \quad (25)$$

and set $\hat{\sigma}^2_{2,G} := \hat{\xi}^2_{\hat{Y}}(\pi) + \hat{\xi}^2_H + \hat{\xi}^2_A(\pi)$. The following theorem establishes the consistency of $\hat{\sigma}^2_{2,G}$ for σ_2^2. In the statement of the theorem, we again make use of the notation introduced preceding theorem 4.

THEOREM 7. Under assumptions 1 and 2,

$$\hat{\sigma}^2_{2,G} \xrightarrow{p} \sigma_2^2$$

as $G \to \infty$. Thus, for $\sigma_2^2 > 0$ and for any $\alpha \in (0, 1)$,

$$P\left\{\theta_2 \in \left[\hat{\theta}_{2,G} \pm \frac{\hat{\sigma}_{2,G}}{\sqrt{G}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]\right\} \to 1 - \alpha$$

as $G \to \infty$.

REMARK 9. It can be shown that a sufficient condition under which $\sigma_2^2 > 0$ holds is that $\text{Var}[\tilde{Y}_g(a) - E[\tilde{Y}_g(a)|S_g]] > 0$ for some $a \in \{0, 1\}$. Similarly to the discussion in remark 1, we expect this to hold outside of pathological cases.

REMARK 10. A natural next step to consider would be the inclusion of additional baseline covariates in the regression specification. To that end, in appendix section A.4, we consider an adjustment strategy based on an alternative estimator of θ_2 given by

$$\hat{\theta}_{2,G}^{\text{sd}} := \frac{(1/G)\sum_{1\le g\le G}\bar{Y}_g N_g A_g}{\bar{N}_G \bar{A}_G} - \frac{(1/G)\sum_{1\le g\le G}\bar{Y}_g N_g (1 - A_g)}{\bar{N}_G (1 - \bar{A}_G)}, \tag{26}$$

where $\bar{N}_G := (1/G)\Sigma_{1\le g\le G} N_g$ and $\bar{A}_G := (1/G)\Sigma_{1\le g\le G} A_g$. Note that $\hat{\theta}_{2,G}^{\text{sd}}$ may be obtained as the estimator of the coefficient of A_g in an ordinary least squares regression of $\Gamma_{g,G} := \bar{Y}_g(N_g/\bar{N}_G)$ on a constant and A_g. Su and Ding (2021) argue in the context of completely randomized experiments that $\hat{\theta}_{2,G}^{\text{sd}}$ is well suited for the inclusion of additional baseline covariates (particularly when N_g is included as a regressor). Moreover, we conjecture in the appendix that a fully nonparametric covariate adjustment strategy based on this estimator is, in fact, efficient.

IV. Simulations

In this section, we illustrate the results in section III with a simulation study. In all cases, data are generated as

$$Y_{i,g}(a) = \eta_g(a) Z_{g,1} + \tilde{m}_a(Z_{g,2}) + U_{i,g}(a), \tag{27}$$

for $a \in \{0, 1\}$, where

- $\eta_g(a)$ are i.i.d. with $\eta_g(0) \sim U[0, 1]$ and $\eta_g(1) \sim U[0, 5]$.
- $U_{i,g}(a)$ are i.i.d. with $U_{i,g}(a) \sim N(0, \sigma(a))$ and $\sigma(1) = \sqrt{2} > \sigma(0) = 1$.
- $\tilde{m}_a(Z_{g,2}) = m_a(Z_{g,2}) - E[m_a(Z_{g,2})]$, where

$$m_1(Z_{g,2}) = Z_{g,2} \quad \text{and} \quad m_0(Z_{g,2}) = -\log(Z_{g,2} + 3) I\left\{Z_{g,2} \le \frac{1}{2}\right\}.$$

- The distribution of $Z_g := (Z_{g,1}, Z_{g,2})$ varies by design as described below.

We consider three alternative distributions of cluster sizes. These distributions are depicted in figure 1. To describe them, let BB(a, b, n_{supp}) be the beta-binomial distribution with parameters a and b and support on $\{0, \ldots, n_{\text{supp}}\}$. We then define

$$N_g = 10(B + 1), \quad \text{where} \quad B \sim \text{BB}(a, b, n_{\text{supp}}),$$

for the following values of (a, b) and n_{supp}:

- $(a, b) = (1, 1)$: Uniform probability mass function (pmf) on 10 to $N_{\max} = 10(n_{\text{supp}} + 1)$.
- $(a, b) = (0.4, 04)$: U-shaped pmf on 10 to $N_{\max} = 10(n_{\text{supp}} + 1)$.
- $(a, b) = (10, 50)$: Bell-shaped pmf on 10 to $N_{\max} = 10(n_{\text{supp}} + 1)$ with a long right tail.

For each of the three distributions of cluster sizes, we consider three alternative ways to draw the observations within the gth cluster that are sampled by the researcher, $\mathcal{M}_g$: (a) $|\mathcal{M}_g| = N_g$, (b) $|\mathcal{M}_g| = 10$, and (c) $|\mathcal{M}_g| = \max\{10, \min\{\gamma N_g, 200\}\}$ with $\gamma = 0.4$.

The combination of the three distributions of N_g and the three distributions of $\mathcal{M}_g$ leads to nine alternative specifications. For each of these specifications, we consider in addition two designs for the distribution of Z_g as follows:

- Design 1: $Z_{g,1} \perp\!\!\!\perp N_g$ with $Z_{g,1} \in \{-1, 1\}$ i.i.d. with $p_z \equiv P\{Z_{g,1} = 1\} = 1/2$.

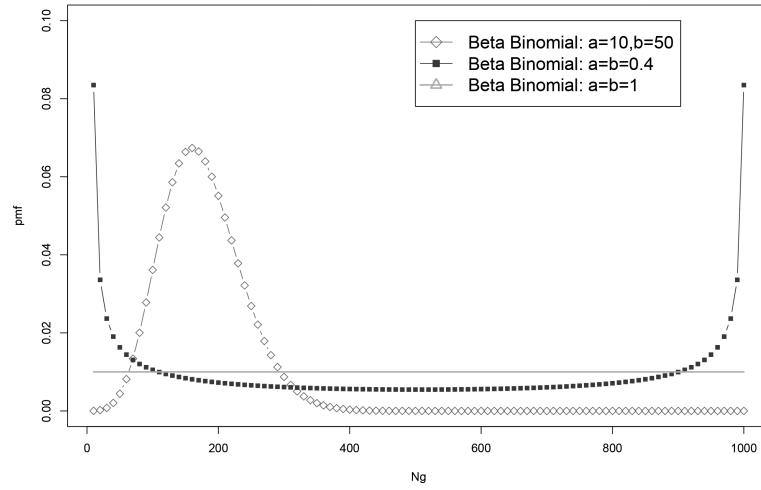

FIG. 1.—Three probability mass functions of N_g when $N_{\max} = 1{,}000$.

- Design 2: $Z_{g,1} = Z_{g,\text{big}} I\{N_g \geq E[N_g]\} + Z_{g,\text{small}} I\{N_g < E[N_g]\}$, where $Z_{g,\text{big}} \in \{-1, 1\}$ with $p_z = 3/4$ and $Z_{g,\text{small}} \in \{-1, 1\}$ i.i.d. with $p_z = 1/4$.
- In both designs, $Z_{g,2} \perp\!\!\!\perp N_g$ with $Z_{g,2} \sim \text{beta}(2, 2)$ (recentered and rescaled by the population mean and variance to have mean zero and variance 1).

Finally, treatment assignment A_g follows a covariate-adaptive randomization (CAR) mechanism based on stratified block randomization with $\pi = 1/2$ within each stratum. Concretely, we stratify the observations as follows:

- CAR-1: $S(\cdot) \perp\!\!\!\perp N_g$, where strata are determined by dividing the support of $Z_{g,2}$ into $|\mathcal{S}| = 10$ intervals of equal length and letting $S(Z_{g,2})$ be the function that returns the interval in which $Z_{g,2}$ lies.
- CAR-2: $S(\cdot) \not\perp\!\!\!\perp N_g$, where strata are determined by the Cartesian product of dividing the support of $Z_{g,2}$ into $|\mathcal{S}|/2$ intervals of equal length and letting $S(Z_{g,2})$ be the function that returns the interval in which $Z_{g,2}$ lies, and dividing the support of N_g by whether N_g is above or below the median of N_g. The total number of strata is $|\mathcal{S}| = |\mathcal{S}|/2 \times 2$ and we set $|\mathcal{S}| = 10$ as in CAR-1.

In principle, we could also consider other assignment mechanisms such as simple random sampling. We decided to focus on stratified block randomization because it is prevalent in practice and our results show that it dominates other mechanisms for which $\tau(s) \neq 0$ in terms of asymptotic efficiency.

The model in (27), as well as the two CAR designs, follows closely the original designs for covariate-adaptive randomization with individual-level data considered in Bugni, Canay, and Shaikh (2018, 2019). Note that for these designs, we obtain that

$$E[Y_{i,g}(1) - Y_{i,g}(0) \,|\, N_g] = 2E[Z_{g,1} \,|\, N_g].$$

In design 1, it follows that $\theta_1 = \theta_2 = 0$. In design 2, on the other hand, it follows that

$$E[Z_{g,1}|N_g] = \begin{cases} E[Z_{g,\text{big}}] = 1/2 & \text{if } N_g \geq E[N_g], \\ E[Z_{g,\text{small}}] = -1/2 & \text{if } N_g < E[N_g], \end{cases}$$

and so

$$\begin{aligned} \theta_1 &= 2P\{N_g \geq E[N_g]\} - 1, \\ \theta_2 &= E\left[\frac{N_g}{E[N_g]} \mid N_g \geq E[N_g]\right] P\{N_g \geq E[N_g]\} \\ &\quad - E\left[\frac{N_g}{E[N_g]} \mid N_g < E[N_g]\right] P\{N_g < E[N_g]\}. \end{aligned}$$

TABLE 1
RESULTS FOR $G = 100$, $N_{\max} = 500$, $Z_g \perp\!\!\!\perp N_g$ (DESIGN 1), $Z_g \not\perp\!\!\!\perp N_g$ (DESIGN 2), AND CAR-1

	True Values		Estimated		Estimated Standard Deviation		Coverage Probability	
$\mathcal{M}_g$, N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
	A. CAR-1: Design 1							
N_g:								
BB(1, 1)	.0000	.0000	−.0016	.0016	4.2885	4.9375	.9440	.9426
BB(.4, .4)	.0000	.0000	−.0080	−.0070	4.2864	5.2952	.9454	.9310
BB(10, 50)	.0000	.0000	−.0001	−.0010	4.2808	4.5852	.9444	.9486
10:								
BB(1, 1)	.0000	.0000	.0007	−.0000	4.3297	4.9780	.9426	.9330
BB(.4, .4)	.0000	.0000	−.0008	.0015	4.3385	5.3582	.9460	.9438
BB(10, 50)	.0000	.0000	.0023	.0063	4.3414	4.6545	.9436	.9440
γN_g:								
BB(1, 1)	.0000	.0000	.0090	.0087	4.2827	4.9091	.9346	.9338
BB(.4, .4)	.0000	.0000	−.0089	−.0044	4.2871	5.3089	.9426	.9376
BB(10, 50)	.0000	.0000	.0049	.0064	4.2994	4.5949	.9436	.9470
	B. CAR-1: Design 2							
N_g:								
BB(1, 1)	.0000	.4900	−.0036	.4819	4.2792	4.7416	.9474	.9454
BB(.4, .4)	.0000	.6581	−.0142	.6387	4.2870	5.0338	.9458	.9424
BB(10, 50)	−.1407	.1625	−.0822	.2172	4.2825	4.5250	.9342	.9440
10:								
BB(1, 1)	.0000	.4900	−.0044	.4844	4.3439	4.8198	.9448	.9454
BB(.4, .4)	.0000	.6581	−.0168	.6366	4.3399	5.1226	.9402	.9426
BB(10, 50)	−.1407	.1625	−.0807	.2173	4.3346	4.5764	.9480	.9526
γN_g:								
BB(1, 1)	.0000	.4900	−.0129	.4673	4.2884	4.7446	.9548	.9428
BB(.4, .4)	.0000	.6581	−.0147	.6366	4.2852	5.0404	.9404	.9462
BB(10, 50)	−.1407	.1625	−.0800	.2201	4.2922	4.5273	.9416	.9398

For each of the above nine specifications and for each design of $Z_{g,1}$ and treatment assignment mechanism, we generate samples using $G \in \{100, 5000\}$ and report the true values of (θ_1, θ_2) defined in (6) and (7), the average across simulations of the estimated values $(\hat{\theta}_{1,G}, \hat{\theta}_{2,G})$ defined in (10) and (18), the average across simulations of the estimated standard deviations $(\hat{\sigma}_{1,G}, \hat{\sigma}_{2,G})$ defined in sections III.B and III.C, and the empirical coverage of the 95% confidence intervals defined in theorems 4 and 7. The results of our simulations are presented in tables 1–6, where in all cases we conducted 5,000 replications. In each table, we find that our empirical coverage probabilities are close to 95% in all cases.

V. Empirical Illustration

In this section, we illustrate our findings by revisiting the empirical application in Celhay et al. (2019). These authors use a field experiment in

TABLE 2
RESULTS FOR $G = 100$, $N_{max} = 1{,}000$, $Z_g \perp\!\!\!\perp N_g$ (Design 1), $Z_g \not\perp\!\!\!\perp N_g$ (Design 2), AND CAR-1

	True Values		Estimated		Estimated Standard Deviation		Coverage Probability	
$\mathcal{M}_g$, N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
	A. CAR-1: Design 1							
N_g:								
BB(1, 1)	.0000	.0000	−.0004	−.0057	4.2824	4.9264	.9408	.9382
BB(.4, .4)	.0000	.0000	.0002	.0008	4.2835	5.3107	.9388	.9388
BB(10, 50)	.0000	.0000	−.0061	−.0080	4.2803	4.5365	.9438	.9436
10:								
BB(1, 1)	.0000	.0000	−.0075	−.0102	4.3407	4.9950	.9374	.9412
BB(.4, .4)	.0000	.0000	.0009	.0052	4.3405	5.3903	.9386	.9426
BB(10, 50)	.0000	.0000	−.0129	−.0147	4.3358	4.5938	.9442	.9490
γN_g:								
BB(1, 1)	.0000	.0000	−.0115	−.0146	4.2837	4.9416	.9438	.9442
BB(.4, .4)	.0000	.0000	−.0095	−.0077	4.2907	5.3277	.9386	.9392
BB(10, 50)	.0000	.0000	−.0022	−.0051	4.2841	4.5295	.9434	.9446
	B. CAR-1: Design 2							
N_g:								
BB(1, 1)	.0000	.4950	.0005	.4895	4.2746	4.7550	.9396	.9410
BB(.4, .4)	.0000	.6690	.0039	.6711	4.2855	5.0689	.9480	.9510
BB(10, 50)	−.0635	.2100	−.0691	.2024	4.2806	4.4677	.9414	.9486
10:								
BB(1, 1)	.0000	.4950	−.0071	.4834	4.3384	4.8310	.9456	.9442
BB(.4, .4)	.0000	.6690	−.0054	.6533	4.3341	5.1559	.9456	.9446
BB(10, 50)	−.0635	.2100	−.0687	.1973	4.3408	4.5444	.9426	.9436
γN_g:								
BB(1, 1)	.0000	.4950	−.0192	.4723	4.2826	4.7515	.9456	.9440
BB(.4, .4)	.0000	.6690	−.0060	.6638	4.2876	5.0702	.9432	.9460
BB(10, 50)	−.0635	.2100	−.0708	.1987	4.2862	4.4822	.9446	.9472

Argentina to study the effects of temporary incentives for medical care providers to adopt early initiation of prenatal care. The medical literature has long recognized the benefits of early initiation of prenatal care. In particular, it allows doctors to detect and treat critical medical conditions, as well as advise mothers on proper nutrition and risk prevention activities in the period in which the fetus is most at risk.

The field experiment in Celhay et al. (2019) took place in Misiones, Argentina, one of the poorest provinces in the country, and with relatively high rate of maternal and child mortality. As part of the national *Plan Nacer* program, the Argentinean government transfers funds to medical care providers in exchange for their patient services. The study selected 37 public primary care facilities (accounting for 70% of the prenatal care visits in the beneficiary population), and randomly assigned 18 to treatment and 19 to control. To the best of our understanding, the treatment assignment was balanced across the 37 clinics and was not stratified. The

TABLE 3
RESULTS FOR $G = 5{,}000$, $N_{\max} = 1{,}000$, $Z_g \perp\!\!\!\perp N_g$ (Design 1), $Z_g \not\perp\!\!\!\perp N_g$ (Design 2), AND CAR-1

	True Values		Estimated		Estimated Standard Deviation		Coverage Probability	
$\mathcal{M}_g$, N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
	A. CAR-1: Design 1							
N_g:								
BB(1, 1)	.0000	.0000	.0003	−.0001	4.3688	5.0513	.9480	.9546
BB(.4, .4)	.0000	.0000	−.0009	−.0010	4.3740	5.4474	.9494	.9530
BB(10, 50)	.0000	.0000	.0008	.0011	4.3690	4.6261	.9532	.9492
10:								
BB(1, 1)	.0000	.0000	.0001	−.0003	4.4315	5.1258	.9542	.9560
BB(.4, .4)	.0000	.0000	.0006	.0013	4.4333	5.5313	.9542	.9510
BB(10, 50)	.0000	.0000	.0015	.0014	4.4330	4.6940	.9538	.9546
γN_g:								
BB(1, 1)	.0000	.0000	.0018	.0024	4.3725	5.0541	.9500	.9468
BB(.4, .4)	.0000	.0000	−.0008	−.0006	4.3785	5.4505	.9548	.9522
BB(10, 50)	.0000	.0000	.0025	.0022	4.3766	4.6319	.9580	.9598
	B. CAR-1: Design 2							
N_g:								
BB(1, 1)	.0000	.4950	.0019	.4964	4.3680	4.8506	.9598	.9606
BB(.4, .4)	.0000	.6690	−.0002	.6675	4.3750	5.1796	.9548	.9552
BB(10, 50)	−.0635	.2100	−.0642	.2088	4.3677	4.5609	.9492	.9508
10:								
BB(1, 1)	.0000	.4950	.0009	.4961	4.4332	4.9315	.9532	.9562
BB(.4, .4)	.0000	.6690	−.0006	.6689	4.4332	5.2654	.9586	.9588
BB(10, 50)	−.0635	.2100	−.0627	.2104	4.4299	4.6277	.9582	.9592
γN_g:								
BB(1, 1)	.0000	.4950	.0002	.4956	4.3732	4.8567	.9586	.9578
BB(.4, .4)	.0000	.6690	−.0000	.6695	4.3784	5.1788	.9586	.9614
BB(10, 50)	−.0635	.2100	−.0635	.2094	4.3744	4.5676	.9538	.9576

intervention was implemented only for 8 months (May 2010 to December 2010), and the clinics were clearly informed of the temporary nature of this intervention. During the intervention period, control group clinics saw no change in their fees for prenatal visits. In contrast, clinics in the treatment group received a threefold increase in payments for any first prenatal visit that occurred before week 13 of pregnancy. Prenatal visits after week 13 or subsequent prenatal visits experienced no change in fees.

REMARK 11. This application features a setting in which all patients from the sampled clinics were included in their study. In terms of our notation, we have $|\mathcal{M}_g| = N_g$ for all $1 \leq g \leq G$. As a consequence, the weights in the weighted average of averages estimator $\hat{\theta}_{2,G}$ equal 1, and the estimator coincides with the difference-in-means estimator; that is, $\hat{\theta}_{2,G} = \hat{\theta}_G^{\text{alt}}$.

TABLE 4
RESULTS FOR $G = 100$, $N_{max} = 500$, $Z_g \perp\!\!\!\perp N_g$ (Design 1), $Z_g \not\perp\!\!\!\perp N_g$ (Design 2), AND CAR-2

	True Values		Estimated		Estimated Standard Deviation		Coverage Probability	
$\mathcal{M}_g$, N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
	A. CAR-2: Design 1							
N_g:								
BB(1, 1)	.0000	.0000	.0043	−.0040	4.2769	4.9142	.9422	.9450
BB(.4, .4)	.0000	.0000	−.0001	−.0029	4.2760	5.2566	.9366	.9372
BB(10, 50)	.0000	.0000	−.0013	−.0021	4.2917	4.5993	.9474	.9486
10:								
BB(1, 1)	.0000	.0000	−.0073	−.0031	4.3318	4.9799	.9456	.9428
BB(.4, .4)	.0000	.0000	−.0071	−.0118	4.3325	5.3390	.9384	.9408
BB(10, 50)	.0000	.0000	.0031	.0063	4.3465	4.6611	.9334	.9432
γN_g:								
BB(1, 1)	.0000	.0000	−.0061	−.0097	4.2783	4.9082	.9460	.9430
BB(.4, .4)	.0000	.0000	−.0006	−.0058	4.2903	5.2822	.9380	.9468
BB(10, 50)	.0000	.0000	−.0078	−.0102	4.3018	4.6021	.9426	.9458
	B. CAR-2: Design 2							
N_g:								
BB(1, 1)	.0000	.4900	−.0027	.4915	4.1664	4.6658	.9540	.9476
BB(.4, .4)	.0000	.6581	−.0029	.6479	4.1693	4.9832	.9518	.9498
BB(10, 50)	−.1407	.1625	−.0897	.2121	4.1669	4.4264	.9462	.9484
10:								
BB(1, 1)	.0000	.4900	−.0015	.4917	4.2266	4.7462	.9516	.9488
BB(.4, .4)	.0000	.6581	−.0006	.6563	4.2239	5.0699	.9544	.9518
BB(10, 50)	−.1407	.1625	−.0795	.2174	4.2293	4.4920	.9464	.9490
γN_g:								
BB(1, 1)	.0000	.4900	−.0035	.4870	4.1682	4.6798	.9496	.9484
BB(.4, .4)	.0000	.6581	−.0036	.6532	4.1737	4.9868	.9578	.9524
BB(10, 50)	−.1407	.1625	−.0877	.2136	4.1861	4.4456	.9516	.9504

Celhay et al. (2019) collected data before, during, and after the intervention period. The preintervention ran between January 2009 and April 2010. During this period, the sample average of the week of the first prenatal visit is 16.97, and only 34.46% of these visits occur before week 13. Figure 2 shows a histogram of this distribution. The aforementioned treatment occurred exclusively during the intervention period, which ran between May 2010 and December 2010. While the treatment and control group affected approximately the same number of facilities, the numbers of treated and control patients are very different due to the unequal number of patients across facilities. Figure 3 provides a histogram of the number of patients attending each clinic for their first prenatal visit during the intervention period. This distribution has a mean and a standard deviation of 33.6 and 16.3 patients per clinic, respectively. Finally, the postintervention period goes between January 2011 and March 2012.

TABLE 5
RESULTS FOR $G = 100$, $N_{\max} = 1{,}000$, $Z_g \perp\!\!\!\perp N_g$ (Design 1), $Z_g \not\perp\!\!\!\perp N_g$ (Design 2), AND CAR-2

	True Values		Estimated		Estimated Standard Deviation		Coverage Probability	
$\mathcal{M}_g$, N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
	A. CAR-2: Design 1							
N_g:								
BB(1, 1)	.0000	.0000	−.0083	−.0120	4.2775	4.9251	.9394	.9402
BB(.4, .4)	.0000	.0000	−.0143	−.0153	4.2801	5.2966	.9434	.9384
BB(10, 50)	.0000	.0000	−.0086	−.0115	4.2897	4.5346	.9422	.9450
10:								
BB(1, 1)	.0000	.0000	−.0152	−.0175	4.3419	4.9979	.9424	.9426
BB(.4, .4)	.0000	.0000	−.0058	−.0066	4.3391	5.3822	.9428	.9410
BB(10, 50)	.0000	.0000	−.0008	−.0005	4.3482	4.6036	.9438	.9480
γN_g:								
BB(1, 1)	.0000	.0000	−.0023	−.0033	4.2840	4.9329	.9440	.9396
BB(.4, .4)	.0000	.0000	.0044	.0003	4.2906	5.3131	.9408	.9424
BB(10, 50)	.0000	.0000	−.0142	−.0118	4.2884	4.5387	.9418	.9404
	B. CAR-2: Design 2							
N_g:								
BB(1, 1)	.0000	.4950	−.0046	.4938	4.1610	4.6755	.9546	.9514
BB(.4, .4)	.0000	.6690	−.0091	.6566	4.1644	4.9964	.9506	.9452
BB(10, 50)	−.0635	.2100	−.0715	.1989	4.1593	4.3726	.9474	.9498
10:								
BB(1, 1)	.0000	.4950	−.0069	.4893	4.2275	4.7582	.9514	.9482
BB(.4, .4)	.0000	.6690	−.0063	.6612	4.2194	5.0787	.9584	.9520
BB(10, 50)	−.0635	.2100	−.0746	.1989	4.2219	4.4408	.9524	.9580
γN_g:								
BB(1, 1)	.0000	.4950	−.0105	.4836	4.1551	4.6726	.9540	.9492
BB(.4, .4)	.0000	.6690	−.0062	.6631	4.1645	4.9924	.9530	.9512
BB(10, 50)	−.0635	.2100	−.0640	.2074	4.1660	4.3863	.9446	.9506

In table 7, we report our estimates of the equally weighted and size-weighted average treatment effect (ATE) and their standard errors.[3] We ran separate estimations using the data in the preintervention, intervention, and postintervention periods. In each case, we considered two possible outcomes: Y_1 = Weeks, which denotes the pregnancy week of the first prenatal visit, and $Y_2 = 1\{\text{Weeks} < 13\}$, which indicates whether the first prenatal visit occurs in pregnancy week 13 or lower. In all periods and for both outcomes, the equally weighted cluster-level ATE, $\hat{\theta}_{1,G}$, appears to be small and statistically insignificant at the usual levels. In words, there seems to be no ATE at the clinic level. These results contrast sharply with those obtained using the size-weighted cluster-level ATE, $\hat{\theta}_{2,G}$, which measures

[3] In the context of the experiment in Celhay et al. (2019), these would be more appropriately labeled as intention-to-treat estimates, given the presence of imperfect compliance in the study.

TABLE 6
RESULTS FOR $G = 5{,}000$, $N_{\max} = 1{,}000$, $Z_g \perp\!\!\!\perp N_g$ (Design 1), $Z_g \not\perp\!\!\!\perp N_g$ (Design 2), AND CAR-2

	True Values		Estimated		Estimated Standard Deviation		Coverage Probability	
$\mathcal{M}_g$, N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
	A. CAR-2: Design 1							
N_g:								
BB(1, 1)	.0000	.0000	.0007	−.0006	4.3737	5.0442	.9570	.9582
BB(.4, .4)	.0000	.0000	.0029	.0033	4.3788	5.4297	.9548	.9568
BB(10, 50)	.0000	.0000	.0001	−.0002	4.3801	4.6374	.9500	.9502
10:								
BB(1, 1)	.0000	.0000	.0009	.0004	4.4386	5.1204	.9548	.9520
BB(.4, .4)	.0000	.0000	.0015	.0026	4.4376	5.5128	.9562	.9550
BB(10, 50)	.0000	.0000	.0007	.0005	4.4435	4.7054	.9550	.9536
γN_g:								
BB(1, 1)	.0000	.0000	.0019	.0020	4.3780	5.0463	.9564	.9540
BB(.4, .4)	.0000	.0000	.0006	−.0002	4.3836	5.4340	.9610	.9566
BB(10, 50)	.0000	.0000	.0005	.0004	4.3865	4.6416	.9496	.9518
	B. CAR-2: Design 2							
N_g:								
BB(1, 1)	.0000	.4950	.0011	.4960	4.2082	4.7586	.9644	.9580
BB(.4, .4)	.0000	.6690	.0001	.6686	4.2122	5.0922	.9680	.9598
BB(10, 50)	−.0635	.2100	−.0638	.2100	4.2169	4.4390	.9624	.9588
10:								
BB(1, 1)	.0000	.4950	.0003	.4959	4.2745	4.8394	.9592	.9582
BB(.4, .4)	.0000	.6690	.0006	.6692	4.2742	5.1821	.9648	.9588
BB(10, 50)	−.0635	.2100	−.0633	.2097	4.2829	4.5094	.9576	.9614
γN_g:								
BB(1, 1)	.0000	.4950	.0010	.4965	4.2125	4.7615	.9622	.9572
BB(.4, .4)	.0000	.6690	.0005	.6687	4.2173	5.0962	.9622	.9618
BB(10, 50)	−.0635	.2100	−.0622	.2111	4.2239	4.4454	.9604	.9586

ATE at the patient level. First, this ATE is not statistically significant during the preintervention period, which we consider a reasonable "placebo-type" finding. Second, during the intervention period, treated clinics had, on average, first prenatal visits 1.4 weeks earlier than control clinics. Moreover, the proportion of prenatal visits before week 13 was 10 percentage points higher in treated clinics than in control clinics during the same time period. These effects are statistically significant with $\alpha = 5\,\%$ and seem economically important relative to the baseline levels. Interestingly, these effects seem to extend quantitatively to the postintervention period, when the treatment incentives were completely removed. These findings demonstrate a statistically significant and economically important ATE at the patient level, both in the long and the short run. From the standpoint of our contribution, these differences between $\hat{\theta}_{1,G}$ and $\hat{\theta}_{2,G}$ point to the fact that cluster sizes are likely nonignorable, as they may reflect essential factors such as the clinic's quality or the nearby population's size. To conclude, it is also worth noting

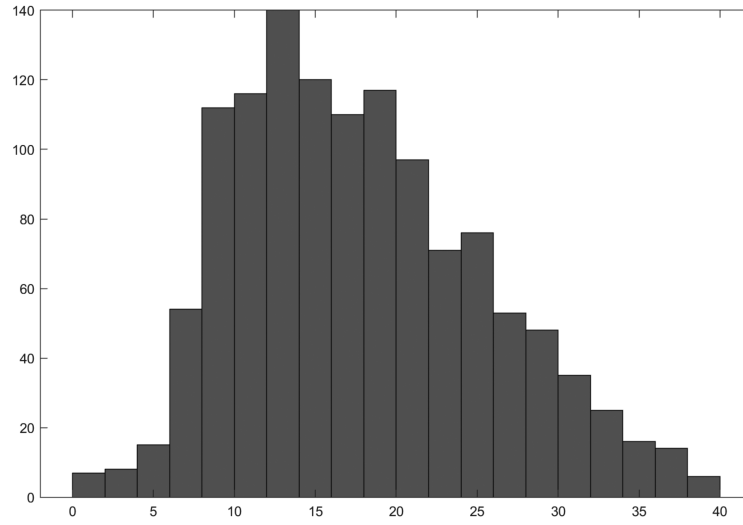

FIG. 2.—Histogram of week of first prenatal visit during preintervention period.

that our results using the size-weighted cluster-level ATE align with the (intention-to-treat) estimates obtained by Celhay et al. (2019); this is expected given that, as explained in remark 11, $|\mathcal{M}_g| = N_g$ for all $1 \leq g \leq G$ in this application and so standard OLS regression recovers $\hat{\theta}_{2,G}$.

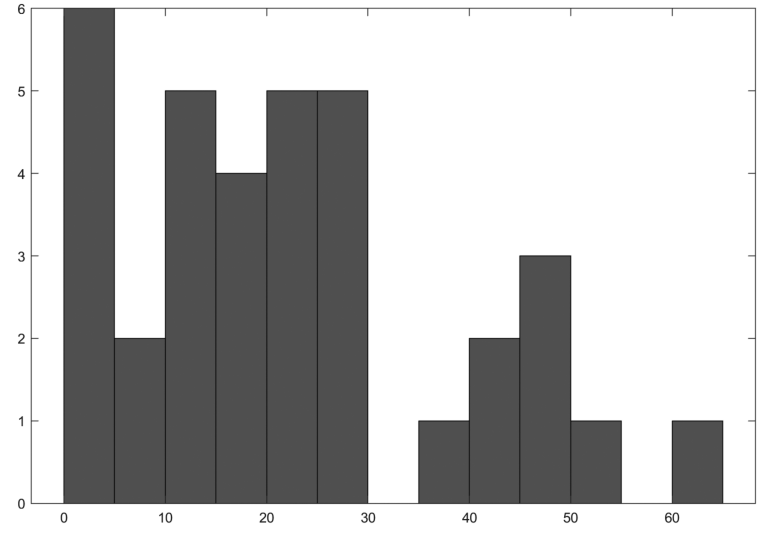

FIG. 3.—Histogram of patients per clinic having first prenatal visit during intervention period.

TABLE 7
ESTIMATION RESULTS BASED ON DATA FROM CELHAY ET AL. (2019)

	Preintervention				Intervention				Postintervention			
	Weeks		Weeks < 13		Weeks		Weeks < 13		Weeks		Weeks < 13	
	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$
Estimate	.09	−.08	.02	.01	−.01	−1.39**	.03	.10***	.11	−1.59**	−.02	.09**
Standard error	.77	.56	.06	.03	.95	.66	.05	.04	.85	.72	.05	.04

NOTE.—We ran our estimation separately for data in the preintervention period (i.e., January 2009 to April 2010), the intervention period (i.e., May 2010 to December 2010), and the postintervention period (i.e., January 2011 to March 2012). The outcome variables are "Weeks," which denotes the pregnancy week of the first prenatal visit, and "Weeks < 13," which indicates whether the first prenatal visit occurred before pregnancy week 13.

** Significant at $\alpha = 5\%$.

*** Significant at $\alpha = 1\%$.

Data Availability

Code and data replicating the tables and figures in this article can be found in Bugni et al. (2025) in the Harvard Dataverse, https://doi.org/10.7910/DVN/C7SDF2.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?" *Q.J.E.* 138:1–35.

Angelucci, M., D. Karlan, and J. Zinman. 2015. "Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco." *American Econ. J. Appl. Econ.* 7:151–82.

Athey, S., and G. W. Imbens. 2017. "The Econometrics of Randomized Experiments." In *Handbook of Economic Field Experiments*, vol. 1, edited by A. V. Banerjee and E. Duflo, 73–140. Amsterdam: Elsevier.

Attanasio, O., B. Augsburg, R. De Haas, E. Fitzsimons, and H. Harmgart. 2015. "The Impacts of Microfinance: Evidence from Joint-Liability Lending in Mongolia." *American Econ. J. Appl. Econ.* 7:90–122.

Bai, Y., J. Liu, A. Shaikh, and M. Tabord-Meehan. 2022. "Inference for Cluster Randomized Experiments with Matched Pairs." Working paper, Dept. Econ., Univ. Michigan.

Bai, Y., J. P. Romano, and A. M. Shaikh. 2021. "Inference in Experiments with Matched Pairs." *J. American Statis. Assoc.* 117:1726–37.

Banerjee, A., E. Duflo, R. Glennerster, and C. Kinnan. 2015. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." *American Econ. J. Appl. Econ.* 7:22–53.

Benhin, E., J. Rao, and A. Scott. 2005. "Mean Estimating Equation Approach to Analysing Cluster-Correlated Data with Nonignorable Cluster Sizes." *Biometrika* 92:435–50.

Bruhn, M., and D. McKenzie. 2008. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Econ. J. Appl. Econ.* 1:200–232.

Bugni, F. A., I. A. Canay, and A. M. Shaikh. 2018. "Inference under Covariate Adaptive Randomization." *J. American Statis. Assoc.* 113:1784–96.

———. 2019. "Inference under Covariate-Adaptive Randomization with Multiple Treatments." *Quantitative Econ.* 10:1747–85.

Bugni, F. A., I. A. Canay, A. M. Shaikh, and M. Tabord-Meehan. 2025. "Replication Data for: 'Inference for Cluster Randomized Experiments with Nonignorable Cluster Sizes.'" Harvard Dataverse, https://doi.org/10.7910/DVN/C7SDF2.

Cameron, A. C., and D. L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *J. Human Resources* 50:317–72.

Celhay, P. A., P. J. Gertler, P. Giovagnoli, and C. Vermeersch. 2019. "Long-Run Effects of Temporary Incentives on Medical Care Productivity." *American Econ. J. Appl. Econ.* 11:92–127.

Chiang, H. D., Y. Sasaki, and Y. Wang. 2023. "On the Inconsistency of Cluster-Robust Inference and How Subsampling Can Fix It." Working paper, Dept. Econ., Univ. Wisconsin.

Crépon, B., F. Devoto, E. Duflo, and W. Parienté. 2015. "Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco." *American Econ. J. Appl. Econ.* 7:123–50.

de Chaisemartin, C., and J. Ramirez-Cuellar. 2020. "At What Level Should One Cluster Standard Errors in Paired Experiments, and in Stratified Experiments with Small Strata?" Working Paper no. 27609 (July), NBER, Cambridge, MA.

Ding, P., X. Li, and L. W. Miratrix. 2017. "Bridging Finite and Super Population Causal Inference." *J. Causal Inference* 5:20160027.

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen. 2019. "Asymptotic Theory and Wild Bootstrap Inference with Clustered Errors." *J. Econometrics* 212:393–412.

Donner, A., and N. Klar. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research.* London: Arnold.

Duflo, E., P. Dupas, and M. Kremer. 2015. "Education, HIV, and Early Fertility: Experimental Evidence from Kenya." *A.E.R.* 105:2757–97.

Duflo, E., R. Glennerster, and M. Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." *Handbook of Development Economics*, vol. 4, edited by T. P. Schultz and J. A. Strauss, 3895–962. Amsterdam: North-Holland.

Hansen, B., and S. Lee. 2019. "Asymptotic Theory for Clustered Samples." *J. Econometrics* 210:268–90.

Hansen, C. B. 2007. "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T Is Large." *J. Econometrics* 141:597–620.

Hayes, R. J., and L. H. Moulton. 2017. *Cluster Randomised Trials.* Boca Raton: CRC.

Kahan, B. C., F. Li, A. J. Copas, and M. O. Harhay. 2023. "Estimands in Cluster-Randomized Trials: Choosing Analyses That Answer the Right Question." *Internat. J. Epidemiology* 52:107–18.

Liang, K.-Y., and S. L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73:13–22.

Liu, J. 2023. "Inference for Two-Stage Experiments under Covariate-Adaptive Randomization." Working paper, Booth School Bus., Univ. Chicago.

Lohr, S. L. 2021. *Sampling: Design and Analysis.* Boca Raton: CRC.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb. 2023. "Cluster-Robust Inference: A Guide to Empirical Practice." *J. Econometrics* 232:272–99.

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

Muralidharan, K., and P. Niehaus. 2017. "Experimentation at Scale." *J. Econ. Perspectives* 31:103–24.

Negi, A., and J. M. Wooldridge. 2021. "Revisiting Regression Adjustment in Experiments with Heterogeneous Treatment Effects." *Econometric Rev.* 40:504–34.

Raudenbush, S. W. 1997. "Statistical Analysis and Optimal Design for Cluster Randomized Trials." *Psychological Methods* 2:173–85.

Raudenbush, S. W., and D. Schwartz. 2020. "Randomized Experiments in Education, with Implications for Multilevel Causal Inference." *Ann. Rev. Statis. and Its Application* 7:177–208.

Rosenberger, W. F., and J. M. Lachin. 2016. *Randomization in Clinical Trials: Theory and Practice*. 2nd ed. New York: Wiley.

Sasaki, Y., and Y. Wang. 2022. "Non-Robustness of the Cluster-Robust Inference: With a Proposal of a New Robust Method." Working paper, Dept. Econ., Vanderbilt Univ.

Schochet, P. Z. 2013. "Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference." *J. Educ. and Behavioral Statis.* 38:219–38.

Schochet, P. Z., N. E. Pashley, L. W. Miratrix, and T. Kautz. 2021. "Design-Based Ratio Estimators and Central Limit Theorems for Clustered, Blocked RCTs." *J. American Statis. Assoc.* 117:2135–46.

Su, F., and P. Ding. 2021. "Model-Assisted Analyses of Cluster-Randomized Experiments." *J. Royal Statis. Soc. Series B* 83:994–1015.

Turner, E. L., F. Li, J. A. Gallis, M. Prague, and D. M. Murray. 2017. "Review of Recent Methodological Developments in Group-Randomized Trials: Part 1—Design." *American J. Public Health* 107:907–15.

Wang, X., E. L. Turner, F. Li, R. Wang, J. Moyer, A. J. Cook, D. M. Murray, and P. J. Heagerty. 2022. "Two Weights Make a Wrong: Cluster Randomized Trials with Variable Cluster Sizes and Heterogeneous Treatment Effects." *Contemporary Clinical Trials* 114:106702.