# Inference for Cluster Randomized Experiments with Non-ignorable Cluster Sizes*

Federico Bugni
Department of Economics
Northwestern University
federico.bugni@northwestern.edu

Ivan A. Canay
Department of Economics
Northwestern University
iacanay@northwestern.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Max Tabord-Meehan
Department of Economics
University of Chicago
maxtm@uchicago.edu

March 20, 2023

**Abstract**

This paper considers the problem of inference in cluster randomized experiments when cluster sizes are non-ignorable. Here, by a cluster randomized experiment, we mean one in which treatment is assigned at the level of the cluster; by non-ignorable cluster sizes we mean that "large" clusters and "small" clusters may be heterogeneous, and, in particular, the effects of the treatment may vary across clusters of differing sizes. In order to permit this sort of flexibility, we consider a sampling framework in which cluster sizes themselves are random. In this way, our analysis departs from earlier analyses of cluster randomized experiments in which cluster sizes are treated as non-random. We distinguish between two different parameters of interest: the equally-weighted cluster-level average treatment effect, and the size-weighted cluster-level average treatment effect. For each parameter, we provide methods for inference in an asymptotic framework where the number of clusters tends to infinity and treatment is assigned using a covariate-adaptive stratified randomization procedure. We additionally permit the experimenter to sample only a subset of the units within each cluster rather than the entire cluster and demonstrate the implications of such sampling for some commonly used estimators. A small simulation study and empirical demonstration show the practical relevance of our theoretical results.

KEYWORDS: Clustered data, randomized experiments, treatment effects, weighted least squares

JEL classification codes: C12, C14

*We would like to thank Eric Auerbach, Joe Romano and conference participants at Cowles, TSE, and ASSA 2023 for helpful comments on this paper. Xun Huang provided excellent research assistance.

# 1 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 et al. (2007) survey a variety of 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., Turner et al., 2017; Donner and Klar, 2000). In this paper, we consider the problem of inference about the effect of a binary treatment on an outcome of interest in such experiments when cluster sizes are non-ignorable, meaning that "large" clusters and "small" clusters may be heterogeneous, and, in particular, the effects of the treatment may vary across clusters of differing sizes.

In order to accommodate this sort of flexibility, we develop a sampling framework in which cluster sizes themselves are permitted to be random. More specifically, in the spirit of the survey sampling literature (see, e.g., Lohr, 2021), we adopt a two-stage sampling design, 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 emphasize, however, that we make no restrictions on the dependence across units within clusters. In the context of this framework, we distinguish between two different parameters of interest: the equally-weighted cluster-level average treatment effect, which corresponds to the average treatment effect for the average outcome within clusters, and the size-weighted cluster-level average treatment effect, which corresponds to the average treatment effect for the aggregate outcome within clusters. In general, when treatment effects are heterogeneous and cluster sizes are non-ignorable, these two parameters differ, but we highlight conditions under which they are equal to one another: for instance, when cluster size is in fact ignorable and treatment effects are suitably homogeneous. Further discussion is provided in Remark 2.2 below.

Our first result establishes that the standard difference-in-means estimator is not generally consistent for either the equally-weighted or size-weighted cluster-level average treatment effects in our framework. As a consequence, for each of these two parameters, 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-"average of averages," i.e., a difference between the average (over clusters) of the average outcome (within clusters) for the treated clusters and the average (over clusters) of the average outcome (within clusters) for the untreated clusters. 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 difference-in-"weighted average of 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. To better understand the empirical relevance of these results, we surveyed all articles involving a cluster randomized experiment published in the American Economic Journal: Applied Economics from 2018 to 2022. We document our findings in Appendix A.3. From this survey, we find that as many as a third of the experiments conduct analyses that, when paired with their corresponding sampling design, may not recover either of the parameters of interest that we define in this paper.

When developing our distributional results we are careful to allow for the possibility that treatment assignment was performed using covariate adaptive randomization. As in Bugni et al. (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 et al. (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 which delivers asymptotically exact tests.

By virtue of its sampling framework, our paper is distinct from a closely related and complimentary 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), Abadie et al. (2017), Hayes and Moulton (2017), de Chaisemartin and Ramirez-Cuellar (2020), Schochet et al. (2021), and Su and Ding (2021). 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 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, in particular, Athey and Imbens (2017, Chapter 8) and Su and Ding (2021, Section 4). These authors additionally provide estimators and methods for inference about each quantity. In this way, our results may be viewed as developing complementary results in a suitably defined "super-population" sampling framework. Further discussion is provided in Remark 3.6 below.

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) and Hansen and Lee (2019). Additional references can be found in the surveys Cameron and Miller (2015) and MacKinnon and Webb (2019). These papers are designed as methods for inference for parameters defined via linear models or estimating equations, rather than parameters like 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 non-random, though we note that in some cases the results are rich enough to permit the distribution of the data to vary across clusters. In fact, the literature has noted that the method described in Liang and Zeger (1986) may fail when cluster sizes are non-ignorable. See, in particular, Benhin et al. (Example 1, 2005). However, to our knowledge none of these papers 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 2 describes our setup and notation, including a formal description of our sampling framework and two parameters of interest. We then propose in Section 3 estimators for each of these two quantities and develop the requisite distributional approximations to use them for inference about each quantity. In Section 4, we demonstrate the finite-sample behavior of our proposed estimators in a small simulation study. Finally, in Section 5 we conduct an empirical exercise to demonstrate the practical relevance of our findings. Proofs of all results are included in the Appendix.

# 2 Setup and Notation

## 2.1 Notation and Sampling Framework

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

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

In addition, define $\mathcal{M}_g$ to be the (possibly random) subset of $\{1, \ldots, N_g\}$ corresponding to the observations within the $g$th 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 $|\mathcal{M}_g| = N_g$. Denote by $P_G$ the distribution of the observed data

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

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 $P_G$ is determined jointly by (1) together with the distribution of $A^{(G)} := (A_g : 1 \leq g \leq G)$ and $Q_G$, so we will state our assumptions below in terms of these two quantities.

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 $s \in \mathcal{S}$.

We begin by describing our assumptions on the distribution of $A^{(G)}$. For $s \in \mathcal{S}$, let

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

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 $\pi$. Our requirements on the treatment assignment mechanism are then summarized as follows:

**Assumption 2.1.** The treatment assignment mechanism is such that

(a) $W^{(G)} \perp\!\!\!\perp A^{(G)} | S^{(G)}$

(b) $\left\{ \left\{ \frac{D_G(s)}{\sqrt{G}} \right\}_{s \in \mathcal{S}} \middle| S^{(G)} \right\} \xrightarrow{d} N(0, \Sigma_D)$ a.s., where

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

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

Assumption 2.1 mirrors the assumption on assignment mechanisms considered in Bugni et al. (2018) for individual-level randomized experiments. Assumption 2.1.(a) requires that the assignment mechanism is a function only of the strata and an exogenous randomization device. Assumption 2.1.(b) requires that the randomization mechanism assigns treatments within each stratum so that the fraction of units being treated has a well behaved limiting distribution centered around the target proportion $\pi$. Bugni et al. (2018) provide several important examples of assignment mechanisms satisfying this assumption which are used routinely in economics. In particular, Assumption 2.1 is satisfied by stratified block randomization (see, e.g., Angelucci et al., 2015; Attanasio et al., 2015; Duflo et al., 2015), which assigns exactly a fraction $\pi$ 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". Despite its broad applicability, Assumption 2.1 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, for $a \in \{0, 1\}$, define

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

Further define $R_G(\mathcal{M}^{(G)}, Z^{(G)}, N^{(G)})$ to be the distribution of

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

where $\mathcal{M}^{(G)} := (\mathcal{M}_g : 1 \le g \le G)$, $Z^{(G)} := (Z_g : 1 \le g \le G)$ and $N^{(G)} := (N_g : 1 \le g \le G)$. Note that $Q_G$ is completely determined by $R_G(\mathcal{M}^{(G)}, Z^{(G)}, N^{(G)})$ and the distribution of $(\mathcal{M}^{(G)}, Z^{(G)}, N^{(G)})$. The following assumption states our requirements on $Q_G$ using this notation.

**Assumption 2.2.** The distribution $Q_G$ is such that

(a) $\{(\mathcal{M}_g, Z_g, N_g), 1 \le g \le G\}$ is an i.i.d. sequence of random variables.

(b) For some family of distributions $\{R(m, z, n) : (m, z, n) \in \text{supp}(\mathcal{M}_g, Z_g, N_g)\}$,

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

(c) $P\{|\mathcal{M}_g| \ge 1\} = 1$ and $E[N_g^2] < \infty$.

(d) For some $C < \infty$, $P\{E[Y_{i,g}^2(a)|N_g, Z_g] \le C$ for all $1 \le i \le N_g\} = 1$ for all $a \in \{0,1\}$ and $1 \le g \le G$.

(e) $\mathcal{M}_g \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \le i \le N_g) \bigm| Z_g, N_g$ for all $1 \le g \le G$.

(f) For $a \in \{0,1\}$ and $1 \le g \le G$,

$$E[\bar{Y}_g(a)|N_g] = E\left[\frac{1}{N_g} \sum_{1 \le i \le N_g} Y_{i,g}(a) \Big| N_g\right] \text{ w.p.1} \ .$$

Assumptions 2.2.(a)–(b) 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

$$\left\{\left(\bar{Y}_g(1), \bar{Y}_g(0), |\mathcal{M}_g|, Z_g, N_g\right), 1 \le g \le G\right\} \tag{3}$$

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

Assumptions 2.2.(c)–(d) 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 2.2.(c) does not rule out the possibility of observing arbitrarily large clusters, but does place restrictions on the heterogeneity of cluster sizes. For instance, two consequences of Assumptions 2.2.(a) and (c) are that[1]

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

and

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

[1] 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 et al. (2021).

which mirror heterogeneity restrictions imposed in the analysis of clustered data when cluster sizes are modeled as non-random (see for example Assumption 2 in Hansen and Lee, 2019).

Assumptions 2.1.(e)–(f) impose high-level restrictions on the two-stage sampling procedure. Assumption 2.1.(e) 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 2.2.(f) is a high-level assumption which guarantees that we can extrapolate from the observations that are sampled to the observations that are not sampled. Note that Assumptions 2.2.(e)–(f) are trivially satisfied whenever $\mathcal{M}_g = \{1, \ldots, N_g\}$ for all $1 \leq g \leq G$ with probability one, i.e., whenever all observations within each cluster are always sampled. Assumption 2.2.(f) is also satisfied whenever Assumption 2.2.(e) holds and there is sufficient homogeneity across the observations within each cluster in the sense that $P\{E[Y_{i,g}(a)|N_g, Z_g] = E[Y_{j,g}(a)|N_g, Z_g]$ for all $1 \leq i, j \leq N_g\} = 1$ for $a \in \{0, 1\}$. Finally, we show in Lemma 2.1 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 2.2.(e)–(f) are also satisfied.

**Lemma 2.1.** *Suppose that* $|\mathcal{M}_g| \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g) \bigm| Z_g, N_g$ *for all* $1 \leq g \leq G$*, 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 2.2.(e)–(f) are satisfied.*

**Remark 2.1.** 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 2.2.(c) at the cost of strengthening Assumption 2.2.(d) 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. ■

## 2.2 Parameters of Interest

In the context of the sampling framework described above, we consider two different parameters of interest. The parameters of interest can both be written in the form

$$E\left[w_g\left(\frac{1}{N_g}\sum_{1\leq i\leq N_g} Y_{i,g}(1) - Y_{i,g}(0)\right)\right]$$

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

$$\theta_1(Q_G) := E\left[\frac{1}{N_g}\sum_{1\leq i\leq N_g} Y_{i,g}(1) - Y_{i,g}(0)\right] . \tag{4}$$

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

$$\theta_2(Q_G) := E\left[\frac{1}{E[N_g]}\sum_{1\le i\le N_g} Y_{i,g}(1) - Y_{i,g}(0)\right] . \tag{5}$$

We refer to this quantity as the size-weighted cluster-level average treatment effect. $\theta_2(Q_G)$ can be thought of as the average treatment effect where individuals are the units of interest. Note that Assumptions 2.2.(a)–(b) 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 $G$. Accordingly, in what follows we simply denote $\theta_1 = \theta_1(Q_G)$, $\theta_2 = \theta_2(Q_G)$.

If treatment effects are heterogeneous and cluster sizes are non-ignorable, then $\theta_1$ and $\theta_2$ are indeed distinct parameters. We illustrate this with a simple numerical exercise in Example 2.1 below.

**Example 2.1.** Suppose clusters represent classrooms, and that we have two types of classrooms: "big" with $N_g = 40$ students and "small" with $N_g = 10$ students. Suppose that $Y_{i,g}(1) - Y_{i,g}(0) = 1$ for all individuals in a "big" classroom and $Y_{i,g}(1) - Y_{i,g}(0) = -2$ for all individuals in a "small" classroom, so that

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

and also

$$E\left[\sum_{1\le i\le N_g}(Y_{i,g}(1) - Y_{i,g}(0)) \,\middle|\, N_g = 40\right] = 40$$
$$E\left[\sum_{1\le i\le N_g}(Y_{i,g}(1) - Y_{i,g}(0)) \,\middle|\, N_g = 10\right] = -20 .$$

Suppose that both types of classrooms are equally likely, i.e.,

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

Given these calculations, it is straightforward to show that the equally-weighted cluster-level average treatment effect is given by $\theta_1 = -1/2$, whereas the size-weighted cluster-level average treatment effect is given by $\theta_2 = 2/5$. In particular, we see in this example that the equally-weighted cluster-level average treatment effect is negative while the size-weighted cluster-level average treatment effect is positive. ■

**Remark 2.2.** While we generally expect $\theta_1$ and $\theta_2$ to be distinct, they are equivalent in some special cases. For example, if all clusters are of the same fixed size $k$, i.e., $P\{N_g = k\} = 1$, 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 $\theta_1 = \theta_2$. Generalizing these two extreme cases, we have that $\theta_1 = \theta_2$ if cluster sizes are ignorable (i.e., $R(m, z, n)$ does not depend on $m$ and $n$) and treatment effects are sufficiently homogeneous in the sense that $P\{E[Y_{i,g}(1) - Y_{i,g}(0)] = E[Y_{j,g}(1) - Y_{j,g}(0)]$ for all $1 \leq i, j \leq N_g\} = 1$. ■

# 3 Main Results

## 3.1 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

$$\hat{\theta}_G^{\text{alt}} := \frac{\sum_{1 \leq g \leq G} \sum_{i \in \mathcal{M}_g} Y_{i,g} A_g}{\sum_{1 \leq g \leq G} |\mathcal{M}_g| A_g} - \frac{\sum_{1 \leq g \leq G} \sum_{i \in \mathcal{M}_g} Y_{i,g}(1 - A_g)}{\sum_{1 \leq g \leq G} |\mathcal{M}_g|(1 - A_g)} . \tag{6}$$

Note that $\hat{\theta}_G^{\text{alt}}$ may be obtained as the estimator of the coefficient on $A_g$ in an ordinary least squares regression of $Y_{i,g}$ on a constant and $A_g$.

The following theorem derives the probability limit of this estimator:

**Theorem 3.1.** *Under Assumptions 2.1 and 2.2,*

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

*as* $G \to \infty$.

The quantity $\vartheta$ corresponds to a *sample*-weighted cluster-level average treatment effect. When treatment effects are heterogeneous and cluster sizes are non-ignorable, $\vartheta$ need not equal either $\theta_1$ defined in (4) or $\theta_2$ defined in (5). We illustrate this in the context of our previous numerical example in Example 3.1 below. As a result, unless the experimenter is interested in a distinct weighting of the cluster-level treatment effects which differs from the population-level weightings proposed in Section 2.2, care must be taken when interpreting $\hat{\theta}_G^{\text{alt}}$. We note, however, that $\vartheta$ is in fact equal to either $\theta_1$ or $\theta_2$ for some sampling designs. Specifically, if $|\mathcal{M}_g| = k$ for all $1 \leq g \leq G$, then $\vartheta$ is equal to $\theta_1$; if $\mathcal{M}_g = \{1, 2, \ldots, N_g\}$ for all $1 \leq g \leq G$ with probability one, then $\vartheta$ is equal to $\theta_2$.

**Example 3.1.** Recall the setting of Example 2.1. Suppose further that the experimenter samples $|\mathcal{M}_g| = 5$ students at random without replacement from each "small" classroom, and $|\mathcal{M}_g| = 10$ students at random without replacement from each "big" classroom. It is now straightforward to show that $\vartheta = 0$, which is not equal to either of the parameters defined in Section 2.2. ■

**Remark 3.1.** Another sampling design for which $\vartheta$ is approximately equal to $\theta_2$ is when $|\mathcal{M}_g| = \lfloor \gamma N_g \rfloor$ for some $0 < \gamma < 1$ and $N_g$ only takes on "large" values. To see this, note in this case it can be shown using

Assumption 2.2(f) and the law of iterated expectations that

$$\vartheta = \frac{E\left[\frac{\lfloor \gamma N_g \rfloor}{N_g} E\left[\sum_{1 \leq i \leq N_g} Y_{i,g}(1) - Y_{i,g}(0) | N_g\right]\right]}{E\left[\frac{\lfloor \gamma N_g \rfloor}{N_g} N_g\right]} ,$$

from which the desired conclusion follows immediately. ■

As a consequence of Theorem 3.1, in what follows we consider alternative estimators which are generally consistent for $\theta_1$ and $\theta_2$ without imposing additional restrictions on the sampling procedure.

## 3.2 Equally-weighted Cluster-level Average Treatment Effect

In this section, we consider estimation of $\theta_1$ defined in (4). To this end, consider the following difference-in-"average of averages" estimator:

$$\hat{\theta}_{1,G} := \frac{\sum_{1 \leq g \leq G} \bar{Y}_g A_g}{\sum_{1 \leq g \leq G} A_g} - \frac{\sum_{1 \leq g \leq G} \bar{Y}_g (1 - A_g)}{\sum_{1 \leq g \leq G} (1 - A_g)} , \tag{7}$$

where

$$\bar{Y}_g = \frac{1}{|\mathcal{M}_g|} \sum_{i \in \mathcal{M}_g} Y_{i,g} . \tag{8}$$

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

$$\begin{aligned} \hat{\mu}^C_{G,a}(s) &:= \frac{1}{\sum_{1 \leq g \leq G} I\{A_g = a, S_g = s\}} \sum_{1 \leq g \leq G} C_g I\{A_g = a, S_g = s\} , \\ \hat{\mu}^C_G(s) &:= \frac{1}{\sum_{1 \leq g \leq G} I\{S_g = s\}} \sum_{1 \leq g \leq G} C_g I\{S_g = s\} , \\ \hat{\mu}^C_{G,a} &:= \frac{1}{\sum_{1 \leq g \leq G} I\{A_g = a\}} \sum_{1 \leq g \leq G} C_g I\{A_g = a\} . \end{aligned}$$

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 an ordinary least squares regression of $\bar{Y}_g$ on a constant and $A_g$. As such, we can view $\hat{\theta}_{1,G}$ as estimating the treatment effect for an individual-level randomized experiment where the clusters are themselves the units of interest. The following theorem derives the asymptotic behavior of this estimator.

**Theorem 3.2.** *Under Assumptions 2.1 and 2.2,*

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

*as $G \to \infty$, where*

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

*with*

$$\begin{aligned} \bar{Y}_g^\dagger(a) &:= \bar{Y}_g(a) - E[\bar{Y}_g(a)|S_g] , \\ \bar{m}_a(S_g) &:= E[\bar{Y}_g(a)|S_g] - E[\bar{Y}_g(a)] , \end{aligned} \tag{9}$$

*and $\pi$, $\tau(\cdot)$ defined as in Assumption 2.1.*

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 all $s$ (i.e., strong balance) and least efficient when $\tau(s) = \pi(1-\pi)$. The next result shows that, as a consequence, the probability limit of the standard heteroskedasticity-robust variance estimator is generally too large relative to $\sigma_1^2$.

**Theorem 3.3.** *Let $\tilde{\sigma}_{1,G}^2$ denote 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*

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

*where*

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

*Then under Assumptions 2.1 and 2.2,*

$$\tilde{\sigma}_{1,G}^2 \xrightarrow{p} \frac{1}{\pi} Var[\bar{Y}_g(1)] + \frac{1}{1-\pi} Var[\bar{Y}_g(0)] \ge \sigma_1^2 , \tag{10}$$

*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 2.1 is satisfied with $\tau(s) = \pi(1-\pi)$ for every $s \in \mathcal{S}$. As such we obtain from Theorem 3.3 that in this case $\tilde{\sigma}_{1,G}^2$ is a consistent estimator of $\sigma_1^2$.

To facilitate the use of Theorem 3.2 for inference about $\theta_1$, we now provide an estimator of $\sigma_1^2$ which is consistent in general. For any $s \in \mathcal{S}$, let

$$G(s) := \sum_{g \in G} 1\{S_g = s\} . \tag{11}$$

Then, define the following estimators:

$$\hat{\zeta}_{\bar{Y}}^2(\pi) := \frac{1}{\pi}\left(\hat{\mu}_{G,1}^{\bar{Y}^2} - \sum_{s\in\mathcal{S}} \frac{G(s)}{G}\hat{\mu}_{G,1}^{\bar{Y}}(s)^2\right) + \frac{1}{1-\pi}\left(\hat{\mu}_{G,0}^{\bar{Y}^2} - \sum_{s\in\mathcal{S}} \frac{G(s)}{G}\hat{\mu}_{G,0}^{\bar{Y}}(s)^2\right) \tag{12}$$

$$\hat{\zeta}_H^2 := \sum_{s\in\mathcal{S}} \frac{G(s)}{G}\left(\left(\hat{\mu}_{G,1}^{\bar{Y}}(s) - \hat{\mu}_{G,1}^{\bar{Y}}\right) - \left(\hat{\mu}_{G,0}^{\bar{Y}}(s) - \hat{\mu}_{G,0}^{\bar{Y}}\right)\right)^2 \tag{13}$$

$$\hat{\zeta}_A^2(\pi) := \sum_{s\in\mathcal{S}} \tau(s)\frac{G(s)}{G}\left(\frac{1}{\pi}\left(\hat{\mu}_{G,1}^{\bar{Y}}(s) - \hat{\mu}_1^{\bar{Y}}\right) + \frac{1}{1-\pi}\left(\hat{\mu}_{G,0}^{\bar{Y}}(s) - \hat{\mu}_{G,0}^{\bar{Y}}\right)\right)^2, \tag{14}$$

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

The following theorem establishes the consistency of $\hat{\sigma}_{1,G}^2$ for $\sigma_1^2$. 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 c.d.f.

**Theorem 3.4.** *Under Assumptions 2.1 and 2.2,*

$$\hat{\sigma}_{1,G}^2 \xrightarrow{p} \sigma_1^2$$

*as $G \to \infty$. Thus, for $\sigma_1^2 > 0$ 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 3.2.** A sufficient condition under which $\sigma_1^2 > 0$ 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 $\sigma_1^2 > 0$ except in pathological cases such as when the distribution of outcomes is degenerate or in cases with perfect negative within-cluster correlation. ■

**Remark 3.3.** An inspection of the proofs of Theorems 3.2–3.4 reveals that the restriction on cluster size heterogeneity implied by Assumption 2.2.(c) is not necessary for establishing these results. Recent work by Sasaki and Wang (2022) exploits a similar observation to study inference for the parameters in a linear model with clustered dependence, where cluster sizes are also modeled as random, but where the distribution of cluster sizes features heavy tails. ■

**Remark 3.4.** 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 A.4 we describe one such adjustment strategy. ■

### 3.3 Size-weighted Cluster-level Average Treatment Effect

In this section, we consider estimation of $\theta_2$ defined in (5). To this end, consider the following difference-in-"weighted average of averages" estimator:

$$\hat{\theta}_{2,G} := \frac{\sum_{1\le g\le G} \bar{Y}_g N_g A_g}{\sum_{1\le g\le G} N_g A_g} - \frac{\sum_{1\le g\le G} \bar{Y}_g N_g (1-A_g)}{\sum_{1\le g\le G} N_g (1-A_g)} , \tag{15}$$

where $\bar{Y}_g$ is defined as in (8). Note that $\hat{\theta}_{2,G}$ may be obtained as the estimator of the coefficient on $A_g$ in a weighted least squares regression of $Y_{i,g}$ on a constant and $A_g$ with weights equal to $\sqrt{N_g/|\mathcal{M}_g|}$. In the special case where $\mathcal{M}_g = \{1,2,\dots,N_g\}$ for all $1 \le g \le G$ with probability one, 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 3.5.** *Under Assumptions 2.1 and 2.2,*

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

*as $G \to \infty$, where*

$$\sigma_2^2 := \frac{1}{\pi} Var[\tilde{Y}_g^\dagger(1)] + \frac{1}{1-\pi} Var[\tilde{Y}_g^\dagger(0)] + E[(\tilde{m}_1(S_g) - \tilde{m}_0(S_g))^2] + E\left[\tau(S_g)\left(\frac{1}{\pi}\tilde{m}_1(S_g) + \frac{1}{1-\pi}\tilde{m}_0(S_g)\right)^2\right] , \tag{16}$$

*with*

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

*and $\pi$, $\tau(\cdot)$ defined as in Assumption 2.1.*

**Remark 3.5.** 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 $N^{(G)}$. As explained in Remark 3.1, however, $\theta_2$ is in some instances approximately equal to $\vartheta$, which may be consistently estimated using $\hat{\theta}_G^{\text{alt}}$. ■

**Remark 3.6.** It is interesting to compare $\sigma_2^2$ to the variance of the difference-in-means estimator from a finite-population analysis. For instance, in the special case where $\mathcal{M}_g = \{1,2,\dots,N_g\}$ and $|\mathcal{S}| = 1$, $\sigma_2^2$ could alternatively be written as

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

with

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

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

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

where

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

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

**Remark 3.7.** As discussed in Remark 2.2, $\theta_1 = \theta_2$ 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 $s$, and show that the probability limit of the cluster-robust variance estimator is generally too large relative to $\sigma^2_2$.

**Theorem 3.6.** *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 $\sqrt{N_g/|\mathcal{M}_g|}$. This estimator can be written as*

$$\tilde{\sigma}^2_{2,G} = \tilde{\sigma}^2_{2,G}(1) + \tilde{\sigma}^2_{2,G}(0) , \tag{17}$$

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

$$\tilde{\sigma}^2_{2,G}(a) := \frac{1}{\left(\frac{1}{G}\sum_{1\le g\le G} N_g I\{A_g = a\}\right)^2}\frac{1}{G}\sum_{1\le g\le G}\left[\left(\frac{N_g}{|\mathcal{M}_g|}\right)^2 I\{A_g = a\}\left(\sum_{i\in\mathcal{M}_g}\hat{\epsilon}_{i,g}(a)\right)^2\right] , \tag{18}$$

*with*

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

*Then, under Assumptions 2.1 and 2.2,*

$$\tilde{\sigma}^2_{2,G} \xrightarrow{p} \frac{1}{\pi} Var[\tilde{Y}_g(1)] + \frac{1}{1-\pi} Var[\tilde{Y}_g(0)] \geq \sigma^2_2 \tag{19}$$

*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 3.5 for inference about $\theta_2$, we now provide an estimator for $\sigma^2_2$ which is consistent in general. Similarly to Bai et al. (2022) and Liu (2023), we construct our estimator using a feasible analog of $\tilde{Y}_g(a)$ given by

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

We then define the following estimators:

$$\hat{\xi}^2_{\tilde{Y}}(\pi) := \frac{1}{\pi} \left( \hat{\mu}^{\hat{Y}^2}_{G,1} - \sum_{s \in \mathcal{S}} \frac{G(s)}{G} \hat{\mu}^{\hat{Y}}_{G,1}(s)^2 \right) + \frac{1}{1-\pi} \left( \hat{\mu}^{\hat{Y}^2}_{G,0} - \sum_{s \in \mathcal{S}} \frac{G(s)}{G} \hat{\mu}^{\hat{Y}}_{G,0}(s)^2 \right) \tag{20}$$

$$\hat{\xi}^2_H := \sum_{s \in \mathcal{S}} \frac{G(s)}{G} \left( \left( \hat{\mu}^{\hat{Y}}_{G,1}(s) - \hat{\mu}^{\hat{Y}}_{G,1} \right) - \left( \hat{\mu}^{\hat{Y}}_{G,0}(s) - \hat{\mu}^{\hat{Y}}_{G,0} \right) \right)^2 \tag{21}$$

$$\hat{\xi}^2_A(\pi) := \sum_{s \in \mathcal{S}} \tau(s) \frac{G(s)}{G} \left( \frac{1}{\pi} \left( \hat{\mu}^{\hat{Y}}_{G,1}(s) - \hat{\mu}^{\hat{Y}}_{G,1} \right) + \frac{1}{1-\pi} \left( \hat{\mu}^{\hat{Y}}_{G,0}(s) - \hat{\mu}^{\hat{Y}}_{G,0} \right) \right)^2 \ , \tag{22}$$

and set $\hat{\sigma}^2_{2,G} := \hat{\xi}^2_{\tilde{Y}}(\pi) + \hat{\xi}^2_H + \hat{\xi}^2_A(\pi)$. The following theorem establishes the consistency of $\hat{\sigma}^2_{2,G}$ for $\sigma^2_2$. In the statement of the theorem, we again make use of the notation introduced preceding Theorem 3.4.

**Theorem 3.7.** *Under Assumptions 2.1 and 2.2,*

$$\hat{\sigma}^2_{2,G} \xrightarrow{p} \sigma^2_2$$

*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 3.8.** 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 3.2, we expect this to hold outside of pathological cases. ■

**Remark 3.9.** An alternative estimator of $\theta_2$ is given by

$$\hat{\theta}_{2,G}^{\text{alt}} := \frac{\frac{1}{G}\sum_{1\le g\le G}\bar{Y}_g N_g A_g}{\bar{N}_G \bar{A}_G} - \frac{\frac{1}{G}\sum_{1\le g\le G}\bar{Y}_g N_g (1-A_g)}{\bar{N}_G(1-\bar{A}_G)} \ , \tag{23}$$

where $\bar{N}_G := \frac{1}{G}\sum_{1\le g\le G} N_g$ and $\bar{A}_G := \frac{1}{G}\sum_{1\le g\le G} A_g$. Note that $\hat{\theta}_{2,G}^{\text{alt}}$ may be obtained as the estimator of the coefficient of $A_g$ in an ordinary least squares regression of $\Psi_{g,G} := \bar{Y}_g(N_g/\bar{N}_G)$ on a constant and $A_g$. By arguing as in the proof of Theorem 3.5, it is possible to derive the asymptotic behavior of this estimator as well, but we do not pursue this further in this paper. ■

**Remark 3.10.** As mentioned earlier, $\hat{\theta}_{2,G}$ can be obtained as the estimator of the coefficient on $A_g$ in a weighted least squares regression of $Y_{i,g}$ on a constant and $A_g$. As in Remark 3.4, a natural next step to consider would be the inclusion additional baseline covariates in this weighted linear regression. Unlike the case of $\hat{\theta}_{1,G}$, an adjustment that ensures efficiency gains cannot be established by immediately appealing to techniques developed in the context of individual-level randomized experiments. For this reason, we leave the study of such an adjustment strategy to future work. ■

# 4 Simulations

In this section, we illustrate the results in Section 3 with a simulation study. In all cases, data are generated as

$$Y_{i,g}(a) = \eta_g(a) Z_{g,1} + \tilde{m}_a(Z_{g,2}) + U_{i,g}(a) \ , \tag{24}$$

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

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

- 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. To describe them, let $BB(a,b,n_{\text{supp}})$ be the Beta-Binomial distribution with parameters $a$ and $b$ and support on $\{0,\ldots,n_{\text{supp}}\}$. We then define

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

for the following values of $(a,b)$ and $n_{\text{supp}}$,

- $(a,b) = (1,1)$: uniform pmf on 10 to $N_{\max} = 10(n_{\text{supp}}+1)$.

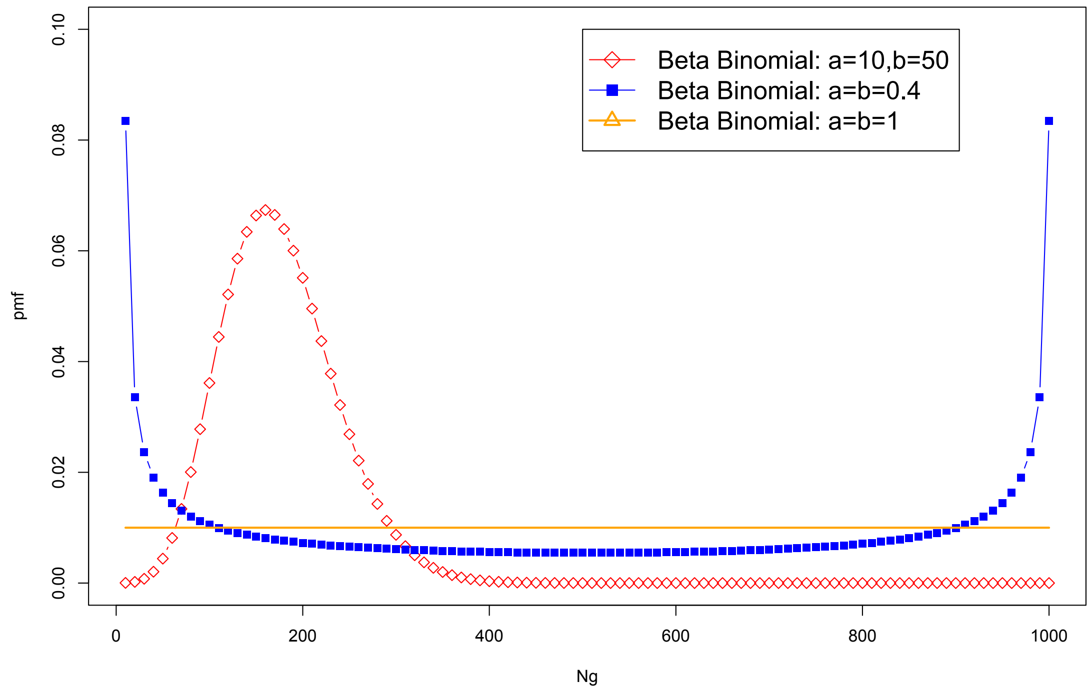


Figure 1: Three probability mass functions (pmf) of $N_g$ when $N_{\max} = 1000$

- $(a, b) = (0.4, 0.4)$: 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 $g$th 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 9 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 $p_z \equiv P\{Z_{g,1} = 1\} = 1/2$.
- **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 $Z_{g,2} \sim \text{Beta}(2, 2)$ (re-centered and re-scaled by the population mean and variance to have mean zero and variance one).

Finally, treatment assignment $A_g$ follows a covariate-adaptive randomization (CAR) mechanism based on stratified block randomization with $\pi = \frac{1}{2}$ 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 (24), as well as the two CAR designs, follow closely the original designs for covariate-adaptive randomization with individual level data considered in Bugni et al. (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

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

For each of the above 9 specifications and for each design of $Z_{g,1}$ and treatment assignment mechanism, we report the true values of $(\theta_1, \theta_2)$ defined in (4) and (5), the average across simulations of the estimated values $(\hat{\theta}_{1,G}, \hat{\theta}_{2,G})$ defined in (7) and (15), the average across simulations of the estimated standard deviations $(\hat{\sigma}_{1,G}, \hat{\sigma}_{2,G})$ defined in Sections 3.2 and 3.3, and the empirical coverage of the 95% confidence intervals defined in Theorems 3.4 and 3.7. The results of our simulations are presented in Tables 1 to 6, where in all cases we conducted 5,000 replications. In each table we find that our empirical coverage probabilties are close to 95% in all cases.

| CAR-1 $\mid$ Design 1 | | True values | | Estimated | | Stand. deviations | | Cov. Prob. | |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{M}_g$ | $N_g$ | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | -0.0016 | 0.0016 | 4.2885 | 4.9375 | 0.9440 | 0.9426 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | -0.0080 | -0.0070 | 4.2864 | 5.2952 | 0.9454 | 0.9310 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | -0.0001 | -0.0010 | 4.2808 | 4.5852 | 0.9444 | 0.9486 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.0000 | 0.0007 | -0.0000 | 4.3297 | 4.9780 | 0.9426 | 0.9330 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | -0.0008 | 0.0015 | 4.3385 | 5.3582 | 0.9460 | 0.9438 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | 0.0023 | 0.0063 | 4.3414 | 4.6545 | 0.9436 | 0.9440 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | 0.0090 | 0.0087 | 4.2827 | 4.9091 | 0.9346 | 0.9338 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | -0.0089 | -0.0044 | 4.2871 | 5.3089 | 0.9426 | 0.9376 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | 0.0049 | 0.0064 | 4.2994 | 4.5949 | 0.9436 | 0.9470 |
| CAR-1 $\mid$ Design 2 | | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.4900 | -0.0036 | 0.4819 | 4.2792 | 4.7416 | 0.9474 | 0.9454 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6581 | -0.0142 | 0.6387 | 4.2870 | 5.0338 | 0.9458 | 0.9424 |
| | $Bb(10,50)$ | -0.1407 | 0.1625 | -0.0822 | 0.2172 | 4.2825 | 4.5250 | 0.9342 | 0.9440 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.4900 | -0.0044 | 0.4844 | 4.3439 | 4.8198 | 0.9448 | 0.9454 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6581 | -0.0168 | 0.6366 | 4.3399 | 5.1226 | 0.9402 | 0.9426 |
| | $Bb(10,50)$ | -0.1407 | 0.1625 | -0.0807 | 0.2173 | 4.3346 | 4.5764 | 0.9480 | 0.9526 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.4900 | -0.0129 | 0.4673 | 4.2884 | 4.7446 | 0.9548 | 0.9428 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6581 | -0.0147 | 0.6366 | 4.2852 | 5.0404 | 0.9404 | 0.9462 |
| | $Bb(10,50)$ | -0.1407 | 0.1625 | -0.0800 | 0.2201 | 4.2922 | 4.5273 | 0.9416 | 0.9398 |

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.

| CAR-1 $\mid$ Design 1 | | True values | | Estimated | | Stand. deviations | | Cov. Prob. | |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{M}_g$ | $N_g$ | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | -0.0004 | -0.0057 | 4.2824 | 4.9264 | 0.9408 | 0.9382 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | 0.0002 | 0.0008 | 4.2835 | 5.3107 | 0.9388 | 0.9388 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | -0.0061 | -0.0080 | 4.2803 | 4.5365 | 0.9438 | 0.9436 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.0000 | -0.0075 | -0.0102 | 4.3407 | 4.9950 | 0.9374 | 0.9412 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | 0.0009 | 0.0052 | 4.3405 | 5.3903 | 0.9386 | 0.9426 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | -0.0129 | -0.0147 | 4.3358 | 4.5938 | 0.9442 | 0.9490 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | -0.0115 | -0.0146 | 4.2837 | 4.9416 | 0.9438 | 0.9442 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | -0.0095 | -0.0077 | 4.2907 | 5.3277 | 0.9386 | 0.9392 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | -0.0022 | -0.0051 | 4.2841 | 4.5295 | 0.9434 | 0.9446 |
| CAR-1 $\mid$ Design 2 | | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.4950 | 0.0005 | 0.4895 | 4.2746 | 4.7550 | 0.9396 | 0.9410 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | 0.0039 | 0.6711 | 4.2855 | 5.0689 | 0.9480 | 0.9510 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0691 | 0.2024 | 4.2806 | 4.4677 | 0.9414 | 0.9486 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.4950 | -0.0071 | 0.4834 | 4.3384 | 4.8310 | 0.9456 | 0.9442 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | -0.0054 | 0.6533 | 4.3341 | 5.1559 | 0.9456 | 0.9446 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0687 | 0.1973 | 4.3408 | 4.5444 | 0.9426 | 0.9436 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.4950 | -0.0192 | 0.4723 | 4.2826 | 4.7515 | 0.9456 | 0.9440 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | -0.0060 | 0.6638 | 4.2876 | 5.0702 | 0.9432 | 0.9460 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0708 | 0.1987 | 4.2862 | 4.4822 | 0.9446 | 0.9472 |

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.

| CAR-1 \| Design 1 | | True values | | Estimated | | Stand. deviations | | Cov. Prob. | |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{M}_g$ | $N_g$ | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | 0.0003 | -0.0001 | 4.3688 | 5.0513 | 0.9480 | 0.9546 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | -0.0009 | -0.0010 | 4.3740 | 5.4474 | 0.9494 | 0.9530 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | 0.0008 | 0.0011 | 4.3690 | 4.6261 | 0.9532 | 0.9492 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.0000 | 0.0001 | -0.0003 | 4.4315 | 5.1258 | 0.9542 | 0.9560 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | 0.0006 | 0.0013 | 4.4333 | 5.5313 | 0.9542 | 0.9510 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | 0.0015 | 0.0014 | 4.4330 | 4.6940 | 0.9538 | 0.9546 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | 0.0018 | 0.0024 | 4.3725 | 5.0541 | 0.9500 | 0.9468 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | -0.0008 | -0.0006 | 4.3785 | 5.4505 | 0.9548 | 0.9522 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | 0.0025 | 0.0022 | 4.3766 | 4.6319 | 0.9580 | 0.9598 |
| CAR-1 \| Design 2 | | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.4950 | 0.0019 | 0.4964 | 4.3680 | 4.8506 | 0.9598 | 0.9606 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | -0.0002 | 0.6675 | 4.3750 | 5.1796 | 0.9548 | 0.9552 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0642 | 0.2088 | 4.3677 | 4.5609 | 0.9492 | 0.9508 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.4950 | 0.0009 | 0.4961 | 4.4332 | 4.9315 | 0.9532 | 0.9562 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | -0.0006 | 0.6689 | 4.4332 | 5.2654 | 0.9586 | 0.9588 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0627 | 0.2104 | 4.4299 | 4.6277 | 0.9582 | 0.9592 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.4950 | 0.0002 | 0.4956 | 4.3732 | 4.8567 | 0.9586 | 0.9578 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | -0.0000 | 0.6695 | 4.3784 | 5.1788 | 0.9586 | 0.9614 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0635 | 0.2094 | 4.3744 | 4.5676 | 0.9538 | 0.9576 |

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.

| CAR-2 \| Design 1 | | True values | | Estimated | | Stand. deviations | | Cov. Prob. | |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{M}_g$ | $N_g$ | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | 0.0043 | -0.0040 | 4.2769 | 4.9142 | 0.9422 | 0.9450 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | -0.0001 | -0.0029 | 4.2760 | 5.2566 | 0.9366 | 0.9372 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | -0.0013 | -0.0021 | 4.2917 | 4.5993 | 0.9474 | 0.9486 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.0000 | -0.0073 | -0.0031 | 4.3318 | 4.9799 | 0.9456 | 0.9428 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | -0.0071 | -0.0118 | 4.3325 | 5.3390 | 0.9384 | 0.9408 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | 0.0031 | 0.0063 | 4.3465 | 4.6611 | 0.9334 | 0.9432 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | -0.0061 | -0.0097 | 4.2783 | 4.9082 | 0.9460 | 0.9430 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | -0.0006 | -0.0058 | 4.2903 | 5.2822 | 0.9380 | 0.9468 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | -0.0078 | -0.0102 | 4.3018 | 4.6021 | 0.9426 | 0.9458 |
| CAR-2 \| Design 2 | | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.4900 | -0.0027 | 0.4915 | 4.1664 | 4.6658 | 0.9540 | 0.9476 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6581 | -0.0029 | 0.6479 | 4.1693 | 4.9832 | 0.9518 | 0.9498 |
| | $Bb(10,50)$ | -0.1407 | 0.1625 | -0.0897 | 0.2121 | 4.1669 | 4.4264 | 0.9462 | 0.9484 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.4900 | -0.0015 | 0.4917 | 4.2266 | 4.7462 | 0.9516 | 0.9488 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6581 | -0.0006 | 0.6563 | 4.2239 | 5.0699 | 0.9544 | 0.9518 |
| | $Bb(10,50)$ | -0.1407 | 0.1625 | -0.0795 | 0.2174 | 4.2293 | 4.4920 | 0.9464 | 0.9490 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.4900 | -0.0035 | 0.4870 | 4.1682 | 4.6798 | 0.9496 | 0.9484 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6581 | -0.0036 | 0.6532 | 4.1737 | 4.9868 | 0.9578 | 0.9524 |
| | $Bb(10,50)$ | -0.1407 | 0.1625 | -0.0877 | 0.2136 | 4.1861 | 4.4456 | 0.9516 | 0.9504 |

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.

| CAR-2 \| Design 1 | | True values | | Estimated | | Stand. deviations | | Cov. Prob. | |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{M}_g$ | $N_g$ | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | -0.0083 | -0.0120 | 4.2775 | 4.9251 | 0.9394 | 0.9402 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | -0.0143 | -0.0153 | 4.2801 | 5.2966 | 0.9434 | 0.9384 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | -0.0086 | -0.0115 | 4.2897 | 4.5346 | 0.9422 | 0.9450 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.0000 | -0.0152 | -0.0175 | 4.3419 | 4.9979 | 0.9424 | 0.9426 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | -0.0058 | -0.0066 | 4.3391 | 5.3822 | 0.9428 | 0.9410 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | -0.0008 | -0.0005 | 4.3482 | 4.6036 | 0.9438 | 0.9480 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | -0.0023 | -0.0033 | 4.2840 | 4.9329 | 0.9440 | 0.9396 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | 0.0044 | 0.0003 | 4.2906 | 5.3131 | 0.9408 | 0.9424 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | -0.0142 | -0.0118 | 4.2884 | 4.5387 | 0.9418 | 0.9404 |
| CAR-2 \| Design 2 | | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.4950 | -0.0046 | 0.4938 | 4.1610 | 4.6755 | 0.9546 | 0.9514 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | -0.0091 | 0.6566 | 4.1644 | 4.9964 | 0.9506 | 0.9452 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0715 | 0.1989 | 4.1593 | 4.3726 | 0.9474 | 0.9498 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.4950 | -0.0069 | 0.4893 | 4.2275 | 4.7582 | 0.9514 | 0.9482 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | -0.0063 | 0.6612 | 4.2194 | 5.0787 | 0.9584 | 0.9520 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0746 | 0.1989 | 4.2219 | 4.4408 | 0.9524 | 0.9580 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.4950 | -0.0105 | 0.4836 | 4.1551 | 4.6726 | 0.9540 | 0.9492 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | -0.0062 | 0.6631 | 4.1645 | 4.9924 | 0.9530 | 0.9512 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0640 | 0.2074 | 4.1660 | 4.3863 | 0.9446 | 0.9506 |

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.

| CAR-2 \| Design 1 | | True values | | Estimated | | Stand. deviations | | Cov. Prob. | |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{M}_g$ | $N_g$ | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | 0.0007 | -0.0006 | 4.3737 | 5.0442 | 0.9570 | 0.9582 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | 0.0029 | 0.0033 | 4.3788 | 5.4297 | 0.9548 | 0.9568 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | 0.0001 | -0.0002 | 4.3801 | 4.6374 | 0.9500 | 0.9502 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.0000 | 0.0009 | 0.0004 | 4.4386 | 5.1204 | 0.9548 | 0.9520 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | 0.0015 | 0.0026 | 4.4376 | 5.5128 | 0.9562 | 0.9550 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | 0.0007 | 0.0005 | 4.4435 | 4.7054 | 0.9550 | 0.9536 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.0000 | 0.0019 | 0.0020 | 4.3780 | 5.0463 | 0.9564 | 0.9540 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.0000 | 0.0006 | -0.0002 | 4.3836 | 5.4340 | 0.9610 | 0.9566 |
| | $Bb(10,50)$ | 0.0000 | 0.0000 | 0.0005 | 0.0004 | 4.3865 | 4.6416 | 0.9496 | 0.9518 |
| CAR-2 \| Design 2 | | $\theta_1$ | $\theta_2$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\sigma}_{1,G}$ | $\hat{\sigma}_{2,G}$ | $CS_{1,G}$ | $CS_{2,G}$ |
| $N_g$ | $Bb(1,1)$ | 0.0000 | 0.4950 | 0.0011 | 0.4960 | 4.2082 | 4.7586 | 0.9644 | 0.9580 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | 0.0001 | 0.6686 | 4.2122 | 5.0922 | 0.9680 | 0.9598 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0638 | 0.2100 | 4.2169 | 4.4390 | 0.9624 | 0.9588 |
| 10 | $Bb(1,1)$ | 0.0000 | 0.4950 | 0.0003 | 0.4959 | 4.2745 | 4.8394 | 0.9592 | 0.9582 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | 0.0006 | 0.6692 | 4.2742 | 5.1821 | 0.9648 | 0.9588 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0633 | 0.2097 | 4.2829 | 4.5094 | 0.9576 | 0.9614 |
| $\gamma N_g$ | $Bb(1,1)$ | 0.0000 | 0.4950 | 0.0010 | 0.4965 | 4.2125 | 4.7615 | 0.9622 | 0.9572 |
| | $Bb(0.4,0.4)$ | 0.0000 | 0.6690 | 0.0005 | 0.6687 | 4.2173 | 5.0962 | 0.9622 | 0.9618 |
| | $Bb(10,50)$ | -0.0635 | 0.2100 | -0.0622 | 0.2111 | 4.2239 | 4.4454 | 0.9604 | 0.9586 |

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.

# 5 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 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 intervention was implemented only for eight months (May 2010 - 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 three-fold 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 5.1.** This application features a setting where 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 one, and the estimator coincides with the difference-in-means estimator, i.e., $\hat{\theta}_{2,G} = \hat{\theta}_G^{\text{alt}}$. ■

Celhay et al. (2019) collected data before, during, and after the intervention period. The pre-intervention 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 number 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. From the standpoint of our contribution, these cluster sizes are likely non-ignorable, as they may reflect essential factors such as the clinic's quality or the nearby population's size. Finally, the post-intervention period goes between January 2011 and March 2012.

In Table 7, we show our estimates of the equally-weighted and size-weighted average treatment effect (ATE) and their standard errors. We ran separate estimations using the data in the pre-intervention, intervention, and post-intervention 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

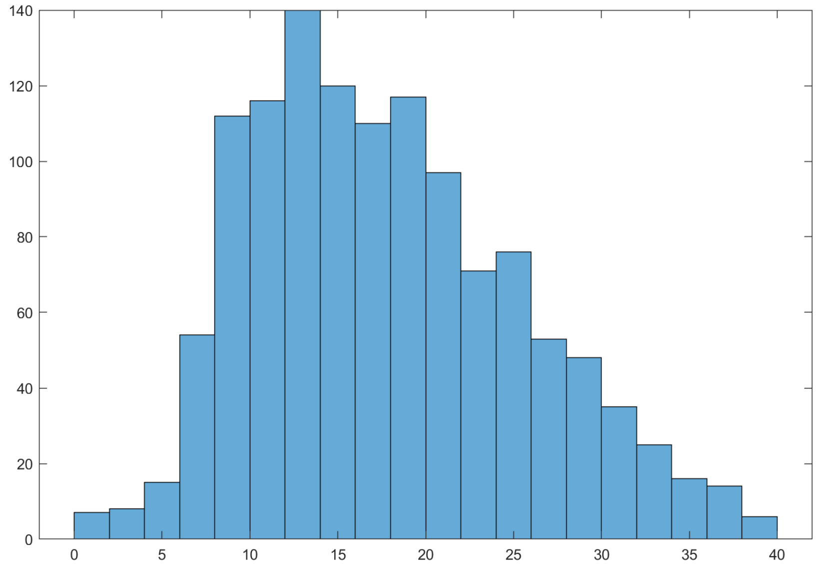


Figure 2: Histogram of the week of the first prenatal visit during the pre-intervention period.

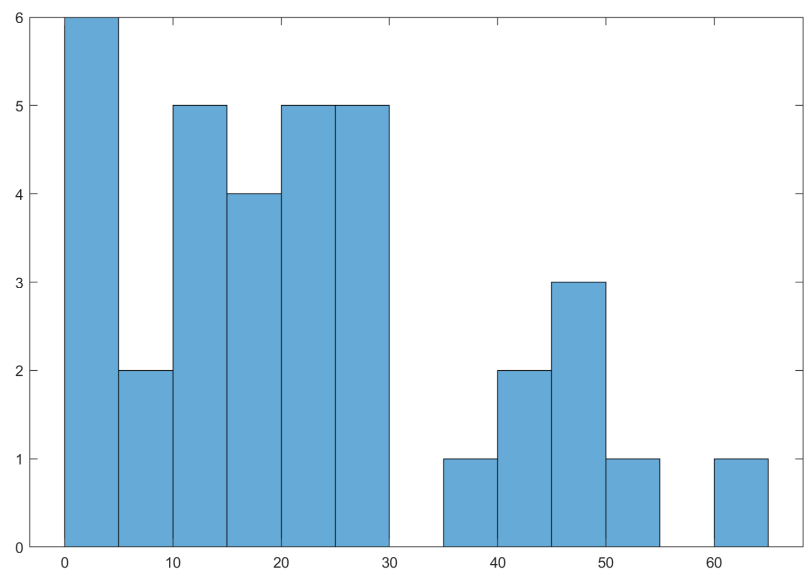


Figure 3: Histogram of the patients per clinic having a first prenatal visit during the intervention period.

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 ATE at the patient level. First, this ATE is not statistically significant during the pre-intervention 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 relative to 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 post-intervention 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. To conclude, it is worth noting that our results using the size-weighted cluster-level ATE align with those obtained by Celhay et al. (2019).

| Period | Pre-intervention | | | | Intervention | | | | Post-intervention | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Outcome | Weeks | | Weeks<13 | | Weeks | | Weeks<13 | | Weeks | | Weeks<13 | |
| Estimator | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ | $\hat{\theta}_{1,G}$ | $\hat{\theta}_{2,G}$ |
| Estimate | 0.09 | -0.08 | 0.02 | 0.01 | -0.01 | -1.39** | 0.03 | 0.10*** | 0.11 | -1.59** | -0.02 | 0.09** |
| s.e. | 0.77 | 0.56 | 0.06 | 0.03 | 0.95 | 0.66 | 0.05 | 0.04 | 0.85 | 0.72 | 0.05 | 0.04 |

Table 7: Estimation results based on data from Celhay et al. (2019). We ran our estimation separately for data in the pre-intervention period (i.e., Jan 2009 - Apr 2010), the intervention period (i.e., May 2010 - Dec 2010), and the post-intervention period (i.e., Jan 2011 - Mar 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. The significance level of the estimators is indicated with stars in the usual manner: "***" means significant at $\alpha = 1\%$, "**" means significant at $\alpha = 5\%$, and"*" means significant at $\alpha = 10\%$.

# A Appendix

Throughout this appendix, we use LIE to denote the law of iterated expectations, LTP to denote the law of total probability, LLN to denote the Kolmogorov's strong law of large numbers, CLT to denote the Lindeberg-Levy central limit theorem, and CMT to denote the continuous mapping theorem.

## A.1 Auxiliary Results

**Lemma A.1.** *Under Assumptions 2.2.(a)-(b),*

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

*is an i.i.d sequence of random variables.*

*Proof.* Let $A_g := (\bar{Y}_g(1), \bar{Y}_g(0))$ and $B_g := (|\mathcal{M}_g|, Z_g, N_g)$. We first show that data are independent, i.e., for arbitrary vectors $a^{(G)}$ and $b^{(G)}$,

$$P\{A^{(G)} \leq a^{(G)}, B^{(G)} \leq b^{(G)}\} \ = \ \prod_{1 \leq g \leq G} P\{A_g \leq a_g, B_g \leq b_g\}\ ,$$

where the inequalities are to be interpreted component-wise. To that end, let $C_g = (Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g)$ and denote by $\Gamma(a_g, N_g, \mathcal{M}_g)$ the (random) set such that $C_g \in \Gamma(a_g, N_g, \mathcal{M}_g)$ if and only if $A_g \leq a_g$. Let $\Gamma^{(G)}$ denote the Cartesian product of $\Gamma(a_g, N_g, \mathcal{M}_g)$ for all $1 \leq g \leq G$. Then,

$$\begin{aligned}
P\{A^{(G)} \leq a^{(G)}, B^{(G)} \leq b^{(G)}\} &\overset{(1)}{=} P\{C^{(G)} \in \Gamma^{(G)}, B^{(G)} \leq b^{(G)}\} \\
&\overset{(2)}{=} E\left[E\left[I\{C^{(G)} \in \Gamma^{(G)}\}|\mathcal{M}^{(G)}, Z^{(G)}, N^{(G)}\right] I\{B^{(G)} \leq b^{(G)}\}\right] \\
&\overset{(3)}{=} E\Bigg[\prod_{1 \leq g \leq G} E\left[C_g \in \Gamma(a_g, N_g, \mathcal{M}_g)|\mathcal{M}_g, Z_g, N_g\right] I\{B_g \leq b_g\}\Bigg] \\
&\overset{(4)}{=} E\Bigg[\prod_{1 \leq g \leq G} E\left[I\{C_g \in \Gamma(a_g, N_g, \mathcal{M}_g)\}I\{B_g \leq b_g\}|\mathcal{M}_g, Z_g, N_g\right]\Bigg] \\
&\overset{(5)}{=} \prod_{1 \leq g \leq G} E\left[I\{C_g \in \Gamma(a_g, N_g, \mathcal{M}_g)\}I\{B_g \leq b_g\}\right] \\
&\overset{(6)}{=} \prod_{1 \leq g \leq G} P\{A_g \leq a_g, B_g \leq b_g\}\ ,
\end{aligned}$$

where (1) and (6) follow from the definition of $\Gamma$, (2) from the LIE and the definition of $B_g$, (3) from Assumption 2.2.(b), (4) from the definition of $B_g$, and (5) from Assumption 2.2.(a) and the LIE.

Next, we show that data are identically distributed. To that end, consider the following derivation for arbitrary vectors $a$ and $b$, and $1 \leq g, g' \leq G$,

$$P\{A_g \leq a, B_g \leq b\} \overset{(1)}{=} P\{C_g \in \Gamma(a, N_g, \mathcal{M}_g), B_g \leq b\}$$

$$\stackrel{(2)}{=} E\left[E\left[I\{C_g \in \Gamma(a, N_g, \mathcal{M}_g)\}|\mathcal{M}_g, Z_g, N_g\right] I\{B_g \le b\}\right]$$
$$\stackrel{(3)}{=} E\left[E\left[I\{C_{g'} \in \Gamma(a, N_{g'}, S_{g'})\}|S_{g'}, Z_{g'}, N_{g'}\right] I\{B_{g'} \le b\}\right]$$
$$\stackrel{(4)}{=} P\{A_{g'} \le a, B_{g'} \le b\} ,$$

where (1) follows from the definition of $\Gamma$, (2) and (4) from the LIE and the definition of $B_g$, and (3) from Assumptions 2.2.(a)–(b). ■

*Proof of Lemma 2.1.* We begin by showing Assumption 2.2.(e). Fix $1 \le g \le G$ arbitrarily. Note that the statement of the lemma indicates that $|\mathcal{M}_g| \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \le i \le N_g)|Z_g, N_g$. Since $\mathcal{M}_g$ is drawn uniformly at random from all possible subsets of size $|\mathcal{M}_g|$ from $\{1, 2, \ldots .N_g\}$ conditionally on $(Z_g, N_g, |\mathcal{M}_g|)$, $\mathcal{M}_g \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \le i \le N_g)|Z_g, N_g, |\mathcal{M}_g|$. The desired result then follows from combining these observations and the LTP.

We next show Assumption 2.2.(f). Fix $a \in \{0, 1\}$ and $1 \le g \le G$ arbitrarily. By the LIE,

$$E[\bar{Y}_g(a)|N_g] \;=\; E\left[E\left[\frac{1}{|\mathcal{M}_g|}\sum_{i\in\mathcal{M}_g} Y_{i,g}(a)\middle| Z_g, N_g, |\mathcal{M}_g|, (Y_{i,g}(a) : 1 \le i \le N_g)\right]\middle| N_g\right] . \tag{25}$$

The inner expectation can be viewed as the expectation of a sample mean of size $|\mathcal{M}_g|$ drawn from the set $(Y_{i,g}(a) : 1 \le i \le N_g)$ without replacement. Hence, by Cochran (2007, Theorem 2.1),

$$E\left[\frac{1}{|\mathcal{M}_g|}\sum_{i\in\mathcal{M}_g} Y_{i,g}(a)\middle| Z_g, N_g, |\mathcal{M}_g|, (Y_{i,g}(a) : 1 \le i \le N_g)\right] \;=\; \frac{1}{N_g}\sum_{1\le i\le N_g} Y_{i,g}(a) . \tag{26}$$

The desired result follows from combining (25) and (26). ■

## A.2 Proof of Theorems

*Proof of Theorem 3.1.* For arbitrary $a \in \{0, 1\}$ and $1 \le g \le G$, consider the following derivation.

$$\begin{aligned} E\left[\left|\sum_{i\in\mathcal{M}_g} Y_{i,g}(a)\right|\right] \;&\le\; E\left[\sum_{i\in\mathcal{M}_g} |Y_{i,g}(a)|\right] \\ &\stackrel{(1)}{=}\; E\left[\sum_{i\in\mathcal{M}_g} E[|Y_{i,g}(a)||N_g, Z_g, \mathcal{M}_g]\right] \\ &\stackrel{(2)}{\le}\; C^{1/2}E[|\mathcal{M}_g|] \le C^{1/2}E[N_g] \;\stackrel{(3)}{<}\; \infty , \end{aligned} \tag{27}$$

where (1) follows from the LIE, (2) from Assumption 2.2.(d) and Jensen's inequality, and (3) from Assumption 2.2.(c) and Jensen's inequality. We further have by Assumption 2.2.(c) that $E[|\mathcal{M}_g|] < \infty$.

Under Assumptions 2.1–2.2, Lemma A.1, (27), and $E[|\mathcal{M}_g|] < \infty$, the desired result follows from Lemma C.4 in Bugni et al. (2019) and the CMT. ■

*Proof of Theorem 3.2.* For any $a \in \{0,1\}$ and $1 \leq g \leq G$, consider the following preliminary derivation.

$$E[\bar{Y}_g(a)^2] \overset{(1)}{\leq} E\left[\frac{1}{|\mathcal{M}_g|}\sum_{i\in\mathcal{M}_g} Y_{i,g}(a)^2\right] \overset{(2)}{=} E\left[\frac{1}{|\mathcal{M}_g|}\sum_{i\in\mathcal{M}_g} E[Y_{i,g}(a)^2|N_g, Z_g, \mathcal{M}_g]\right] \overset{(3)}{\leq} C < \infty , \quad (28)$$

where (1) follows from Jensen's inequality, (2) from the LIE, and (3) from Assumptions 2.2.(d)-(e). Next, note that by Assumption 2.2.(f) and the LIE,

$$\theta_1 = E\left[\frac{1}{N_g}\sum_{1\leq i\leq N_g} Y_{i,g}(1) - Y_{i,g}(0)\right] = E\left[\bar{Y}_g(1) - \bar{Y}_g(0)\right] . \quad (29)$$

Under (28), (29), and Assumptions 2.1–2.2, the desired result follows immediately from the proof of Theorem 4.1 in Bugni et al. (2018) where the clusters are viewed as the experimental units with potential outcomes given by $(\bar{Y}_g(0), \bar{Y}_g(1))$. ■

*Proof of Theorem 3.3.* Fix $a \in \{0,1\}$ and $r \in \{0,1,2\}$ arbitrarily. For any $1 \leq g \leq G$, we can repeat arguments in the proof of Theorem 3.2 to show that

$$E[\bar{Y}_g(a)^r] < \infty . \quad (30)$$

Under Assumptions 2.1–2.2, Lemma A.1, and (30), Lemma C.4 in Bugni et al. (2019) implies that

$$\frac{1}{G}\sum_{1\leq g\leq G} \bar{Y}_g^r I\{A_g = a\} \xrightarrow{p} P\{A_g = a\}E[\bar{Y}_g(a)^r] . \quad (31)$$

From this and the CMT, we conclude that

$$\widehat{\mathrm{Var}}_G[\bar{Y}_g(a)] \xrightarrow{p} \mathrm{Var}[\bar{Y}_g(a)] . \quad (32)$$

To conclude the proof, we note that the convergence in (10) follows from (31) (for $r = 0$ and $a \in \{0,1\}$), (32) (for $a \in \{0,1\}$), $\pi \in (0,1)$, and the CMT. Also, the inequality in (10) follows from the fact that $\mathrm{Var}[\bar{Y}_g^\dagger(a)] = \mathrm{Var}[\bar{Y}_g(a)] - \mathrm{Var}[E[\bar{Y}_g(a)|S_g]]$ and $\mathrm{Var}[E[\bar{Y}_g(a)|S_g]] = E[\bar{m}_a(S_g)^2]$ for $a \in \{0,1\}$. Some additional algebra will confirm the necessary and sufficient conditions for the inequality to become an equality. ■

*Proof of Theorem 3.4.* The result follows immediately from the proof of Theorem 4.2 in Bugni et al. (2018) where the clusters are viewed as the experimental units with potential outcomes given by $(\bar{Y}_g(0), \bar{Y}_g(1))$. ■

*Proof of Theorem 3.5.* We follow the general strategy in the proof of Theorem 4.1 in Bugni et al. (2018), as extended in Liu (2023). By the LIE and Assumption 2.2.(f), $E[\sum_{1\leq i\leq N_g} Y_{i,g}(a)] = E[\bar{Y}_g(a)N_g]$ for $a \in \{0,1\}$, and thus

$$\theta_2 = \frac{E\left[\bar{Y}_g(1)N_g\right]}{E[N_g]} - \frac{E\left[\bar{Y}_g(0)N_g\right]}{E[N_g]} .$$

As a consequence,

$$\sqrt{G}(\hat{\theta}_{2,G} - \theta_2) = \sqrt{G}(h(\hat{\Theta}_G) - h(\Theta)) , \quad (33)$$

where the function $h : \mathbb{R}^4 \to \mathbb{R}$ defined as $h(w,x,y,z) := \frac{w}{x} - \frac{y}{z}$, $G_1 := \frac{1}{G}\sum_{1\le g\le G} I\{A_g = 1\}$, and

$$\hat{\Theta}_G := \begin{pmatrix} \frac{1}{G_1}\sum_{1\le g\le G} \bar{Y}_g(1)N_g I\{A_g=1\} \\ \frac{1}{G_1}\sum_{1\le g\le G} N_g I\{A_g=1\} \\ \frac{1}{G-G_1}\sum_{1\le g\le G} \bar{Y}_g(0)N_g I\{A_g=0\} \\ \frac{1}{G-G_1}\sum_{1\le g\le G} N_g I\{A_g=0\} \end{pmatrix},$$
$$\Theta := (E[\bar{Y}_g(1)N_g], E[N_g], E[\bar{Y}_g(0)N_g], E[N_g])' .$$

By (33), we derive our result by characterizing the asymptotic distribution of $\sqrt{G}(\hat{\Theta}_G - \Theta)$ and applying the Delta method. To this end, note that

$$\sqrt{G}\left(\hat{\Theta}_G - \Theta\right) \overset{(1)}{=} \begin{pmatrix} \frac{\pi}{G_1/G} & 0 & 0 & 0 \\ 0 & \frac{\pi}{G_1/G} & 0 & 0 \\ 0 & 0 & \frac{1-\pi}{1-G_1/G} & 0 \\ 0 & 0 & 0 & \frac{1-\pi}{1-G_1/G} \end{pmatrix} \sqrt{G}\mathbb{L}_G \overset{(2)}{=} (I + o_p(1))\sqrt{G}\mathbb{L}_G , \quad (34)$$

where (1) uses that $\mathbb{L}_G := \left(L_G^{\mathrm{YN1}}, L_G^{\mathrm{N1}}, L_G^{\mathrm{YN0}}, L_G^{\mathrm{N0}}\right)'$ with

$$L_G^{\mathrm{YN1}} := \frac{1}{\pi}\left(\frac{1}{G}\sum_{1\le g\le G}\left(\bar{Y}_g(1)N_g - E\left[\bar{Y}_g(1)N_g\right]\right) I\{A_g=1\}\right),$$
$$L_G^{\mathrm{N1}} := \frac{1}{\pi}\left(\frac{1}{G}\sum_{1\le g\le G}\left(N_g - E\left[N_g\right]\right) I\{A_g=1\}\right),$$
$$L_G^{\mathrm{YN0}} := \frac{1}{1-\pi}\left(\frac{1}{G}\sum_{1\le g\le G}\left(\bar{Y}_g(0)N_g - E\left[\bar{Y}_g(0)N_g\right]\right) I\{A_g=0\}\right),$$
$$L_G^{\mathrm{N0}} := \frac{1}{1-\pi}\left(\frac{1}{G}\sum_{1\le g\le G}\left(N_g - E\left[N_g\right]\right) I\{A_g=0\}\right),$$

and (2) follows from Assumption 2.1, as it implies that $G_1/G = \sum_{s\in S} D_G(s)/G + \pi \xrightarrow{p} \pi$.

By (34), the next step to characterize the asymptotic distribution of $\sqrt{G}(\hat{\Theta}_G - \Theta)$ is to find the asymptotic distribution of $\sqrt{G}\mathbb{L}_G$. To this end, for any $a \in \{0,1\}$, we define

$$\mathbf{d} := \left(D_G(s)/\sqrt{G} \ : \ s\in\mathcal{S}\right)'$$
$$\mathbf{p} := \left(\sqrt{G}\left(G(s)/G - p(s)\right) \ : \ s\in\mathcal{S}\right)'$$
$$\mathbf{m}_a^{\mathrm{Y}} := \left(E\left[\bar{Y}_g(a)N_g \mid S_g = s\right] - E\left[\bar{Y}_g(a)N_g\right] \ : \ s\in\mathcal{S}\right)'$$
$$\mathbf{m}^{\mathrm{N}} := \left(E\left[N_g \mid S_g = s\right] - E\left[N_g\right] \ : \ s\in\mathcal{S}\right)' ,$$

where $G(s)$ is as in (11). By some algebra, we find that

$$\sqrt{G}\mathbb{L}_G \;=\; \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & \frac{1}{\pi}\left(\mathbf{m}_1^{\mathrm{Y}}\right)' & \left(\mathbf{m}_1^{\mathrm{Y}}\right)' \\ 0 & 1 & 0 & 0 & \frac{1}{\pi}\left(\mathbf{m}^{\mathrm{N}}\right)' & \left(\mathbf{m}^{\mathrm{N}}\right)' \\ 0 & 0 & 1 & 0 & -\frac{1}{1-\pi}\left(\mathbf{m}_0^{\mathrm{Y}}\right)' & \left(\mathbf{m}_0^{\mathrm{Y}}\right)' \\ 0 & 0 & 0 & 1 & -\frac{1}{1-\pi}\left(\mathbf{m}^{\mathrm{N}}\right)' & \left(\mathbf{m}^{\mathrm{N}}\right)' \end{pmatrix}}_{=:\,B'} \underbrace{\begin{pmatrix} \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\frac{1}{\pi}(\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|S_g])I\{A_g=1\} \\ \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\frac{1}{\pi}(N_g - E[N_g|S_g])I\{A_g=1\} \\ \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\frac{1}{1-\pi}(\bar{Y}_g(0)N_g - E[\bar{Y}_g(0)N_g|S_g])I\{A_g=0\} \\ \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\frac{1}{1-\pi}(N_g - E[N_g|S_g])I\{A_g=0\} \\ \mathbf{d} \\ \mathbf{p} \end{pmatrix}}_{=:\,\mathbf{y}_G} . \tag{35}$$

Under Assumptions 2.1–2.2, one can follow the partial sum and decomposition arguments developed in Lemma B.2 of Bugni et al. (2018) (or, equivalently, Lemma C.1 in Bugni et al. (2019)), to obtain

$$\mathbf{y}_G \;\xrightarrow{d}\; \mathcal{N}(0,\Sigma)\ , \tag{36}$$

where

$$\Sigma \;=\; \begin{pmatrix} \Sigma_1 & 0 & 0 & 0 \\ 0 & \Sigma_0 & 0 & 0 \\ 0 & 0 & \Sigma_D & 0 \\ 0 & 0 & 0 & \Sigma_G \end{pmatrix}\ ,$$

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

$$\begin{aligned} \Sigma_a &:= \frac{1}{P\{A_g = a\}}\begin{pmatrix} E\left[\mathrm{Var}\left[\bar{Y}_g(a)N_g|S_g\right]\right] & E\left[\mathrm{Cov}\left[\bar{Y}_g(a)N_g, N_g|S_g\right]\right] \\ E\left[\mathrm{Cov}\left[\bar{Y}_g(a)N_g, N_g|S_g\right]\right] & E\left[\mathrm{Var}\left[N_g|\, S_g\right]\right] \end{pmatrix}\ , \\ \Sigma_D &:= \mathrm{diag}\left(p(s)\tau(s) : s \in \mathcal{S}\right)\ , \\ \Sigma_G &:= \mathrm{diag}\left(p(s) : s \in \mathcal{S}\right) - \left(p(s) : s \in \mathcal{S}\right)\left(p(s) : s \in \mathcal{S}\right)'\ . \end{aligned}$$

By combining (34), (35), and (36), we conclude that

$$\sqrt{G}(\hat{\Theta}_G - \Theta) \;=\; \sqrt{G}\mathbb{L}_G + o_P(1) \;\xrightarrow{d}\; \mathcal{N}(0,\mathbb{V})\ , \tag{37}$$

where $\mathbb{V} := B'\Sigma B$. By additional calculations involving the law of total variance/covariance, $\mathbb{V}$ is a 4x4 symmetric matrix whose components are given by

$$\begin{aligned} \mathbb{V}_{11} \;&=\; \frac{1}{\pi}\mathrm{Var}\left[\bar{Y}_g(1)N_g\right] - \frac{1-\pi}{\pi}\mathrm{Var}\left[E\left[\bar{Y}_g(1)N_g \mid S_g\right]\right] + E\left[\frac{\tau(S_g)}{\pi^2}\left(E[\bar{Y}_g(1)N_g \mid S_g] - E[\bar{Y}_g(1)N_g]\right)^2\right] \\ \mathbb{V}_{12} \;&=\; \frac{1}{\pi}\mathrm{Cov}[\bar{Y}_g(1)N_g, N_g] - \frac{1-\pi}{\pi}\mathrm{Cov}[E[\bar{Y}_g(1)N_g|S_g], E[N_g|S_g]] \\ &\quad + E\left[\frac{\tau(S_g)}{\pi^2}\left(E[\bar{Y}_g(1)N_g \mid S_g] - E[\bar{Y}_g(1)N_g]\right)\left(E[N_g \mid S_g] - E[N_g]\right)\right] \\ \mathbb{V}_{13} \;&=\; \mathrm{Cov}[E[\bar{Y}_g(1)N_g|S_g], E[\bar{Y}_g(0)N_g|S_g]] \\ &\quad - E\left[\frac{\tau(S_g)}{\pi(1-\pi)}\left(E[\bar{Y}_g(1)N_g \mid S_g] - E[\bar{Y}_g(1)N_g]\right)\left(E[\bar{Y}_g(0)N_g \mid S_g] - E[\bar{Y}_g(0)N_g]\right)\right] \\ \mathbb{V}_{14} \;&=\; \mathrm{Cov}[E[\bar{Y}_g(1)N_g|S_g], E[N_g|S_g]] - E\left[\frac{\tau(S_g)}{\pi(1-\pi)}\left(E[\bar{Y}_g(1)N_g \mid S_g] - E[\bar{Y}_g(1)N_g]\right)\left(E[N_g \mid S_g] - E[N_g]\right)\right] \end{aligned}$$

$$\mathbb{V}_{22} = \frac{1}{\pi}\operatorname{Var}[N_g] - \frac{1-\pi}{\pi}\operatorname{Var}[E[N_g|S_g]] + E\left[\frac{\tau(S_g)}{\pi^2}\left(E[N_g \mid S_g] - E[N_g]\right)^2\right]$$

$$\mathbb{V}_{23} = \operatorname{Cov}[E[N_g|S_g], E[\bar{Y}_g(0)N_g|S_g]] - E\left[\frac{\tau(S_g)}{\pi(1-\pi)}\left(E[N_g \mid S_g] - E[N_g]\right)\left(E[\bar{Y}_g(0)N_g \mid S_g] - E[\bar{Y}_g(0)N_g]\right)\right]$$

$$\mathbb{V}_{24} = \operatorname{Var}[E[N_g|S_g]] - E\left[\frac{\tau(S_g)}{\pi(1-\pi)}\left(E[N_g \mid S_g] - E[N_g]\right)^2\right]$$

$$\mathbb{V}_{33} = \frac{1}{1-\pi}\operatorname{Var}[\bar{Y}_g(0)N_g] - \frac{\pi}{1-\pi}\operatorname{Var}[E[\bar{Y}_g(0)N_g|S_g]] + E\left[\frac{\tau(S_g)}{(1-\pi)^2}\left(E[\bar{Y}_g(0)N_g \mid S_g] - E[\bar{Y}_g(0)N_g]\right)^2\right]$$

$$\begin{aligned}\mathbb{V}_{34} &= \frac{1}{1-\pi}\operatorname{Cov}[\bar{Y}_g(0)N_g, N_g] - \frac{\pi}{1-\pi}\operatorname{Cov}[E[\bar{Y}_g(0)N_g|S_g], E[N_g|S_g]] \\ &\quad + E\left[\frac{\tau(S_g)}{(1-\pi)^2}\left(E[\bar{Y}_g(0)N_g \mid S_g] - E[\bar{Y}_g(0)N_g]\right)\left(E[N_g \mid S_g] - E[N_g]\right)\right]\end{aligned}$$

$$\mathbb{V}_{44} = \frac{1}{1-\pi}\operatorname{Var}[N_g] - \frac{\pi}{1-\pi}\operatorname{Var}[E[N_g|S_g]] + E\left[\frac{\tau(S_g)}{(1-\pi)^2}\left(E[N_g \mid S_g] - E[N_g]\right)^2\right] .$$

By (33), (37), and the Delta method, we get

$$\sqrt{G}(\hat{\theta}_2 - \theta_2) = \sqrt{G}(h(\hat{\Theta}) - h(\Theta)) \xrightarrow{d} \mathcal{N}\big(\,0,\ (\nabla h_0)'\mathbb{V}(\nabla h_0)\,\big) ,$$

where

$$\nabla h_0 := \left(\frac{1}{E[N_g]}, -\frac{E[\bar{Y}_g(1)N_g]}{E[N_g]^2}, -\frac{1}{E[N_g]}, \frac{E[\bar{Y}_g(0)N_g]}{E[N_g]^2}\right)' .$$

To conclude the proof, we need to show that $\sigma_2^2$ in (16) is equal to $(\nabla h_0)'\mathbb{V}(\nabla h_0)$. This follows from additional algebraic calculations similar to those in the proof of Theorem 3.1 in Bai et al. (2022). ■

*Proof of Theorem 3.6.* First, we verify that $\tilde{\sigma}_2^2$ can be written as in (17) with (18). To that end, let $\mathbf{1}_K$ denote a column vector of ones of length $K$. The cluster-robust variance estimator can then be written as

$$G\left(\sum_{1\le g\le G} X_g' X_g\right)^{-1}\left(\sum_{1\le g\le G} X_g'\hat{\epsilon}_g\hat{\epsilon}_g' X_g\right)\left(\sum_{1\le g\le G} X_g' X_g\right)^{-1},$$

where

$$\begin{aligned} X_g &:= \left(\ \mathbf{1}_{|\mathcal{M}_g|}\cdot\sqrt{\tfrac{N_g}{|\mathcal{M}_g|}} \qquad \mathbf{1}_{|\mathcal{M}_g|}\cdot\sqrt{\tfrac{N_g}{|\mathcal{M}_g|}}A_g\ \right) \\ \hat{\epsilon}_g &:= \left(\hat{\epsilon}_{i,g}(1)\sqrt{\tfrac{N_g}{|\mathcal{M}_g|}}A_g + \hat{\epsilon}_{i,g}(0)\sqrt{\tfrac{N_g}{|\mathcal{M}_g|}}(1-A_g)\ :\ i\in\mathcal{M}_g\right)' .\end{aligned}$$

By doing some algebra, it follows that

$$\sum_{1\le g\le G} X_g' X_g = \begin{pmatrix}\sum_{1\le g\le G} N_g & \sum_{1\le g\le G} N_g A_g \\ \sum_{1\le g\le G} N_g A_g & \sum_{1\le g\le G} N_g A_g\end{pmatrix},$$

and

$$\sum_{1\le g\le G} X_g'\hat{\epsilon}_g\hat{\epsilon}_g'X_g =$$

$$\sum_{1\le g\le G} A_g\left(\frac{N_g}{|\mathcal{M}_g|}\right)^2\Big(\sum_{i\in\mathcal{M}_g}\hat{\epsilon}_{i,g}(1)\Big)^2\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}+\sum_{1\le g\le G}(1-A_g)\left(\frac{N_g}{|\mathcal{M}_g|}\right)^2\Big(\sum_{i\in\mathcal{M}_g}\hat{\epsilon}_{i,g}(0)\Big)^2\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}.$$

The desired result then follows from further algebraic manipulations based on these expressions.

Next, we prove (19). To this end, fix $a\in\{0,1\}$, $r\in\{0,1,2\}$, and $l\in\{1,2\}$ arbitrarily. For any $1\le g\le G$, we can repeat arguments in the proof of Theorem 3.2 to show that

$$E[N_g^l\bar{Y}_g(a)^r] \;<\; \infty\ . \tag{38}$$

Under Assumptions 2.1–2.2, Lemma A.1, and (38), Lemma C.4. in Bugni et al. (2019) implies that

$$\frac{1}{G}\sum_{1\le g\le G} N_g^l\bar{Y}_g^r(a)I\{A_g=a\} \;\xrightarrow{p}\; E[N_g^l\bar{Y}_g(a)^r]P\{A_g=a\}\ . \tag{39}$$

From these results, consider the following derivation.

$$\begin{aligned}\tilde{\sigma}^2_{2,G}(a) \;&=\; \frac{\frac{1}{G}\sum_{1\le g\le G}N_g^2 I\{A_g=a\}(\frac{1}{|S_g|}\sum_{i\in S_g}\hat{\epsilon}_{i,g}(a))^2}{(\frac{1}{G}\sum_{1\le g\le G}N_g I\{A_g=a\})^2}\\
&\overset{(1)}{=}\; \frac{\frac{1}{G}\sum_{1\le g\le G}N_g^2\bar{Y}_g(a)^2 I\{A_g=a\}}{(\frac{1}{G}\sum_{1\le g\le G}N_g I\{A_g=a\})^2} + \frac{(\frac{1}{G}\sum_{1\le g\le G}N_g^2 I\{A_g=a\})(\frac{1}{G}\sum_{1\le g\le G}N_g\bar{Y}_g(a)I\{A_g=a\})^2}{(\frac{1}{G}\sum_{1\le g\le G}N_g I\{A_g=a\})^4}\\
&\quad -\; 2\frac{(\frac{1}{G}\sum_{1\le g\le G}N_g^2 I\{A_g=a\}\bar{Y}_g(a))\frac{1}{G}\sum_{1\le g\le G}N_g\bar{Y}_g(a)I\{A_g=a\}}{(\frac{1}{G}\sum_{1\le g\le G}N_g I\{A_g=a\})^3}\\
&\overset{(2)}{\xrightarrow{p}}\; \frac{E[N_g^2\bar{Y}_g(a)^2]+\frac{E[N_g^2](E[N_g\bar{Y}_g(a)])^2}{(E[N_g])^2}-2\frac{E[N_g^2\bar{Y}_g(a)]E[N_g\bar{Y}_g(a)]}{E[N_g]}}{E[N_g]^2P\{A_g=a\}} \;\overset{(3)}{=}\; \frac{1}{P\{A_g=a\}}\mathrm{Var}[\tilde{Y}_g(a)]\ ,\end{aligned} \tag{40}$$

where (1) follows from the definition of $\hat{\epsilon}_{i,g}(a)$, (2) from (39) (with $r\in\{0,1,2\}$, and $l\in\{1,2\}$), $E[N_g]>0$, $P\{A_g=a\}>0$, and the CMT, and (3) from the definition of $\tilde{Y}_g(a)$.

To conclude the proof, we note that the convergence in (19) follows from (40) (for $a\in\{0,1\}$) and the CMT. Also, the inequality in (19) follows from the fact that $\mathrm{Var}[\tilde{Y}_g^\dagger(a)]=\mathrm{Var}[\tilde{Y}_g(a)]-\mathrm{Var}[E[\tilde{Y}_g(a)|S_g]]$ and $\mathrm{Var}[E[\tilde{Y}_g(a)|S_g]]=E[\tilde{m}_a(S_g)^2]$ for $a\in\{0,1\}$. Some additional algebra will confirm the necessary and sufficient conditions for the inequality to become an equality. ■

*Proof of Theorem 3.7.* The result follows by studying the probability limit of each component and then applying the CMT. Specifically, it will follow once we establish that, for $a\in\{0,1\}$,

$$\hat{\mu}^{\hat{Y}}_{G,a}(s) \;\xrightarrow{p}\; E[\tilde{Y}_g(a)|S_g=s]\ , \tag{41}$$

$$\hat{\mu}^{\hat{Y}^2}_{G,a} \;\xrightarrow{p}\; E[\tilde{Y}_g^2(a)]\ . \tag{42}$$

We only show (41), as (42) can be shown by similar arguments. To this end, fix $a \in \{0,1\}$ and $r \in \{0,1\}$ arbitrarily. Under Assumptions 2.1–2.2, Lemma C.4 in Bugni et al. (2019) implies that

$$\frac{1}{G}\sum_{1\le g\le G}\tilde{Y}_g(a)^r I\{A_g=a,S_g=s\} \xrightarrow{p} p(s)P\{A_g=a\}E[\tilde{Y}_g(a)^r|S_g=s]~. \quad (43)$$

In turn, (43) (for $r \in \{0,1\}$), $p(s)P\{A_g=a\}>0$, and the CMT imply that

$$\hat{\mu}^{\tilde{Y}}_{G,a}(s) ~:=~ \frac{\sum_{1\le g\le G}\tilde{Y}_g(a)I\{A_g=a,S_g=s\}}{\sum_{1\le g\le G}I\{A_g=a,S_g=s\}} \xrightarrow{p} E[\tilde{Y}_g(a)|S_g=s]~. \quad (44)$$

Next, consider the following derivation.

$$\begin{aligned}
&\frac{1}{G}\sum_{1\le g\le G}(\hat{Y}_g-\tilde{Y}_g(a))I\{A_g=a,S_g=s\}\\
&\overset{(1)}{=} \frac{1}{\frac{1}{G}\sum_{1\le j\le G}N_j}\frac{1}{G}\sum_{1\le g\le G}N_g\bar{Y}_g(a)I\{A_g=a,S_g=s\} - \frac{1}{E[N_g]}\frac{1}{G}\sum_{1\le g\le G}N_g\bar{Y}_g(a)I\{A_g=a,S_g=s\}\\
&\quad - \frac{1}{\frac{1}{G}\sum_{1\le j\le G}N_j}\left(\frac{\frac{1}{G}\sum_{1\le j\le G}\bar{Y}_j(a)I\{A_j=a\}N_j}{\frac{1}{G}\sum_{1\le j\le G}I\{A_j=a\}N_j}\right)\frac{1}{G}\sum_{1\le g\le G}N_gI\{A_g=a,S_g=s\}\\
&\quad + \frac{1}{E[N_g]}\frac{E[N_g\bar{Y}_g(a)]}{E[N_g]}\frac{1}{G}\sum_{1\le g\le G}N_gI\{A_g=a,S_g=s\} \overset{(2)}{\xrightarrow{p}} 0~,
\end{aligned} \quad (45)$$

where (1) follows from the definitions of $\hat{Y}_g$ and $\tilde{Y}_g(a)$, and (2) from repeated applications of the LLN, Lemma C.4 in Bugni et al. (2019), and the CMT.

The desired result holds by the following derivation.

$$\hat{\mu}^{\hat{Y}}_{G,a}(s) \overset{(1)}{=} \hat{\mu}^{\tilde{Y}}_{G,a}(s) + o_P(1) \overset{(2)}{\xrightarrow{p}} E[\tilde{Y}_g(a)|S_g=s]~,$$

where (1) follows from (43) (for $r=0$), $p(s)P\{A_g=a\}>0$, (45), and the CMT, and (2) from (44). ■

### A.3 Survey of Articles Published in AEJ:Applied 2018-2022

In this section, we document some relevant features of the sampling design and subsequent analyses from every article involving a cluster randomized experiment published in the AEJ: Applied from 2018 to 2022. Table 8 summarizes our findings. For each article, we record:

1. The sampling design used within clusters. We find that 6/15 papers employ a sampling design such that $|\mathcal{M}_g| = \lfloor \gamma N_g \rfloor$ (all of these use $\gamma = 1$), one paper employs a sampling design such that $|\mathcal{M}_g| = k$ for $k$ fixed, and one paper features both of these designs. 7/15 papers employ a sampling design that does not correspond to either of these designs (denoted "other").
2. Whether the analyses in the paper discuss the use of sampling weights in their regressions and/or the

| Article | Sampling Design | Discussion of Weighting? |
|---|---|---|
| Giné and Mansuri (2018) | other | no |
| Lafortune et al. (2018) | $\lvert\mathcal{M}_g\rvert = N_g$ | no |
| McIntosh et al. (2018) | other | yes |
| Björkman Nyqvist et al. (2019) | other | no |
| Bolhaar et al. (2019) | $\lvert\mathcal{M}_g\rvert = N_g$ | no |
| Busso and Galiani (2019) | other | no |
| Celhay et al. (2019) | $\lvert\mathcal{M}_g\rvert = N_g$ | no |
| Deserranno et al. (2019) | other | yes |
| Loyalka et al. (2019) | other | no |
| Bandiera et al. (2020) | $\lvert\mathcal{M}_g\rvert = k$ | no |
| Banerjee et al. (2020) | $\lvert\mathcal{M}_g\rvert = N_g$ or $\lvert\mathcal{M}_g\rvert = k$ (depending on dep. variable) | no |
| McKenzie and Puerto (2021) | $\lvert\mathcal{M}_g\rvert = N_g$ | no |
| Mohanan et al. (2021) | other | no |
| Muralidharan et al. (2021) | $\lvert\mathcal{M}_g\rvert = N_g$ | yes |
| Wheeler et al. (2022) | $\lvert\mathcal{M}_g\rvert = N_g$ | no |

Table 8: Summary of sampling designs and subsequent analyses for papers in AEJ:applied from 2018 to 2022. The second column describes the sampling scheme within cluster, where "other" indicates a sampling scheme other than $|\mathcal{M}_g| = \lfloor \gamma N_g \rfloor$ for some constant $\gamma$ or $|\mathcal{M}_g| = k$ for some constant $k$. The third column indicates whether the paper includes a discussion about the use of sampling weights and/or the effect that this may have on the parameter of interest.

effect that this may have on the parameter of interest. We find that 11/14 papers do not feature such a discussion (however, in some cases it could be argued that an explicit discussion is not required).

### A.4 Covariate Adjustment

In this section, we consider further adjusting the estimator $\hat{\theta}_{1,G}$ of $\theta_1$ to exploit additional information in $Z_g$ and $N_g$ beyond that contained by $S_g$. We take care to do so in a way that ensures that the limiting variance of the covariate-adjusted estimator is weakly lower than its unadjusted counterpart (a detailed study of such adjustment strategies beyond what we consider here is pursued in a follow-up paper Aitken et al., 2023). In what follows, we denote by $X_g$ a subvector of $(Z_g', N_g)'$.

In order to define our covariate-adjusted estimator of $\theta_1$, we require some further notation. To this end, define

$$\begin{aligned}
\hat{\gamma}_{G,a}(s) &:= \left(\frac{1}{G_a(s)} \sum_{1 \le g \le G} (X_g - \hat{\mu}^X_{G,a}(s))(X_g - \hat{\mu}^X_{G,a}(s))' I\{S_g = s, A_g = a\}\right)^{-1} \\
&\qquad \times \left(\frac{1}{G_a(s)} \sum_{1 \le g \le G} (X_g - \hat{\mu}^X_{G,a}(s))(\bar{Y}_g(a) - \hat{\mu}^{\bar{Y}}_{G,a}(s))' I\{S_g = s, A_g = a\}\right) \\
\gamma_a(s) &:= \text{Var}[X_g | S_g = s]^{-1} \text{Cov}[X_g, \bar{Y}_g(a) | S_g = s] \ .
\end{aligned}$$

In other words, $\hat{\gamma}_{G,a}(s)$ is the ordinary least squares estimator of the coefficient on $X_g$ in a regression of $\bar{Y}_G$ on a constant and $X_g$ using only observations with $S_g = s$ and $A_g = s$, and $\gamma_a(s)$ is its population counterpart. Using this notation, further define

$$\begin{aligned}\bar{V}_g(a) &:= \bar{Y}_g(a) - (X_g - E[X_g|S_g])'\gamma_a(S_g) \\ \widehat{\bar{V}}_g(a) &:= \bar{Y}_g - (X_g - \hat{\mu}^X_G(S_g))'\hat{\gamma}_{G,a}(S_g) \ .\end{aligned}$$

In terms of this notation, the covariate-adjusted estimator of $\theta_1$ we consider is given by:

$$\hat{\theta}^{\text{adj}}_{1,G} := \frac{\sum_{1\le g\le G}\widehat{\bar{V}}_g(1)A_g}{\sum_{1\le g\le G}A_g} - \frac{\sum_{1\le g\le G}\widehat{\bar{V}}_g(0)(1-A_g)}{\sum_{1\le g\le G}(1-A_g)} \ . \tag{46}$$

Following the proof strategy in Theorem 3.2 we should obtain the following limiting distribution under appropriate assumptions:

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

as $G \to \infty$, where

$$\begin{aligned}\sigma^2_{1,\text{adj}} \quad := \quad & \frac{1}{\pi}\text{Var}[\bar{V}^\dagger_g(1)] + \frac{1}{1-\pi}\text{Var}[\bar{V}^\dagger_g(0)] + E[(\gamma_1(S_g) - \gamma_0(S_g))'\text{Var}[X_g|S_g](\gamma_1(S_g) - \gamma_0(S_g))] \\ & + E[(\bar{m}_1(S_g) - \bar{m}_0(S_g))^2] + E\left[\tau(S_g)\left(\frac{1}{\pi}\bar{m}_1(S_g) + \frac{1}{1-\pi}\bar{m}_0(S_g)\right)^2\right] \ ,\end{aligned}$$

with

$$\bar{V}^\dagger_g(a) := \bar{V}_g(a) - E[\bar{V}_g(a)|S_g] \ ,$$

$\bar{m}_a(S_g)$ defined as in (9), and $\pi$, $\tau(\cdot)$ defined as in Assumption 2.1. Furthermore, following similar arguments to those used when analyzing regression adustment for individual-level randomized experiments, it should be the case that

$$\sigma^2_1 - \sigma^2_{1,\text{adj}} = E[\Xi(S_g)'\text{Var}[X_g|S_g]\Xi(S_g)] \ge 0 \ ,$$

where

$$\Xi(S_g) := \sqrt{\frac{1-\pi}{\pi}}\gamma_0(S_g) + \sqrt{\frac{\pi}{1-\pi}}\gamma_1(S_g) \ ,$$

so that we ensure that the limiting variance of this estimator is weakly lower than its unadjusted counterpart. Next, to facilitate the use of this result for inference about $\theta_1$, we suggest an estimator of $\sigma^2_{1,\text{adj}}$. The estimator is given by

$$\hat{\sigma}^2_{1,\text{adj},G} := \hat{\zeta}^2_{\widehat{\bar{V}}}(\pi) + \hat{\Delta}_1 + \hat{\zeta}^2_H + \hat{\zeta}^2_A(\pi) \ ,$$

where

$$\hat{\zeta}^2_{\widehat{\bar{V}}}(\pi) := \frac{1}{\pi}\left(\hat{\mu}^{\widehat{\bar{V}}^2_g(1)}_{G,1} - \sum_{s\in\mathcal{S}}\frac{G(s)}{G}\hat{\mu}^{\widehat{\bar{V}}_g(1)}_{G,1}(s)^2\right) + \frac{1}{1-\pi}\left(\hat{\mu}^{\widehat{\bar{V}}^2_g(0)}_{G,0} - \sum_{s\in\mathcal{S}}\frac{G(s)}{G}\hat{\mu}^{\widehat{\bar{V}}_g(0)}_{G,0}(s)^2\right) \ ,$$

$\hat{\zeta}^2_H$ is defined as in (13), $\hat{\zeta}^2_A(\pi)$ is defined as in (14), and

$$\hat{\Delta}_1 := \sum_{s\in\mathcal{S}}\frac{G(s)}{G}(\hat{\gamma}_{G,1}(s) - \hat{\gamma}_{G,0}(s))'\left(\hat{\mu}^{XX'}_G(s) - \hat{\mu}^X_G(s)\hat{\mu}^{X'}_G(s)\right)(\hat{\gamma}_{G,1}(s) - \hat{\gamma}_{G,0}(s)) \ .$$

Following the proof strategy of Theorem 3.4, we should obtain under appropriate assumptions that

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

as $G \to \infty$. Thus, for $\sigma^2_{1,\mathrm{adj}} > 0$ and for any $\alpha \in (0,1)$,

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

as $G \to \infty$.

## References


Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017): "When should you adjust standard errors for clustering?" Tech. rep., National Bureau of Economic Research.

Aitken, C., F. Bugni, I. Canay, A. Shaikh, and M. Tabord-Meehan (2023): "Regression Adjustment for Stratified Cluster-Randomized Experiments," .

Angelucci, M., D. Karlan, and J. Zinman (2015): "Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco," *American Economic Journal: Applied Economics*, 7, 151–82.

Athey, S. and G. W. Imbens (2017): "The econometrics of randomized experiments," in *Handbook of economic field experiments*, Elsevier, vol. 1, 73–140.

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 Economic Journal: Applied Economics*, 7, 90–122.

Bai, Y., J. Liu, A. Shaikh, and M. Tabord-Meehan (2022): "Inference for Cluster Randomized Experiments with Matched Pairs," Working paper.

Bai, Y., J. P. Romano, and A. M. Shaikh (2021): "Inference in experiments with matched pairs," *Journal of the American Statistical Association*, 1–12.

Bandiera, O., N. Buehren, R. Burgess, M. Goldstein, S. Gulesci, I. Rasul, and M. Sulaiman (2020): "Women's empowerment in action: evidence from a randomized control trial in Africa," *American Economic Journal: Applied Economics*, 12, 210–59.

Banerjee, A., E. Duflo, R. Glennerster, and C. Kinnan (2015): "The miracle of microfinance? Evidence from a randomized evaluation," *American economic journal: Applied economics*, 7, 22–53.

Banerjee, A., E. Duflo, C. Imbert, S. Mathew, and R. Pande (2020): "E-governance, accountability, and leakage in public programs: Experimental evidence from a financial management reform in india," *American Economic Journal: Applied Economics*, 12, 39–72.

Benhin, E., J. Rao, and A. Scott (2005): "Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes," *Biometrika*, 92, 435–450.

Björkman Nyqvist, M., A. Guariso, J. Svensson, and D. Yanagizawa-Drott (2019): "Reducing child mortality in the last mile: experimental evidence on community health promoters in Uganda," *American Economic Journal: Applied Economics*, 11, 155–92.

Bolhaar, J., N. Ketel, and B. van Der Klaauw (2019): "Job search periods for welfare applicants: Evidence from a randomized experiment," *American Economic Journal: Applied Economics*, 11, 92–125.

Bruhn, M. and D. McKenzie (2008): "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 1, 200–232.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018): "Inference under Covariate Adaptive Randomization," *Journal of the American Statistical Association (Theory & Methods)*, 113, 1741–1768.

——— (2019): "Inference under Covariate-Adaptive Randomization with Multiple Treatments," *Quantitative Economics*, 10, 1741–1768.

Busso, M. and S. Galiani (2019): "The causal effect of competition on prices and quality: Evidence from a field experiment," *American Economic Journal: Applied Economics*, 11, 33–56.

Cameron, A. C. and D. L. Miller (2015): "A practitioner's guide to cluster-robust inference," *Journal of human resources*, 50, 317–372.

Celhay, P. A., P. J. Gertler, P. Giovagnoli, and C. Vermeersch (2019): "Long-run effects of temporary incentives on medical care productivity," *American Economic Journal: Applied Economics*, 11, 92–127.

Cochran, W. G. (2007): *Sampling techniques*, John Wiley & Sons.

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 Economic Journal: Applied Economics*, 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?" Tech. rep., National Bureau of Eco nomic Research.

Deserranno, E., M. Stryjan, and M. Sulaiman (2019): "Leader selection and service delivery in community groups: Experimental evidence from uganda," *American Economic Journal: Applied Economics*, 11, 240–67.

Ding, P., X. Li, and L. W. Miratrix (2017): "Bridging finite and super population causal inference," *Journal of Causal Inference*, 5.

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," *American Economic Review*, 105, 2757–2797.

Duflo, E., R. Glennerster, and M. Kremer (2007): "Using randomization in development economics research: A toolkit," *Handbook of development economics*, 4, 3895–3962.

Giné, X. and G. Mansuri (2018): "Together we will: experimental evidence on female voting behavior in Pakistan," *American Economic Journal: Applied Economics*, 10, 207–35.

Hansen, B. E. and S. Lee (2019): "Asymptotic theory for clustered samples," *Journal of econometrics*, 210, 268–290.

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.

Hayes, R. J. and L. H. Moulton (2017): *Cluster randomised trials*, Chapman and Hall/CRC.

Lafortune, J., J. Riutort, and J. Tessada (2018): "Role models or individual consulting: The impact of personalizing micro-entrepreneurship training," *American Economic Journal: Applied Economics*, 10, 222–45.

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," *arXiv preprint arXiv:2301.09016*.

Lohr, S. L. (2021): *Sampling: design and analysis*, Chapman and Hall/CRC.

Loyalka, P., A. Popova, G. Li, and Z. Shi (2019): "Does teacher training actually work? Evidence from a large-scale randomized evaluation of a national teacher training program," *American Economic Journal: Applied Economics*, 11, 128–54.

MacKinnon, J. G. and M. Webb (2019): "When and how to deal with clustered errors in regression models," Tech. rep., Queen's Economics Department Working Paper.

McIntosh, C., T. Alegría, G. Ordóñez, and R. Zenteno (2018): "The neighborhood impacts of local infrastructure investment: Evidence from urban Mexico," *American Economic Journal: Applied Economics*, 10, 263–86.

McKenzie, D. and S. Puerto (2021): "Growing markets through business training for female entrepreneurs: a market-level randomized experiment in Kenya," *American Economic Journal: Applied Economics*, 13, 297–332.

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.

Mohanan, M., K. Donato, G. Miller, Y. Truskinovsky, and M. Vera-Hernández (2021): "Different Strokes for Different Folks? Experimental Evidence on the Effectiveness of Input and Output Incentive Contracts for Health Care Providers with Varying Skills," *American Economic Journal: Applied Economics*, 13, 34–69.

Muralidharan, K., P. Niehaus, S. Sukhtankar, and J. Weaver (2021): "Improving last-mile service delivery using phone-based monitoring," *American Economic Journal: Applied Economics*, 13, 52–82.

Negi, A. and J. M. Wooldridge (2021): "Revisiting regression adjustment in experiments with heterogeneous treatment effects," *Econometric Reviews*, 40, 504–534.

Raudenbush, S. W. (1997): "Statistical analysis and optimal design for cluster randomized trials." *Psychological methods*, 2, 173.

Raudenbush, S. W. and D. Schwartz (2020): "Randomized experiments in education, with implications for multilevel causal inference," *Annual review of statistics and its application*, 7, 177–208.

Rosenberger, W. F. and J. M. Lachin (2016): *Randomization in Clinical Trials: Theory and Practice*, John Wiley & Sons, Inc., second ed.

Sasaki, Y. and Y. Wang (2022): "Non-Robustness of the Cluster-Robust Inference: with a Proposal of a New Robust Method," *arXiv preprint arXiv:2210.16991*.

Schochet, P. Z. (2013): "Estimators for clustered education RCTs using the Neyman model for causal inference," *Journal of Educational and Behavioral Statistics*, 38, 219–238.

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," *Journal of the American Statistical Association*, 1–12.

Su, F. and P. Ding (2021): "Model-assisted analyses of cluster-randomized experiments," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

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 journal of public health*, 107, 907–915.

Wheeler, L., R. Garlick, E. Johnson, P. Shaw, and M. Gargano (2022): "LinkedIn (to) job opportunities: Experimental evidence from job readiness training," *American Economic Journal: Applied Economics*, 14, 101–25.