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

April 19, 2022

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 simple random sampling. 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 shows 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 and Joe Romano 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 simple random sampling. 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 be 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.

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.

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). In addition, 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. Finally, in Section 4, we demonstrate the practical relevance of our theoretical results with a small simulation study.

2 Setup and Notation

2.1 Notation and Sampling Framework

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

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

In addition, define S_g to be the (possibly random) subset of $\{1, \ldots, N_g\}$ corresponding to the observations within the gth cluster that are sampled by the researcher. We emphasize that a realization of S_g is a *set* whose cardinality we denote by $|S_g|$, whereas a realization of N_g is a positive integer. For example, in the event that all observations in a cluster are sampled, $S_g = \{1, \ldots, N_g\}$ and $|S_g| = N_g$. Denote by P_G the distribution of the observed data

$$X^{(G)} := (((Y_{i,g} : i \in S_g), A_g, Z_g, N_g) : 1 \leq g \leq G)$$

and by Q_G the distribution of

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

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.

We begin by describing our assumptions on the distribution of $A^{(G)}$. In short, we assume that treatment is assigned using cluster-level simple random sampling. Formally, we impose the following assumption:

Assumption 2.1. The treatment assignment mechanism satisfies

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

(b) $A^{(G)} \sim \text{Binomial}(G, \pi)$, where $0 < \pi < 1$.

Despite its simplicity, this treatment assignment scheme remains widely used. This assumption, however, precludes many other popular treatment assignment schemes, including "matched pairs" (see, e.g., Banerjee et al., 2015; Crépon et al., 2015), and stratified block randomization (see, e.g. Attanasio et al., 2015; Angelucci et al., 2015). The analysis of these more complicated treatment assignment schemes can be found in companion papers Bugni et al. (2022) and 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

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

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

$$\{(Y_{i,g}(1), Y_{i,g}(0) : 1 \le i \le N_g) : 1 \le g \le G\} \bigm| S^{(G)}, Z^{(G)}, N^{(G)} \ ,$$

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

Assumption 2.2. The distribution Q_G is such that

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

(b) For some family of distributions $\{R(s,z,n) : (s,z,n) \in \text{supp}(S_g, Z_g, N_g)\}$,

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

(c) $P\{N_g \ge 1\} = 1$ 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) $S_g \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \le i \le N_g) \bigm| Z_g, N_g$ 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

$$\left\{ \left(\bar{Y}_g(1), \bar{Y}_g(0), |S_g|, Z_g, N_g\right), 1 \le g \le G \right\} \tag{2}$$

is an i.i.d. sequence of random variables, as established by Lemma 5.1 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]

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

and

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

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 $S_g = \{1, \ldots, N_g\}$ 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 S_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 $|S_g| \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g) \bigm| Z_g, N_g$ for all $1 \leq g \leq G$, and that, conditionally on $(Z_g, N_g, |S_g|)$, S_g is drawn uniformly at random from all possible subsets of size $|S_g|$ 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

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

[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).

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

$$\theta_1(Q_G) := E\left[\frac{1}{N_g}\sum_{1\le i\le N_g} Y_{i,g}(1) - Y_{i,g}(0)\right] . \tag{3}$$

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 $\omega_g = N_g/E[N_g]$, thus weighting the average effect of the treatment across clusters in proportion to their size:

$$\theta_2(Q_G) := E\left[\frac{1}{E[N_g]}\sum_{1\le i\le N_g} Y_{i,g}(1) - Y_{i,g}(0)\right] . \tag{4}$$

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 (S_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 θ_1 and θ_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 θ_1 and θ_2 to be distinct, they are equivalent in some special cases. For example, if all clusters are of the same fixed size k, i.e., $P\{N_g = k\} = 1$, then it follows immediately that $\theta_1 = \theta_2$. Alternatively, if treatment effects are constant, so that $P\{Y_{i,g}(1) - Y_{i,g}(0) = \tau \text{ for all } 1 \leq i \leq N_g\} = 1$, then $\theta_1 = \theta_2$. Generalizing these two extreme cases, we have that $\theta_1 = \theta_2$ if cluster sizes are ignorable (i.e., $R(s, z, n)$ does not depend on s and n) and treatment effects are sufficiently homogeneous in the sense that $P\{E[Y_{i,g}(1) - Y_{i,g}(0)] = E[Y_{j,g}(1) - Y_{j,g}(0)] \text{ for all } 1 \leq i, j \leq N_g\} = 1$. ■

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

$$\hat{\theta}_G^{\text{alt}} := \frac{\sum_{1 \leq g \leq G} \sum_{i \in S_g} Y_{i,g} A_g}{\sum_{1 \leq g \leq G} |S_g| A_g} - \frac{\sum_{1 \leq g \leq G} \sum_{i \in S_g} Y_{i,g}(1 - A_g)}{\sum_{1 \leq g \leq G} |S_g|(1 - A_g)} . \tag{5}$$

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,*

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

as $G \to \infty$.

The quantity ϑ corresponds to a *sample*-weighted cluster-level average treatment effect. When treatment effects are heterogeneous and cluster sizes are non-ignorable, ϑ need not equal either θ_1 defined in (3) or θ_2 defined in (4). 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 ϑ is in fact equal to either θ_1 or θ_2 for some sampling designs. Specifically, if $|S_g| = k$ for all $1 \leq g \leq G$, then ϑ is equal to θ_1; if $S_g = \{1, 2, \ldots, N_g\}$ for all $1 \leq g \leq G$ with probability one, then ϑ is equal to θ_2.

Example 3.1. Recall the setting of Example 2.1. Suppose further that the experimenter samples $|S_g| = 5$ students at random without replacement from each "small" classroom, and $|S_g| = 10$ 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 ϑ is approximately equal to θ_2 is when $|S_g| = \lfloor \gamma N_g \rfloor$ 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

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

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 θ_1 and θ_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 θ_1 defined in (3). To this end, consider the following difference-in-"average of averages" estimator:

$$\hat{\theta}_{1,G} := \frac{\sum_{1 \le g \le G} \bar{Y}_g A_g}{\sum_{1 \le g \le G} A_g} - \frac{\sum_{1 \le g \le G} \bar{Y}_g (1 - A_g)}{\sum_{1 \le g \le G} (1 - A_g)} , \tag{6}$$

where

$$\bar{Y}_g = \frac{1}{|S_g|} \sum_{i \in S_g} Y_{i,g} . \tag{7}$$

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. 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

$$\sigma_1^2 := \frac{1}{\pi} Var[\bar{Y}_g(1)] + \frac{1}{1-\pi} Var[\bar{Y}_g(0)] ,$$

with π defined as in Assumption 2.1.

To facilitate the use of Theorem 3.2 for inference about θ_1, we now provide a consistent estimator of σ_1^2. To this end, for $a \in \{0, 1\}$, define

$$\widehat{Var}_G[\bar{Y}_g(a)] := \frac{\sum_{1 \le g \le G} \bar{Y}_g^2 I\{A_g = a\}}{\sum_{1 \le g \le G} I\{A_g = a\}} - \left(\frac{\sum_{1 \le g \le G} \bar{Y}_g I\{A_g = a\}}{\sum_{1 \le g \le G} I\{A_g = a\}}\right)^2 .$$

Using this notation, define

$$\hat{\sigma}^2_{1,G} := \frac{1}{\pi}\widehat{Var}_G[\bar{Y}_g(1)] + \frac{1}{1-\pi}\widehat{Var}_G[\bar{Y}_g(0)] \ . \tag{8}$$

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

Theorem 3.3. *Under Assumptions 2.1 and 2.2,*

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

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

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

as $G \to \infty$.

Remark 3.2. Theorem 3.3 remains true if $\hat{\sigma}^2_{1,G}$ is replaced with

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

It is straightforward to show that $\tilde{\sigma}^2_{1,G}$ can be obtained as the usual 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. ■

Remark 3.3. A sufficient condition under which $\sigma^2_1 > 0$ holds is that $\text{Var}[\bar{Y}_g(a)] > 0$ for some $a \in \{0,1\}$. More generally, we expect $\sigma^2_1 > 0$ except in pathological cases such as when the distribution of outcomes is degenerate or in cases with perfect negative within-cluster correlation. ■

3.3 Size-weighted Cluster-level Average Treatment Effect

In this section, we consider estimation of θ_2 defined in (4). To this end, consider the following difference-in-"weighted average of averages" estimator:

$$\hat{\theta}_{2,G} := \frac{\sum_{1\le g\le G} \bar{Y}_g N_g A_g}{\sum_{1\le g\le G} N_g A_g} - \frac{\sum_{1\le g\le G} \bar{Y}_g N_g (1-A_g)}{\sum_{1\le g\le G} N_g (1-A_g)} \ , \tag{9}$$

where $\bar{Y}_g$ is defined as in (7). 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 $\sqrt{N_g/|S_g|}$. Note that, in the special case where $S_g = \{1, 2, \ldots, N_g\}$ for all $1 \le g \le G$ with probability one, we have $\hat{\theta}_{2,G} = \hat{\theta}^{\text{alt}}_G$ (i.e. the weights collapse to 1). The following theorem derives the asymptotic behavior of this estimator.

Theorem 3.4. *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*

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

with

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

Remark 3.4. It is interesting to compare σ_2^2 to the variance of the difference-in-means estimator from a finite-population analysis. For instance, in the special case where $S_g = \{1, 2, \ldots, N_g\}$, it follows from Theorem 1 of Su and Ding (2021) that the finite-population design-based variance is given by (in our notation):

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

where

$$\begin{aligned}N \quad &:= \quad \sum_{1\leq g\leq G} N_g\\ \tilde{\epsilon}_{i,g}(a) \quad &:= \quad Y_{i,g}(a) - \frac{1}{N}\sum_{1\leq g\leq G}\sum_{1\leq i\leq N_g} Y_{i,g}(a) .\end{aligned}$$

We emphasize that in the finite-population framework adopted by Su and Ding (2021), all of the above quantities are non-random. From this we see that the comparison between σ_2^2 and $\sigma^2_{2,G,\text{finpop}}$ 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 σ_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 3.5. As discussed in Remark 2.2, $\theta_1 = \theta_2$ whenever $N_g = k$ for all $1 \leq g \leq G$. Furthermore, in this case we have $\hat{\theta}_{1,G} = \hat{\theta}_{2,G}$ and thus $\sigma_1^2 = \sigma_2^2$ as well. ■

In parallel with our development in the preceding section, we now provide a consistent estimator of σ_2^2. To this end, define

$$\hat{\sigma}^2_{2,G} := \hat{\sigma}^2_{2,G}(1) + \hat{\sigma}^2_{2,G}(0) ,$$

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

$$\hat{\sigma}^2_{2,G}(a) := \frac{1}{\left(\frac{1}{G}\sum_{1\le g\le G} N_g I\{A_g = a\}\right)^2} \frac{1}{G} \sum_{1\le g\le G} \left[\left(\frac{N_g}{|S_g|}\right)^2 I\{A_g = a\} \left(\sum_{i\in S_g} \hat{\epsilon}_{i,g}(a) \right)^2 \right] , \quad (10)$$

where

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

The following theorem establishes the consistency of $\hat{\sigma}^2_{2,G}$ for σ^2_2. In the statement of the theorem, we again make use of the notation introduced preceding Theorem 3.3.

Theorem 3.5. *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 3.6. In Lemma 5.2 in the Appendix we establish that $\hat{\sigma}^2_{2,G}$ can be obtained as the cluster-robust variance 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 $\sqrt{N_g/|S_g|}$. ■

Remark 3.7. It can be shown that a sufficient condition under which σ^2_2 holds is that $\text{Var}[\bar{Y}_g(a)] > 0$ for some $a \in \{0,1\}$. As in our discussion in Remark 3.3, we expect this to hold outside of pathological cases. ■

Remark 3.8. Note further that A_g is independent of N_g under Assumption 2.1.(a). This observation motivates an alternative estimator

$$\hat{\theta}^{\text{alt}}_{2,G} := \frac{\frac{1}{G}\sum_{1\le g\le G} \bar{Y}_g N_g A_g}{\bar{N}_G \bar{A}_G} - \frac{\frac{1}{G}\sum_{1\le g\le G}\sum_{i\in S_g} \bar{Y}_g N_g (1-A_g)}{\bar{N}_G(1-\bar{A}_G)} , \quad (11)$$

where $\bar{N}_G := \frac{1}{G}\sum_{1\le g\le G} N_g$ and $\bar{A}_G := \frac{1}{G}\sum_{1\le g\le G} A_g$. By arguing as in the proof of Theorem 3.4, it is possible to derive the asymptotic behavior of this estimator as well, but we do not pursue this further in this paper. ■

4 Simulations

In this section, we illustrate the results presented in Section 3 with a simulation study. In all cases, potential outcomes are generated according to the equation:

$$Y_{i,g}(a) := \mu_g(a)\eta_{1,g} + \eta_{2,g}(a) Z_g + U_{i,g}(a) , \quad (12)$$

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

- $(\eta_{1,g}, \eta_{2,g}(a))$ are i.i.d. with $\eta_{1,g} \sim U[0,1]$, $\eta_{2,g}(0) \sim U[0,1]$, and $\eta_{2,g}(1) \sim U[0,5]$.
- $U_{i,g}(a)$ are i.i.d. with $U_{i,g}(a) \sim N(0, \sigma(a))$ and $\sigma(1) = \sqrt{2} > \sigma(0) = 1$.
- $\mu_g(a)$ are constants.
- The distribution of Z_g varies by design as described below.

We consider four alternative specifications for the distribution of cluster sizes N_g. Let $BB(a, b, n_{\text{supp}})$ be the Beta-Binomial distribution with parameters a and b and support on $\{0, \dots, n_{\text{supp}}\}$. The first three cases are given by

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

for the following values of (a, b) and $N_{\max} := 10(n_{\text{supp}} + 1)$,

- $(a, b) = (1, 1)$: results in a uniform probability mass function (p.m.f.) on $\{10, \dots, N_{\max}\}$.
- $(a, b) = (0.4, 0.4)$: results in a "U-shaped" p.m.f. on $\{10, \dots, N_{\max}\}$.
- $(a, b) = (10, 90)$: results in a "bell-shaped" p.m.f. with a long right tail on $\{10, \dots, N_{\max}\}$.

These three cases satisfy Assumption 2.2.(c). For the fourth and final case,

$$N_g = 10\zeta \quad \text{where} \quad \zeta \sim \text{ zeta distribution with parameter } s = 1.5 \ .$$

The last design is one where ζ has a finite mean ($s > 1$) but has infinite variance ($s \leq 2$), thus violating $E[N_g^2] < \infty$ in Assumption 2.2.(c). Figure 1 shows the p.m.f. of N_g for each of these four designs.

For each of the four distributions of cluster sizes, S_g is drawn uniformly at random from $\{1, 2, \dots, N_g\}$ as in Lemma 2.1 with three alternative sample sizes $|S_g|$: (a) $|S_g| = N_g$, (b) $|S_g| = 10$, and (c) $|S_g| = \max\{10, \min\{\gamma N_g, 200\}\}$ with $\gamma = 0.4$.

The combination of the four distributions of N_g and three sample sizes $|S_g|$ leads to 12 alternative specifications. For each of these specifications, we consider in addition two designs for Z_g:

- **Design 1**: $Z_g \perp\!\!\!\perp N_g$ where $Z_g \in \{-1, 1\}$ i.i.d. with $P\{Z_g = 1\} = 1/2$.
- **Design 2**: $Z_g = Z_{g,\text{big}} I\{N_g \geq E[N_g]\} + Z_{g,\text{small}} I\{N_g < E[N_g]\}$ where $Z_{g,\text{big}} \in \{-1, 1\}$ with $P\{Z_{g,\text{big}} = 1\} = 3/4$ and, independently, $Z_{g,\text{small}} \in \{-1, 1\}$ i.i.d. with $P\{Z_{g,\text{small}} = 1\} = 1/4$.

Finally, in all cases $\mu_g(0) = \mu_g(1) = 0$ and treatment assignment A_g is i.i.d. Bernoulli with probability 1/2 of success.

We note that for these designs we obtain that

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

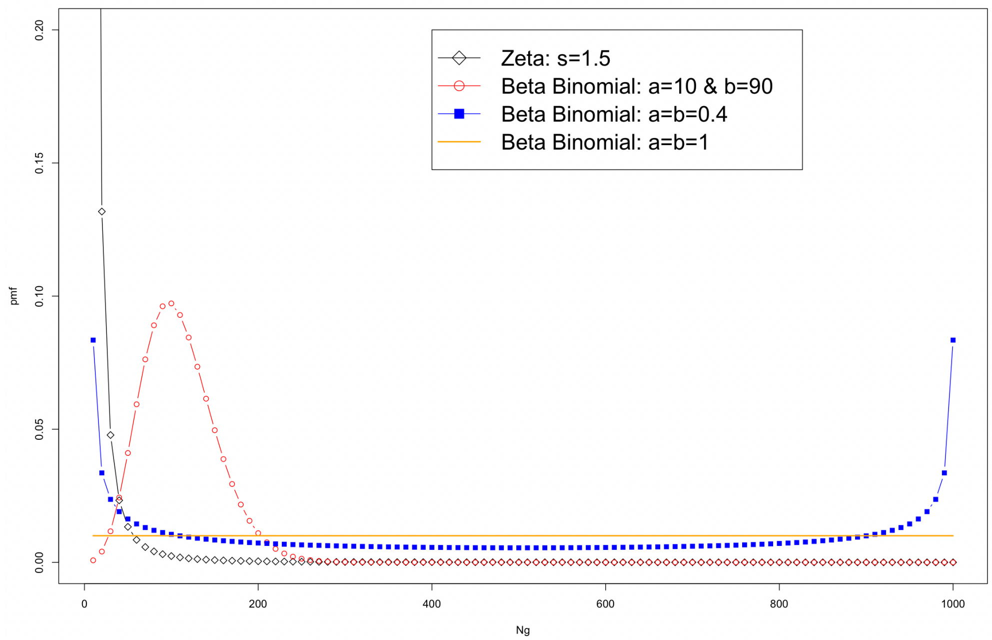

Figure 1: Four probability mass functions of N_g when $N_{\max} = 1000$

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

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

and so

$$\theta_1 = P\{N_g \geq E[N_g]\} - (1 - P\{N_g \geq E[N_g]\})$$
$$\theta_2 = E\left[\frac{N_g}{E[N_g]} \mid N_g \geq E[N_g]\right] P\{N_g \geq E[N_g]\} - E\left[\frac{N_g}{E[N_g]} \mid N_g < E[N_g]\right] P\{N_g < E[N_g]\} .$$

For each of the above 12 specifications and for each design of Z_g, we report the true values of (θ_1, θ_2) defined in (3) and (4), the average across simulations of the estimated values $(\hat{\theta}_{1,G}, \hat{\theta}_{2,G})$ defined in (6) and (9), the average across simulations of the estimated standard errors $(\hat{\sigma}_{1,G}, \hat{\sigma}_{2,G})$ defined in (8) and (10), and the empirical coverage of the 95% confidence intervals defined in Theorems 3.3 and 3.5. The results of our simulations are presented in Tables 2 to 6, where in all cases we conducted 5,000 replications. Table 7, in turn, reports summary statistics for the maximum value of N_g across simulations when $N_g = 10\zeta$, and Table 8 reports coverage probabilities for different values of G for Design 2 when $N_{\max} = 500$.

Tables 1 and 2 present results for Design 1, where $Z_g \perp\!\!\!\perp N_g$, for $N_{\max} = 500$ and $N_{\max} = 1000$. Despite the increased heterogeneity in cluster sizes that are allowed in Table 2 due to the larger vale of $N_{\max}$, coverage probabilities are mostly unaffected and close to the nominal levels when Assumption 2.2.(c) is satisfied. When $N_g = 10\zeta$ so that Assumption 2.2.(c) is violated, the confidence interval for θ_1 is not affected while the one for θ_2 undercovers in all cases.

Tables 3 and 4 present the same comparison for Design 2, where $Z_g \not\perp\!\!\!\perp N_g$. In this design, coverage mildly deteriorates moving from $N_{\max} = 500$ and $N_{\max} = 1000$, but still remains close to the nominal levels. Comparing the results from the previous tables with those in Tables 5 and 6, where $G = 5,000$ as opposed to $G = 100$, we see that the coverage of the confidence intervals in all cases where Assumption 2.2.(c) is satisfied improves, whereas the coverage of the confidence intervals when Assumption 2.2.(c) fails, either remaining at similar levels below the nominal levels (as is the case for θ_2) or deteriorating (as is the case for θ_1). A further inspection of this case shows that larger values of G increase the chances of getting unreasonably large numbers for N_g in the case where $N_g = 10\zeta$. Table 7 illustrates this, where we can see that in Table 1 the largest draw of N_g was $103,200$, while the largest draw in Table 5 was $2,853,880$. Since when $N_g = 10\zeta$ most of the clusters have size $N_g = 10$, this distribution creates situations where the largest cluster is around $200,000$ times larger than the "modal" cluster and at that point our asymptotic approximations may not provide an accurate representation of the finite-sample properties.

Finally, Table 8 compares coverage probabilities of the confidence intervals for θ_1 and θ_2 in the case of Design 2 with $N_{\max} = 500$, as G increases from 100 to $5,000$. We are particularly interested in the case where $N_g = 10(B + 1)$ and $(a, b) = (10, 90)$, since in this case the p.m.f. of N_g exhibits a long right tail as illustrated in Figure 1. This feature again creates a situation where there are a few clusters that are much larger than most other clusters. The results show that the empirical coverage deteriorates going from $G = 100$ to $G = 500$, but that eventually all empirical coverage probabilities approach their nominal levels.

Design 1		True values		Estimated		Stand. errors		Cov. Prob.	
$\|S_g\|$	N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
Ng	$BB(1,1)$	0.0000	0.0000	0.0009	0.0041	4.1806	4.7330	0.9444	0.9384
	$BB(0.4,0.4)$	0.0000	0.0000	-0.0052	0.0013	4.1888	5.0641	0.9450	0.9346
	$BB(10,90)$	0.0000	0.0000	-0.0025	-0.0006	4.1962	4.5060	0.9472	0.9434
	zeta(1.5)	0.0000	0.0000	0.0087	0.0011	4.2200	5.8472	0.9512	0.8816
10	$BB(1,1)$	0.0000	0.0000	0.0072	0.0071	4.2349	4.7956	0.9424	0.9362
	$BB(0.4,0.4)$	0.0000	0.0000	0.0051	0.0086	4.2458	5.1546	0.9480	0.9344
	$BB(10,90)$	0.0000	0.0000	0.0062	0.0031	4.2393	4.5552	0.9474	0.9408
	zeta(1.5)	0.0000	0.0000	-0.0004	-0.0064	4.2390	5.8859	0.9462	0.8852
γNg	$BB(1,1)$	0.0000	0.0000	-0.0043	-0.0040	4.1889	4.7375	0.9432	0.9382
	$BB(0.4,0.4)$	0.0000	0.0000	0.0021	0.0020	4.1914	5.0687	0.9460	0.9344
	$BB(10,90)$	0.0000	0.0000	-0.0034	-0.0058	4.1991	4.5108	0.9480	0.9396
	zeta(1.5)	0.0000	0.0000	-0.0022	0.0011	4.2359	5.7975	0.9450	0.8816

Table 1: Results for the simulation design in (12) for $G = 100$, $N_{\max} = 500$, and $Z_g \perp\!\!\!\perp N_g$

Design 1		True values		Estimated		Stand. errors		Cov. Prob.	
$\|S_g\|$	N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
Ng	$BB(1,1)$	0.0000	0.0000	0.0011	0.0010	4.1674	4.7913	0.9462	0.9448
	$BB(0.4,0.4)$	0.0000	0.0000	-0.0026	-0.0037	4.1753	5.1591	0.9524	0.9490
	$BB(10,90)$	0.0000	0.0000	-0.0003	-0.0009	4.1735	4.4638	0.9508	0.9542
	zeta(1.5)	0.0000	0.0000	-0.0002	0.0010	4.2236	9.4473	0.9562	0.9234
10	$BB(1,1)$	0.0000	0.0000	-0.0016	-0.0025	4.2338	4.8668	0.9536	0.9492
	$BB(0.4,0.4)$	0.0000	0.0000	0.0026	0.0010	4.2345	5.2418	0.9462	0.9464
	$BB(10,90)$	0.0000	0.0000	-0.0015	-0.0027	4.2371	4.5345	0.9498	0.9502
	zeta(1.5)	0.0000	0.0000	-0.0001	-0.0113	4.2351	9.5222	0.9534	0.9284
γNg	$BB(1,1)$	0.0000	0.0000	0.0000	0.0001	4.1728	4.7930	0.9546	0.9546
	$BB(0.4,0.4)$	0.0000	0.0000	0.0004	-0.0016	4.1806	5.1624	0.9532	0.9542
	$BB(10,90)$	0.0000	0.0000	-0.0004	-0.0010	4.1841	4.4730	0.9520	0.9552
	zeta(1.5)	0.0000	0.0000	0.0011	0.0008	4.2262	9.4149	0.9456	0.9170

Table 2: Results for the simulation design in (12) for $G = 100$, $N_{\max} = 1000$, and $Z_g \perp\!\!\!\perp N_g$

Design 2		True values		Estimated		Stand. errors		Cov. Prob.	
$\|S_g\|$	N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
Ng	$BB(1,1)$	0.0002	0.4902	0.0034	0.4941	4.1905	4.5179	0.9492	0.9384
	$BB(0.4,0.4)$	0.0003	0.6582	-0.0073	0.6484	4.1855	4.7295	0.9486	0.9418
	$BB(10,90)$	0.0420	0.3841	-0.0826	0.2545	4.1834	4.4077	0.9320	0.9312
	zeta(1.5)	-0.4905	0.2353	-0.5786	0.0562	4.1042	5.8882	0.9258	0.8510
10	$BB(1,1)$	0.0002	0.4902	0.0106	0.4946	4.2459	4.5969	0.9474	0.9416
	$BB(0.4,0.4)$	0.0003	0.6582	0.0009	0.6409	4.2470	4.8158	0.9522	0.9382
	$BB(10,90)$	0.0420	0.3841	-0.0566	0.2837	4.2420	4.4707	0.9332	0.9264
	zeta(1.5)	-0.4905	0.2353	-0.5556	0.0800	4.1169	5.9297	0.9334	0.8640
γNg	$BB(1,1)$	0.0002	0.4902	-0.0012	0.4846	4.1968	4.5290	0.9526	0.9390
	$BB(0.4,0.4)$	0.0003	0.6582	-0.0001	0.6529	4.1990	4.7378	0.9486	0.9388
	$BB(10,90)$	0.0420	0.3841	-0.0676	0.2725	4.1912	4.4071	0.9358	0.9342
	zeta(1.5)	-0.4905	0.2353	-0.5549	0.0736	4.1086	5.8543	0.9296	0.8608

Table 3: Results for the simulation design in (12) for $G = 100$, $N_{\max} = 500$, and $Z_g \not\perp N_g$

Design 2		True values		Estimated		Stand. errors		Cov. Prob.	
$\|S_g\|$	N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
Ng	$BB(1,1)$	-0.0001	0.4948	-0.0001	0.4942	4.1718	4.5694	0.9588	0.9516
	$BB(0.4,0.4)$	0.0003	0.6690	0.0010	0.6676	4.1765	4.8015	0.9464	0.9474
	$BB(10,90)$	0.0003	0.3047	-0.0403	0.2645	4.1702	4.3745	0.9134	0.9160
	zeta(1.5)	-0.4909	0.2308	-0.5581	0.1420	4.0984	9.3269	0.8330	0.8808
10	$BB(1,1)$	-0.0001	0.4948	-0.0021	0.4930	4.2357	4.6502	0.9474	0.9442
	$BB(0.4,0.4)$	0.0003	0.6690	-0.0036	0.6650	4.2339	4.8905	0.9474	0.9468
	$BB(10,90)$	0.0003	0.3047	-0.0406	0.2643	4.2363	4.4510	0.9146	0.9182
	zeta(1.5)	-0.4909	0.2308	-0.5533	0.1459	4.1100	9.3977	0.8284	0.8790
γNg	$BB(1,1)$	-0.0001	0.4948	-0.0012	0.4934	4.1730	4.5699	0.9518	0.9436
	$BB(0.4,0.4)$	0.0003	0.6690	-0.0003	0.6700	4.1804	4.8040	0.9476	0.9460
	$BB(10,90)$	0.0003	0.3047	-0.0404	0.2649	4.1830	4.3858	0.9146	0.9172
	zeta(1.5)	-0.4909	0.2308	-0.5577	0.1486	4.1009	9.3446	0.8332	0.8826

Table 4: Results for the simulation design in (12) for $G = 100$, $N_{\max} = 1000$, and $Z_g \not\perp N_g$

Design 1		True values		Estimated		Stand. errors		Cov. Prob.	
$\|S_g\|$	N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
Ng	$BB(1,1)$	0.0000	0.0000	0.0033	0.0038	4.1718	4.7753	0.9518	0.9480
	$BB(0.4,0.4)$	0.0000	0.0000	0.0000	-0.0005	4.1818	5.1298	0.9516	0.9432
	$BB(10,90)$	0.0000	0.0000	0.0017	0.0018	4.1777	4.5343	0.9520	0.9526
	zeta(1.5)	0.0000	0.0000	0.0015	-0.0017	4.2253	8.0718	0.9508	0.9198
10	$BB(1,1)$	0.0000	0.0000	0.0045	0.0063	4.2393	4.8619	0.9490	0.9522
	$BB(0.4,0.4)$	0.0000	0.0000	-0.0015	-0.0012	4.2361	5.2108	0.9464	0.9458
	$BB(10,90)$	0.0000	0.0000	0.0034	0.0037	4.2384	4.6089	0.9490	0.9504
	zeta(1.5)	0.0000	0.0000	0.0050	0.0089	4.2376	8.1965	0.9486	0.9108
γNg	$BB(1,1)$	0.0000	0.0000	0.0003	0.0006	4.1786	4.7806	0.9478	0.9432
	$BB(0.4,0.4)$	0.0000	0.0000	-0.0043	-0.0061	4.1815	5.1248	0.9438	0.9456
	$BB(10,90)$	0.0000	0.0000	0.0033	0.0038	4.1941	4.5515	0.9478	0.9508
	zeta(1.5)	0.0000	0.0000	-0.0003	0.0015	4.2298	8.0831	0.9440	0.9142

Table 5: Results for the simulation design in (12) for $G = 5000$, $N_{\max} = 500$, and $Z_g \perp\!\!\!\perp N_g$

Design 2		True values		Estimated		Stand. errors		Cov. Prob.	
$\|S_g\|$	N_g	θ_1	θ_2	$\hat{\theta}_{1,G}$	$\hat{\theta}_{2,G}$	$\hat{\sigma}_{1,G}$	$\hat{\sigma}_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
Ng	$BB(1,1)$	0.0002	0.4902	0.0001	0.4890	4.1759	4.5597	0.9502	0.9502
	$BB(0.4,0.4)$	0.0003	0.6582	-0.0017	0.6538	4.1768	4.7740	0.9500	0.9484
	$BB(10,90)$	0.0420	0.3841	-0.0101	0.3317	4.1776	4.4148	0.9068	0.9136
	zeta(1.5)	-0.4905	0.2353	-0.5591	0.1205	4.0987	8.0310	0.8816	0.8738
10	$BB(1,1)$	0.0002	0.4902	0.0035	0.4921	4.2384	4.6345	0.9528	0.9524
	$BB(0.4,0.4)$	0.0003	0.6582	0.0027	0.6558	4.2346	4.8630	0.9532	0.9480
	$BB(10,90)$	0.0420	0.3841	-0.0124	0.3296	4.2319	4.4802	0.9098	0.9188
	zeta(1.5)	-0.4905	0.2353	-0.5521	0.1365	4.1138	8.0774	0.8786	0.8718
γNg	$BB(1,1)$	0.0002	0.4902	-0.0029	0.4837	4.1775	4.5581	0.9470	0.9440
	$BB(0.4,0.4)$	0.0003	0.6582	0.0019	0.6619	4.1840	4.7747	0.9608	0.9480
	$BB(10,90)$	0.0420	0.3841	-0.0134	0.3293	4.1925	4.4311	0.9120	0.9176
	zeta(1.5)	-0.4905	0.2353	-0.5538	0.1217	4.0995	7.9542	0.8858	0.8778

Table 6: Results for the simulation design in (12) for $G = 5000$, $N_{\max} = 500$, and $Z_g \not\perp\!\!\!\perp N_g$

Table	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1 ($G = 100, N_{\max} = 500$)	40	110	170	407	300	103,200
4 ($G = 100, N_{\max} = 1000$)	30	110	170	2000	310	7,969,670
5 ($G = 1000, N_{\max} = 500$)	430	1470	2360	4590	4332	286,130
6 ($G = 5000, N_{\max} = 1000$)	440	1460	2290	5267	4062	2,853,880

Table 7: Summary statistics for the maximum of N_g across simulation draws when $N_g \sim$ zeta(1.5)

Design 2		$G=100$		$G=500$		$G=1000$		$G=5000$	
$\lvert S_g \rvert$	N_g	$CS_{1,G}$	$CS_{2,G}$	$CS_{1,G}$	$CS_{2,G}$	$CS_{1,G}$	$CS_{2,G}$	$CS_{1,G}$	$CS_{2,G}$
Ng	$BB(1,1)$	0.9492	0.9384	0.9502	0.9502	0.9574	0.9532	0.9488	0.9530
	$BB(0.4,0.4)$	0.9486	0.9418	0.9500	0.9484	0.9516	0.9482	0.9492	0.9482
	$BB(10,90)$	0.9320	0.9312	0.9068	0.9136	0.9018	0.9072	0.9496	0.9492
	zeta(1.5)	0.9258	0.8510	0.8816	0.8738	0.8348	0.8918	0.7564	0.8722
10	$BB(1,1)$	0.9474	0.9416	0.9528	0.9524	0.9532	0.9472	0.9488	0.9526
	$BB(0.4,0.4)$	0.9522	0.9382	0.9532	0.9480	0.9476	0.9494	0.9520	0.9490
	$BB(10,90)$	0.9332	0.9264	0.9098	0.9188	0.9026	0.9098	0.9476	0.9434
	zeta(1.5)	0.9334	0.8640	0.8786	0.8718	0.8496	0.8846	0.7562	0.8784
γN_g	$BB(1,1)$	0.9526	0.9390	0.9470	0.9440	0.9568	0.9510	0.9568	0.9556
	$BB(0.4,0.4)$	0.9486	0.9388	0.9608	0.9480	0.9536	0.9484	0.9512	0.9502
	$BB(10,90)$	0.9358	0.9342	0.9120	0.9176	0.8994	0.9070	0.9426	0.9478
	zeta(1.5)	0.9296	0.8608	0.8858	0.8778	0.8340	0.8844	0.7570	0.8764

Table 8: Coverage probabilities across G for $N_{\max}=500$ and $Z_g \not\perp N_g$

5 Appendix

5.1 Auxiliary Results

Lemma 5.1. *Under Assumptions 2.2(a)-(b),*

$$\{(\bar{Y}_g(1), \bar{Y}_g(0), |S_g|, Z_g, N_g), 1 \le g \le G\} ,$$

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

Proof. Let $A_g = (\bar{Y}_g(1), \bar{Y}_g(0))$, $B_g = (|S_g|, Z_g, N_g)$. Then our first goal is to show that for arbitrary vectors $a^{(G)}$ and $b^{(G)}$,

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

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

$$\begin{aligned}
P\{A^{(G)} \le a^{(G)}, B^{(G)} \le b^{(G)}\} &= P\{C^{(G)} \in \Gamma^{(G)}, B^{(G)} \le b^{(G)}\} \\
&= E\left[E\left[I\{C^{(G)} \in \Gamma^{(G)}\}|S^{(G)}, Z^{(G)}, N^{(G)}\right] I\{B^{(G)} \le b^{(G)}\}\right] \\
&= E\left[\prod_{1 \le g \le G} E\left[C_g \in \Gamma(a_g, N_g, S_g)|S_g, Z_g, N_g\right] I\{B_g \le b_g\}\right] \\
&= E\left[\prod_{1 \le g \le G} E\left[I\{C_g \in \Gamma(a_g, N_g, S_g)\}I\{B_g \le b_g\}|S_g, Z_g, N_g\right]\right] \\
&= \prod_{1 \le g \le G} E\left[I\{C_g \in \Gamma(a_g, N_g, S_g)\}I\{B_g \le b_g\}\right] \\
&= \prod_{1 \le g \le G} P\{A_g \le a_g, B_g \le b_g\} ,
\end{aligned}$$

where the first equality follows from the definition of Γ, the second equality follows from the law of iterated expectations, the third equality follows from Assumption 2.2.(b), the fourth equality follows from standard properties of conditional expectations, the fifth equality follows from Assumption 2.2.(a), and the sixth equality follows from the definition of Γ. Next, we show that for arbitrary vectors a and b

$$P\{A_g \le a, B_g \le b\} = P\{A_{g'} \le a, B_{g'} \le b\} ,$$

for any $1 \le g, g' \le G$. To that end,

$$\begin{aligned}
P\{A_g \le a, B_g \le b\} &= P\{C_g \in \Gamma(a, N_g, S_g), B_g \le b\} \\
&= E\left[E\left[I\{C_g \in \Gamma(a, N_g, S_g)\}|S_g, Z_g, N_g\right] I\{B_g \le b\}\right] \\
&= E\left[E\left[I\{C_{g'} \in \Gamma(a, N_{g'}, S_{g'})\}|S_{g'}, Z_{g'}, N_{g'}\right] I\{B_{g'} \le b\}\right]
\end{aligned}$$

$$= P\{A_{g'} \leq a, B_{g'} \leq b\} ,$$

where the first inequality follows from the definition of Γ, the second from the law of iterated expectations, the third from Assumptions 2.2.(a)–(b), and the fourth from the law of iterated expectations and the definition of Γ. ■

Lemma 5.2. *$\hat{\sigma}^2_{2,G}$ can be obtained as the cluster-robust variance estimator of the variance of the coefficient of A_g in a weighted-least squares regression of $Y_{i,g}$ on a constant and A_g, with weights equal to $\sqrt{N_g/|S_g|}$.*

Proof. Let $\mathbf{1}_K$ denote a column vector of ones of length K. The cluster-robust variance estimator can then be written as:

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

where

$$X_g = \begin{pmatrix} \mathbf{1}_{|S_g|} \cdot \sqrt{\frac{N_g}{|S_g|}} & \mathbf{1}_{|S_g|} \cdot \sqrt{\frac{N_g}{|S_g|}} A_g \end{pmatrix}$$

and

$$\hat{\epsilon}_g = \left(\hat{\epsilon}_{i,g}(1) \sqrt{\frac{N_g}{|S_g|}} A_g + \hat{\epsilon}_{i,g}(0) \sqrt{\frac{N_g}{|S_g|}} (1 - A_g) : i \in S_g \right)' .$$

Expanding and simplifying gives us our result. ■

5.2 Proof of Lemma 2.1

Proof. First we show that Assumption 2.2.(e) is satisfied. By the assumptions of the proposition we have $|S_g| \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g) | Z_g, N_g$ and by our specification of the sampling procedure for S_g we have that $S_g \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g) | Z_g, N_g, |S_g|$. Hence $S_g \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g) | Z_g, N_g$ by the properties of conditional expectations. Next we show that Assumption 2.2.(f) is satisfied. To that end, note that by the law of iterated expectations

$$E[\bar{Y}_g(a) | N_g] = E\left[E\left[\frac{1}{|S_g|} \sum_{i \in S_g} Y_{i,g}(a) \middle| Z_g, N_g, |S_g|, (Y_{i,g}(a) : 1 \leq i \leq N_g) \right] \middle| N_g \right] .$$

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

$$E\left[\frac{1}{|S_g|} \sum_{i \in S_g} Y_{i,g}(a) \middle| Z_g, N_g, |S_g|, (Y_{i,g}(a) : 1 \leq i \leq N_g) \right] = \frac{1}{N_g} \sum_{1 \leq i \leq N_g} Y_{i,g}(a) ,$$

and hence it follows that

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

as desired. ■

5.3 Proof of Theorem 3.1

Proof. By Assumption 2.1 and Assumptions 2.2.(c)–(e),

$$\begin{aligned} E\left[\left|\sum_{i\in S_g} Y_{i,g} I\{A_g = a\}\right|\right] &= E\left[\left|\sum_{i\in S_g} Y_{i,g}(a) I\{A_g = a\}\right|\right] \\ &\lesssim E\left[\sum_{i\in S_g} E[|Y_{i,g}(a)||N_g, Z_g, S_g]\right] \\ &\lesssim E\left[|S_g|\right] \le E[N_g] < \infty ,\end{aligned}$$

and a similar argument establishes that $E[|S_g|I\{A_g = a\}] < \infty$. It thus follows by Lemma 5.1, the law of large numbers, and the continuous mapping theorem that

$$\hat{\theta}_G^{\text{alt}} \xrightarrow{P} \frac{E[A_g \sum_{i\in S_g} Y_{i,g}(1)]}{E[|S_g|A_g]} - \frac{E[(1-A_g)\sum_{i\in S_g} Y_{i,g}(0)]}{E[|S_g|(1-A_g)]} .$$

The result then follows by Assumption 2.1.(a). ■

5.4 Proof of Theorem 3.2

Proof. By Assumption 2.2.(f),

$$\hat{\theta}_{1,G} - \theta_1 = \frac{\sum_{1\le g\le G}\bar{Y}_g(1)A_g}{\sum_{1\le g\le G} A_g} - \frac{\sum_{1\le g\le G}\bar{Y}_g(0)(1-A_g)}{\sum_{1\le g\le G}(1-A_g)} - \left(E[\bar{Y}_g(1)] - E[\bar{Y}_g(0)]\right) .$$

Re-grouping,

$$\sqrt{G}(\hat{\theta}_{1,G} - \theta_1) = \frac{1}{\frac{1}{G}\sum_{g=1}^G A_g} L_{1,G} - \frac{1}{\frac{1}{G}\sum_{g=1}^G (1-A_g)} L_{0,G} ,$$

where

$$L_{1,G} = \frac{1}{\sqrt{G}}\sum_{g=1}^G \left(\bar{Y}_g(1) - E[\bar{Y}_g(1)]\right) A_g ,$$

$$L_{0,G} = \frac{1}{\sqrt{G}}\sum_{g=1}^G \left(\bar{Y}_g(0) - E[\bar{Y}_g(0)]\right)(1-A_g) .$$

By Assumption 2.1 and Assumptions 2.2.(c)–(e),

$$E[(\bar{Y}_g(a) - E[\bar{Y}_g(a)])^2 I\{A_g = a\}] \lesssim E[\bar{Y}_g(a)^2]$$

$$\le E\left[\frac{1}{|S_g|}\sum_{i\in S_g} Y_{i,g}(a)^2\right]$$

$$= E\left[\frac{1}{|S_g|}\sum_{i\in S_g} E[Y_{i,g}(a)^2|N_g, Z_g, S_g]\right]$$

$$\lesssim E\left[\frac{|S_g|}{|S_g|}\right] < \infty ,$$

where the second inequality follows from Cauchy-Schwarz. It follows by Lemma 5.1 and the Lindeberg-Levy CLT that

$$\begin{pmatrix} L_{1,G} \\ L_{0,G} \end{pmatrix} \xrightarrow{d} N\left(0_2, \begin{pmatrix} \mathrm{Var}[\bar{Y}_g(1)]\pi & 0 \\ 0 & \mathrm{Var}[\bar{Y}_g(0)](1-\pi) \end{pmatrix}\right) ,$$

where we have used Assumption 2.1 to simplify the expression for the covariance matrix. By Assumption 2.1.(b), the law of large numbers and Slutsky's theorem,

$$\sqrt{G}(\hat{\theta}_2 - \theta_2) \xrightarrow{d} N(0, \sigma_1^2) ,$$

where

$$\sigma_1^2 = \begin{pmatrix} \frac{1}{\pi} & -\frac{1}{1-\pi} \end{pmatrix} \begin{pmatrix} \mathrm{Var}[\bar{Y}_g(1)]\pi & 0 \\ 0 & \mathrm{Var}[\bar{Y}_g(0)](1-\pi) \end{pmatrix} \begin{pmatrix} \frac{1}{\pi} \\ -\frac{1}{1-\pi} \end{pmatrix} .$$

Expanding and simplifying this expression gives us our result. ■

5.5 Proof of Theorem 3.3

Proof. Using Assumption 2.1 and Assumptions 2.2.(c)–(e), it can be shown using similar arguments to those used in the proof of Theorem 3.2 that

$$E\left[\bar{Y}_g^r I\{A_g = a\}\right] = E\left[\bar{Y}_g(a)^r I\{A_g = a\}\right] < \infty ,$$

for $r \in \{1, 2\}$. Hence by Lemma 5.1 and the law of large numbers,

$$\frac{1}{G}\sum_{g=1}^{G} \bar{Y}_g^2 I\{A_g = a\} \xrightarrow{P} E[\bar{Y}_g(a)^2 I\{A_g = a\}] ,$$

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

Then by the continuous mapping theorem and the law of large numbers,

$$\widehat{\mathrm{Var}}_G[\bar{Y}_g(1)] \xrightarrow{P} \frac{E[\bar{Y}_g(1)^2 A_g]}{\pi} - \left(\frac{E[\bar{Y}_g(1)A_g]}{\pi}\right)^2 = \mathrm{Var}[\bar{Y}_g(1)] ,$$

where the last equality follows from Assumption 2.1. Similarly,

$$\widehat{\text{Var}}_G[\bar{Y}_g(0)] \xrightarrow{P} \text{Var}[\bar{Y}_g(0)] \ .$$

By the continuous mapping theorem it then follows that

$$\hat{\sigma}^2_{1,G} \xrightarrow{P} \frac{1}{\pi}\text{Var}\left[\bar{Y}_g\left(1\right)\right] + \frac{1}{1-\pi}\text{Var}\left[\bar{Y}_g\left(0\right)\right] = \sigma_1^2 \ ,$$

as desired. ■

5.6 Proof of Theorem 3.4

Proof. By Assumption 2.2.(f),

$$\hat{\theta}_{2,G} - \theta_2 = \frac{\sum_{1\le g\le G} N_g\bar{Y}_g(1)A_g}{\sum_{1\le g\le G} N_gA_g} - \frac{\sum_{1\le g\le G} N_g\bar{Y}_g(0)(1-A_g)}{\sum_{1\le g\le G} N_g(1-A_g)} - \left(\frac{E[N_g\bar{Y}_g(1)]}{E[N_g]} - \frac{E[N_g\bar{Y}_g(0)]}{E[N_g]}\right) \ .$$

Re-grouping,

$$\begin{aligned}\sqrt{G}(\hat{\theta}_{2,G} - \theta_2) = &\frac{1}{\frac{1}{G}\sum_{1\le g\le G} N_gA_g}L_{1,G} - \frac{1}{\frac{1}{G}\sum_{1\le g\le G} N_g(1-A_g)}L_{0,G} - \frac{E[N_g\bar{Y}_g(1)]}{E[N_g]}\frac{E[N_gA_g]}{\frac{1}{G}\sum_{1\le g\le G} N_gA_g}L_{N1,G} \\ &+ \frac{E[N_g\bar{Y}_g(0)]}{E[N_g]}\frac{E[N_g(1-A_g)]}{\frac{1}{G}\sum_{1\le g\le G} N_g(1-A_g)}L_{N0,G}\end{aligned}$$

where

$$\begin{aligned} L_{1,G} &:= \frac{1}{\sqrt{G}}\sum_{1\le g\le G}\left(N_g\bar{Y}_g(1)A_g - E[N_g\bar{Y}_g(1)A_g]\right) \\ L_{0,G} &:= \frac{1}{\sqrt{G}}\sum_{1\le g\le G}\left(N_g\bar{Y}_g(0)(1-A_g) - E[N_g\bar{Y}_g(0)(1-A_g)]\right) \\ L_{N1,G} &:= \frac{1}{\sqrt{G}}\sum_{1\le g\le G}\left(\frac{N_gA_g}{E[N_gA_g]} - 1\right) \\ L_{N0,G} &:= \frac{1}{\sqrt{G}}\sum_{1\le g\le G}\left(\frac{N_g(1-A_g)}{E[N_g(1-A_g)]} - 1\right) \ . \end{aligned}$$

Using Assumption 2.1 and Assumptions 2.2.(c)–(e), it can be shown using similar arguments to those used in the proof of Theorem 3.2 that

$$E[(N_g\bar{Y}_g(a)I\{A_g = a\})^2] < \infty \ ,$$

$$E\left[\left(\frac{N_gI\{A_g = a\}}{E[N_gI\{A_g = a\}]}\right)^2\right] < \infty \ .$$

It thus follows from Lemma 5.1 and the Linderberg-Levy CLT that

$$\begin{pmatrix} L_{1,G} \\ L_{0,G} \\ L_{N1,G} \\ L_{N0,G} \end{pmatrix} \xrightarrow{d} N\left(0_4, \begin{pmatrix} \sigma^2(1) & \rho_{10} & \rho_{1N_1} & \rho_{1N_0} \\ \rho_{10} & \sigma^2(0) & \rho_{0N_1} & \rho_{0N_0} \\ \rho_{1N_1} & \rho_{0N_1} & \sigma_N^2(1) & \rho_{N_1N_0} \\ \rho_{1N_0} & \rho_{0N_0} & \rho_{N_1N_0} & \sigma_N^2(0) \end{pmatrix}\right) ,$$

where

$$\begin{aligned} \sigma^2(1) &= \text{Var}(N_g\bar{Y}_g(1)A_g), & \sigma^2(0) &= \text{Var}(N_g\bar{Y}_g(0)(1-A_g)), \\ \sigma_N^2(1) &= \text{Var}\left(\frac{N_gA_g}{E[N_gA_g]}\right), & \sigma_N^2(0) &= \text{Var}\left(\frac{N_g(1-A_g)}{E[N_g(1-A_g)]}\right), \\ \rho_{10} &= \text{Cov}(N_g\bar{Y}_g(1)A_g, N_g\bar{Y}_g(0)(1-A_g)), & \rho_{0N_1} &= \text{Cov}\left(N_g\bar{Y}_g(0)(1-A_g), \frac{N_gA_g}{E[N_gA_g]}\right) \\ \rho_{1N_1} &= \text{Cov}\left(N_g\bar{Y}_g(1)A_g, \frac{N_gA_g}{E[N_gA_g]}\right), & \rho_{0N_0} &= \text{Cov}\left(N_g\bar{Y}_g(0)(1-A_g), \frac{N_g(1-A_g)}{E[N_g(1-A_g)]}\right), \\ \rho_{1N_0} &= \text{Cov}\left(N_g\bar{Y}_g(1)A_g, \frac{N_g(1-A_g)}{E[N_g(1-A_g)]}\right), & \rho_{N_1N_0} &= \text{Cov}\left(\frac{N_gA_g}{E[N_gA_g]}, \frac{N_g(1-A_g)}{E[N_g(1-A_g)]}\right). \end{aligned}$$

By Assumption 2.1, Lemma 5.1, the law of large numbers, and Slutsky's theorem, we thus have that

$$\sqrt{G}(\hat{\theta}_2 - \theta_2) \xrightarrow{d} N(0, V) ,$$

where

$$V = \begin{pmatrix} \frac{1}{E[N_g]\pi} & -\frac{1}{E[N_g](1-\pi)} & -\frac{E[N_g\bar{Y}_g(1)]}{E[N_g]} & \frac{E[N_g\bar{Y}_g(0)]}{E[N_g]} \end{pmatrix} \begin{pmatrix} \sigma^2(1) & \rho_{10} & \rho_{1N_1} & \rho_{1N_0} \\ \rho_{10} & \sigma^2(0) & \rho_{0N_1} & \rho_{0N_0} \\ \rho_{1N_1} & \rho_{0N_1} & \sigma_N^2(1) & \rho_{N_1N_0} \\ \rho_{1N_0} & \rho_{0N_0} & \rho_{N_1N_0} & \sigma_N^2(0) \end{pmatrix} \begin{pmatrix} \frac{1}{E[N_g]\pi} \\ -\frac{1}{E[N_g](1-\pi)} \\ -\frac{E[N_g\bar{Y}_g(1)]}{E[N_g]} \\ \frac{E[N_g\bar{Y}_g(0)]}{E[N_g]} \end{pmatrix} .$$

Expanding and simplifying this expression gives us our result. ■

5.7 Proof of Theorem 3.5

Proof. Using Assumptions 2.1, 2.2, and Lemma 5.1, it can be shown by repeated applications of the law of large numbers, the continuous mapping theorem, and a bit of algebra that

$$\hat{\sigma}_{2,G}^2(a) \xrightarrow{P} \frac{1}{E[N_g]^2 P\{A_g = a\}} E\left[\left(\frac{N_g}{|S_g|}\right)^2 \left(\sum_{i\in S_g} \epsilon_{i,g}(a)\right)^2\right] .$$

The result then follows by the continuous mapping theorem. ■

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge (2017), "When should you adjust standard errors for clustering?" Technical report, National Bureau of Economic Research.

Angelucci, Manuela, Dean Karlan, and Jonathan Zinman (2015), "Microcredit impacts: Evidence from a randomized microcredit program placement experiment by compartamos banco." *American Economic Journal: Applied Economics*, 7, 151–82.

Athey, Susan and Guido W Imbens (2017), "The econometrics of randomized experiments." In *Handbook of economic field experiments*, volume 1, 73–140, Elsevier.

Attanasio, Orazio, Britta Augsburg, Ralph De Haas, Emla Fitzsimons, and Heike Harmgart (2015), "The impacts of microfinance: Evidence from joint-liability lending in mongolia." *American Economic Journal: Applied Economics*, 7, 90–122.

Bai, Yuehao, Joe Romano, Azeem Shaikh, and Max Tabord-Meehan (2022), "Inference for cluster randomized experiments with matched pairs." Working paper.

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

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan (2015), "The miracle of microfinance? evidence from a randomized evaluation." *American economic journal: Applied economics*, 7, 22–53.

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

Bugni, Federico, Ivan Canay, Vishal Kamat, Azeem Shaikh, and Max Tabord-Meehan (2022), "Inference under cluster-level covariate-adaptive randomization." Working paper.

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

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

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William 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, Clément and Jaime Ramirez-Cuellar (2020), "At what level should one cluster standard errors in paired experiments, and in stratified experiments with small strata?" Technical report, National Bureau of Economic Research.

Ding, Peng, Xinran Li, and Luke W Miratrix (2017), "Bridging finite and super population causal inference." *Journal of Causal Inference*, 5.

Donner, Allan and Neil Klar (2000), "Design and analysis of cluster randomization trials in health research."

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

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

Hansen, Christian B (2007), "Asymptotic properties of a robust variance matrix estimator for panel data when t is large." *Journal of Econometrics*, 141, 597–620.

Hayes, Richard J and Lawrence H Moulton (2017), *Cluster randomised trials*. Chapman and Hall/CRC.

Liang, Kung-Yee and Scott L Zeger (1986), "Longitudinal data analysis using generalized linear models." *Biometrika*, 73, 13–22.

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

MacKinnon, James G and Matthew Webb (2019), "When and how to deal with clustered errors in regression models." Technical report, Queen's Economics Department Working Paper.

Middleton, Joel A and Peter M Aronow (2015), "Unbiased estimation of the average treatment effect in cluster-randomized experiments." *Statistics, Politics and Policy*, 6, 39–75.

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

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

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

Schochet, Peter Z, Nicole E Pashley, Luke W Miratrix, and Tim Kautz (2021), "Design-based ratio estimators and central limit theorems for clustered, blocked rcts." *Journal of the American Statistical Association*, 1–12.

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

Turner, Elizabeth L, Fan Li, John A Gallis, Melanie Prague, and David M Murray (2017), "Review of recent methodological developments in group-randomized trials: part 1—design." *American journal of public health*, 107, 907–915.