THE UNIVERSITY OF CHICAGO

ESSAYS ON EXPERIMENTAL DESIGN UNDER COVARIATE-ADAPTIVE RANDOMIZATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
JIZHOU LIU

CHICAGO, ILLINOIS
AUGUST 2024

To my parents, Yuhui and Zhihua.

“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.”

— Ronald A. Fisher

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am deeply grateful to my advisors, Azeem Shaikh, Christian Hansen, Max Tabord-Meehan, and Tetsuya Kaji, for their invaluable insight and unwavering support. Their guidance has profoundly shaped both this dissertation and my development as a researcher.

Throughout the process of writing this dissertation, I have received support from numerous professors, colleagues, and friends. I am especially grateful to Yuehao Bai, who introduced me to the world of academic research in econometrics and has been an exceptional friend, coauthor, and mentor. I am also deeply thankful to Panos Toulis, who first introduced me to randomization tests and has provided tremendous help and patience. I thank all my friends, colleagues, and staff in the Booth Ph.D. program for helping me navigate this journey. Finally, I am extremely grateful to my parents for their endless love and support.

ABSTRACT

This dissertation studies statistical inference in randomized experiments, extending important covariate-adaptive randomization tools to three commonly used experimental designs. Each chapter addresses a distinct experimental design, contributing to the broader field of design and analysis of experiments.

Chapter 1 investigates inference in randomized controlled trials with multiple treatments, specifically under a "matched tuples" design. Here, units are grouped into homogeneous blocks, and each treatment is randomly assigned within these blocks. The study establishes conditions for the asymptotic normality of a sample analogue estimator and constructs a consistent estimator of its asymptotic variance. It also compares the asymptotic properties of the fully-blocked 2^K factorial design with stratified factorial designs, demonstrating the efficiency of the former. Simulation studies and empirical applications highlight the practical implications of these results.

Chapter 2 explores inference in cluster randomized trials using a "matched pairs" design, where clusters are paired based on baseline covariates and one cluster in each pair is randomly assigned to treatment. This chapter presents the large-sample behavior of a weighted difference-in-means estimator and proposes a unified variance estimator consistent under different matching regimes. It also evaluates common t-tests and a randomization test within this framework, establishing their validity. Additionally, a covariate-adjusted estimator is proposed, showing precision improvements under certain conditions. Theoretical findings are supported by a simulation study.

Chapter 3 addresses inference in two-stage randomized experiments under covariate-adaptive randomization. In this design, clusters are first stratified and assigned to treatment or control, followed by a second stage where units within treated clusters are further randomized. The chapter develops difference-in-"average of averages" estimators for primary and spillover effects, proving their consistency and asymptotic normality. It also demonstrates the efficiency of using covariate information in the design stage and the pitfalls of ignoring it. Finally, it studies optimal use of covariate information under covariate-adaptive randomization in large samples. The theoretical results are validated through simulations and an empirical application.

Together, these chapters advance the understanding of statistical inference in complex experimental designs, offering robust methods for empirical researchers dealing with stratified experiments.

CHAPTER 1

INFERENCE FOR MATCHED TUPLES AND FULLY BLOCKED FACTORIAL DESIGNS

1.1 Introduction

This paper studies inference in randomized controlled trials with multiple treatments, where treatment status is determined according to a "matched tuples" design. If there are $|\mathcal{D}|$ possible treatments, then by a matched tuples design, we mean an experimental design where units are sampled i.i.d. from the population of interest, grouped into "homogeneous" blocks of size $|\mathcal{D}|$, and finally, within each block, exactly one individual is randomly assigned to each of the $|\mathcal{D}|$ treatments. As such, matched tuples designs generalize the concept of matched pairs designs to settings with more than two treatments. Matched tuples designs are commonly used in the social sciences: see Bold et al. (2018), Brown and Andrabi (2020), de Mel et al. (2013), and Fafchamps et al. (2014) for examples in economics, and are often motivated using the simulation evidence presented in Bruhn and McKenzie (2009a). However, we are not aware of any formal results which establish valid asymptotically exact methods of inference for matched tuples designs. Accordingly, in this paper we establish general results about estimation and inference for matched tuples designs, and then apply these results to study the asymptotic properties of what we call "fully-blocked" 2^K factorial designs.

We first study estimation and inference for matched tuples designs in the general setting where the parameter of interest is a vector of linear contrasts over the collection of average outcomes for each treatment. Parameters of this form include standard average treatment effects (ATEs) used to compare one treatment relative to another, but as we explain below also include more complicated parameters which may be of interest, for instance, in the analysis of factorial designs. We first establish conditions under which a sample analogue estimator is asymptotically normal and construct a consistent estimator of its corresponding asymptotic variance. Combining these results establishes the asymptotic validity of tests based on these estimators. We then consider the asymptotic properties of two commonly recommended inference procedures. The first is based on a linear regression with block fixed effects. Importantly, we find the t-test based on such a regression is in general not valid for testing the null hypothesis that a pairwise ATE is equal to a prespecified value. The second is based on a linear regression with cluster-robust standard errors, where clusters are defined at the block level. Here we find that the corresponding t-test is generally valid but conservative, and that this conservativeness increases in the number of treatments.

Next, we apply our results to study the asymptotic properties of "fully-blocked" 2^K factorial designs. Factorial designs are classical experimental designs (see Wu and Hamada, 2011, for a textbook treatment) which are increasingly being used in the social sciences (see for instance Alatas et al., 2012; Besedeš et al., 2012; DellaVigna et al., 2016; Kaur et al., 2015; Karlan et al., 2014). In a 2^K factorial design, each treatment is a combination of multiple "factors," where each factor can take two distinct values, or "levels." As a consequence, a full 2^K factorial design can be thought of as a randomized experiment with 2^K distinct treatments (importantly however, the analysis of factorial designs typically considers factorial effects as the parameters of interest: see Section 1.3.3 for a definition). A fully-blocked factorial design is then simply a matched tuples design with blocks of size 2^K. Leveraging our previous results, we establish that our estimator achieves a lower asymptotic variance under the fully-blocked design than under any stratified factorial design which stratifies the experimental sample into a finite number of "large" strata (such designs include complete randomization as a special case). We also consider settings where only one factor may be of primary interest, and establish that even in such cases it is more efficient to perform a fully-blocked design than to perform a matched pairs design which exclusively focuses on the primary factor of interest.

In a simulation study, we find that although our inference results are asymptotically exact, our proposed tests may be conservative in finite samples when the experiment features many treatments or many blocking variables. Accordingly, we also study the behavior of a matched tuples design with "replicates," where we form blocks of size *two* times the number of treatments, and each treatment is assigned exactly *twice* at random within each block. Although we find that such a design results in an estimator with slightly larger mean-squared error, the rejection probabilities of our proposed tests become much closer to the nominal level, which may result in improved power. Further discussion is provided in Section 1.3.2 below.

Although the analysis of matched tuples designs has to our knowledge not received much attention, there are large literatures on both the analysis of matched pairs designs and the analysis of factorial designs. Recent papers which have analyzed the properties of matched pairs designs include Athey and Imbens (2017a), Bai et al. (2021a), Bai (2022a), de Chaisemartin and Ramirez-Cuellar (2022a), Cytrynbaum (2021), Imai et al. (2009), Jiang et al. (2020), Fogarty (2018), and van der Laan et al. (2012). Our analysis builds directly on the framework developed in Bai et al. (2021a), and our Theorems 1.3.1 and 1.3.2 nest some of their results when specialized to the setting of a binary treatment. Cytrynbaum (2021) considers a generalization of matched pairs designs, a special case of which he refers to as a matched tuples design. However, his design groups units into homogeneous blocks in order to assign a binary treatment with unequal treatment fractions. In contrast, we consider a setting where units are grouped into homogeneous blocks in order to assign multiple treatments.

Recent papers which have analyzed factorial designs include Branson et al. (2016), Dasgupta et al. (2015), Li et al. (2020), Muralidharan et al. (2019), Pashley and Bind (2019), and Liu et al. (2022). Our setup and notation for 2^K factorial designs mirrors the framework introduced in Dasgupta et al. (2015), although our setup differs in that we consider a "super-population" framework where potential outcomes are modeled as random, whereas they maintain a finite population framework where potential outcomes are modeled as fixed.[1] Borrowing the framework from Dasgupta et al. (2015), Branson et al. (2016) and Li et al. (2020) propose re-randomization designs for factorial experiments which are shown to have favorable efficiency properties relative to a completely randomized design. Although we do not provide formal results comparing our fully-blocked design to these re-randomization designs, our simulation evidence suggests that, at least in the inferential framework considered in our paper, the fully-blocked design can improve efficiency relative to these re-randomization designs. Also closely related to our paper is Liu et al. (2022), who extend the results in Dasgupta et al. (2015) to general stratified randomized designs. Their results on variance estimation specifically exclude the setting where each treatment is assigned exactly once per block, which is the primary setting that we consider in this paper.

The rest of the paper is organized as follows. In Section 1.2 we describe our setup and notation. Section 1.3 presents the main results. In Section 1.4, we examine the finite sample behavior of various experimental designs via simulation in the context of 2^K factorial experiments. Finally, in Section 1.5 we illustrate our proposed inference methods in an empirical application based on the experiment conducted in Fafchamps et al. (2014). We conclude with recommendations for empirical practice in Section 1.6.

1.2 Setup and Notation

Let $Y_i \in \mathbf{R}$ denote the observed outcome of interest for the ith unit. Let $D_i \in \mathcal{D}$ denote treatment status for the ith unit, where $\mathcal{D}$ denotes a finite set of values of the treatment. We assume $\mathcal{D} = \{1, \ldots, |\mathcal{D}|\}$. Generally, we use $D_i = 1$ to indicate the ith unit is untreated,

[1] The finite population "design-based" perspective may be particularly attractive in settings where the experimental sample is not explicitly drawn from a larger population. In Appendix A.4.2 we provide some preliminary simulation evidence that our proposed estimators may be relevant in such a setting as well, however, given the simulation evidence in de Chaisemartin and Ramirez-Cuellar (2022a) and our currently incomplete understanding of the design-based properties of our estimators, we do not make any general claims in this paper.

but such a restriction is not necessary for our results. Let X_i denote the observed baseline covariates for the ith unit, and denote its dimension by $\dim(X_i)$. For $d \in \mathcal{D}$, let $Y_i(d)$ denote the potential outcome for the ith unit if its treatment status were d. The observed outcome and potential outcomes are related to treatment status by the expression

$$Y_i = \sum_{d \in \mathcal{D}} Y_i(d) I\{D_i = d\} \ . \tag{1.1}$$

We suppose our sample consists of $J_n := (|\mathcal{D}|)n$ i.i.d. units. For any random variable indexed by i, for example D_i, we denote by $D^{(n)}$ the random vector $(D_1, D_2, \ldots, D_{J_n})$. Let P_n denote the distribution of the observed data $Z^{(n)}$ where $Z_i = (Y_i, D_i, X_i)$, and Q_n denote the distribution of $W^{(n)}$, where $W_i = (Y_i(1), Y_i(2), \ldots, Y_i(|\mathcal{D}|), X_i)$. We assume that $W^{(n)}$ consists of J_n i.i.d observations, so that $Q_n = Q^{J_n}$, where Q is the marginal distribution of W_i. Given Q_n, P_n is then determined by (1.1) and the mechanism for determining treatment assignment. We thus state our assumptions in terms of assumptions on Q and the treatment assignment mechanism.

Our object of interest will generically be defined as a vector of linear contrasts over the collection of expected potential outcomes across treatments. Formally, let

$$\Gamma(Q) := (\Gamma_1(Q), \ldots, \Gamma_{|\mathcal{D}|}(Q))' \ ,$$

where $\Gamma_d(Q) := E_Q[Y_i(d)]$ for $d \in \mathcal{D}$. Let ν be a real-valued $m \times |\mathcal{D}|$ matrix. We define

$$\Delta_\nu(Q) := \nu \Gamma(Q) \in \mathbf{R}^m \ ,$$

as our generic parameter of interest. For example, in the special case where $\mathcal{D} = \{1, 2\}$ and $\nu = (-1, 1)$, $\Delta_\nu(Q) = E_Q[Y_i(2) - Y_i(1)]$ corresponds to the familiar average treatment effect for a binary treatment. Further examples of $\Delta_\nu(Q)$ are provided in Examples 1.3.1 and 1.3.2

below.

We now describe our assumptions on Q. Our first assumption imposes restrictions on the (conditional) moments of the potential outcomes:

Assumption 1.2.1. The distribution Q is such that

(a) $0 < E[\text{Var}[Y_i(d)|X_i]]$ for $d \in \mathcal{D}$.

(b) $E[Y_i^2(d)] < \infty$ for $d \in \mathcal{D}$.

(c) $E[Y_i(d)|X_i = x]$, $E[Y_i^2(d)|X_i = x]$, and $\text{Var}[Y_i(d)|X_i]$ are Lipschitz for $d \in \mathcal{D}$.

Assumption 1.2.1(a) is a mild restriction imposed to rule out degenerate situations and Assumption 1.2.1(b) is another mild restriction that permits the application of suitable laws of large numbers and central limit theorems. Assumption 1.2.1(c), on the other hand, is a smoothness requirement that ensures that units that are "close" in terms of their baseline covariates are also "close" in terms of their potential outcomes. Assumption 1.2.1(c) is a key assumption for establishing the asymptotic exactness of our proposed tests, since it allows us to argue that certain intermediate quantities in the derivations of our variance estimators vanish asymptotically (see for instance the proof of Lemma A.3.2). Similar smoothness requirements are also imposed in Bai et al. (2021a).

Next, we specify our assumptions on the mechanism determining treatment status. In words, we consider treatment assignments which first stratify the experimental sample into n blocks of size $|\mathcal{D}|$ using the observed baseline covariates $X^{(n)}$, and then assign one unit to each treatment uniformly at random within each block. We call such a design a *matched tuples* design. Formally, let

$$\lambda_j = \lambda_j(X^{(n)}) \subseteq \{1, \ldots, J_n\},\ 1 \leq j \leq n$$

denote n sets each consisting of $|\mathcal{D}|$ elements that form a partition of $\{1, \ldots, J_n\}$.

We assume treatment is assigned as follows:

Assumption 1.2.2. Treatments are assigned so that $\{Y^{(n)}(d) : d \in \mathcal{D}\} \perp\!\!\!\perp D^{(n)}|X^{(n)}$ and, conditional on $X^{(n)}$,

$$\{(D_i : i \in \lambda_j) : 1 \le j \le n\}$$

are i.i.d. and each uniformly distributed over all permutations of $(1, 2, \dots, |\mathcal{D}|)$.

We further require that the units in each block be "close" in terms of their baseline covariates in the following sense:

Assumption 1.2.3. The blocks satisfy

$$\frac{1}{n} \sum_{1 \le j \le n} \max_{i,k \in \lambda_j} ||X_i - X_k||^2 \xrightarrow{P} 0 \ .$$

We will also sometimes require that the distances between units in adjacent blocks be "close" in terms of their baseline covariates:

Assumption 1.2.4. The blocks satisfy

$$\frac{1}{n} \sum_{1 \le j \le \lfloor n/2 \rfloor} \max_{i \in \lambda_{2j-1}, k \in \lambda_{2j}} ||X_i - X_k||^2 \xrightarrow{P} 0 \ .$$

We provide three examples of blocking algorithms which satisfy Assumptions 1.2.3–1.2.4:

1. Univariate covariate: When $\dim(X_i) = 1$, we can order units from smallest to largest according to X_i and then block adjacent units into blocks of size $|\mathcal{D}|$. It then follows from Theorem 4.1 of Bai et al. (2021a) that Assumptions 1.2.3–1.2.4 are satisfied as long as $E[X_i^2] < \infty$.

2. Pre-stratification: Suppose we have a covariate vector $\tilde{X}_i = (\tilde{X}_{1i}, \tilde{X}_{2i})$, where $\dim(\tilde{X}_{2i}) = 1$. Let S be a function that maps from the support of $\tilde{X}_{1i}$ to a discrete set $\mathcal{S} =$

$\{1, \ldots, |\mathcal{S}|\}$. Define $S_{1i} = S(\tilde{X}_{1i})$. For all units with the same value of S_i, order the units from smallest to largest according to $\tilde{X}_{2i}$ and then block adjacent units into blocks of size $|\mathcal{D}|$.[2] It follows from Theorem 4.1 of Bai et al. (2021a) that the resulting blocks satisfy Assumptions 1.2.3–1.2.4 with $X_i = (S_{1i}, \tilde{X}_{2i})$ as long as $E[\tilde{X}_{2i}^2] < \infty$. As an example, suppose $\tilde{X}_1 =$ (gender, education level) and $\tilde{X}_2 =$ income. In this case, the blocks could be formed by first stratifying according to gender and education level and then blocking on income. A similar blocking procedure is used in the experiment conducted by Fafchamps et al. (2014) which we revisit in our empirical application in Section 1.5.

3. Recursive pairing: When $\dim(X_i) > 1$ and $|\mathcal{D}| = 2^K$ for some K, we could form blocks by repeatedly implementing the "pairs of pairs" algorithm in Section 4 of Bai et al. (2021a) to successively larger groups of size 2^k for $k = 0, 1, \ldots, K$. To do this, units would first be matched into pairs (using for instance the non-bipartite matching algorithm from the `R` package `nbpMatching`). Next, these matched pairs would themselves be matched into "pairs of pairs" using the average value of the covariates in each pair, in order to generate groups of size four. Continuing in this fashion, we would match pairs of groups until obtaining groups of size 2^K. This is the algorithm we employ in our simulation designs. Such an algorithm could again be shown to satisfy Assumptions 1.2.3–1.2.4.

[2] If the number of units in a stratum is not divisible by $|\mathcal{D}|$, we could simply assign the remaining units at random or drop them from the experiment.

1.3 Main Results

1.3.1 Inference for Matched Tuples Designs

In this section, we study estimation and inference for a general m-dimensional parameter $\Delta_\nu(Q)$ under a matched tuples design. For a pre-specified $\ell \times 1$ column vector Δ_0 and $\ell \times m$ matrix Ψ of rank ℓ, the testing problem of interest is

$$H_0 : \Psi\Delta_\nu(Q) = \Delta_0 \text{ versus } H_1 : \Psi\Delta_\nu(Q) \neq \Delta_0 \tag{1.2}$$

at level $\alpha \in (0, 1)$. First we describe our estimator of $\Delta_\nu(Q)$. For $d \in \mathcal{D}$, define

$$\hat{\Gamma}_n(d) := \frac{1}{n} \sum_{1 \leq i \leq J_n} I\{D_i = d\}Y_i \ ,$$

and let $\hat{\Gamma}_n = (\hat{\Gamma}_n(1), \ldots, \hat{\Gamma}_n(|\mathcal{D}|))'$. In words, $\hat{\Gamma}_n(d)$ is simply the sample mean of the observations with treatment status $D_i = d$, and $\hat{\Gamma}_n$ is the vector of sample means across all treatments $d \in \mathcal{D}$. With $\hat{\Gamma}_n$ in hand, our estimator of $\Delta_\nu(Q)$ is then given by

$$\hat{\Delta}_{\nu,n} := \nu\hat{\Gamma}_n \ .$$

In what follows, it will be useful to define $\Gamma_d(X_i) := E[Y_i(d)|X_i]$. Our first result derives the limiting distribution of $\hat{\Delta}_{\nu,n}$ under our maintained assumptions.

Theorem 1.3.1. *Suppose Q satisfies Assumption 1.2.1 and the treatment assignment mechanism satisfies Assumptions 1.2.2–3.4.3. Then,*

$$\sqrt{n}(\hat{\Delta}_{\nu,n} - \Delta_\nu(Q)) \xrightarrow{d} N(0, \mathbb{V}_\nu) \ ,$$

where $\mathbb{V}_\nu := \nu \mathbb{V} \nu'$*, with*

$$\mathbb{V} := \mathbb{V}_1 + \mathbb{V}_2 \ , \tag{1.3}$$

$$\mathbb{V}_1 := \operatorname{diag}(E[\operatorname{Var}[Y_i(d)|X_i]] : d \in \mathcal{D}) \ ,$$

$$\mathbb{V}_2 := \left[\frac{1}{|\mathcal{D}|}\operatorname{Cov}[\Gamma_d(X_i), \Gamma_{d'}(X_i)]\right]_{d,d' \in \mathcal{D}} \ .$$

To construct our test, we next define a consistent estimator for the asymptotic variance matrix $\mathbb{V}_\nu$. To begin, note by the law of total variance that

$$E[\operatorname{Var}[Y_i(d)|X_i]] = \operatorname{Var}[Y_i(d)] - E[E[Y_i(d)|X_i]^2] + E[Y_i(d)]^2 \ .$$

Therefore, in order to estimate $\mathbb{V}_1$ consistently, it suffices to provide consistent estimators for $E[E[Y_i(d)|X_i]^2]$, $E[Y_i(d)]$, and $\operatorname{Var}[Y_i(d)]$. A similar remark applies to $\mathbb{V}_2$. In light of this, define

$$\begin{aligned}
\hat{\rho}_n(d,d) &:= \frac{2}{n} \sum_{1 \le j \le \lfloor n/2 \rfloor} \Big(\sum_{i \in \lambda_{2j-1}} Y_i I\{D_i = d\}\Big)\Big(\sum_{i \in \lambda_{2j}} Y_i I\{D_i = d\}\Big) \\
\hat{\rho}_n(d,d') &:= \frac{1}{n} \sum_{1 \le j \le n} \Big(\sum_{i \in \lambda_j} Y_i I\{D_i = d\}\Big)\Big(\sum_{i \in \lambda_j} Y_i I\{D_i = d'\}\Big) \text{ if } d \ne d' \\
\hat{\sigma}_n^2(d) &:= \frac{1}{n} \sum_{1 \le i \le J_n} (Y_i - \hat{\Gamma}_n(d))^2 I\{D_i = d\} \ .
\end{aligned}$$

To understand the construction, note that in order to estimate $E[E[Y_i(d)|X_i]^2]$ consistently, we would ideally average over the products of the outcomes of two units with similar values of X_i and both with treatment status d. By construction, however, only one unit in each block has treatment status d. To overcome this problem, note that Assumption 1.2.4 ensures that in the limit units in adjacent blocks also have similar values of X_i. Therefore, to construct our estimator of $E[E[Y_i(d)|X_i]^2]$, denoted by $\hat{\rho}_n(d,d)$, we average over the product of the outcomes of the units with treatment status d in two adjacent blocks. $\hat{\rho}_n(d,d)$ is analogous to the “pairs of pairs" variance estimator in Bai et al. (2021a). A similar construction has also been used in Abadie and Imbens (2008) in a related setting. On the other hand, for $d \neq d'$, we have distinct units with treatment status d and d' within each block, and therefore our estimator of $E[E[Y_i(d)|X_i]E[Y_i(d')|X_i]]$, denoted $\hat{\rho}_n(d, d')$, can be estimated using units within the same block.

Our estimator for $\mathbb{V}_\nu$ is then given by $\hat{\mathbb{V}}_{\nu,n} := \nu\hat{\mathbb{V}}_n\nu'$, where

$$\begin{aligned}
\hat{\mathbb{V}}_n &:= \hat{\mathbb{V}}_{1,n} + \hat{\mathbb{V}}_{2,n} \\
\hat{\mathbb{V}}_{1,n} &:= \operatorname{diag}\left(\hat{\mathbb{V}}_{1,n}(d) : d \in \mathcal{D}\right) \\
\hat{\mathbb{V}}_{2,n} &:= \left[\hat{\mathbb{V}}_{2,n}(d, d')\right]_{d,d' \in \mathcal{D}} ,
\end{aligned}$$

with

$$\begin{aligned}
\hat{\mathbb{V}}_{1,n}(d) &:= \hat{\sigma}_n^2(d) - (\hat{\rho}_n(d, d) - \hat{\Gamma}_n^2(d)) \\
\hat{\mathbb{V}}_{2,n}(d, d') &:= \frac{1}{|\mathcal{D}|}(\hat{\rho}_n(d, d') - \hat{\Gamma}_n(d)\hat{\Gamma}_n(d')) .
\end{aligned}$$

Given this estimator, our test is given by

$$\phi_n^\nu(Z^{(n)}) = I\{T_n^\nu(Z^{(n)}) > c_{1-\alpha}\} ,$$

where

$$T_n^\nu(Z^{(n)}) = n(\Psi\hat{\Delta}_{\nu,n} - \Psi\Delta_0)'(\Psi\hat{\mathbb{V}}_{\nu,n}\Psi')^{-1}(\Psi\hat{\Delta}_{\nu,n} - \Psi\Delta_0) ,$$

and $c_{1-\alpha}$ is the $1 - \alpha$ quantile of the χ_ℓ^2 distribution. Our next result establishes the consistency of $\hat{\mathbb{V}}_n$ for $\mathbb{V}$ and the asymptotic validity of the above test.

Theorem 1.3.2. *Suppose Q satisfies Assumption 1.2.1 and the treatment assignment mech-*

anism satisfies Assumptions 1.2.2–1.2.4. Then,

$$\hat{\mathbb{V}}_n \xrightarrow{P} \mathbb{V} \ .$$

Therefore, for the problem of testing (1.2) at level $\alpha \in (0,1)$, $\phi_n^\nu(Z^{(n)})$ satisfies

$$\lim_{n \to \infty} E[\phi_n^\nu(Z^{(n)})] = \alpha \ ,$$

under the null hypothesis.

Example 1.3.1. (Inference for Matched Triples) Consider the setting where $\mathcal{D} = \{1,2,3\}$, where we consider $d = 1$ as a control arm and $d = 2,3$ as treatment sub-arms. See, for example, Bold et al. (2018) and Brown and Andrabi (2020). Suppose our parameter of interest is the vector of average treatment effects for the treatments $d = 2,3$ versus control $d = 1$. In this case, the parameter of interest is given by $\Delta_\nu(Q)$, where

$$\nu = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \ .$$

It follows from Theorem 1.3.1 that

$$\sqrt{n}(\hat{\Delta}_{\nu,n} - \Delta_\nu(Q)) \xrightarrow{d} N(0, \mathbb{V}_\nu) \ ,$$

where

$$\mathbb{V}_\nu = \begin{pmatrix} \sigma^2_{\nu,1,1} & \sigma^2_{\nu,1,2} \\ \sigma^2_{\nu,1,2} & \sigma^2_{\nu,2,2} \end{pmatrix} \ ,$$

and

$$
\begin{aligned}
\sigma^2_{\nu,1,1} &= E[\mathrm{Var}[Y_i(1)|X_i]] + E[\mathrm{Var}[Y_i(2)|X_i]] + \frac{1}{3}E\left[((\Gamma_1(X_i)-\Gamma_1)-(\Gamma_2(X_i)-\Gamma_2))^2\right] \\
\sigma^2_{\nu,2,2} &= E[\mathrm{Var}[Y_i(1)|X_i]] + E[\mathrm{Var}[Y_i(3)|X_i]] + \frac{1}{3}E\left[((\Gamma_1(X_i)-\Gamma_1)-(\Gamma_3(X_i)-\Gamma_3))^2\right] \\
\sigma^2_{\nu,1,2} &= E[\mathrm{Var}[Y_i(1)|X_i]] \\
&\quad + \frac{1}{3}E\left[((\Gamma_1(X_i)-\Gamma_1)-(\Gamma_2(X_i)-\Gamma_2))\,((\Gamma_1(X_i)-\Gamma_1)-(\Gamma_3(X_i)-\Gamma_3))\right]~,
\end{aligned}
$$

where we recall $\Gamma_d(X_i) = E[Y_i(d)|X_i]$. These variance formulas imply the following two observations: first, by decomposing $\sigma^2_{\nu,1,1}$ using the law of total variance, we can show that the commonly-used two-sample t-test is conservative when testing the null hypothesis on the contrast of any two treatment levels in a matched tuples design. A similar observation was made in the special case of a matched-pair design in Bai et al. (2021a). Second, the adjusted t-test developed in Bai et al. (2021a) is also conservative for testing such hypotheses. Specifically, Bai et al. (2021a) study inference for $E[Y(2) - Y(1)]$ in a matched-pair design when $|\mathcal{D}| = 2$ and the sample size is $2n$. In a matched triples experiment with $|\mathcal{D}| = 3$ and sample size $3n$, researchers may be tempted to apply the variance estimator from Theorem 3.3 in Bai et al. (2021a) to the subsample with $D_i \in \{1, 2\}$. However, it can be shown in our framework that the limit of the variance estimator from Bai et al. (2021a) is given by replacing $\frac{1}{3}$ in the last term of $\sigma^2_{\nu,1,1}$ with $\frac{1}{2}$. Therefore, the test which studentizes using the variance estimator from Bai et al. (2021a) would be asymptotically conservative in our setting. ■

Next, we study the properties of two commonly recommended inference procedures in the analysis of matched tuple designs. The first procedure is a t-test obtained from a linear regression of outcomes on treatment indicators and block fixed effects. Specifically, we consider a t-test obtained from the following regression:

$$Y_i = \sum_{d \in \mathcal{D} \backslash \{1\}} \beta(d) I\{D_i = d\} + \sum_{1 \leq j \leq n} \delta_j I\{i \in \lambda_j\} + \epsilon_i \ , \tag{1.4}$$

which we interpret as the projection of Y on the indicators for treatment status and block fixed effects. Let $\hat{\beta}_n(d)$, $d \in \mathcal{D} \backslash \{1\}$ and $\hat{\delta}_{j,n}$, $1 \leq j \leq n$ denote the OLS estimators of $\beta(d)$, $d \in \mathcal{D} \backslash \{1\}$ and δ_j, $1 \leq j \leq n$. It is common in practice to use $\hat{\beta}_n(d)$ as an estimator for the pairwise average treatment effect between treatment d and treatment 1. See, for instance, de Mel et al. (2013) and Fafchamps et al. (2014). Furthermore, researchers often conduct inference on the pairwise ATEs using the heteroskedasticity-robust variance estimator obtained from (1.4). Formally, for $d \in \mathcal{D} \backslash \{1\}$ and $\Delta_0 \in \mathbf{R}$, consider the problem of testing

$$E_Q[Y_i(d)] - E_Q[Y_i(1)] = \Delta_0 \text{ versus } H_1 : E_Q[Y_i(d)] - E_Q[Y_i(1)] \neq \Delta_0 \tag{1.5}$$

at level $\alpha \in (0, 1)$. Let $\kappa_j \cdot \hat{\mathbb{V}}_n^{\text{sfe}}(d, 1)$ denote the "HCj" heteroskedasticity-robust variance estimator of $\hat{\beta}_n(d)$ from the linear regression in (1.4), where κ_j for $j \in \{0, 1\}$ corresponds to one of two common degrees of freedom corrections (see MacKinnon and White, 1985):

$$\kappa_j = \begin{cases} 1 & \text{if } j = 0 \\ \frac{|\mathcal{D}|n}{|\mathcal{D}|n - (|\mathcal{D}| - 1 + n)} & \text{if } j = 1 \ . \end{cases}$$

The test is then defined as

$$\phi_n^{\text{sfe}}(Z^{(n)}) = I\{|T_n^{\text{sfe}}(Z^{(n)})| > z_{1-\frac{\alpha}{2}}\} \ , \tag{1.6}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-th quantile of the standard normal distribution and

$$T_n^{\text{sfe}}(Z^{(n)}) = \frac{\hat{\beta}_n(d) - \Delta_0}{\sqrt{\kappa_j \cdot \hat{\mathbb{V}}_n^{\text{sfe}}(d,1)}} \ . \tag{1.7}$$

The following theorem shows that the OLS estimator $\hat{\beta}_n(d)$ is numerically equivalent to the standard difference-in-means estimator. However, it shows that the t-test defined in (1.6) is not generally valid for testing the null hypothesis defined in (1.5).

Theorem 1.3.3. *Suppose Q satisfies Assumption 1.2.1 and the treatment assignment mechanism satisfies Assumptions 1.2.2–1.2.4. Then,*

$$\hat{\beta}_n(d) = \hat{\Gamma}_n(d) - \hat{\Gamma}_n(1) \textit{ for } d \in \mathcal{D}\backslash\{1\} \ .$$

Moreover,

- *Using estimator* HC0*, the limiting rejection probability of the test defined in* (1.6) *could be strictly larger than α.*
- *Using estimator* HC1*, the limiting rejection probability of the test defined in* (1.6) *could be strictly larger than α for $|\mathcal{D}| > 2$.*

Bai et al. (2021a) remark that the test defined in (1.6) is conservative in the context of a matched-pair design when using HC1. Theorem 1.3.3 shows that, when considering a matched tuples design with more than two treatments, this is no longer necessarily the case.

Remark 1.3.1. An inspection of the proof of Theorem 1.3.3 reveals that the probability limit of $n \cdot \kappa_1 \hat{\mathbb{V}}_n^{\rm sfe}(d,1)$ is given by

$$\begin{aligned}\frac{|\mathcal{D}|}{|\mathcal{D}|-1}\Bigg(& \operatorname{Var}\left[\Gamma_1(X_i) - \frac{1}{|\mathcal{D}|}\sum_{d' \in \mathcal{D}} \Gamma_{d'}(X_i)\right] \\ & + \left(1 - \frac{1}{|\mathcal{D}|}\right)^2 E[\operatorname{Var}[Y_i(1)|X_i]] + \frac{1}{|\mathcal{D}|^2}\sum_{d' \in \mathcal{D}\backslash\{1\}} E[\operatorname{Var}[Y_i(d')|X_i]] \\ & + \operatorname{Var}\left[\Gamma_d(X_i) - \frac{1}{|\mathcal{D}|}\sum_{d' \in \mathcal{D}} \Gamma_{d'}(X_i)\right] + \left(1 - \frac{1}{|\mathcal{D}|}\right)^2 E[\operatorname{Var}[Y_i(d)|X_i]] \\ & + \frac{1}{|\mathcal{D}|^2}\sum_{d' \in \mathcal{D}\backslash\{d\}} E[\operatorname{Var}[Y_i(d')|X_i]]\Bigg) ~,\end{aligned}$$

whereas the true asymptotic variance of $\hat{\beta}_n(d)$ is given by

$$E\left[\operatorname{Var}[Y_i(d)|X_i]\right] + E[\operatorname{Var}[Y_i(1)|X_i]] + \frac{1}{|\mathcal{D}|}E\left[((\Gamma_d(X_i) - \Gamma_d) - (\Gamma_1(X_i) - \Gamma_1))^2\right] ~.$$

From these expressions, we can conclude that when $|\mathcal{D}|$ is large it is likely that $\kappa_1 \hat{\mathbb{V}}_n^{\rm sfe}(d,1)$ is conservative. However, as shown in the proof of Theorem 1.3.3, this cannot be guaranteed for finite $|\mathcal{D}| > 2$ in general. ■

The second procedure is a block-cluster robust t-test which modifies a recent proposal in de Chaisemartin and Ramirez-Cuellar (2022a) to the setting with multiple treatments. Specifically, we consider a cluster-robust t-test constructed from a regression of outcomes on a constant and treatment indicators:

$$Y_i = \gamma(1) + \sum_{d \in \mathcal{D}\backslash\{1\}} \gamma(d) I\{D_i = d\} + \epsilon_i ~,$$

where clusters are defined at the level of *blocks* of units $\{\lambda_j\}_{1 \le j \le \mathcal{D}}$. Let $\hat{\gamma}_n(d)$, $d \in \mathcal{D}\backslash\{1\}$ denote the OLS estimator of $\gamma(d)$, it then follows immediately that $\hat{\gamma}_n(d) = \hat{\Gamma}_n(d) - \hat{\Gamma}_n(1)$.

We then consider the problem of testing (1.5) at level $\alpha \in (0,1)$ using a test defined by

$$\phi_n^{\text{bcve}}(Z^{(n)}) = I\{|T_n^{\text{bcve}}(Z^{(n)})| > z_{1-\frac{\alpha}{2}}\} ,$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-th quantile of the standard normal distribution and

$$T_n^{\text{bcve}}(Z^{(n)}) = \frac{\hat{\gamma}_n(d) - \Delta_0}{\sqrt{\hat{\mathbb{V}}_n^{\text{bcve}}(d)}} , \tag{1.8}$$

with $\hat{\mathbb{V}}_n^{\text{bcve}}(d)$ denoting the d-th diagonal element of the block-cluster variance estimator defined as:

$$\begin{aligned}&\hat{\mathbb{V}}_n^{\text{bcve}} \\ &= \left(\sum_{1\le j\le n}\sum_{i\in\lambda_j} C_i C_i'\right)^{-1} \left(\sum_{1\le j\le n}\left(\sum_{i\in\lambda_j}\hat{\epsilon}_i C_i\right)\left(\sum_{i\in\lambda_j}\hat{\epsilon}_i C_i\right)'\right)\left(\sum_{1\le j\le n}\sum_{i\in\lambda_j} C_i C_i'\right)^{-1} ,\end{aligned} \tag{1.9}$$

where $C_i = (1, I\{D_i = 2\}, \dots, I\{D_i = |\mathcal{D}|\})'$ and $\hat{\epsilon}_i = \sum_{d\in\mathcal{D}\setminus\{1\}}(Y_i - \hat{\gamma}_n(d))I\{D_i = d\} + Y_i I\{D_i = 1\} - \hat{\gamma}_n(1)$.

The following theorem shows that the t-test defined in (1.8) is generally conservative for testing the null hypothesis defined in (1.5).

Theorem 1.3.4. *Consider the block-cluster variance estimator $\hat{\mathbb{V}}_n^{\text{bcve}}$ as defined in* (1.9) *in the Appendix. Then the d-th diagonal element of this estimator is equal to*

$$n \cdot \hat{\mathbb{V}}_n^{\text{bcve}}(d) = \frac{1}{n}\sum_{1\le j\le n}\left(\sum_{i\in\lambda_j} Y_i I\{D_i = d\} - \sum_{i\in\lambda_j} Y_i I\{D_i = 1\}\right)^2 - (\hat{\Gamma}_n(d) - \hat{\Gamma}_n(1))^2 .$$

Moreover, under Assumptions 1.2.1–1.2.3,

$$n\cdot\hat{\mathbb{V}}_n^{\mathrm{bcve}}(d) \xrightarrow{p} E[\mathrm{Var}[Y_i(d)|X_i]]+E[\mathrm{Var}[Y_i(1)|X_i]]+E\left[((\Gamma_d(X_i)-\Gamma_d)-(\Gamma_1(X_i)-\Gamma_1))^2\right] .$$

It thus follows that the test defined in (1.8) *is conservative for testing the null hypothesis defined in* (1.5) *unless*

$$E\left[((\Gamma_d(X_i)-\Gamma_d)-(\Gamma_1(X_i)-\Gamma_1))^2\right]=0 . \tag{1.10}$$

Remark 1.3.2. An inspection of the proof of Theorem 1.3.4 reveals that, unless (1.10) holds, the difference between the probability limit of $n\cdot\hat{\mathbb{V}}_n^{\mathrm{bcve}}(d)$ and the asymptotic variance of $\hat{\Gamma}_n(d)-\hat{\Gamma}_n(1)$ is equal to

$$\left(1-\frac{1}{|\mathcal{D}|}\right)E\left[((\Gamma_d(X_i)-\Gamma_d)-(\Gamma_1(X_i)-\Gamma_1))^2\right] .$$

It thus follows that the test defined in (1.8) in fact becomes more conservative for testing (1.5) as the number of treatments $|\mathcal{D}|$ increases. ■

1.3.2 Inference for "Replicate" Designs

Our analysis so far has focused on the setting where $J_n=|\mathcal{D}|n$ units are blocked into n blocks of size $|\mathcal{D}|$, and each treatment $d\in\mathcal{D}$ is assigned exactly once in each block. In this section, we consider a modification of this design where units are grouped into blocks of size $2|\mathcal{D}|$ and each treatment status $d\in\mathcal{D}$ is assigned exactly *twice* in each block. Formally, for the remainder of this section suppose n is even, and let

$$\tilde{\lambda}_j=\tilde{\lambda}_j(X^{(n)})\subseteq\{1,\dots,J_n\},\ 1\le j\le n/2$$

denote $n/2$ sets each consisting of $2|\mathcal{D}|$ elements that form a partition of $\{1,\dots,J_n\}$.

We assume treatment is assigned as follows:

Assumption 1.3.1. Treatments are assigned so that $\{Y^{(n)}(d) : d \in \mathcal{D}\} \perp\!\!\!\perp D^{(n)} | X^{(n)}$ and, conditional on $X^{(n)}$,

$$\{(D_i : i \in \tilde{\lambda}_j) : 1 \leq j \leq n/2\}$$

are i.i.d. and each uniformly distributed over all permutations of $(1, 1, 2, 2, \ldots, |\mathcal{D}|, |\mathcal{D}|)$.

We further require that the units in each block be "close" in terms of their baseline covariates in the following sense:

Assumption 1.3.2. The blocks satisfy

$$\frac{1}{n} \sum_{1 \leq j \leq n/2} \max_{i,k \in \tilde{\lambda}_j} ||X_i - X_k||^2 \xrightarrow{P} 0 \ .$$

We first establish that the limiting distribution of $\hat{\Delta}_{\nu,n}$ for such a "replicate" design is the same as that for the matched tuples design considered in Theorem 1.3.1.

Theorem 1.3.5. *Suppose Q satisfies Assumption 1.2.1 and the treatment assignment mechanism satisfies Assumptions 1.3.1–1.3.2. Then,*

$$\sqrt{n}(\hat{\Delta}_{\nu,n} - \Delta_\nu(Q)) \xrightarrow{d} N(0, \mathbb{V}_\nu) \ ,$$

with $\mathbb{V}_\nu$ as defined in Theorem 1.3.1.

Although the limiting distribution of $\hat{\Delta}_{\nu,n}$ for the standard matched tuples and replicate designs are identical, variance estimation in the replicate design is often understood to be conceptually simpler, because each treatment status is assigned *twice* in each block (see for instance the discussion of variance estimation in Athey and Imbens, 2017a, in the context of matched pair designs). Indeed, in this case an alternative variance estimator can be constructed which is identical to the estimator proposed in Section 1.3.1 except that we replace $\hat{\rho}_n(d, d)$ by

$$\tilde{\rho}_n(d, d) = \frac{2}{n} \sum_{1 \leq j \leq \lfloor n/2 \rfloor} \Big(\prod_{i \in \lambda_j} Y_i I\{D_i = d\} \Big) ,$$

which no longer requires averaging over the product of outcomes of units in adjacent blocks. The following theorem establishes the consistency of $\tilde{\rho}_n(d, d)$, where importantly we note that Assumption 1.2.4, which maintains that adjacent blocks be “close", is no longer required. It is then straightforward to show the consistency of the corresponding variance estimator for $\hat{\Delta}_{\nu,n}$ constructed by replacing $\hat{\rho}_n(d, d)$ in $\hat{\mathbb{V}}_n$ with $\tilde{\rho}_n(d, d)$.

Theorem 1.3.6. *Suppose Q satisfies Assumption 1.2.1 and the treatment assignment mechanism satisfies Assumptions 1.3.1–1.3.2. Then,*

$$\tilde{\rho}_n(d, d) \xrightarrow{P} E[E[Y_i(d)|X_i]^2] \ . \tag{1.11}$$

We remark that Theorems 1.3.1–1.3.2 and Theorems 1.3.5–1.3.6, yielding identical conclusions, do not allow us to effectively compare the properties of the standard matched tuples design and matched tuples with replicates. In order to compare these designs, we evaluate their finite sample properties via simulation in Section 1.4. There, we find that the mean squared error of $\hat{\Delta}_{\nu,n}$ under the replicate design is typically larger than under the standard non-replicate design. However, we also find that the rejection probabilities of our proposed tests under the replicate design are much closer to the nominal level relative to the non-replicate design, which can sometimes exhibit rejection probabilities strictly smaller than the nominal level when matching on multiple covariates. As a result, the replicate design is sometimes able to achieve better power relative to the non-replicate design. We emphasize, however, that our current asymptotic framework is not precise enough to capture these differences. One possible conjecture is that since replicate designs could be thought of as convex combinations of matched tuples designs (see Lemma 2 in Bai, 2022a), it is as if we are averaging over multiple matched tuples designs when we estimate the limiting variance. However, we leave a detailed theoretical comparison of these two designs to future work.

1.3.3 Asymptotic Properties of Fully-Blocked 2^K Factorial Designs

In this section we apply the results derived in Sections 1.3.1–1.3.2 to study the asymptotic properties of what we call “fully-blocked" 2^K factorial designs.

First, we describe the setup of a 2^K factorial experiment, the resulting parameters of interest, and their corresponding estimators (see Wu and Hamada, 2011, for a textbook treatment). A 2^K factorial design assigns treatments which are combinations of multiple “factors," where each factor can take two distinct values, or “levels.” For instance, Karlan et al. (2014) study the effect of capital constraints and uninsured risk on the investment decisions of farmers in Ghana. In their setting, each treatment consists of two factors: whether or not a household receives a cash grant, and whether or not a household receives an insurance grant. Our setup and notation mirror the framework introduced in Dasgupta et al. (2015) and Li et al. (2020). Given K factors each with two treatment levels $\{-1, +1\}$, our set of treatments $\mathcal{D}$ now consists of all possible 2^K factor combinations. For a factor combination $d \in \mathcal{D}$, define $\iota_k(d) \in \{-1, +1\}$ to be the level of factor k under treatment d. The vector $\iota(d) := (\iota_1(d), \iota_2(d), \ldots, \iota_K(d))$ then describes the levels of all K factors associated with factor combination d. This notation allows us to define *factorial effects* as parameters of the form $\Delta_\nu(Q)$ for appropriately constructed contrast vectors ν. For instance, consider the contrast vector defined as

$$\nu_k := (\iota_k(1), \iota_k(2), \ldots, \iota_k(|\mathcal{D}|)) \ .$$

Then, the parameter $\Delta_{\nu_k}(Q)$ obtained from this contrast can be written as

$$\Delta_{\nu_k}(Q) = \sum_{d \in \mathcal{D}} I\{\iota_k(d) = +1\}\Gamma_d(Q) - \sum_{d \in \mathcal{D}} I\{\iota_k(d) = -1\}\Gamma_d(Q) \ .$$

We define the *main effect* of factor k as $2^{-(K-1)}\Delta_{\nu_k}(Q)$. In words, the main effect of factor k measures the average difference between the outcomes of factor combinations under which the kth factorial effect is 1 versus the outcomes of factor combinations under which the kth factorial effect is -1. The re-scaling $2^{-(K-1)}$ is introduced because there are 2^{K-1} possible values for all the factor combinations when fixing the kth factor. We call ν_k the *generating vector* for the main effect of factor k.

We can subsequently build on the generating vectors of the main effects in order to define the *interaction effects* between various factors. The interaction effect between a given set of factors is defined using the contrast obtained from taking the element-wise product of the generating vectors for the relevant factors. For instance, the two-factor interaction between factors k and k' is defined as $2^{-(K-1)}\Delta_{\nu_{k,k'}}(Q)$, where $\nu_{k,k'} := \nu_k \odot \nu_{k'}$ and $\odot$ denotes element-wise multiplication. Similarly, the three-factor interaction $2^{-(K-1)}\Delta_{\nu_{k,k',k''}}(Q)$ is defined using the contrast vector $\nu_{k,k',k''} := \nu_k \odot \nu_{k'} \odot \nu_{k''}$. We illustrate these definitions in the special case of a 2^2 factorial design in Example 1.3.2 below. For simplicity, in what follows, we omit the re-scaling by $2^{-(K-1)}$ in our discussions and results.

Example 1.3.2. Here we illustrate the concept of main and interaction effects in the case of a 2^2 factorial design. Table 1.1 depicts the 4 factor combinations and their corresponding factor levels.

From the column labeled Factor 1 we observe that the generating vector for the main effect of factor one, ν_1, is given by

$$\nu_1 = (-1, -1, +1, +1) \ ,$$

Factor Combination	Factor 1	Factor 2	Factor 1/2 Interaction
1	-1	-1	+1
2	-1	+1	-1
3	+1	-1	-1
4	+1	+1	+1

Table 1.1: Example of a 2^2 factorial design

so that the main effect of factor one is given by (up to re-scaling)

$$\Delta_{\nu_1}(Q) = E_Q[Y_i(+1,+1) + Y_i(+1,-1)] - E_Q[Y_i(-1,+1) + Y_i(-1,-1)] \ ,$$

where here we have indexed potential outcomes explicitly by their factor levels. Similarly, the column labeled Factor 2 corresponds to the generating vector for the main effect of factor two, ν_2. To define the interaction effect between factors one and two, we construct the relevant contrast by taking the element-wise product of ν_1 and ν_2:

$$\nu_{1,2} = \nu_1 \odot \nu_2 = (+1,-1,-1,+1) \ ,$$

this produces the column labeled Factor 1/2 Interaction. Accordingly, the interaction effect between factors one and two is given by (up to re-scaling)

$$\Delta_{\nu_{1,2}}(Q) = E_Q[Y_i(+1,+1) - Y_i(-1,+1)] - E_Q[Y_i(+1,-1) - Y_i(-1,-1)] \ .$$

In words, $\Delta_{\nu_{1,2}}(Q)$ measures the difference in the the average difference in potential outcomes over factor one when factor two is set to 1 versus the average difference in potential outcomes over factor one when factor two is set to -1. ■

Given the above setup, we estimate the factorial effect given by $\Delta_\nu(Q)$ using the estimator $\hat{\Delta}_{\nu,n}$ defined in Section 1.3.1. Wu and Hamada (2011) and Dasgupta et al. (2015) explain that $\hat{\Delta}_{\nu,n}$ is a standard estimator in this context. For instance, the estimator of the main effect of factor k, $2^{-(K-1)}\hat{\Delta}_{\nu_k,n}$, is in fact the difference-in-means estimator over the k-th factor:

$$\begin{aligned}2^{-(K-1)}\hat{\Delta}_{\nu_k,n} &= \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}\hat{\Gamma}_n(d) - \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}} I\{\iota_k(d)=-1\}\hat{\Gamma}_n(d) \\ &= \frac{1}{n2^{K-1}}\sum_{1\le i\le J_n}\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}I\{D_i=d\}Y_i \\ &\quad - \frac{1}{n2^{K-1}}\sum_{1\le i\le J_n}\sum_{d\in\mathcal{D}} I\{\iota_k(d)=-1\}I\{D_i=d\}Y_i \\ &= \frac{1}{n2^{K-1}}\sum_{1\le i\le J_n} I\{\iota_k(D_i)=+1\}Y_i - \frac{1}{n2^{K-1}}\sum_{1\le i\le J_n} I\{\iota_k(D_i)=-1\}Y_i\ .\end{aligned}$$

Then, we compare the asymptotic variance of the estimator $\hat{\Delta}_{\nu,n}$ under what we call a “fully-blocked" factorial design relative to some alternative designs. A fully-blocked factorial design first blocks the experimental sample into n blocks of size 2^K based on the observable characteristics $X^{(n)}$, and then assigns each of the 2^K factor combinations exactly once in each block. Formally, a fully-blocked factorial design is simply a matched tuples design as defined in Section 1.2, where $\mathcal{D}$ consists of the set of all possible factor combinations.

Our first result compares the fully-blocked factorial design to completely randomized and stratified factorial designs. Given a 2^K factorial experiment and a sample of size $J_n = n2^K$, a completely randomized factorial design simply assigns n individuals to each of the 2^K factor combinations at random. A stratified factorial design first partitions the covariate space into a finite number of groups, or “strata", and then performs a completely randomized factorial design within each stratum. Formally, let $h : \text{supp}(X) \to \{1, \ldots, S\}$ be a function which maps covariate values into a set of discrete strata labels. Then, a stratified factorial design performs a completely randomized factorial design within each stratum produced by $h(\cdot)$. Note that a completely randomized design is a special case of the stratified factorial design where the co-domain of $h(\cdot)$ is a singleton. See Branson et al. (2016) and Li et al. (2020) for further discussion of these designs. Theorem 1.3.7 shows that the asymptotic variance of

$\hat{\Delta}_{\nu,n}$ is weakly smaller under a fully-blocked factorial design than that under *any* stratified factorial design as defined above, as long as the potential outcomes satisfy the smoothness assumptions described in Assumption 1.2.1(c).

Theorem 1.3.7. *Suppose Assumptions 1.2.1(a)-(b) hold and let $h : \operatorname{supp}(X) \to \{1, \ldots, S\}$ be any measurable function which maps covariate values into a set of discrete strata labels. Let $\Delta_\nu(Q)$ be a factorial effect for some 1×2^K contrast vector ν. Then under a stratified factorial design with strata defined by $h(\cdot)$,*

$$\sqrt{n}(\hat{\Delta}_{\nu,n} - \Delta_\nu(Q)) \xrightarrow{d} N(0, \sigma^2_{h,\nu}) \ ,$$

where $\sigma^2_{h,\nu} = \nu \mathbb{V}_h \nu'$, with

$$\begin{aligned}
\mathbb{V}_h &:= \mathbb{V}_{h,1} + \mathbb{V}_{h,2} \\
\mathbb{V}_{h,1} &:= \operatorname{diag}(E[\operatorname{Var}[Y_i(d)|h(X_i)]] : d \in \mathcal{D}) \\
\mathbb{V}_{h,2} &:= \left[\frac{1}{|\mathcal{D}|}\operatorname{Cov}[E[Y_i(d)|h(X_i)], E[Y_i(d')|h(X_i)]]\right]_{d,d' \in \mathcal{D}} \ .
\end{aligned}$$

Moreover,

$$\sigma^2_\nu \le \sigma^2_{h,\nu} \ ,$$

where $\sigma^2_\nu = \mathbb{V}_\nu$ (as defined in Theorem 1.3.1) is the asymptotic variance of $\hat{\Delta}_{\nu,n}$ (under Assumptions 1.2.1–1.2.3) for a fully-blocked factorial design.

Remark 1.3.3. Branson et al. (2016) and Li et al. (2020) propose re-randomization designs in the context of factorial experiments which are also shown to have favorable properties relative to complete and stratified factorial designs. In Section 1.4.1, we compare the mean-squared error of the fully-blocked design to a re-randomized design via Monte Carlo simulation. ■

Our next result considers settings where only a subset of the factors are of primary interest to the researcher. For instance, Besedeš et al. (2012) use a factorial design to study how the number of options in an agent's choice set affects their ability to make optimal decisions. Here the primary factor of interest is the number of options (four or thirteen), but the design also features other secondary factors. In such a case we might imagine that a matched pairs design which focuses on the factor of primary interest and assigns the other factors by i.i.d. coin flips may be more efficient for estimating the primary factorial effect than the fully-blocked design which treats all the factors symmetrically. In particular, we consider a setting where we are interested in the average main effect on the kth factor, $\Delta_{\nu_k}(Q)$, and compare the performance of the fully-blocked design to a design which performs matched pairs over the kth factor while assigning the other factors to individuals at random using i.i.d. Bernoulli(1/2) assignment. We call such a design the "factor k specific" matched pairs design. Formally, let

$$\zeta_j = \zeta_j(X^{(n)}) \subset \{1, \ldots, 2^K n\},\ 1 \leq j \leq 2^{K-1}n$$

denote a partition of the set of indices such that each ζ_j contains two units. The "factor k specific" matched pairs design satisfies the following assumption:

Assumption 1.3.3. Treatment status is assigned so that $\{Y^{(n)}(d) : d \in \mathcal{D}\} \perp\!\!\!\perp D^{(n)} | X^{(n)}$ and, conditional on $X^{(n)}$,

$$\{(\iota_k(D_i) : i \in \zeta_j) : 1 \leq j \leq 2^{K-1}n\}$$

are i.i.d. and each uniformly distributed over $\{(-1, +1), (+1, -1)\}$. Furthermore, independently of $X^{(n)}$ and independently across $1 \leq j \leq K, j \neq k$, $\iota_j(D_i)$ is i.i.d. across $1 \leq i \leq 2^K n$ and $P\{\iota_j(D_i) = -1\} = P\{\iota_j(D_i) = +1\} = \frac{1}{2}$.

Theorem 1.3.8 shows that the asymptotic variance of $\hat{\Delta}_{\nu_1,n}$ is weakly smaller under a fully-blocked design than that under the factor specific matched pairs design.

Theorem 1.3.8. *Suppose Assumptions 1.2.1–1.2.3 hold and the treatment assignment mechanism satisfies Assumption 1.3.3. Then,*

$$\sqrt{n}(\hat{\Delta}_{\nu_k,n} - \Delta_{\nu_k}(Q)) \xrightarrow{d} N(0, \mathbb{V}_{\nu_k} + \xi_1 + \xi_0) \ ,$$

where $\mathbb{V}_{\nu_k}$ *is defined in Theorem 1.3.1, and*

$$\xi_1 = \sum_{d \in \mathcal{D}: \iota_k(d) = +1} E\left[\left(\Gamma_d(X_i) - \frac{1}{2^{K-1}} \sum_{d' \in \mathcal{D}: \iota_k(d') = +1} \Gamma_{d'}(X_i)\right)^2\right]$$
$$\xi_0 = \sum_{d \in \mathcal{D}: \iota_k(d) = -1} E\left[\left(\Gamma_d(X_i) - \frac{1}{2^{K-1}} \sum_{d' \in \mathcal{D}: \iota_k(d') = -1} \Gamma_{d'}(X_i)\right)^2\right] .$$

Remark 1.3.4. In this section we have presented results for “full" factorial designs, which assign individuals to every possible combination of factors. This is in contrast to “fractional" factorial designs, which assign only a subset of the possible factor combinations (see for example Wu and Hamada, 2011; Pashley and Bind, 2019). We leave possible extensions of our procedure to the fractional case for future work. ■

1.4 Simulations

In this section we examine the finite sample performance of the estimator $\hat{\Delta}_{\nu,n}$ and the test $\phi_n^{\nu}(Z^{(n)})$ in the context of a 2^K factorial experiment, under various alternative experimental designs. In Sections 1.4.1 and 1.4.2 the data generating processes are as specified below (in Section 1.4.3 we study an alternative design with multiple covariates and factors). For $d = (d^{(1)}, d^{(2)}) \in \{-1, 1\}^2$ and $1 \leq i \leq 4n$, the potential outcomes are generated according to the equation:

$$Y_i(d) = \mu_d + \mu_d(X_i) + \sigma_d(X_i)\epsilon_i \ .$$

In each of the specifications, $((X_i, \epsilon_i) : 1 \leq i \leq 4n)$ are i.i.d; for $1 \leq i \leq 4n$, X_i and ϵ_i are independent.

Model 1: $\mu_{1,a}(X_i) = \mu_{-1,a}(X_i) = \gamma X_i$ for $a \in \{-1, 1\}$, where $\gamma = 1$. $\mu_{1,1} = 2\mu_{1,-1} = 4\mu_{-1,1} = 2\tau$ for a parameter $\tau \in \{0, 0.2\}$, $\mu_{-1,-1} = 0$, $\epsilon_i \sim N(0, 1)$ and $X_i \sim N(0, 1)$ for all $d \in \{-1, 1\}^2$ and $\sigma_d(X_i) = 1$.

Model 2: As in Model 1, but $\mu_d(X_i) = X_i + (X_i^2 - 1)/3$.

Model 3: As in Model 1, but $\mu_d(X_i) = \gamma_d X_i + (X_i^2 - 1)/3$. $\gamma_{1,1} = 2$, $\gamma_{-1,1} = 1$, $\gamma_{1,-1} = 1/2$ and $\gamma_{-1,-1} = -1$.

Model 4: As in Model 3, but $\mu_d(X_i) = \sin(\gamma_d X_i)$.

Model 5: As in Model 3, $\mu_d(X_i) = \sin(\gamma_d X_i) + \gamma_d X_i/10 + (X_i^2 - 1)/3$.

Model 6: As in Model 3, but $\sigma_d(X_i) = (1 + d^{(1)} + d^{(2)})X_i^2$.

We consider five parameters of interest as listed in Table 3.1. $\Delta_{\nu_1}(Q)$ and $\Delta_{\nu_2}(Q)$ correspond to the main factorial effects for the two factors. $\Delta_{\nu_{1,2}}(Q)$ corresponds to the interaction effect between the two factors, as discussed in Example 1.3.2. $\Delta_{\nu_1^1}(Q)$ and $\Delta_{\nu_{-1}^1}(Q)$ denote the average effect of one factor, keeping the value of the other factor fixed at 1 or -1. All simulations are performed with a sample of size $4n = 1000$.

Parameter of interest	Formula
$\frac{1}{2}\Delta_{\nu_1}(Q)$	$\frac{1}{2}E[Y_i(1,1) - Y_i(-1,1)] + \frac{1}{2}E[Y_i(1,-1) - Y_i(-1,-1)]$
$\frac{1}{2}\Delta_{\nu_2}(Q)$	$\frac{1}{2}E[Y_i(1,1) - Y_i(1,-1)] + \frac{1}{2}E[Y_i(-1,1) - Y_i(-1,-1)]$
$\frac{1}{2}\Delta_{\nu_{1,2}}(Q)$	$\frac{1}{2}E[Y_i(1,1) - Y_i(-1,1)] - \frac{1}{2}E[Y_i(1,-1) - Y_i(-1,-1)]$
$\Delta_{\nu_1^1}(Q)$	$E[Y_i(1,1) - Y_i(-1,1)]$
$\Delta_{\nu_{-1}^1}(Q)$	$E[Y_i(1,-1) - Y_i(-1,-1)]$

Table 1.2: Parameters of interest

1.4.1 MSE Properties of the Matched Tuples Design

In this section, we study the mean-squared-error performance of $\hat{\Delta}_{\nu,n}$ across several experimental designs. We analyze and compare the MSE for all five parameters of interest for the following seven experimental designs:

1. **(B-B)** $(D_i^{(1)}, D_i^{(2)})$ are i.i.d. across $1 \leq i \leq 4n$ and the two entries are independently distributed as $2A - 1$, where A follows Bernoulli(1/2).

2. **(C)** $(D_i^{(1)}, D_i^{(2)})$ are jointly drawn from a completely randomized design. We uniformly at random divide the experimental sample of size $4n$ into four groups of size n and assign a different $d \in \{-1, 1\}^2$ for each group.

3. **(MP-B)** A matched-pair design for $D^{(1)}$, where units are ordered and paired according to X_i. For each pair, uniformly at random assign $D_i^{(1)} = 1$ to one of the units. Independently, $(D_i^{(2)} : 1 \leq i \leq 4n)$ are i.i.d. with the distribution of $2A - 1$, where $A \sim$ Bernoulli(1/2).

4. **(MT)** Matched tuples design where units are ordered according to X_i.

5. **(Large-2)** A stratified design, where the experimental sample is divided into two strata using the median of X_i as the cutoff. In each stratum, treatment is assigned as in **C**.

6. **(Large-4)** As in **(Large-2)**, but with four strata.

7. **(RE)** A re-randomization design using a Mahalanobis balance function. As outlined in Branson et al. (2016), we select the main-effect threshold criterion to be the $100(0.01^{1/K})$ percentile of a χ_p^2 distribution with $p = \dim(X_i)$, and select the interaction-effect threshold criterion to be $100(0.01^{1/L})$, where L is the number of interaction effects.

Table 3.2 displays the ratio of the MSE of each design relative to the MSE of **MT**, computed across 4,000 Monte Carlo replications. In each of the designs, we set treatment effects to zero by setting $\tau = 0$. As expected from Theorems 1.3.7 and 1.3.8, **MT** outperforms **B-B**, **C**, **MP-B**, **Large-2**, and **Large-4** in every model specification. We also find that **MT** compares favorably to **RE**, with **RE** slightly outperforming **MT** in some cases, but with **MT** outperforming in general. Although we do not have formal results comparing the matched tuples design to re-randomization, we note that re-randomization redraws treatments until the distances between certain features of the covariate distribution across treatment statuses are below certain pre-specified thresholds. In contrast, the matched tuples design attempts to *minimize* these distances by blocking units finely based on the covariates. See also Remark 3 of Bai (2022a) for a related observation in the binary treatment setting.

1.4.2 Inference

In this section, we study the finite sample properties of several different tests of the null hypothesis $H_0 : \Delta_\nu = 0$ for various choices of ν, against the alternative hypotheses implied by setting $\tau = 0.2$. In this section we restrict our attention to five assignment mechanisms: **B-B**, **C**, **MT**, **Large-2** and **Large-4**. We exclude **MP-B** because it is a non-standard experimental design for which we have not developed an inference procedure. We also exclude the re-randomization design (**RE**) because, although it is a widely studied design, the inferential results in Li et al. (2020) are derived in a finite population framework which is distinct from our super-population framework, and their resulting limiting distribution is non-normal.

In each case we perform our hypothesis tests at a significance level of 0.05. For design **B-B**, tests are performed using a standard t-test. For designs **C**, **Large-2** and **Large-4** the tests are constructed using the asymptotic normality result from Theorem 1.3.7 combined with variance estimators constructed using the same plug-in method as in Bugni et al. (2018a) and Bugni et al. (2019a). For design **MT** the test is constructed as described in Theorem 1.3.2. Table 1.4 displays the rejection probabilities under the null and alternative

Model	Parameter	**B-B**	**C**	**MP-B**	**MT**	**Large-2**	**Large-4**	**RE**
1	Δ_{ν_1}	2.099	1.948	1.045	1.000	1.335	1.138	1.031
	Δ_{ν_2}	2.036	2.015	2.113	1.000	1.407	1.179	0.988
	$\Delta_{\nu_{1,2}}$	2.008	2.044	2.016	1.000	1.423	1.091	1.014
	$\Delta_{\nu^1_1}$	2.051	2.014	1.563	1.000	1.402	1.134	1.029
	$\Delta_{\nu^1_{-1}}$	2.057	1.978	1.498	1.000	1.357	1.095	1.017
2	Δ_{ν_1}	2.327	2.168	1.044	1.000	1.546	1.249	1.232
	Δ_{ν_2}	2.254	2.259	2.355	1.000	1.619	1.312	1.209
	$\Delta_{\nu_{1,2}}$	2.249	2.287	2.173	1.000	1.646	1.225	1.250
	$\Delta_{\nu^1_1}$	2.285	2.265	1.634	1.000	1.599	1.260	1.227
	$\Delta_{\nu^1_{-1}}$	2.291	2.190	1.585	1.000	1.593	1.215	1.255
3	Δ_{ν_1}	2.042	1.996	1.792	1.000	1.422	1.206	1.124
	Δ_{ν_2}	1.576	1.527	1.480	1.000	1.221	1.140	1.109
	$\Delta_{\nu_{1,2}}$	3.113	2.982	1.943	1.000	1.900	1.337	1.187
	$\Delta_{\nu^1_1}$	3.401	3.351	2.237	1.000	1.979	1.410	1.225
	$\Delta_{\nu^1_{-1}}$	1.899	1.802	1.619	1.000	1.388	1.166	1.103
4	Δ_{ν_1}	1.311	1.305	1.252	1.000	1.100	1.070	1.194
	Δ_{ν_2}	1.218	1.210	1.167	1.000	1.063	1.064	1.057
	$\Delta_{\nu_{1,2}}$	1.296	1.289	1.152	1.000	1.184	1.084	1.191
	$\Delta_{\nu^1_1}$	1.416	1.401	1.259	1.000	1.158	1.080	1.249
	$\Delta_{\nu^1_{-1}}$	1.201	1.202	1.150	1.000	1.128	1.075	1.140
5	Δ_{ν_1}	1.603	1.606	1.315	1.000	1.280	1.169	1.375
	Δ_{ν_2}	1.444	1.458	1.378	1.000	1.225	1.173	1.235
	$\Delta_{\nu_{1,2}}$	1.607	1.598	1.351	1.000	1.370	1.184	1.390
	$\Delta_{\nu^1_1}$	1.802	1.797	1.415	1.000	1.353	1.192	1.441
	$\Delta_{\nu^1_{-1}}$	1.434	1.434	1.262	1.000	1.301	1.164	1.332
6	Δ_{ν_1}	1.119	1.122	1.116	1.000	1.055	1.021	1.065
	Δ_{ν_2}	1.051	1.042	1.056	1.000	1.026	0.991	0.989
	$\Delta_{\nu_{1,2}}$	1.107	1.104	1.077	1.000	1.074	0.994	1.018
	$\Delta_{\nu^1_1}$	1.096	1.100	1.088	1.000	1.058	1.005	1.051
	$\Delta_{\nu^1_{-1}}$	1.197	1.177	1.137	1.000	1.092	1.017	0.996

Table 1.3: Ratio of MSEs relative to MT

hypotheses, computed from 2,000 Monte Carlo replications. The results show that the rejection probabilities are universally around 0.05 under the null hypothesis, which verifies the validity of our tests across all the designs. Under the alternative hypotheses implied by

$\tau = 0.2$, the rejection probabilities vary substantially across the different designs, outcome models and parameters. However, our matched tuples design displays the highest power for almost all parameters and model specifications.

		Under H_0					Under H_1				
Model	Parameter	**B-B**	**C**	**MT**	**Large-2**	**Large-4**	**B-B**	**C**	**MT**	**Large-2**	**Large-4**
	Δ_{ν_1}	0.057	0.049	0.051	0.050	0.046	0.790	0.803	0.977	0.915	0.963
	Δ_{ν_2}	0.052	0.059	0.046	0.060	0.058	0.371	0.403	0.675	0.534	0.593
1	$\Delta_{\nu_{1,2}}$	0.049	0.059	0.049	0.059	0.043	0.081	0.093	0.126	0.100	0.106
	$\Delta_{\nu_1^1}$	0.052	0.043	0.048	0.064	0.040	0.646	0.656	0.921	0.816	0.884
	$\Delta_{\nu_{-1}^1}$	0.056	0.051	0.044	0.057	0.048	0.361	0.333	0.594	0.499	0.545
	Δ_{ν_1}	0.053	0.043	0.049	0.048	0.045	0.738	0.737	0.976	0.875	0.951
	Δ_{ν_2}	0.056	0.061	0.046	0.059	0.056	0.341	0.377	0.670	0.483	0.551
2	$\Delta_{\nu_{1,2}}$	0.052	0.065	0.050	0.060	0.044	0.082	0.091	0.126	0.101	0.095
	$\Delta_{\nu_1^1}$	0.049	0.051	0.046	0.057	0.036	0.597	0.610	0.919	0.758	0.840
	$\Delta_{\nu_{-1}^1}$	0.056	0.051	0.046	0.054	0.048	0.340	0.310	0.598	0.436	0.500
	Δ_{ν_1}	0.054	0.056	0.050	0.053	0.052	0.571	0.570	0.837	0.705	0.787
	Δ_{ν_2}	0.056	0.057	0.056	0.057	0.059	0.235	0.259	0.361	0.286	0.323
3	$\Delta_{\nu_{1,2}}$	0.051	0.051	0.052	0.062	0.047	0.060	0.064	0.116	0.091	0.082
	$\Delta_{\nu_1^1}$	0.048	0.051	0.046	0.061	0.035	0.402	0.421	0.885	0.624	0.762
	$\Delta_{\nu_{-1}^1}$	0.061	0.047	0.060	0.056	0.057	0.255	0.234	0.374	0.310	0.340
	Δ_{ν_1}	0.049	0.051	0.045	0.045	0.050	0.908	0.905	0.968	0.956	0.957
	Δ_{ν_2}	0.051	0.052	0.051	0.051	0.058	0.488	0.520	0.604	0.569	0.559
4	$\Delta_{\nu_{1,2}}$	0.056	0.052	0.049	0.065	0.045	0.092	0.102	0.126	0.117	0.111
	$\Delta_{\nu_1^1}$	0.050	0.048	0.051	0.054	0.045	0.762	0.785	0.908	0.865	0.886
	$\Delta_{\nu_{-1}^1}$	0.044	0.055	0.048	0.052	0.046	0.498	0.472	0.544	0.528	0.523
	Δ_{ν_1}	0.054	0.054	0.045	0.045	0.043	0.844	0.847	0.964	0.912	0.937
	Δ_{ν_2}	0.053	0.056	0.051	0.048	0.053	0.416	0.445	0.589	0.491	0.505
5	$\Delta_{\nu_{1,2}}$	0.052	0.054	0.049	0.059	0.049	0.092	0.099	0.124	0.110	0.099
	$\Delta_{\nu_1^1}$	0.051	0.052	0.049	0.058	0.043	0.674	0.688	0.911	0.810	0.847
	$\Delta_{\nu_{-1}^1}$	0.050	0.062	0.049	0.056	0.049	0.416	0.403	0.523	0.461	0.474
	Δ_{ν_1}	0.050	0.050	0.043	0.058	0.043	0.129	0.128	0.122	0.115	0.130
	Δ_{ν_2}	0.053	0.059	0.057	0.057	0.051	0.074	0.086	0.088	0.079	0.080
6	$\Delta_{\nu_{1,2}}$	0.047	0.046	0.052	0.053	0.044	0.052	0.046	0.052	0.057	0.050
	$\Delta_{\nu_1^1}$	0.049	0.046	0.049	0.051	0.043	0.082	0.083	0.077	0.082	0.081
	$\Delta_{\nu_{-1}^1}$	0.059	0.056	0.058	0.059	0.056	0.140	0.113	0.125	0.131	0.135

Table 1.4: Rejection probabilities under the null and alternative hypothesis

1.4.3 Experiments with More Factors and Covariates

In this section we repeat the previous simulation exercises while varying the number of factors K and the number of observed covariates $\dim(X_i)$. The data generating process is constructed as follows:

$$Y_i(d) = \begin{cases} \tau d^{(1)} + \tilde{X}_i'\beta + \epsilon_i, & \text{if } K = 1 \\ \tau \cdot \left(d^{(1)} + \frac{\sum_{k=2}^{K} d^{(k)}}{K-1} \right) + \gamma_d \tilde{X}_i'\beta + \epsilon_i, & \text{if } K \geq 2 \end{cases}$$

where $\tau \in \{0, 0.1\}$, $d = (d^{(1)}, \ldots, d^{(K)})$ and $d^{(k)} \in \{-1, 1\}$ represents the treatment status of the k-th factor. We set $\gamma_d = 1$ if $d^{(2)} = 1$, $\gamma_d = -1$ otherwise, in order to ensure the conditional means are heterogeneous in the second factor. $\tilde{X}_i$ contains 9 covariates, out of which the first $\dim(X_i)$ covariates are observed and used for the experimental designs. The distributions of $\tilde{X}_i, \epsilon_i$ and the values of β are calibrated using data obtained from Branson et al. (2016), who study the covariate balancing properties of 2^K factorial re-randomization designs using data from the New York Department of Education (NYDE). Details on the empirical context and construction of the data generating process are provided in Appendix A.4.3.

To construct our matched tuples of size 2^K when $\dim(X_i) > 1$, we employ the recursive pairing algorithm described in Section 1.2 using the Mahalanobis distance. We emphasize, however, that this approach is not guaranteed to be optimal, and we leave the study of potentially more effective matching algorithms to future work.

In addition to the standard matched tuples design (**MT**), we also include a matched tuples design with a *replicate* for each treatment as described in Section 1.3.2, denoted by **MT2**. For example, in the **MT2** design with two factors, units are matched into groups of *eight*, and two units receive each factor combination. We also continue to consider the alternative designs (**C**, **Large-4**, **MP-B** and **RE**) from Section 1.4.1. When constructing the strata for **Large-4**, we stratify on one covariate drawn at random from the set of available covariates.

In Table 1.5 we report the ratio of the MSE of each design relative to the MSE of **MT** when $\dim(X_i) = 1$ and $K = 1$ (computed from 4,000 Monte Carlo replications). For all experiments in this section, the number of observations is fixed to be 1,280 so that we have 20 matched tuples of size 64 when $K = 6$. Our simulation results are consistent with those in Section 1.4.1: **MT** displays the lowest MSE across almost all model specifications. Although **MT2** generally produces larger MSEs than **MT**, it still performs favorably relative to the other designs. For methods that use an increasing number of covariates when $\dim(X_i)$ increases (**MT**, **MT2**, **MP-B** and **RE**), we observe that the MSE in fact *increases* with the number of available covariates. We expect this is because (as shown in Appendix A.4.3) the first covariate is a much stronger predictor of the control outcome than the other available covariates, which are relatively uninformative.

$\dim(X_i)$	Method	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	Method	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
1	**MT**	1.000	1.003	1.006	1.113	1.297	1.945	**C**	9.151	8.554	8.642	8.939	9.015	9.181
2		1.027	1.052	1.107	1.180	1.463	2.293		9.120	8.528	8.568	8.867	9.053	9.114
4		1.043	1.130	1.420	1.687	2.170	3.338		8.968	8.364	8.569	8.868	8.949	8.765
6		1.192	1.495	1.763	2.241	3.097	4.304		8.945	8.327	8.588	8.994	9.081	8.853
9		1.284	1.702	2.047	2.781	3.337	4.081		8.934	8.309	8.600	8.788	8.915	8.526
1	**MT2**	1.017	1.049	1.074	1.297	1.916	2.903	**Large-4**	4.393	4.605	4.674	4.634	4.393	4.381
2		1.044	1.086	1.212	1.547	2.200	3.585		6.523	6.926	6.745	6.704	6.521	6.367
4		1.224	1.332	1.620	2.231	3.379	4.799		7.321	8.100	7.407	7.559	7.542	7.399
6		1.451	1.901	2.339	3.061	4.020	5.721		8.143	8.137	7.644	7.801	8.288	7.906
9		1.609	2.140	2.693	3.231	4.387	6.903		8.093	8.075	8.170	7.799	8.129	8.402
1	**MP-B**	0.991	8.693	8.807	8.964	8.991	8.829	**RE**	1.073	1.091	1.296	2.032	3.040	3.640
2		0.978	8.854	8.897	8.863	8.811	9.072		1.090	1.069	1.955	3.284	4.282	5.094
4		0.967	8.970	8.711	9.020	8.855	8.749		1.320	1.410	3.278	4.640	5.504	6.270
6		1.175	9.148	8.753	8.941	8.774	8.596		1.961	1.886	3.976	5.648	6.223	6.759
9		1.227	8.793	8.989	9.444	9.227	8.273		2.515	2.566	4.957	6.265	6.676	7.455

Table 1.5: Ratio of MSEs relative to MT using a single factor and covariate

In Table 1.6, we compute the rejection probabilities when testing the null hypothesis $H_0 : \Delta_{\nu_1} = 0$ against the alternative implied by setting $\tau = 0.1$, for various choices of K and $\dim(X_i)$ (computed from 1,000 Monte Carlo replications). Under the null hypothesis, we observe that our tests under design **MT** become conservative as $\dim(X_i)$ and K increase. In particular, we notice a large difference between $K = 4$ and $K = 5$. However, despite being conservative, **MT** still displays favorable power properties relative to **C** and **Large-4** for all but the largest choices of K.

Our next observation is that our tests under design **MT2** remain exact even as $\dim(X_i)$ and K both increase. As we explain in Section 1.3.2, we suspect that our challenges for inference using **MT** come from poor estimation of the variance, which seems to be alleviated in **MT2**, where the number of observations receiving each treatment within a tuple are doubled. As a result of this exactness, **MT2** achieves higher power than **MT** when $\dim(X_i)$ and K are large. To further explore these power improvements, Figure 1.1 presents power plots for three specific choices of K and $\dim(X_i)$ with τ ranging from 0 to 0.1 (Figure A.1 in the appendix presents power plots for alternatives implied by larger values than $\tau = 0.1$). First, when $\dim(X_i)$ and K are small, for instance $\dim(X_i) = K = 1$, we observe no significant difference between the power plots generated by **MT** and **MT2**. However, when the dimension of the covariates and factors are both large, for instance $\dim(X_i) = 6, K = 4$, **MT2** dominates **MT** for all alternative hypotheses. Therefore, our recommendation to practitioners is to consider a matched tuples design when working with few treatments and covariates, but to consider the replicated design when dealing with a large number of treatments and/or covariates.

1.5 Empirical Application

In this section, we illustrate the inference procedures introduced in Section 1.3 using the data collected in Fafchamps et al. (2014)[3]. Fafchamps et al. (2014) conduct a randomized

[3] The original paper features six rounds of surveys which were pooled in the final analysis. We perform our analysis exclusively on the data obtained in the sixth round in order to avoid complications related to time-series dependence across rounds. For simplicity, we additionally drop quadruplets with missing values, and 4 "leftover" groups whose sizes range from 5 to 8 firms. This results in a final sample of 120 quadruplets, or $4n = 480$. Further results on the long-run effects (collected in a seventh survey wave) are contained in

		Under H_0						Under H_1					
Method	$\dim(X_i)$	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
MT	1	0.049	0.045	0.033	0.023	0.009	0.008	0.998	1.000	1.000	0.997	0.980	0.837
	2	0.047	0.043	0.041	0.018	0.008	0.002	0.999	0.998	0.997	0.997	0.935	0.732
	4	0.040	0.029	0.031	0.011	0.009	0.008	1.000	1.000	0.979	0.946	0.794	0.583
	6	0.037	0.018	0.010	0.022	0.010	0.007	0.999	0.989	0.936	0.870	0.668	0.479
	9	0.041	0.026	0.016	0.019	0.014	0.003	0.988	0.961	0.895	0.810	0.674	0.319
MT2	1	0.054	0.054	0.044	0.059	0.047	0.052	1.000	0.999	1.000	0.996	0.973	0.858
	2	0.048	0.053	0.041	0.058	0.039	0.055	1.000	0.999	1.000	0.985	0.943	0.784
	4	0.075	0.048	0.054	0.056	0.060	0.046	0.996	0.993	0.981	0.951	0.843	0.673
	6	0.053	0.067	0.046	0.054	0.045	0.046	0.988	0.967	0.926	0.857	0.744	0.579
	9	0.065	0.050	0.053	0.059	0.060	0.047	0.983	0.944	0.872	0.840	0.704	0.494
C	1	0.062	0.054	0.041	0.056	0.059	0.069	0.437	0.449	0.410	0.445	0.463	0.459
	2	0.063	0.049	0.038	0.051	0.065	0.068	0.434	0.450	0.410	0.442	0.459	0.459
	4	0.064	0.050	0.038	0.048	0.055	0.057	0.425	0.448	0.400	0.443	0.457	0.468
	6	0.066	0.052	0.045	0.048	0.054	0.055	0.430	0.437	0.409	0.436	0.437	0.463
	9	0.063	0.042	0.050	0.033	0.054	0.048	0.417	0.439	0.420	0.433	0.433	0.448
Large-4	1	0.050	0.044	0.059	0.061	0.053	0.057	0.685	0.699	0.701	0.683	0.730	0.770
	2	0.046	0.050	0.043	0.052	0.044	0.065	0.560	0.564	0.575	0.585	0.582	0.634
	4	0.053	0.064	0.039	0.059	0.056	0.062	0.497	0.490	0.486	0.527	0.521	0.577
	6	0.055	0.053	0.049	0.057	0.059	0.071	0.462	0.444	0.495	0.519	0.520	0.553
	9	0.044	0.041	0.056	0.051	0.049	0.076	0.457	0.451	0.493	0.490	0.511	0.571

Table 1.6: Rejection probabilities when testing $H_0 : \Delta_{\nu_1} = 0$ under the null and alternative hypothesis

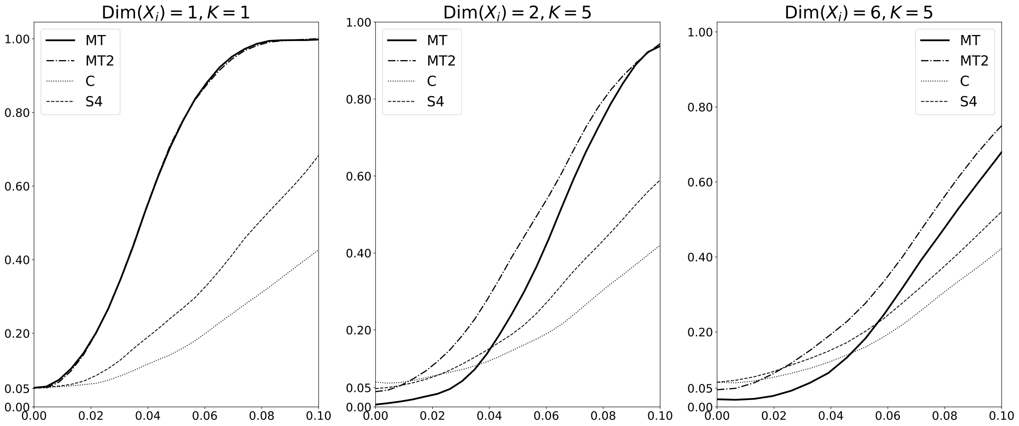

Figure 1.1: Rejection probability under various choices of τ

Table A.4 in Section A.4 of the appendix.

experiment in order to investigate the effects of several capital aid programs on the profits of small businesses in Ghana. In their experiment, there are three treatment arms, where (in our notation) $D_i = 1$ indicates that the ith firm is untreated, $D_i = 2$ indicates being offered cash, and $D_i = 3$ indicates being offered in-kind grants. The null hypotheses of interest are

$$H_0^d : E[Y_i(1)] = E[Y_i(d)] \text{ versus } H_1 : E[Y_i(1)] \neq E[Y_i(d)] \tag{1.12}$$

for $d \in \{2, 3\}$, as well as

$$H_0^{2,3} : E[Y_i(2)] = E[Y_i(3)] \text{ versus } H_1 : E[Y_i(2)] \neq E[Y_i(3)] \ . \tag{1.13}$$

In their experimental design, blocks are defined by quadruplets, where each quadruplet contains *two* untreated units with $D_i = 1$, one treated unit with $D_i = 2$, and one treated unit with $D_i = 3$. Despite the slight departure from the framework presented in Sections 1.2–1.3, in that there are two untreated units in each quadruplet, we show in Appendix A.1.1 that a slight modification of the variance estimator in Theorem 1.3.2 produces a valid test for (1.12)–(1.13). Specifically, we pretend that there are four treatment levels in each quadruplet, while the first two are in fact controls. Then, by setting generating vectors $\nu^2 = (-1/2, -1/2, 1, 0)$, $\nu^3 = (-1/2, -1/2, 0, 1)$, and $\nu^{2,3} = (0, 0, -1, 1)$ and proceeding with the testing procedure in Theorem 1.3.2, we obtain valid tests for H_0^d and $H_0^{2,3}$. For each of the hypotheses in (1.12)–(1.13), we implement the following tests:

— A t-test based on the OLS estimator in a linear regression of Y on 1, $I\{D_i = 2\}$, and $I\{D_i = 3\}$, together with the usual heteroskedasticity-robust variance estimator.

— The test introduced in Proposition A.1.1, which implements the test from Theorem 1.3.2 as described above to accommodate for the fact that there are two untreated units in each block.

We note that Fafchamps et al. (2014) test (1.12) and (1.13) using a t-test obtained from a linear regression of outcomes on treatment indicators and block fixed effects. However, as was shown in Theorem 1.3.3, such a procedure is not guaranteed to be valid. On the other hand, we expect that the t-test obtained from a linear regression without block fixed effects should be conservative for testing (1.12)–(1.13) in light of the observations made in Example 1.3.1 and the fact that this test coincides with a standard two-sample t-test.

Our results are presented in Table 1.7. The point estimates of the two methods are identical because the OLS estimator coincides with the difference-in-means estimator. However, the standard errors obtained from our variance estimator are always smaller than the heteroskedasticy-robust standard errors. For example, when testing (1.12) for $d = 3$ among the female subsample, the standard error produced from our variance estimator is 15.21 whereas the heteroskedasticy robust standard error is 18.13. We note that overall the improvements are modest; this suggests that the conditional expectation of the outcomes does not vary substantially with the observable characteristics in this survey wave. This is further corroborated by the calibrated simulations presented in Table A.2 in Appendix A.4.

1.6 Recommendations for Empirical Practice

We conclude with some recommendations for empirical practice based on our theoretical results as well as the simulation study above. For inference about the linear contrast of expected outcomes given by Δ_ν in a matched tuples design, we recommend the test ϕ_n^ν defined in Section 1.3.1: our simulations results show that this test does a good job of controlling size in large samples (approximately 80 blocks). We have shown that tests based on the heteroskedasticity-robust variance estimator from a linear regression of outcomes on treatment and block fixed effects may be *invalid*, in the sense of having rejection probability strictly greater than the nominal level under the null hypothesis. Tests based on the heteroskedasticity-robust variance or block-cluster variance estimators from a linear regres-

Table 1.7: Point estimates and standard errors for testing the treatment effects of cash and in-kind grants using different methods (wave 6)

					High Initial	Low Initial
		All Firms	Males	Females	Profit Women	Profit Women
		(1)	(2)	(3)	(4)	(5)
	Cash treatment	19.64	24.84	16.30	33.09	7.01
OLS		(15.42)	(27.29)	(18.13)	(42.56)	(11.58)
(standard t-test)	In-kind treatment	20.26	4.48	30.42	65.36	11.10
		(15.67)	(18.42)	(22.83)	(53.28)	(15.31)
	Cash=in-kind (p-val)	0.975	0.493	0.600	0.610	0.817
	Cash treatment	19.64	24.84	16.30	33.09	7.01
Difference-in-means		(14.24)	(26.05)	(15.21)	(39.27)	(11.15)
(adjusted t-test)	In-kind treatment	20.26	4.48	30.42	65.36	11.10
		(15.24)	(17.79)	(21.97)	(48.27)	(14.99)
	Cash=in-kind (p-val)	0.974	0.468	0.567	0.576	0.815

Note: The results in this table are based on the data from the sixth wave of data collection. For each treatment and each subsample, the number in the first row is the point estimate and that in the second row is the standard error. For testing the equality of the average potential outcomes under the two values of treatment, we report the p-values as in Fafchamps et al. (2014).

sion of outcomes on treatment are valid but potentially conservative, which would result in a loss of power relative to our proposed test.

We also find that matched tuples designs have favorable efficiency properties relative to other popular designs (with a specific illustration in the setting of 2^K factorial designs). However, this comes with the caveat that when dealing with a large number of treatments (in our simulations, this translated to having fewer than 80 blocks) and/or large number of covariates, practitioners may want to consider the replicated matched tuples design introduced in Section 1.3.2, as our simulations suggest that this design may have more robust size control, which translates to better power in such cases.

CHAPTER 2

INFERENCE IN CLUSTER RANDOMIZED TRIALS WITH MATCHED PAIRS

2.1 Introduction

This paper studies the problem of inference in cluster randomized experiments where treatment status is determined according to a “matched pairs” design. Here, by a cluster randomized experiment, we mean one in which treatment is assigned at the level of the cluster; by a “matched pairs” design we mean that the sample of clusters is paired according to baseline, cluster-level covariates and, within each pair, one cluster is selected at random for treatment. Cluster matched pair designs feature prominently in all parts of the sciences: examples in economics include Banerjee et al. (2015) and Crépon et al. (2015).

Following recent work in Bugni et al. (2022a), we develop our results in a sampling framework where clusters are realized as a random sample from a population of clusters. Importantly, in this framework cluster sizes are modeled as random and “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. The framework additionally allows for the possibility of two-stage sampling, in which a subset of units is sampled from the set of units within each sampled cluster.

We first study the large-sample behavior of a weighted difference-in-means estimator under two distinct sets of assumptions on the matching procedure. Specifically, we distinguish between settings where the matching procedure does or does not match on a function of cluster size. For both cases, we establish conditions under which our estimator is asymptotically normal and derive simple, closed-form expressions for the asymptotic variance. Using these results, we establish formally that employing cluster size as a matching variable in addition to baseline covariates delivers a weak (and often strict) improvement in asymptotic efficiency relative to matching on baseline covariates alone. We then propose a variance estimator which is consistent for either asymptotic variance depending on the nature of the matching procedure. Combining these results establishes the asymptotic exactness of tests based on our estimators.

We then consider the asymptotic properties of two commonly recommended inference procedures based on linear regressions of the individual-level outcomes on a constant and cluster-level treatment. The first inference procedure clusters at the level of treatment assignment. The second inference procedure clusters at the level of assignment pairs, as recently recommended in de Chaisemartin and Ramirez-Cuellar (2019). We establish that both procedures are generally conservative in our framework.

Next, we study the behavior of a randomization test which permutes the treatment status for clusters within pairs. We establish the finite-sample validity of such a test for testing a certain null hypothesis related to the equality of potential outcome distributions under treatment and control, and then establish asymptotic validity for testing null hypotheses about the size-weighted average treatment effect. We emphasize, however, that the latter result relies heavily on our choice of test statistic, which is studentized using our novel variance estimator. In simulations, we find that this randomization test controls size more reliably than any of the other inference procedures we consider in the paper, while delivering comparable power.

Finally, we derive large-sample results for a covariate-adjusted version of our estimator, which is designed to improve precision by exploiting additional baseline covariates which were not used for treatment assignment. As discussed in Bai et al. (2023a) and Cytrynbaum (2023a), standard covariate adjustments based on a regression using treatment-covariate interactions (see, for instance, Negi and Wooldridge, 2021, for a succinct treatment) are not guaranteed to improve efficiency when treatment assignment is not completely randomized. For this reason, we consider a modified version of the estimator developed in Bai et al. (2023a) for individual-level matched pair experiments. Our results show that our covariate-adjusted estimator is guaranteed to improve asymptotic efficiency relative to the unadjusted estimator, whenever the matching procedure matches on cluster size. Interestingly, we also find that this improvement in efficiency is *not* guaranteed when cluster size is excluded as a matching variable, and document in a simulation study that in fact such covariate adjustments may increase variance.

The analysis of data from cluster randomized experiments and data from experiments with matched pairs has received considerable attention (see Donner and Klar, 2000; Athey and Imbens, 2017a; Hayes and Moulton, 2017, for general overviews), but most recent work has focused on only one of these two features at a time. Recent work on the analysis of cluster randomized experiments includes Middleton and Aronow (2015), Su and Ding (2021), Schochet et al. (2021), and Wang et al. (2022) (see Bugni et al., 2022a, for a general discussion of this literature as well as further references). Recent work on the analysis of matched pairs experiments includes Jiang et al. (2020), Cytrynbaum (2021), Bai et al. (2023c), and Bai (2022a) (see Bai et al., 2022c, for a general discussion of this literature as well as further references). Two papers which focus specifically on the analysis of cluster randomized experiments with matched pairs are Imai et al. (2009) and de Chaisemartin and Ramirez-Cuellar (2019). Both papers maintain a finite-population perspective, where the primary source of uncertainty is "design-based," stemming from the randomness in treatment assignment. In such a framework, both papers study the finite and large-sample behavior of difference-in-means type estimators and propose corresponding variance estimators which are shown to be conservative. In contrast, our paper maintains a "super-population" sampling framework and proposes a novel variance estimator which is shown to be asymptotically exact in our setting.

The remainder of the paper is organized as follows. In Section 2.2 we describe our setup and notation. Section 2.3-2.6 present our main results. Section 2.7 studies the finite-sample behavior of our proposed tests via a simulation study. We conclude with recommendations for empirical practice in Section 2.8.

2.2 Setup and Notation

In this section we introduce the notation and assumptions which are common to both matching procedures considered in Section 2.3. We broadly follow the setup and notation developed in Bugni et al. (2022a). Let $Y_{i,g} \in \mathbf{R}$ denote the (observed) outcome of interest for the ith unit in the gth cluster, $D_g \in \{0,1\}$ denote the treatment received by the gth cluster, $X_g \in \mathbf{R}^k$ the observed, baseline covariates for the gth cluster, and $N_g \in \mathbf{Z}_+$ the size of the gth cluster. In what follows we sometimes refer to the vector (X_g, N_g) as W_g. Further denote by $Y_{i,g}(d)$ the potential outcome of the ith unit in cluster g, when all units in the gth cluster receive treatment $d \in \{0,1\}$. As usual, the observed outcome and potential outcomes are related to treatment assignment by the relationship

$$Y_{i,g} = Y_{i,g}(1)D_g + Y_{i,g}(0)(1 - D_g) \ . \tag{2.1}$$

In addition, define $\mathcal{M}_g$ to be the (possibly random) subset of $\{1, 2, \ldots, N_g\}$ corresponding to the observations within the gth cluster that are sampled by the researcher. We emphasize that a realization of $\mathcal{M}_g$ is a *set* whose cardinality we denote by $|\mathcal{M}_g|$, whereas a realization of N_g is a positive integer. For example, in the event that all observations in a cluster are sampled, $\mathcal{M}_g = \{1, \ldots, N_g\}$ and $|\mathcal{M}_g| = N_g$. We assume throughout that our sample consists of $2G$ clusters and denote by P_G the distribution of the observed data

$$Z^{(G)} := (((Y_{i,g} : i \in \mathcal{M}_g), D_g, X_g, N_g) : 1 \leq g \leq 2G) \ ,$$

and by Q_G the distribution of

$$(((Y_{i,g}(1), Y_{i,g}(0) : 1 \le i \le N_g), \mathcal{M}_g, X_g, N_g) : 1 \le g \le 2G) \ .$$

Note that P_G is determined jointly by (2.1) together with the distribution of $D^{(G)} := (D_g : 1 \le g \le 2G)$ and Q_G, so we will state our assumptions below in terms of these two quantities.

We now describe some preliminary assumptions on Q_G that we maintain throughout the paper. In order to do so, it is useful to introduce some further notation. To this end, for $d \in \{0, 1\}$, define

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

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

$$((Y_{i,g}(1), Y_{i,g}(0) : 1 \le i \le N_g) : 1 \le g \le 2G) \bigm| \mathcal{M}_g^{(G)}, X^{(G)}, N^{(G)} \ ,$$

where $\mathcal{M}_g^{(G)} := (\mathcal{M}_g : 1 \le g \le 2G)$, $X^{(G)} := (X_g : 1 \le g \le 2G)$ and $N^{(G)} := (N_g : 1 \le g \le 2G)$. Note that Q_G is completely determined by $R_G(\mathcal{M}_g^{(G)}, X^{(G)}, N^{(G)})$ and the distribution of $(\mathcal{M}_g^{(G)}, X^{(G)}, N^{(G)})$. The following assumption states our main requirements on Q_G using this notation.

Assumption 2.2.1. The distribution Q_G is such that

(a) $\{(\mathcal{M}_g, X_g, N_g), 1 \le g \le 2G\}$ is an i.i.d. sequence of random variables.

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

$$R_G(\mathcal{M}_g^{(G)}, X^{(G)}, N^{(G)}) = \prod_{1 \le g \le 2G} R(\mathcal{M}_g, X_g, N_g) \ .$$

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

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

(e) $\mathcal{M}_g \perp\!\!\!\perp (Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g) \bigm| X_g, N_g$ for all $1 \leq g \leq 2G$.

(f) For $d \in \{0, 1\}$ and $1 \leq g \leq 2G$,

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

For completeness, we reproduce some of the observations from Bugni et al. (2022a) regarding these assumptions. As shown in Bugni et al. (2022a), an important implication of Assumptions 2.2.1(a)–(b) for our purposes is that

$$\left\{\left(\bar{Y}_g(1), \bar{Y}_g(0), |\mathcal{M}_g|, X_g, N_g\right), 1 \leq g \leq 2G\right\} , \tag{2.2}$$

is an i.i.d. sequence of random variables. Assumptions 2.2.1.(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 2.2.1.(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 2.2.1.(a) and (c) are that

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

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 2.2.1(e)–(f) impose high-level restrictions on the two-stage sampling procedure. Assumption 2.2.1(e) allows the subset of observations sampled by the experimenter to depend on X_g and N_g, but rules out dependence on the potential outcomes within the cluster itself. Assumption 2.2.1(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. It can be shown that Assumptions 2.2.1(e)–(f) are satisfied if $\mathcal{M}_g$ is drawn as a random sample without replacement from $\{1, 2, \dots, N_g\}$ in an appropriate sense (see Lemma 2.1 in Bugni et al., 2022a).

Our object of interest is the size-weighted cluster-level average treatment effect, which may be expressed in our notation as

$$\Delta(Q_G) = E\left[\frac{N_g}{E[N_g]}\left(\frac{1}{N_g}\sum_{i=1}^{N_g}(Y_{i,g}(1) - Y_{i,g}(0))\right)\right] = E\left[\frac{1}{E[N_g]}\sum_{i=1}^{N_g}(Y_{i,g}(1) - Y_{i,g}(0))\right] .$$

This parameter, which weights the cluster-level average treatment effects proportional to cluster size, can be thought of as the average treatment effect where individuals are the unit of interest. Note that Assumptions 2.2.1(a)–(b) imply that we may express $\Delta(Q_G)$ as a function of R and the common distribution of $(\mathcal{M}_g, X_g, N_g)$. In particular, this implies that $\Delta(Q_G)$ does not depend on G. Accordingly, in what follows we simply denote $\Delta = \Delta(Q_G)$.

In Sections 2.3–2.5, we study the asymptotic behavior of the following size-weighted difference-in-means estimator:

$$\hat{\Delta}_G := \hat{\mu}_G(1) - \hat{\mu}_G(0) , \tag{2.3}$$

where

$$\hat{\mu}_G(d) := \frac{1}{N(d)}\sum_{g=1}^{2G} I\{D_g = d\}\frac{N_g}{|\mathcal{M}_g|}\sum_{i\in\mathcal{M}_g} Y_{i,g} ,$$

with

$$N(d) := \sum_{g=1}^{2G} N_g I\{D_g = d\} \ .$$

Note that this estimator may be obtained as the estimator of the coefficient of D_g in a weighted least squares regression of $Y_{i,g}$ on a constant and D_g with weights equal to $\sqrt{N_g/|\mathcal{M}_g|}$. In the special case that all observations in each cluster are sampled, so that $\mathcal{M}_g = \{1, 2, \ldots, N_g\}$ for all $1 \leq g \leq G$ with probability one, this estimator collapses to the standard difference-in-means estimator. In Section 2.6 we consider a covariate-adjusted modification of $\hat{\Delta}_G$ which is designed to incorporate additional baseline covariates which were not used for treatment assignment.

Remark 2.2.1. Following the recommendations in Bruhn and McKenzie (2009a) and Glennerster and Takavarasha (2013), it is common practice to conduct inference in matched pair experiments using the standard errors obtained from a regression of individual level outcomes on treatment and a collection of pair-level fixed effects. We do not analyze the asymptotic properties of such an approach for two reasons. First, in the context of individual-level randomized experiments, Bai et al. (2022c) and Bai et al. (2023c) argue that such a regression estimator is in fact numerically equivalent to the simple difference-in-means estimator, but that the resulting standard errors are generally conservative (and in some cases possibly invalid). This result generalizes immediately to the clustered setting in the special case where all clusters are the same size and $\mathcal{M}_g = \{1, 2, \ldots, N_g\}$. Second, when cluster sizes vary, this numerical equivalence no longer holds, and in such cases de Chaisemartin and Ramirez-Cuellar (2019) argue (in an alternative inferential framework) that the corresponding regression estimator may no longer be consistent for the average treatment effect of interest. ■

Remark 2.2.2. Bugni et al. (2022a) also define an alternative treatment effect parameter given by

$$\Delta^{\text{eq}}(Q_G) = E\left[\frac{1}{N_g}\sum_{i=1}^{N_g}(Y_{i,g}(1) - Y_{i,g}(0))\right] .$$

This parameter, which weights the cluster-level average treatment effects equally regardless of cluster size, can be thought of as the average treatment effect where the clusters themselves are the units of interest. For this parameter, the analysis of matched-pair designs for individual-level treatments developed in Bai et al. (2022c) applies directly to the data obtained from the cluster-level averages $\{(\bar{Y}_g, D_g, X_g, N_g) : 1 \leq g \leq 2G\}$, where $\bar{Y}_g = \frac{1}{|\mathcal{M}_g|}\sum_{i\in\mathcal{M}_g} Y_{i,g}$. As a result, we do not pursue a detailed description of inference for this parameter in the paper. ■

Remark 2.2.3. In Appendix B.3, we consider a generalization of our main results to settings with multiple treatments (i.e. “matched-tuples" designs) as considered in Bai et al. (2023c). ■

2.3 Asymptotic Behavior of $\hat{\Delta}_G$ for Cluster-Matched Pair Designs

In this section, we consider the asymptotic behavior of $\hat{\Delta}_G$ for two distinct types of cluster-matched pair designs. Section 2.3.1 studies a setting where cluster size is *not* used as a matching variable when forming pairs. Section 2.3.2 considers the setting where we do allow for pairs to be matched based on cluster size in an appropriate sense made formal below.

2.3.1 Not Matching on Cluster Size

In this section, we consider a setting where cluster size is not used as a matching variable. First, we describe our formal assumptions on the mechanism determining treatment assignment. The G pairs of clusters may be represented by the sets

$$\{\pi(2g-1), \pi(2g)\} \text{ for } g = 1, ..., G ,$$

where $\pi = \pi_G(X^{(G)})$ is a permutation of $2G$ elements. Given such a π, we assume that treatment status is assigned as follows:

Assumption 2.3.1. Treatment status is assigned so that

$$\left\{\left((Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g), N_g, \mathcal{M}_g\right)\right\}_{g=1}^{2G} \perp\!\!\!\perp D^{(G)}|X^{(G)} .$$

Conditional on $X^{(G)}$, $(D_{\pi(2g-1)}, D_{\pi(2g)})$, $g = 1, ..., G$ are i.i.d. and each uniformly distributed over $\{(0,1),(1,0)\}$.

We further require that the clusters in each pair be “close" in terms of their baseline covariates in the following sense:

Assumption 2.3.2. The pairs used in determining treatment assignment satisfy

$$\frac{1}{G}\sum_{g=1}^{G}\left|X_{\pi(2g)} - X_{\pi(2g-1)}\right|^r \xrightarrow{P} 0 ,$$

for $r \in \{1, 2\}$.

Bai et al. (2022c) provide results which facilitate the construction of pairs which satisfy Assumption 2.3.2. For instance, if $\dim(X_g) = 1$ and we order clusters from smallest to largest according to X_g and then pair adjacent units, it follows from Theorem 4.1 in Bai et al. (2022c) that Assumption 2.3.2 is satisfied if $E[X_g^2] < \infty$. Next, we state the additional assumptions on Q_G we require beyond those stated in Assumption 2.2.1:

Assumption 2.3.3. The distribution Q_G is such that

(a) $E[\bar{Y}_g^r(d)N_g^\ell|X_g = x]$, are Lipschitz for $d \in \{0,1\}$, $r, \ell \in \{0,1,2\}$,

(b) For some $C < \infty$, $P\{E[N_g|X_g] \leq C\} = 1$.

Assumption 2.3.3(a) is a smoothness requirement analogous to Assumption 2.1(c) in Bai et al. (2022c) that ensures that units within clusters which are “close" in terms of their baseline covariates are suitably comparable. Assumption 2.3.3(b) imposes an additional restriction on the distribution of cluster sizes beyond what is stated in Assumption 2.2.1(c). Under these assumptions, we obtain the following result:

Theorem 2.3.1. *Under Assumptions 2.2.1 and 2.3.1–2.3.3,*

$$\sqrt{G}(\hat{\Delta}_G - \Delta) \xrightarrow{d} N(0, \omega^2) ~,$$

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

$$\omega^2 = E[\tilde{Y}_g^2(1)] + E[\tilde{Y}_g^2(0)] - \frac{1}{2}E[(E[\tilde{Y}_g(1) + \tilde{Y}_g(0)|X_g])^2] ~,$$

with

$$\tilde{Y}_g(d) = \frac{N_g}{E[N_g]}\left(\bar{Y}_g(d) - \frac{E[\bar{Y}_g(d)N_g]}{E[N_g]}\right) ~.$$

Note that the asymptotic variance we obtain in Theorem 2.3.1 corresponds exactly to the asymptotic variance of the difference-in-means estimator for matched pairs designs with individual-level assignment (as derived in Bai et al., 2022c), but with transformed cluster-level potential outcomes given by $\tilde{Y}_g(d)$. Accordingly, our result collapses exactly to theirs when $P\{N_g = 1\} = 1$. Theorem 2.3.1 also quantifies the gain in precision obtained from using a matched pairs design versus complete randomization (i.e., assigning half of the clusters to treatment at random): it can be shown that the limiting distribution of $\hat{\Delta}_G$ under complete randomization is given by

$$\sqrt{G}(\hat{\Delta}_G - \Delta) \xrightarrow{d} N(0, \omega_0^2) ~,$$

where $\omega_0^2 = E[\tilde{Y}_g^2(1)] + E[\tilde{Y}_g^2(0)]$. We thus immediately obtain that $\omega^2 \leq \omega_0^2$. Moreover, this inequality is strict unless $E[\tilde{Y}_g(1) + \tilde{Y}_g(0)|X_g] = 0$.

2.3.2 Matching on Cluster Size

In this section, we repeat the exercise in Section 2.3.1 in a setting where the assignment mechanism matches on baseline characteristics *and* (some function of) cluster size in an appropriate sense to be made formal below. First, we describe how to modify our assumptions on the mechanism determining treatment assignment. The G pairs of clusters are still represented by the sets

$$\{\pi(2g-1), \pi(2g)\} \text{ for } g = 1, ..., G \ ,$$

however, now we allow the permutation $\pi = \pi_G(X^{(G)}, N^{(G)}) = \pi_G(W^{(G)})$ to be a function of cluster size. Given such a π, we assume that treatment status is assigned as follows:

Assumption 2.3.4. Treatment status is assigned so that

$$\{((Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g), \mathcal{M}_g)\}_{g=1}^{2G} \perp\!\!\!\perp D^{(G)} | W^{(G)} \ .$$

Conditional on $W^{(G)}$, $(D_{\pi(2g-1)}, D_{\pi(2g)})$, $g = 1, ..., G$ are i.i.d. and each uniformly distributed over $\{(0,1),(1,0)\}$.

We also modify the assumption on how pairs are formed:

Assumption 2.3.5. The pairs used in determining treatment assignment satisfy

$$\frac{1}{G}\sum_{g=1}^{G} N_{\pi(2g)}^{\ell} \left|W_{\pi(2g)} - W_{\pi(2g-1)}\right|^r \xrightarrow{P} 0 \ ,$$

for $\ell \in \{0,1,2\}$, $r \in \{1,2\}$.

Unlike for Assumption 2.3.2, the discussion in Bai et al. (2022c) does not provide conditions for matching algorithms which guarantee that Assumption 2.3.5 holds. Accordingly, in Proposition 2.3.1 we provide lower-level sufficient conditions for Assumption 2.3.5 which can be verified using the results in Bai et al. (2022c).

Proposition 2.3.1. *Suppose* $E[N_g^4] < \infty$ *and*

$$\frac{1}{G}\sum_{g=1}^{G}|W_{\pi(2g)} - W_{\pi(2g-1)}|^r \xrightarrow{P} 0 \ ,$$

for $r \in \{1,2,3,4\}$*, then Assumption 2.3.5 holds.*

We also modify the smoothness requirement as follows:

Assumption 2.3.6. The distribution Q_G is such that $E[\bar{Y}_g^r(d)|W_g = w]$ are Lipschitz for $d \in \{0,1\}$, $r \in \{1,2\}$.

We then obtain the following analogue to Theorem 2.3.1:

Theorem 2.3.2. *Under Assumptions 2.2.1 and 2.3.4–2.3.6,*

$$\sqrt{G}(\hat{\Delta}_G - \Delta) \xrightarrow{d} N(0,\nu^2) \ ,$$

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

$$\nu^2 = E[\tilde{Y}_g^2(1)] + E[\tilde{Y}_g^2(0)] - \frac{1}{2}E[(E[\tilde{Y}_g(1) + \tilde{Y}_g(0)|X_g, N_g])^2] \ , \tag{2.4}$$

with

$$\tilde{Y}_g(d) = \frac{N_g}{E[N_g]}\left(\bar{Y}_g(d) - \frac{E[\bar{Y}_g(d)N_g]}{E[N_g]}\right) \ .$$

Note that the asymptotic variance ν^2 has exactly the same form as ω^2 from Section 2.3.1, with the only difference being that the final term of the expression conditions on both cluster characteristics X_g and cluster size N_g. From this result it then follows that matching on cluster size in addition to cluster characteristics leads to a weakly lower asymptotic variance. To see this, note that by comparing ω^2 and ν^2 we obtain that

$$\omega^2 - \nu^2 = -\frac{1}{2}\left(E[E[\tilde{Y}_g(1) + \tilde{Y}_g(0)|X_g]^2] - E[E[\tilde{Y}_g(1) + \tilde{Y}_g(0)|X_g, N_g]^2]\right) \ .$$

It then follows by the law of iterated expectations and Jensen's inequality that $\omega^2 \geq \nu^2$.

2.4 Variance Estimation

In this section, we construct variance estimators for the asymptotic variances ω^2 and ν^2 obtained in Section 2.3. In fact, we propose a *single* variance estimator that is consistent for *both* ω^2 and ν^2 depending on the nature of the matching procedure. As noted in the discussion following Theorem 2.3.1, the expressions for ω^2 and ν^2 correspond exactly to the asymptotic variance obtained in Bai et al. (2022c) with the individual-level outcome replaced by a cluster-level transformed outcome. We thus follow the variance construction from Bai et al. (2022c), but replace the individual outcomes with feasible versions of these transformed outcomes. To that end, consider the observed adjusted outcome defined as:

$$\hat{Y}_g = \frac{N_g}{\frac{1}{2G}\sum_{1\leq j\leq 2G} N_j}\left(\bar{Y}_g - \frac{\frac{1}{G}\sum_{1\leq j\leq 2G}\bar{Y}_j I\{D_j = D_g\}N_j}{\frac{1}{G}\sum_{1\leq j\leq 2G} I\{D_j = D_g\}N_j}\right) ,$$

where

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

We then propose the following variance estimator:

$$\hat{v}^2_G = \hat{\tau}^2_G - \frac{1}{2}\hat{\lambda}^2_G , \tag{2.5}$$

where

$$\begin{aligned}\hat{\tau}_G^2 =& \frac{1}{G} \sum_{1 \leq j \leq G} \left(\hat{Y}_{\pi(2j)} - \hat{Y}_{\pi(2j-1)}\right)^2 \\ \hat{\lambda}_G^2 =& \frac{2}{G} \sum_{1 \leq j \leq \lfloor G/2 \rfloor} \left(\hat{Y}_{\pi(4j-3)} - \hat{Y}_{\pi(4j-2)}\right) \left(\hat{Y}_{\pi(4j-1)} - \hat{Y}_{\pi(4j)}\right) \\ & \qquad \times (D_{\pi(4j-3)} - D_{\pi(4j-2)})(D_{\pi(4j-1)} - D_{\pi(4j)}) ~.\end{aligned}$$

Note that the construction of $\hat{v}_G^2$ can be motivated using the same intuition as the variance estimators studied in Bai et al. (2022c) and Bai et al. (2023c): to consistently estimate quantities like (for instance) $E[E[\tilde{Y}_g(1)|X_g]E[\tilde{Y}_g(0)|X_g]]$ which appear in ω^2, we average across "pairs of pairs" of clusters. As a consequence, we will additionally require that the matching algorithm satisfy the condition that "pairs of pairs" of clusters are sufficiently close in terms of their baseline covariates/cluster size, as formalized in the following two assumptions:

Assumption 2.4.1. The pairs used in determining treatment status satisfy

$$\frac{1}{G} \sum_{1 \leq j \leq \left\lfloor \frac{G}{2} \right\rfloor} \left|X_{\pi(4j-k)} - X_{\pi(4j-\ell)}\right|^2 \xrightarrow{P} 0$$

for any $k \in \{2,3\}$ and $\ell \in \{0,1\}$.

Assumption 2.4.2. The pairs used in determining treatment status satisfy

$$\frac{1}{G} \sum_{1 \leq j \leq \left\lfloor \frac{G}{2} \right\rfloor} N_{\pi(4j-k)}^2 \left|W_{\pi(4j-k)} - W_{\pi(4j-\ell)}\right|^2 \xrightarrow{P} 0$$

for any $k \in \{2,3\}$ and $\ell \in \{0,1\}$.

As noted in Bai et al. (2022c), given pairs which satisfy Assumptions 2.3.2 or 2.3.5, it is frequently possible to reorder the pairs so that Assumptions 2.4.1 or 2.4.2 are satisfied. We then obtain the following two consistency results for the estimator $\hat{v}_G^2$:

Theorem 2.4.1. *Suppose Assumption 2.2.1 holds. If additionally Assumptions 2.3.1–2.3.3 and 2.4.1 hold, then*

$$\hat{v}_G^2 \xrightarrow{P} \omega^2 \ .$$

Alternatively, if Assumptions 2.3.4–2.3.6 and 2.4.2 hold, then

$$\hat{v}_G^2 \xrightarrow{P} \nu^2 \ .$$

Next, we derive the limits in probability of two commonly recommended variance estimators obtained from a (weighted) linear regression of the individual-level outcomes $Y_{i,g}$ on a constant and cluster-level treatment D_g. The first variance estimator we consider, which we denote by $\hat{\omega}^2_{\text{CR,G}}$, is simply the cluster-robust variance estimator of the coefficient of D_g as defined in equation (B.8) in the appendix. Theorem 2.4.2 derives the limit in probability of $\hat{\omega}^2_{\text{CR,G}}$ under a matched pair design which matches on baseline covariates as defined in Section 2.3.1, and shows that it is generally too large relative to ω^2.

Theorem 2.4.2. *Under Assumptions 2.2.1 and 2.3.1–2.3.3,*

$$\hat{\omega}^2_{\text{CR,G}} \xrightarrow{P} E[\tilde{Y}_g(1)^2] + E[\tilde{Y}_g(0)^2] \geq \omega^2 \ ,$$

with equality if and only if

$$E[\tilde{Y}_g(1) + \tilde{Y}_g(0)|X_g] = 0 \ . \tag{2.6}$$

The next variance estimator we consider, which we denote by $\hat{\omega}^2_{\text{PCVE,G}}$, is the variance estimator of the coefficient of D_g obtained from clustering on the assignment *pairs* of clusters as defined in equation (B.9) in the appendix. de Chaisemartin and Ramirez-Cuellar (2019) call this the pair-cluster variance estimator (PCVE)[1]. Theorem 2.4.3 derives the limit in probability of $\hat{\omega}^2_{\text{PCVE,G}}$ in the special case where $N_g = k$ for $g = 1, \ldots, 2G$ for some fixed k and $|\mathcal{M}_g| = N_g$, and shows that it is generally too large relative to ω^2.

Theorem 2.4.3. *Suppose Assumptions 2.2.1 and 2.3.1–2.3.3 hold. If in addition we impose that $N_g = k$ for $g = 1, \ldots, 2G$ for some fixed positive integer k and that $|\mathcal{M}_g| = N_g$, then*

$$\hat{\omega}^2_{\text{PCVE,G}} \xrightarrow{P} \omega^2 + \frac{1}{2}E\left[(E[\tilde{Y}_g(1) - \tilde{Y}_g(0)|X_g])^2\right] \geq \omega^2 \ ,$$

with equality if and only if

$$E[\tilde{Y}_g(1) - \tilde{Y}_g(0)|X_g] = 0 \ . \tag{2.7}$$

Although we do not derive the limit in probability of $\hat{\omega}^2_{\text{PCVE,G}}$ in the general case, our simulation evidence in Section 2.7 suggests that the limit of $\hat{\omega}^2_{\text{PCVE,G}}$ remains conservative, and that the conditions under which it is consistent for ω^2 are the same as those in equation (2.7). From Theorems 2.4.2 and 2.4.3 we obtain that neither cluster-robust standard error is consistent for ω^2 unless the baseline covariates are irrelevant for the potential outcomes in an appropriate sense. In particular, equation (2.7) holds when the average treatment difference for the sampled units in a cluster are homogeneous, in the sense that $\bar{Y}_g(1) - \bar{Y}_g(0)$ is constant. We further note that the conditions under which $\hat{\omega}^2_{\text{CR,G}}$ and $\hat{\omega}^2_{\text{PCVE,G}}$ are consistent for ω^2 are exactly analogous to the conditions under which Bai et al. (2022c) derive (in the setting of an individual-level matched pairs experiment) that the two-sample t-test and matched pairs t-test are asymptotically exact, respectively.

[1] We emphasize, however, that de Chaisemartin and Ramirez-Cuellar (2019) propose their variance estimator in a finite population “design-based" inferential framework, which is distinct from the superopopulation framework we consider here.

2.5 Randomization Tests

In this section, we study the properties of a randomization test based on the idea of permuting the treatment assignments for clusters within pairs. In Section 2.5.1 we present some finite-samples properties of our proposed test, and in Section 2.5.2 we establish its large sample validity for testing the null hypothesis $H_0 : \Delta(Q_G) = 0$.

First, we construct the test. Denote by $\mathbf{H}_G$ the group of all permutations on $2G$ elements and by $\mathbf{H}_G(\pi)$ the subgroup that only permutes elements within pairs defined by π:

$$\mathbf{H}_G(\pi) = \{h \in \mathbf{H}_G : \{\pi(2g-1), \pi(2g)\} = \{h(\pi(2j-1)), h(\pi(2j))\} \text{ for } 1 \leq g \leq G\} \ .$$

Define the action of $h \in \mathbf{H}_G(\pi)$ on $Z^{(G)}$ as follows:

$$hZ^{(G)} = \{((Y_{i,g} : i \in \mathcal{M}_g), D_{h(g)}, X_g, N_g) : 1 \leq g \leq 2G\} \ .$$

The randomization test we consider is then given by

$$\phi_G^{\text{rand}}(Z^{(G)}) = I\{T_G(Z^{(G)}) > \hat{R}_G^{-1}(1-\alpha)\} \ ,$$

where

$$\hat{R}_G(t) = \frac{1}{|\mathbf{H}_G(\pi)|} \sum_{h \in \mathbf{H}_G(\pi)} I\{T_G(hZ^{(G)}) \leq t\} \ ,$$

with

$$T_G(Z^{(G)}) = \left| \frac{\sqrt{G}\hat{\Delta}_G}{\hat{v}_G} \right| \ .$$

Remark 2.5.1. As is often the case for randomization tests, $\hat{R}_G(t)$ may be difficult to compute in situations where $|\mathbf{H}_G(\pi)| = 2^G$ is large. In such cases, we may replace $\mathbf{H}_G(\pi)$ with a stochastic approximation $\hat{\mathbf{H}}_G = \{h_1, h_2, \ldots, h_B\}$, where h_1 is the identity transformation and $h_2, \ldots, h_B$ are i.i.d. uniform draws from $\mathbf{H}_G(\pi)$. The results in Section 2.5.1 continue to hold with such an approximation; the results in Section 2.5.2 continue to hold provided $B \to \infty$ as $G \to \infty$. ■

2.5.1 Finite-Sample Results

In this section we present some finite-sample properties of the proposed test. Consider testing the null hypothesis that the distribution of potential outcomes within a cluster are equal across treatment and control conditional on observable characteristics and cluster size:

$$H_0^{X,N} : (Y_{i,g}(1) : 1 \leq i \leq N_g)|(X_g, N_g) \overset{d}{=} (Y_{i,g}(0) : 1 \leq i \leq N_g)|(X_g, N_g) \ . \tag{2.8}$$

We then establish the following result on the finite sample validity of our randomization test for testing (2.8):

Theorem 2.5.1. *Suppose Assumption 2.2.1 holds and that the treatment assignment mechanism satisfies Assumption 2.3.1 or 2.3.4. Then, for the problem of testing* (2.8) *at level* $\alpha \in (0, 1)$, $\phi_G^{\text{rand}}(Z^{(G)})$ *satisfies*

$$E[\phi_G^{\text{rand}}(Z^{(G)})] \leq \alpha \ ,$$

under the null hypothesis.

Remark 2.5.2. The proof of Theorem 2.5.1 follows classical arguments that underlie the finite sample validity of randomization tests more generally. Accordingly, as in those arguments, the result continues to hold if the test statistic T_G is replaced by any other test statistic which is a function of $Z^{(G)}$. ■

2.5.2 Large-Sample Results

In this section, we establish the large-sample validity of the randomization test ϕ_G^{rand} for testing the null hypothesis

$$H_0 : \Delta(Q_G) = 0 \ . \tag{2.9}$$

In Remark 2.5.3 we describe how to modify the test for testing non-zero null hypotheses.

Theorem 2.5.2. *Suppose Q_G satisfies Assumption 2.2.1, and either*

- *Assumption 2.3.3 with treatment assignment mechanism satisfying Assumption 2.3.1 and 2.4.1 ,*
- *Assumption 2.3.6 with treatment assignment mechanism satisfying Assumptions 2.3.4 and 2.4.2 .*

Further, suppose that the probability limit of $\hat{v}_G^2$ is positive, then

$$\sup_{t \in \mathbf{R}} |\hat{R}_G(t) - (\Phi(t) - \Phi(-t))| \xrightarrow{P} 0 \ ,$$

where $\Phi(\cdot)$ *is the standard normal CDF. Thus, for the problem of testing* (2.9) *at level* $\alpha \in (0,1)$, $\phi_G^{\text{rand}}(Z^{(G)})$ *satisfies*

$$\lim_{G \to \infty} E[\phi_G^{\text{rand}}(Z^{(G)})] = \alpha \ ,$$

under the null hypothesis.

Theorems 2.5.1 and 2.5.2 highlight that the randomization test $\phi_G^{\text{rand}}(Z^{(G)})$ is asymptotically valid for testing (2.9) while additionally retaining the finite-sample validity described in Section 2.5.1 under the null hypothesis (2.8). In Section 2.7.1 we illustrate the benefit of this additional robustness on the small-sample behavior of $\phi_G^{\text{rand}}(Z^{(G)})$ relative to tests constructed using Gaussian critical values. We note that, unlike for the null hypothesis considered in Section 2.5.1, the choice of test statistic T_G is crucial for establishing Theorem 2.5.2. Similar observations have been made in related contexts in Janssen (1997), Chung and Romano (2013), Bugni et al. (2018b) and Bai et al. (2022c).

Remark 2.5.3. We briefly describe how to modify the test ϕ_G^{rand} for testing general null hypotheses of the form

$$H_0 : \Delta(Q_G) = \Delta_0 \ .$$

To this end, let

$$\tilde{Z}^{(G)} := (((Y_{i,g} - D_g \Delta_0 : i \in \mathcal{M}_g), D_g, X_g, N_g) : 1 \leq g \leq 2G) \ ,$$

then it can be shown that under the assumptions given in Theorem 2.5.2, the test $\phi_G^{\text{rand}}(\tilde{Z}^{(G)})$ obtained by replacing $Z^{(G)}$ with $\tilde{Z}^{(G)}$ satisfies

$$\lim_{G \to \infty} E[\phi_G^{\text{rand}}(\tilde{Z}^{(G)})] = \alpha \ ,$$

under the null hypothesis. ■

2.6 Covariate Adjustment

In this section, we consider a linearly covariate-adjusted modification of $\hat{\Delta}_G$ that is designed to improve precision by exploiting additional observed baseline covariates that were not used for treatment assignment. To that end, we consider a setting in which we observe two sets of baseline covariates, X_g and C_g, where $X_g \in \mathbf{R}^k$ denotes the original set of baseline covariates used for treatment assignment, and $C_g \in \mathbf{R}^\ell$ denotes the covariates in addition to X_g that were not used for treatment assignment. Note that C_g could also include cluster-level aggregates of individual-level outcomes, including intracluster means and quantiles.

Before proceeding, we note that for the remainder of Section 2.6, the assumptions in Section 2.2 should be modified such that X_g is replaced by (X_g, C_g) throughout. In particular, references to Assumption 2.2.1 below should be understood to hold with (X_g, C_g) in place of X_g.

Our primary focus will be on settings in which the cluster size N_g is used in determining the pairs. We comment on the case when N_g is not used in determining pairs in Remark 2.6.1, and, importantly, note that in such settings the adjustments we consider here are *not* guaranteed to improve precision). As in Section 2.3.2, let $\pi = \pi_G(X^{(G)}, N^{(G)})$ denote the permutation that determines the pairs. We then assume that treatment status is assigned as follows:

Assumption 2.6.1. Treatment status is assigned so that

$$\{((Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g), \mathcal{M}_g, C_g)\}_{g=1}^{2G} \perp\!\!\!\perp D^{(G)} | (X^{(G)}, N^{(G)}) \ .$$

Conditional on $(X^{(G)}, N^{(G)})$, $(D_{\pi(2g-1)}, D_{\pi(2g)})$, $g = 1, ..., G$ are i.i.d. and each uniformly distributed over $\{(0,1), (1,0)\}$.

We consider a linearly covariate-adjusted estimator of $\Delta(Q)$ based on a set of regressors generated by X_g, N_g, C_g. To this end, define $\psi_g = \psi(X_g, N_g, C_g)$, where $\psi : \mathbf{R}^k \times \mathbf{R} \times \mathbf{R}^\ell \to \mathbf{R}^p$. We impose the following assumptions on ψ:

Assumption 2.6.2. The function ψ is such that

(a) No component of ψ is a constant and $E[\text{Var}[\psi_g | X_g, N_g]]$ is nonsingular.

(b) $\text{Var}[\psi_g] < \infty$.

(c) $E[\psi_g | W_g = w]$, $E[\psi_g \psi_g' | W_g = w]$, and $E[\psi_g \bar{Y}_g^r(d) | W_g = w]$ for $d \in \{0, 1\}$ and $r \in \{1, 2\}$ are Lipschitz.

(d) For some $c < \infty$, $P\{E[\|\psi_g\|^2 \bar{Y}_g^2(d) | X_g, N_g] \leq c\} = 1$ for $d \in \{0, 1\}$.

As discussed in Bai et al. (2023a) and Cytrynbaum (2023a), standard covariate adjustments based on a regression using treatment-covariate interactions (see, for instance, Negi and Wooldridge, 2021, for a succinct treatment) are not guaranteed to improve efficiency when treatment assignment is not completely randomized. For this reason, we consider a modified version of the adjusted estimator developed in Bai et al. (2023a) for individual-level matched pair experiments. Let $\hat{\beta}_G$ denote the OLS estimator of the slope coefficient in the linear regression of of $(\bar{Y}_{\pi(2g-1)} N_{\pi(2g-1)} - \bar{Y}_{\pi(2g)} N_{\pi(2g)})(D_{\pi(2g-1)} - D_{\pi(2g)})$ on a constant and $(\psi_{\pi(2g-1)} - \psi_{\pi(2g)})(D_{\pi(2g-1)} - D_{\pi(2g)})$. We then define our covariate-adjusted estimator as

$$\hat{\Delta}_G^{\text{adj}} = \frac{\frac{1}{G} \sum_{1 \leq g \leq 2G} (\bar{Y}_g N_g - (\psi_g - \bar{\psi}_G)' \hat{\beta}_G) D_g}{\frac{1}{G} \sum_{1 \leq g \leq 2G} N_g D_g} - \frac{\frac{1}{G} \sum_{1 \leq g \leq 2G} (\bar{Y}_g N_g - (\psi_g - \bar{\psi}_G)' \hat{\beta}_G)(1 - D_g)}{\frac{1}{G} \sum_{1 \leq g \leq 2G} N_g (1 - D_g)} , \tag{2.10}$$

where

$$\bar{\psi}_G = \frac{1}{2G} \sum_{1 \leq g \leq 2G} \psi_g .$$

Theorem 2.6.1 derives the limiting distribution of $\hat{\Delta}_G^{\text{adj}}$, and, importantly, it shows that the limiting variance of $\hat{\Delta}_G^{\text{adj}}$ is no larger than that of $\hat{\Delta}_G$ in (2.3) and can be strictly smaller.

Theorem 2.6.1. *Under Assumptions 2.2.1, 2.3.5, 2.3.6, 2.6.1, and 2.6.2,*

$$\sqrt{G}(\hat{\Delta}_G^{\text{adj}} - \Delta) \xrightarrow{d} N(0, \varsigma^2)$$

as $G \to \infty$, where

$$\varsigma^2 = E[\text{Var}[Y_g^*(1) | X_g, N_g]] + E[\text{Var}[Y_g^*(0) | X_g, N_g]] + \frac{1}{2} E[(E[Y_g^*(1) - Y_g^*(0) | X_g, N_g] - \Delta)^2] ,$$

with

$$\begin{aligned} Y_g^*(d) &= \frac{\bar{Y}_g(d)N_g - (\psi_g - E[\psi_g])'\beta^*}{E[N_g]} - \frac{N_g}{E[N_g]}\frac{E[\bar{Y}_g(d)N_g - (\psi_g - E[\psi_g])'\beta^*]}{E[N_g]} \\ &= \tilde{Y}_g(d) - \frac{(\psi_g - E[\psi_g])'\beta^*}{E[N_g]} \ , \end{aligned}$$

and

$$\beta^* = (2E[\mathrm{Var}[\psi_g|X_g, N_g]])^{-1}E[\mathrm{Cov}[\psi_g, \bar{Y}_g(1)N_g + \bar{Y}_g(0)N_g|X_g, N_g]] \ . \tag{2.11}$$

Moreover,

$$\varsigma^2 = \nu^2 - \kappa^2 \ , \tag{2.12}$$

where ν^2 is as in (2.4) *and*

$$\kappa^2 = \frac{E[((\psi_g - E[\psi_g|X_g, N_g])'\beta^*)^2]}{E[N_g]^2} \ .$$

As a consequence, $\varsigma^2 \leq \nu^2$, with equality if and only if $\kappa^2 = 0$.

Note that the asymptotic variance ς^2 has the same form as the variance ν^2, but with new transformed outcomes $Y_g^*(d)$ which can be expressed as covariate-adjusted versions of the original transformed outcomes $\tilde{Y}_g(d)$. Exploiting this observation is what allows us to establish that $\varsigma^2 = \nu^2 - \kappa^2$. As a consequence, we find that the asymptotic variance of $\hat{\Delta}_G^{\mathrm{adj}}$ is lower than that of $\hat{\Delta}_G$ whenever the adjustment is appropriately “relevant," in the sense that $\kappa^2 \neq 0$.

Remark 2.6.1. In order to guarantee that $\varsigma^2 \leq \nu^2$ in Theorem 2.6.1, it was crucial to assume that N_g is contained in the set of matching variables. If instead clusters are only matched according to X_g as in Section 2.3.1, then under suitable modifications of Assumptions 2.6.1

and 2.6.2 it can be shown that the limiting variance of $\hat{\Delta}_G^{\mathrm{adj}}$ is given by

$$E[\mathrm{Var}[Y_g^*(1)|X_g]] + E[\mathrm{Var}[Y_g^*(0)|X_g]] + \frac{1}{2}E[(E[Y_g^*(1) - Y_g^*(0)|X_g] - \Delta)^2] \ ,$$

where in this case $Y_g^*(d) = \tilde{Y}_g(d) - \frac{(\psi_g - E[\psi_g])'\beta^*}{E[N_g]}$, with

$$\beta^* = (2E[\mathrm{Var}[\psi_g|X_g]])^{-1}E[\mathrm{Cov}[\psi_g, \bar{Y}_g(1)N_g + \bar{Y}_g(0)N_g|X_g]] \ .$$

Note that this expression mirrors the expression for ς^2 but removes the conditioning on N_g throughout. It can then be shown that the decomposition obtained in (2.12) no longer applies, and in general the covariate-adjusted estimator is no longer guaranteed to have a smaller limiting variance than the unadjusted estimator $\hat{\Delta}_G$. We illustrate this point via simulation in Section 2.7.2. ■

Remark 2.6.2. Although the estimator in (2.10) is closely related to the class of covariate-adjusted estimators in Bai et al. (2023a), a direct application of the results therein is prohibited because the two denominators in (2.10) are the average cluster sizes of treated and untreated clusters and are therefore random. As a result, unlike in Bai et al. (2023a), the demeaning of ψ in (2.10) is crucial for the results in Theorem 2.6.1 to hold. In particular, some remainder terms in the proof of Theorem 2.6.1 are no longer $o_P(1)$ without the demeaning. Moreover, unlike for individual-level experiments, $\hat{\Delta}_G^{\mathrm{adj}}$ cannot be interpreted as the intercept of a linear regression as in Bai et al. (2023a). ■

For variance estimation, define

$$\mathring{Y}_g = \frac{1}{\frac{1}{2G}\sum_{1 \le j \le 2G} N_j}\left(N_g\bar{Y}_g - N_g\frac{\frac{1}{G}\sum_{1 \le j \le 2G}\bar{Y}_j I\{D_j = D_g\}N_j}{\frac{1}{G}\sum_{1 \le j \le 2G} I\{D_j = D_g\}N_j} - \psi_g'\hat{\beta}_G\right) \ .$$

We then propose the following variance estimator:

$$\mathring{\varsigma}_G^2 = \mathring{\tau}_G^2 - \frac{1}{2}\mathring{\lambda}_G^2 \ , \tag{2.13}$$

where

$$\begin{aligned}
\mathring{\tau}_G^2 =& \frac{1}{G} \sum_{1 \leq j \leq G} \left(\mathring{Y}_{\pi(2j)} - \mathring{Y}_{\pi(2j-1)} \right)^2 \\
\mathring{\lambda}_G^2 =& \frac{2}{G} \sum_{1 \leq j \leq \lfloor G/2 \rfloor} \left(\mathring{Y}_{\pi(4j-3)} - \mathring{Y}_{\pi(4j-2)} \right) \left(\mathring{Y}_{\pi(4j-1)} - \mathring{Y}_{\pi(4j)} \right) \\
& \qquad \times (D_{\pi(4j-3)} - D_{\pi(4j-2)})(D_{\pi(4j-1)} - D_{\pi(4j)}) \ .
\end{aligned}$$

The following theorem establishes the consistency of the variance estimator:

Theorem 2.6.2. *Under Assumptions 2.3.5, 2.3.6, 2.4.2, 2.6.1, and 2.6.2,*

$$\mathring{\varsigma}_G^2 \xrightarrow{P} \varsigma^2 \ .$$

2.7 Simulations

2.7.1 Unadjusted Estimation

In this section, we examine the finite-sample behavior of the estimation and inference procedures considered in Sections 2.3-2.5. We further compare these procedures to tests and confidence intervals constructed using the standard cluster-robust variance estimator (CR) and the pair cluster variance estimator (PCVE) proposed in de Chaisemartin and Ramirez-Cuellar (2019). For $d \in \{0, 1\}$, $1 \leq g \leq 2G$, the potential outcomes are generated according to the equation

$$Y_{i,g}(d) = \mu_d(X_g, X_g^{(N)}) + 2\epsilon_{d,i,g} \ .$$

Where, in each specification, $(X_g, X_g^{(N)})$, $g = 1, \ldots, 2G$ are i.i.d. with $X_g, X_g^{(N)} \sim \ Beta(2,4)$, and $(\epsilon_{0,i,g}, \epsilon_{1,i,g})$, $g = 1, \ldots, 2G$, $i = 1, \ldots, N_g$ are i.i.d. with $\epsilon_{0,i,g}, \epsilon_{1,i,g} \sim N(0,1)$ independently. We consider the following two specifications for μ_d:

Model 1: $\mu_1(X_g, X_g^{(N)}) = \mu_0(X_g, X_g^{(N)}) = 10(X_g - 1/3) + 6(X_g^{(N)} - 1/3) + 2$.

Model 2: $\mu_1(X_g, X_g^{(N)}) = 10(X_g^2 - 1/7) + 6(X_g^{(N)} - 1/3) + 2$ and $\mu_0(X_g, X_g^{(N)}) = 0$.

Note that Model 1 satisfies the homogeneity condition in (2.7) whereas Model 2 does not. In both cases, N_g, $g = 1, \ldots, 2G$ are i.i.d. with $N_g \sim \ Binomial(R, X_g^{(N)}) + (500 - R)$, where R determines the difference in maximum and minimum cluster sizes. In particular R satisfies the property that $N_g \in [N_{min}, N_{max}]$ with $N_{max} - N_{min} = R$ and we consider $R \in \{49, 149, 249, 349, 449\}$ with $N_{max} = 500$ fixed. For each model and distribution of cluster sizes, we consider two alternative pair-matching procedures. First, we consider a design which matches clusters using X_g only. To construct these pairs, we sort the clusters according to X_g and pair adjacent clusters. Next, we consider a design which matches clusters using both X_g and N_g. To construct these pairs, we match the clusters according to their Mahalanobis distance using the non-bipartite matching algorithm from the R package `nbpMatching`.

Tables 2.1–2.4 report the coverage and average length of 95% confidence intervals constructed using our variance estimator as well as the CR and PCVE estimators. For Model 1 in Table 2.1, we find that, in accordance with Theorems 3.4.2–2.4.3, the CR variance estimator is extremely conservative, whereas our proposed variance estimator (denoted $\hat{v}_G^2$) and the PCVE variance estimator have exact coverage asymptotically. This feature translates to significantly smaller confidence intervals: on average the confidence intervals constructed using $\hat{v}_G^2$ or PCVE are almost half the length of those constructed using CR when $G \geq 50$. However, the confidence intervals constructed using $\hat{v}_G^2$ or PCVE undercover when $G < 50$. We find similar results when matching on both X_g and N_g in Table 2.2. Comparing across

Tables 2.1 and 2.2 we find that, in line with the discussions following Theorems 2.3.1 and 2.3.2, matching on N_g in addition to X_g results in a large reduction in the average length of confidence intervals constructed using $\hat{v}_G^2$ (or PCVE), but no change in the average length of confidence intervals constructed using CR.

Moving to Model 2 in Tables 2.3 and 2.4, here we find that confidence intervals constructed using CR continue to be conservative, but now the confidence intervals constructed using PCVE are *also* conservative, and numerically very similar to those constructed using CR. In contrast, the confidence intervals constructed using $\hat{v}_G^2$ remain exact asymptotically. Once again this translates to smaller confidence intervals for $\hat{v}_G^2$: on average the confidence intervals constructed using $\hat{v}_G^2$ are approximately 25% smaller than those constructed using CR or PCVE when $G \geq 50$. However, once again we find that the confidence intervals constructed using $\hat{v}_G^2$ can undercover when $G < 50$, with the size of the distortion growing as a function of the cluster size heterogeneity.

Next, to further address the small-sample coverage distortions observed in Tables 2.1-2.4, we study the size and power of 0.05-level hypothesis tests conducted using our proposed randomization test, as well as standard t-tests constructed using the CR and PCVE estimators, in Tables 2.5–2.6 below.[2] In Table 2.5 we find that tests based on the CR variance estimator are extremely conservative, and this translates to having essentially no power against our chosen alternative. Tests based on the PCVE estimator produce non-trivial power, but also size-distortions in small samples. In contrast, since Model 1 satisfies the null hypothesis considered in (2.8), our randomization test is valid in finite samples by construction, and displays comparable power to the PCVE-based test even when the latter does not control size. When moving to Model 2 in Table 2.6 we are only guaranteed that the randomization test is asymptotically valid, but we find that the test is still able to control size in small

[2] Here we move to studying the properties of hypothesis tests instead of confidence intervals to avoid having to perform test-inversion for our randomization test, but we expect that similar results would continue to hold for confidence intervals as well.

samples as long as cluster-size heterogeneity is not too large. Importantly, in such cases, both the CR and PCVE-based tests also fail to control size. Finally, the randomization test displays favorable power relative to both the CR and PCVE-based tests throughout Table 2.6 except for some cases when $G = 12$.

2.7.2 Covariate-Adjusted Estimation

In this section, we examine the finite-sample behavior of the covariate-adjusted estimator considered in Section 2.6. In particular, we contrast the efficiency properties of $\hat{\Delta}_G^{\text{adj}}$ when matching versus not matching on cluster size. We consider the following modification of Model 2:

Model Adj.: $\mu_1(X_g, X_g^{(N)}) = 10(X_g^2 - 1/7) + 6(X_g^{(N)} - 1/3) + 25$ and $\mu_0(X_g, X_g^{(N)}) = 0$.

In addition, we introduce a new matching variable H_g, $g = 1, \ldots, 2G$, i.i.d. with $H_g \sim U[0,1]$ generated independently of all other variables, and modify the distribution of N_g so that $N_g \sim Binomial(R, 1 - X_g^{(N)}) + (500 - R)$.

Table 2.7 reports the coverage and average length of 95% confidence intervals constructed using our variance estimators when matching using both H_g and N_g, for $\hat{\Delta}_G$ versus $\hat{\Delta}_G^{\text{adj}}$ with $\psi_g = (X_g, X_g^{(N)})$. In accordance with Theorem 2.6.1, we find that for moderate to large samples ($G \geq 50$), covariate adjustment leads to smaller average CI lengths even as we increase the amount of cluster size heterogeneity. In contrast, Table 2.8 reports the coverage and average lengths of 95% confidence intervals (CIs) constructed using our variance estimators when matching using *only* H_g, for $\hat{\Delta}_G$ versus $\hat{\Delta}_G^{\text{adj}}$ with $\psi_g = (X_g, X_g^{(N)})$. In general, we find that when cluster-size heterogeneity is low, covariate adjustment leads to smaller average CI lengths. However, as the amount of heterogeneity increases, the average CI length for the adjusted estimator rapidly overtakes the length for the unadjusted estimator. We emphasize that this does not seem to be a small-sample issue: even with

$G = 250$, the average CI length for the adjusted estimator is over two times larger than for the unadjusted estimator in the most extreme case.

2.8 Recommendations for Empirical Practice

Based on our theoretical results as well as the simulation study above, we conclude with some recommendations for practitioners when conducting inference about the size-weighted ATE in our super-population framework. Our recommendations below depend on whether the number of clusters is moderately large (e.g., at least 50 pairs) or small (e.g., less than 50 pairs).

If the number of clusters is moderately large, then our recommendation is that practitioners should employ either the covariate-adjusted tests based on the covariate-adjusted estimator $\hat{\Delta}_G^{\mathrm{adj}}$ defined in Section 2.6 paired with its corresponding variance estimator $\mathring{\varsigma}_G^2$ and a normal critical value or the unadjusted tests based on the unadjusted estimator $\hat{\Delta}_G$ introduced in Section 2.2 paired with its corresponding variance estimator $\hat{v}_G^2$ and a normal critical value. Whenever cluster size is used in determining the pairs, our results show that covariate-adjusted tests are more powerful in large samples than unadjusted tests; in practice, this feature continues to hold even when cluster size was not used in determining the pairs, provided that cluster-size heterogeneity is not too great (i.e., in our simulations, a ratio of largest to smallest cluster size of less than 2). Outside of these circumstances, we recommend that practitioners employ the unadjusted tests.

If, on the other hand, the number of clusters is small, then we recommend instead that practitioners use the randomization test based on the un-adjusted estimator $\hat{\Delta}_G$ paired with its corresponding variance estimator $\hat{v}_G^2$ outlined in Section 2.5. In our simulations, this test controlled size more reliably than any of the other inference procedures we considered in the paper, while delivering comparable power. Note that by modifying the test as in Remark 2.5.3, the test could also be inverted to construct confidence intervals if desired.

Table 2.1: Model 1 - Matching on X_g[*]

N_{max}/N_{min}		$G = 12$	$G = 26$	$G = 50$	$G = 100$	$G = 150$	$G = 200$	$G = 250$
Coverage								
1.11	$\hat{v}_G^2$	0.9185	0.9290	0.9420	0.9465	0.9375	0.9460	0.9515
	CR	0.9985	0.9990	0.9995	1	1	1	1
	PCVE	0.9230	0.9310	0.9385	0.9405	0.9395	0.9480	0.9520
1.42	$\hat{v}_G^2$	0.9005	0.9345	0.9345	0.9480	0.9490	0.9545	0.9615
	CR	0.9980	0.9995	0.9985	0.9995	0.9995	1	1
	PCVE	0.9035	0.9380	0.9375	0.9490	0.9495	0.9550	0.9595
1.99	$\hat{v}_G^2$	0.9130	0.9330	0.9380	0.9385	0.9490	0.9455	0.9365
	CR	0.9985	0.9985	1	1	1	1	0.9995
	PCVE	0.9095	0.9230	0.9420	0.9420	0.9495	0.9460	0.9350
3.31	$\hat{v}_G^2$	0.9065	0.9180	0.9340	0.9415	0.9470	0.9450	0.9520
	CR	0.9950	0.9980	0.9980	0.9985	1	0.9985	0.9995
	PCVE	0.8980	0.9155	0.9330	0.9380	0.9465	0.9470	0.9500
9.80	$\hat{v}_G^2$	0.9035	0.9230	0.9420	0.9340	0.9440	0.9415	0.9495
	CR	0.9925	0.9940	0.9970	0.9985	0.9975	0.9995	0.9990
	PCVE	0.8925	0.9100	0.9365	0.9330	0.9425	0.9385	0.9475
Average Length								
1.11	$\hat{v}_G^2$	1.72150	1.16078	0.84582	0.59830	0.48784	0.42466	0.37936
	CR	3.20593	2.21689	1.61886	1.15015	0.94053	0.81591	0.73010
	PCVE	1.69494	1.15171	0.84119	0.59746	0.48744	0.42415	0.37895
1.42	$\hat{v}_G^2$	1.75019	1.18859	0.86476	0.61378	0.50112	0.43567	0.38917
	CR	3.21821	2.22957	1.62982	1.15829	0.94732	0.82180	0.73543
	PCVE	1.72075	1.17840	0.86140	0.61286	0.50024	0.43527	0.38897
1.99	$\hat{v}_G^2$	1.80502	1.23175	0.89937	0.63958	0.52250	0.45322	0.40566
	CR	3.24165	2.25077	1.64811	1.17207	0.95862	0.83166	0.74408
	PCVE	1.77287	1.21936	0.89602	0.63843	0.52133	0.45352	0.40524
3.31	$\hat{v}_G^2$	1.90111	1.30589	0.96060	0.68446	0.55910	0.48664	0.43505
	CR	3.27892	2.28895	1.68064	1.19654	0.97928	0.84959	0.76030
	PCVE	1.85679	1.29128	0.95566	0.68299	0.55824	0.48568	0.43437
9.80	$\hat{v}_G^2$	2.09510	1.45719	1.08057	0.77340	0.63320	0.55071	0.49226
	CR	3.35580	2.36729	1.75068	1.24963	1.02275	0.88759	0.79443
	PCVE	2.03228	1.43576	1.07565	0.77259	0.63171	0.54976	0.49203

[*] Number of clusters= $2G$ with $G = 12, 26, 50, 100, 150, 200, 250$. Number of replications for each G is 2000. $N_{max} = 500$.

Table 2.2: Model 1 - Matching on X_g and N_g*

N_{max}/N_{min}		$G = 12$	$G = 26$	$G = 50$	$G = 100$	$G = 150$	$G = 200$	$G = 250$
				Coverage				
	$\hat{v}_G^2$	0.9105	0.9285	0.9345	0.9430	0.9470	0.9495	0.9565
1.11	CR	1	1	1	1	1	1	1
	PCVE	0.9100	0.9260	0.9360	0.9460	0.9460	0.9480	0.9555
	$\hat{v}_G^2$	0.9210	0.9410	0.9400	0.9510	0.9490	0.9300	0.9445
1.42	CR	1	1	1	1	1	1	1
	PCVE	0.9215	0.9405	0.9425	0.9555	0.9465	0.9325	0.9425
	$\hat{v}_G^2$	0.9170	0.9460	0.9420	0.9505	0.9485	0.9495	0.9570
1.99	CR	1	1	1	1	1	1	1
	PCVE	0.9110	0.9440	0.9395	0.9520	0.9490	0.9510	0.9555
	$\hat{v}_G^2$	0.9220	0.9280	0.9295	0.9430	0.9440	0.9480	0.9390
3.31	CR	1	1	1	1	1	1	1
	PCVE	0.9150	0.9290	0.9325	0.9470	0.9435	0.9510	0.9405
	$\hat{v}_G^2$	0.9015	0.9260	0.9320	0.9505	0.9485	0.9405	0.9435
9.80	CR	1	1	1	1	1	1	1
	PCVE	0.8860	0.9225	0.9380	0.9495	0.9485	0.9420	0.9475
				Average Length				
	$\hat{v}_G^2$	1.20496	0.64428	0.39514	0.24765	0.19157	0.16045	0.14069
1.11	CR	3.21594	2.22170	1.62079	1.15081	0.94092	0.81621	0.73031
	PCVE	1.18192	0.63873	0.39376	0.24689	0.19111	0.16028	0.14062
	$\hat{v}_G^2$	1.16805	0.58866	0.34117	0.19821	0.14670	0.12020	0.10335
1.42	CR	3.23229	2.23499	1.63182	1.15901	0.94776	0.82214	0.73561
	PCVE	1.14574	0.58388	0.34065	0.19783	0.14622	0.12000	0.10327
	$\hat{v}_G^2$	1.18988	0.60685	0.34699	0.19474	0.14244	0.11466	0.09729
1.99	CR	3.25786	2.25761	1.65083	1.17312	0.95917	0.83201	0.74440
	PCVE	1.16373	0.59889	0.34582	0.19426	0.14229	0.11456	0.09728
	$\hat{v}_G^2$	1.27089	0.64963	0.37337	0.20857	0.15167	0.12110	0.10157
3.31	CR	3.29929	2.29885	1.68464	1.19841	0.98016	0.85013	0.76067
	PCVE	1.23316	0.64188	0.37129	0.20767	0.15108	0.12084	0.10134
	$\hat{v}_G^2$	1.41981	0.75053	0.43329	0.24285	0.17464	0.13851	0.11558
9.80	CR	3.38816	2.38329	1.75642	1.25248	1.02442	0.88868	0.79508
	PCVE	1.36449	0.73612	0.42992	0.24197	0.17401	0.13826	0.11549

* Number of clusters= $2G$ with $G = 12, 26, 50, 100, 150, 200, 250$. Number of replications for each G is 2000. $N_{max} = 500$.

Table 2.3: Model 2 - Matching on ${X_g}^*$

N_{max}/N_{min}		$G = 12$	$G = 26$	$G = 50$	$G = 100$	$G = 150$	$G = 200$	$G = 250$
				Coverage				
1.11	$\hat{v}_G^2$	0.9260	0.9375	0.9420	0.9420	0.9460	0.9465	0.9510
	CR	0.9570	0.9635	0.9755	0.9790	0.9825	0.9835	0.9800
	PCVE	0.9560	0.9645	0.9750	0.9785	0.9825	0.9835	0.9805
1.42	$\hat{v}_G^2$	0.9280	0.9395	0.9455	0.9405	0.9490	0.9495	0.9490
	CR	0.9525	0.9705	0.9705	0.9715	0.9795	0.9860	0.9820
	PCVE	0.9535	0.9710	0.9705	0.9735	0.9795	0.9860	0.9820
1.99	$\hat{v}_G^2$	0.9180	0.9325	0.9385	0.9455	0.9480	0.9420	0.9465
	CR	0.9415	0.9595	0.9680	0.9765	0.9770	0.9805	0.9800
	PCVE	0.9415	0.9605	0.9675	0.9770	0.9780	0.9800	0.9805
3.31	$\hat{v}_G^2$	0.8965	0.9290	0.9390	0.9480	0.9440	0.9400	0.9495
	CR	0.9325	0.9615	0.9700	0.9750	0.9775	0.9750	0.9765
	PCVE	0.9315	0.9615	0.9685	0.9755	0.9780	0.9745	0.9770
9.80	$\hat{v}_G^2$	0.8850	0.9085	0.9295	0.9380	0.9360	0.9375	0.9445
	CR	0.9155	0.9460	0.9640	0.9660	0.9660	0.9685	0.9755
	PCVE	0.9175	0.9450	0.9635	0.9660	0.9665	0.9680	0.9755
				Average Length				
1.11	$\hat{v}_G^2$	1.64579	1.11414	0.80852	0.57317	0.46677	0.40525	0.36269
	CR	1.88285	1.31397	0.96438	0.68747	0.56044	0.48713	0.43634
	PCVE	1.88367	1.31373	0.96432	0.68752	0.56044	0.48718	0.43636
1.42	$\hat{v}_G^2$	1.67055	1.13171	0.81934	0.58015	0.47436	0.41154	0.36739
	CR	1.90602	1.32885	0.97303	0.69262	0.56755	0.49258	0.44032
	PCVE	1.90579	1.32897	0.97283	0.69257	0.56751	0.49262	0.44026
1.99	$\hat{v}_G^2$	1.67377	1.14094	0.83413	0.59068	0.48377	0.41909	0.37493
	CR	1.90337	1.33455	0.98635	0.70162	0.57506	0.49879	0.44584
	PCVE	1.90395	1.33471	0.98606	0.70146	0.57506	0.49874	0.44586
3.31	$\hat{v}_G^2$	1.69386	1.16940	0.85636	0.61062	0.49954	0.43424	0.38770
	CR	1.91395	1.35515	1.00133	0.71846	0.58755	0.51145	0.45702
	PCVE	1.91241	1.35461	1.00137	0.71861	0.58755	0.51149	0.45699
9.80	$\hat{v}_G^2$	1.74999	1.23124	0.90607	0.64424	0.52971	0.45990	0.41091
	CR	1.95803	1.40591	1.04446	0.74668	0.61421	0.53318	0.47665
	PCVE	1.95767	1.40633	1.04420	0.74671	0.61422	0.53315	0.47665

[*] Number of clusters= $2G$ with $G = 12, 26, 50, 100, 150, 200, 250$. Number of replications for each G is 2000. $N_{max} = 500$.

Table 2.4: Model 2 - Matching on X_g and N_g [*]

N_{max}/N_{min}		$G=12$	$G=26$	$G=50$	$G=100$	$G=150$	$G=200$	$G=250$
				Coverage				
1.11	$\hat{v}_G^2$	0.9420	0.9480	0.9545	0.9495	0.9455	0.9530	0.9530
	CR	0.9670	0.9845	0.9875	0.9900	0.9915	0.9950	0.9935
	PCVE	0.9680	0.9850	0.9865	0.9900	0.9910	0.9950	0.9935
1.42	$\hat{v}_G^2$	0.9315	0.9475	0.9515	0.9530	0.9515	0.9580	0.9510
	CR	0.9665	0.9850	0.9850	0.9895	0.9915	0.9955	0.9955
	PCVE	0.9660	0.9850	0.9845	0.9900	0.9915	0.9960	0.9955
1.99	$\hat{v}_G^2$	0.9270	0.9430	0.9510	0.9520	0.9480	0.9575	0.9520
	CR	0.9650	0.9825	0.9885	0.9905	0.9930	0.9970	0.9945
	PCVE	0.9670	0.9815	0.9880	0.9900	0.9930	0.9970	0.9945
3.31	$\hat{v}_G^2$	0.9160	0.9365	0.9525	0.9480	0.9510	0.9525	0.9485
	CR	0.9580	0.9795	0.9890	0.9885	0.9930	0.9955	0.9940
	PCVE	0.9580	0.9800	0.9890	0.9890	0.9930	0.9955	0.9940
9.80	$\hat{v}_G^2$	0.9065	0.9330	0.9430	0.9510	0.9515	0.9495	0.9510
	CR	0.9410	0.9765	0.9845	0.9890	0.9880	0.9955	0.9915
	PCVE	0.9430	0.9755	0.9830	0.9890	0.9875	0.9955	0.9915
				Average Length				
1.11	$\hat{v}_G^2$	1.57502	1.02869	0.73036	0.51031	0.41388	0.35765	0.31902
	CR	1.89796	1.31976	0.96665	0.68810	0.56233	0.48793	0.43636
	PCVE	1.89800	1.31982	0.96657	0.68813	0.56236	0.48790	0.43634
1.42	$\hat{v}_G^2$	1.58361	1.03237	0.73193	0.50975	0.41335	0.35758	0.31856
	CR	1.91602	1.33100	0.97594	0.69418	0.56753	0.49302	0.44052
	PCVE	1.91549	1.33128	0.97597	0.69423	0.56756	0.49301	0.44049
1.99	$\hat{v}_G^2$	1.61080	1.04567	0.74313	0.51722	0.41903	0.36217	0.32297
	CR	1.93406	1.34395	0.98875	0.70392	0.57534	0.49967	0.44684
	PCVE	1.93403	1.34409	0.98881	0.70388	0.57529	0.49964	0.44680
3.31	$\hat{v}_G^2$	1.63660	1.07550	0.76774	0.53170	0.43114	0.37227	0.33175
	CR	1.94629	1.37114	1.01341	0.72038	0.58976	0.51183	0.45771
	PCVE	1.94802	1.37098	1.01337	0.72047	0.58984	0.51198	0.45771
9.80	$\hat{v}_G^2$	1.70687	1.13039	0.80947	0.55966	0.45337	0.39151	0.34801
	CR	1.98400	1.41410	1.05392	0.75111	0.61528	0.53484	0.47768
	PCVE	1.98403	1.41488	1.05356	0.75103	0.61532	0.53482	0.47769

[*] Number of clusters= $2G$ with $G = 12, 26, 50, 100, 150, 200, 250$. Number of replications for each G is 2000. $N_{max} = 500$.

Table 2.5: Model 1 - Randomization Test (RT) vs. CR/PCVE [*]

		Size under H_0			Power under $H_1 : \Delta_0 + 1/4$		
N_{max}/N_{min}		$G = 12$	$G = 26$	$G = 50$	$G = 12$	$G = 26$	$G = 50$
			Matching on X_g				
	RT	0.0395	0.0560	0.0505	0.0755	0.1220	0.2030
1.11	CR	0.0015	0.0010	0.0005	0.0095	0.0105	0.0160
	PCVE	0.0770	0.0690	0.0615	0.1195	0.1410	0.1995
	RT	0.0610	0.0445	0.0540	0.0935	0.1055	0.1970
1.42	CR	0.0020	0.0005	0.0015	0.0105	0.0105	0.0210
	PCVE	0.0965	0.0620	0.0625	0.1365	0.1220	0.1955
	RT	0.0505	0.0505	0.0505	0.0770	0.1130	0.1820
1.99	CR	0.0015	0.0015	0	0.0130	0.0100	0.0195
	PCVE	0.0905	0.0770	0.0580	0.1195	0.1260	0.1825
	RT	0.0570	0.0595	0.0555	0.0745	0.1130	0.1670
3.31	CR	0.0050	0.0020	0.0020	0.0145	0.0190	0.0270
	PCVE	0.1020	0.0845	0.0670	0.1220	0.1340	0.1760
	RT	0.0455	0.0500	0.0475	0.0715	0.1105	0.1410
9.80	CR	0.0075	0.0060	0.0030	0.0280	0.0230	0.0305
	PCVE	0.1075	0.0900	0.0635	0.1335	0.1380	0.1605
			Matching on X_g and N_g				
	RT	0.0490	0.0535	0.0585	0.1165	0.3050	0.6760
1.11	CR	0	0	0	0	0	0
	PCVE	0.0900	0.0740	0.0640	0.1540	0.2395	0.5015
	RT	0.0440	0.0475	0.0480	0.1290	0.3595	0.7820
1.42	CR	0	0	0	0	0	0
	PCVE	0.0785	0.0595	0.0575	0.1635	0.2810	0.5705
	RT	0.0510	0.0400	0.0480	0.1255	0.3380	0.7795
1.99	CR	0	0	0	0	0	0
	PCVE	0.0890	0.0560	0.0605	0.1580	0.2630	0.5785
	RT	0.0440	0.0500	0.0555	0.1185	0.3370	0.7075
3.31	CR	0	0	0	0	0	0
	PCVE	0.0850	0.0710	0.0675	0.1590	0.2825	0.5220
	RT	0.0525	0.0550	0.0500	0.1180	0.2780	0.5965
9.80	CR	0	0	0	0.0005	0	0
	PCVE	0.1140	0.0775	0.0620	0.1750	0.2540	0.4625

[*] Number of clusters= $2G$ with $G = 12, 26, 50$. Number of replications for each G is 2000. $N_{max} = 500$.

Table 2.6: Model 2 - Randomization Test (RT) vs. CR/PCVE*

		Size under H_0			Power under $H_1 : \Delta_0 + 1/4$		
N_{max}/N_{min}		$G = 12$	$G = 26$	$G = 50$	$G = 12$	$G = 26$	$G = 50$
		Matching on X_g					
	RT	0.0345	0.0425	0.0480	0.0305	0.0790	0.1650
1.11	CR	0.0430	0.0365	0.0245	0.0540	0.0645	0.1120
	PCVE	0.0440	0.0355	0.0250	0.0550	0.0655	0.1115
	RT	0.0370	0.0365	0.0445	0.0370	0.0675	0.1685
1.42	CR	0.0475	0.0295	0.0295	0.0575	0.0560	0.1125
	PCVE	0.0465	0.0290	0.0295	0.0560	0.0540	0.1145
	RT	0.0465	0.0445	0.0490	0.0385	0.0785	0.1485
1.99	CR	0.0585	0.0405	0.0320	0.0620	0.0675	0.1005
	PCVE	0.0585	0.0395	0.0325	0.0615	0.0675	0.1005
	RT	0.0565	0.0495	0.0520	0.0390	0.0660	0.1360
3.31	CR	0.0675	0.0385	0.0300	0.0610	0.0620	0.1010
	PCVE	0.0685	0.0385	0.0315	0.0595	0.0625	0.1025
	RT	0.0700	0.0660	0.0600	0.0405	0.0550	0.1140
9.80	CR	0.0845	0.0540	0.0360	0.0585	0.0600	0.0895
	PCVE	0.0825	0.0550	0.0365	0.0595	0.0580	0.0895
		Matching on X_g and N_g					
	RT	0.0250	0.0310	0.0370	0.0195	0.0735	0.1800
1.11	CR	0.0330	0.0155	0.0125	0.0240	0.0365	0.0765
	PCVE	0.0320	0.0150	0.0135	0.0235	0.0360	0.0790
	RT	0.0295	0.0290	0.0345	0.0205	0.0730	0.1740
1.42	CR	0.0335	0.0150	0.0150	0.0245	0.0385	0.0640
	PCVE	0.0340	0.0150	0.0155	0.0250	0.0365	0.0675
	RT	0.0345	0.0325	0.0415	0.0200	0.0665	0.1655
1.99	CR	0.0350	0.0175	0.0115	0.0225	0.0310	0.0600
	PCVE	0.0330	0.0185	0.0120	0.0230	0.0320	0.0610
	RT	0.0390	0.0390	0.0340	0.0150	0.0590	0.1415
3.31	CR	0.0420	0.0205	0.0110	0.0220	0.0295	0.0610
	PCVE	0.0420	0.0200	0.0110	0.0210	0.0310	0.0595
	RT	0.0555	0.0445	0.0415	0.0260	0.0405	0.1180
9.80	CR	0.0590	0.0235	0.0155	0.0295	0.0270	0.0505
	PCVE	0.0570	0.0245	0.0170	0.0295	0.0265	0.0510

* Number of clusters= $2G$ with $G = 12, 26, 50$. Number of replications for each G is 2000. $N_{max} = 500$.

Table 2.7: Covariate Adjustment - Matching on H_g and N_g*

N_{max}/N_{min}	ψ_g	$G = 12$	$G = 26$	$G = 50$	$G = 100$	$G = 150$	$G = 200$	$G = 250$
				Coverage				
1.11	-	0.9120	0.9275	0.9475	0.9395	0.9425	0.9510	0.9425
	$(X_g, X_g^{(N)})$	0.8625	0.8970	0.9360	0.9405	0.9440	0.9495	0.9495
1.42	-	0.9135	0.9245	0.9415	0.9445	0.9495	0.9425	0.9425
	$(X_g, X_g^{(N)})$	0.8990	0.9195	0.9375	0.9515	0.9470	0.9515	0.9455
1.99	-	0.9085	0.9250	0.9420	0.9470	0.9455	0.9545	0.9520
	$(X_g, X_g^{(N)})$	0.9175	0.9355	0.9500	0.9520	0.9505	0.9505	0.9470
3.31	-	0.9090	0.9265	0.9340	0.9515	0.9465	0.9465	0.9535
	$(X_g, X_g^{(N)})$	0.9335	0.9365	0.9480	0.9515	0.9510	0.9525	0.9550
9.80	-	0.9070	0.9245	0.9330	0.9375	0.9510	0.9455	0.9440
	$(X_g, X_g^{(N)})$	0.9325	0.9340	0.9475	0.9470	0.9575	0.9500	0.9555
				Average Length				
1.11	-	1.77556	1.21499	0.88201	0.62584	0.51123	0.44346	0.39699
	$(X_g, X_g^{(N)})$	1.30671	0.93116	0.68816	0.49242	0.40372	0.35104	0.31400
1.42	-	1.74117	1.20501	0.87067	0.62002	0.50712	0.43888	0.39274
	$(X_g, X_g^{(N)})$	1.46021	0.96656	0.69879	0.49479	0.40412	0.35025	0.31292
1.99	-	1.72916	1.19588	0.86887	0.61669	0.50509	0.43677	0.39112
	$(X_g, X_g^{(N)})$	1.81983	1.09008	0.74580	0.50919	0.41110	0.35398	0.31603
3.31	-	1.71004	1.19463	0.86708	0.61577	0.50301	0.43573	0.39127
	$(X_g, X_g^{(N)})$	2.36813	1.30774	0.83203	0.54137	0.42815	0.36460	0.32354
9.80	-	1.72505	1.19952	0.86484	0.61768	0.50429	0.43672	0.39197
	$(X_g, X_g^{(N)})$	3.06889	1.60986	0.97620	0.59917	0.46025	0.38545	0.33953

* Number of clusters= $2G$ with $G = 12, 26, 50, 100, 150, 200, 250$. Number of replications for each G is 2000. $N_{max} = 500$.

Table 2.8: Covariate Adjustment - Matching on H_g[*]

N_{max}/N_{min}	ψ_g	$G = 12$	$G = 26$	$G = 50$	$G = 100$	$G = 150$	$G = 200$	$G = 250$
				Coverage				
1.11	-	0.9015	0.9235	0.9435	0.9395	0.9365	0.9445	0.9485
	$(X_g, X_g^{(N)})$	0.8485	0.9060	0.9275	0.9425	0.9420	0.9510	0.9430
1.42	-	0.9070	0.9315	0.9365	0.9405	0.9455	0.9490	0.9525
	$(X_g, X_g^{(N)})$	0.9005	0.9230	0.9465	0.9510	0.9430	0.9475	0.9520
1.99	-	0.9050	0.9310	0.9450	0.9450	0.9480	0.9530	0.9465
	$(X_g, X_g^{(N)})$	0.9190	0.9395	0.9485	0.9470	0.9520	0.9495	0.9515
3.31	-	0.9100	0.9340	0.9410	0.9535	0.9520	0.9490	0.9485
	$(X_g, X_g^{(N)})$	0.9155	0.9325	0.9475	0.9485	0.9435	0.9535	0.9510
9.80	-	0.8975	0.9305	0.9410	0.9435	0.9420	0.9430	0.9545
	$(X_g, X_g^{(N)})$	0.9190	0.9440	0.9345	0.9455	0.9405	0.9490	0.9410
				Average Length				
1.11	-	1.86744	1.31289	0.95830	0.68388	0.55761	0.48368	0.43289
	$(X_g, X_g^{(N)})$	1.30222	0.94977	0.70427	0.50804	0.41405	0.36055	0.32280
1.42	-	1.86822	1.30105	0.95121	0.67677	0.55462	0.48111	0.43046
	$(X_g, X_g^{(N)})$	1.76667	1.22571	0.89458	0.63665	0.52213	0.45247	0.40482
1.99	-	1.85639	1.29289	0.94626	0.67421	0.55160	0.47822	0.42849
	$(X_g, X_g^{(N)})$	2.54781	1.72304	1.25092	0.87988	0.72210	0.62598	0.55911
3.31	-	1.83716	1.29155	0.94173	0.67099	0.54871	0.47588	0.42645
	$(X_g, X_g^{(N)})$	3.56010	2.39697	1.73381	1.22024	0.99619	0.86635	0.77370
9.80	-	1.83555	1.28894	0.93697	0.66756	0.54602	0.47402	0.42411
	$(X_g, X_g^{(N)})$	4.86067	3.24720	2.34399	1.64604	1.34678	1.16835	1.04106

[*] Number of clusters= $2G$ with $G = 12, 26, 50, 100, 150, 200, 250$. Number of replications for each G is 2000. $N_{max} = 500$.

CHAPTER 3

INFERENCE FOR TWO-STAGE EXPERIMENTS UNDER COVARIATE-ADAPTIVE RANDOMIZATION

3.1 Introduction

This paper considers the problem of inference in two-stage randomized experiments under covariate-adaptive randomization. Here, a two-stage randomized experiment refers to a design where clusters (e.g., households, schools, or graph partitions) are initially randomly assigned to either a control or treatment group. Subsequently, random assignment of units within each treated cluster to either treatment or control is carried out based on a pre-determined treated fraction. Covariate-adaptive randomization refers to randomization schemes that first stratify according to baseline covariates and then assign treatment status so as to achieve "balance" within each stratum. Two-stage randomized experiments are widely used in social science (see for example Duflo and Saez (2003); Haushofer and Shapiro (2016); McKenzie and Puerto (2021)), and discussed by statisticians (see for example Hudgens and Halloran (2008)), as a general approach to causal inference with interference; that is, when one individual's treatment status affects outcomes of other individuals. Moreover, practitioners often use covariate information to design more efficient two-stage experiments (see for example Duflo and Saez, 2003; Beuermann et al., 2015; Ichino and Schündeln, 2012; Aramburu et al., 2019; Hidrobo et al., 2016; Kinnan et al., 2020; Malani et al., 2021; Muralidharan and Sundararaman, 2015; Banerjee et al., 2021; Rogers and Feller, 2018). However, to the best of my knowledge, there has not yet been any formal analysis on covariate-adaptive randomization in two-stage randomized experiments. Accordingly, this paper establishes general results about estimation and inference for two-stage designs under covariate-adaptive randomization. Subsequently, I propose and examine the optimality of two-stage designs with "matched tuples", i.e. a generalized matched-pair design (see Bai (2022b) and Bai et al.

(2022b)).

This paper first classifies experiments under covariate-adaptive randomization into two categories: “large strata” and “small strata”, and then comes up with two different asymptotic regimes to study the large sample properties of such designs. In the “large strata” case, clusters are divided into a fixed number of large strata according to baseline cluster-level covariates. In each stratum, the number of clusters grows to infinity as the total number of clusters grows to infinity. Conversely, in the case of “small strata”, clusters are divided into small strata of fixed size, and thus the number of strata grows to infinity as the total number of clusters grows to infinity. Such asymptotic regimes are also manifested in previous works on covariate-adaptive experiments with individual-level assignments (see Bugni et al. (2018a) for “large strata”, Bai et al. (2021b) for “small strata”, and Cytrynbaum (2023b) for both). Adopting this classification enables the development of asymptotically-exact statistical inference methods for a wide range of covariate-adaptive designs found in the empirical literature. Moreover, separating the two asymptotic regimes facilitates the design of variance estimators suitable for each scenario.[1]

This paper then considers the asymptotic properties of a commonly recommended inference procedure based on a linear regression with cluster-robust standard errors. My findings suggest that the corresponding t-test is generally valid but conservative. I also demonstrate that in the first stage of cluster-level assignment, covariate information about clusters is important for both designing efficient experiments and consistently estimating variances under covariate-adaptive randomization. However, in the second stage of unit-level assignment, while individual-level covariate information is useful for improving efficiency, it is not required for the proposed inference method. Specifically, I show that consistent variance estimators can be constructed using only the cluster-level covariates from the first stage

[1] I acknowledge that the terms “cluste” and “stratum” are both used in the literature to describe groupings of units, which can lead to confusion. Here, I define a cluster as a pre-determined group of units (e.g., households, schools, or graph partitions) and a stratum as a group of clusters that share similar baseline cluster-level covariates.

design, regardless of the use of individual-level covariates in the second stage.

Next, I apply the results to study optimal use of covariate information in two-stage designs. Here, by “optimal”, I mean designs that achieve the minimum asymptotic variances within the class of designs considered in the paper. For all estimands of interest, the designs in the first and second stage affect the efficiency independently. Thus, I am able to identify optimal designs in the first and second stage separately and use them together as the optimal two-stage design. My result shows that, at each stage, the asymptotically optimal design is a “matched tuples” design where clusters or units are matched based on an index function (similar to Bai (2022b)) that is specific to the given estimator. In a simulation study, the results demonstrate that properly designed two-stage experiments utilizing the optimality results outperform other designs. However, the efficiency gain achieved through proper second-stage randomization is significantly lower compared to the first stage under my simulation specifications.

Finally, this paper evaluates the proposed inference method against various regression-based methods commonly used in empirical literature in a simulation study and empirical application. The simulation study confirms the asymptotic exactness of the inference results and highlights that statistical inference based on various ordinary least squares regressions could either be too conservative or invalid. Specifically, my result verifies that the commonly used regression with cluster-robust standard errors is conservative, while the other regression-based methods examined in the paper, such as regressions with strata fixed effects or heteroskedasticity-robust standard errors, are generally invalid. In the empirical application, I demonstrate the proposed inference method based on the experiment conducted in Ichino and Schündeln (2012) and compare it with regression-based methods. The empirical findings are consistent with the results of the simulation study.

The analysis of data from two-stage randomized experiments and experiments under covariate-adaptive randomization has received considerable attention, but most work has focused on only one of these two features at a time. Previous work on the analysis of two-stage randomized experiments includes Liu and Hudgens (2014), Rigdon and Hudgens (2015), Basse and Feller (2018), Basse et al. (2019), Vazquez-Bare (2022), Cruces et al. (2022), Imai et al. (2021) and Jiang et al. (2022b). Recent work on the analysis of covariate-adaptive experiments includes Bugni et al. (2018a), Cytrynbaum (2023b), Jiang et al. (2021), Jiang et al. (2022a), Bai et al. (2021b), Bai (2022b), Bai et al. (2022b), Bai et al. (2022a) and Bai et al. (2023b). In fact, both Basse and Feller (2018) and Imai et al. (2021) applied their inference methods, which do not account for covariate information, to two-stage experiments under covariate-adaptive randomization.[2] My framework of analysis follows closely Bugni et al. (2022b), in which they formalize cluster randomized experiments in a super population framework.

This paper contributes to the methodology for a growing number of empirical papers using two-stage experiments with covariate-adaptive randomization. For instance, Hidrobo et al. (2016), Banerjee et al. (2021), Rogers and Feller (2018) and Foos and de Rooij (2017) conducted two-stage randomized experiments that stratify clusters or units into a small number of large strata according to their baseline covariates, typically known as stratification design. Duflo and Saez (2003), Beuermann et al. (2015), Ichino and Schündeln (2012) and Kinnan et al. (2020) conducted two-stage randomized experiments in which clusters or units are matched into small strata according to their baseline covariates, commonly known as matched pairs, matched triplets or matched tuples designs.

The rest of the paper is organized as follows. Section 3.2 describes the setup and notation. Section 3.3 and 3.4 present the main results under the two asymptotic regimes. Section 3.5 discusses the optimality of matched tuples designs. Section 3.6 examines the finite sample behavior of various experimental designs through simulations. Section 3.7 illustrates the

[2] Basse and Feller (2018) analyzes the empirical application from Rogers and Feller (2018), whose design involves stratification on school, grade, and prior-year absences. Imai et al. (2021) analyzes the empirical application from Kinnan et al. (2020), whose design involves matching villages (clusters) and households into small blocks.

proposed inference methods in an empirical application based on the experiment conducted in Ichino and Schündeln (2012). Finally, I conclude with recommendations for empirical practice in Section 3.8.

3.2 Setup and Notation

Let $Y_{i,g}$ and $X_{i,g}$ denote the observed outcome and individual baseline covariates of the ith unit in the gth cluster, respectively. Denote by $Z_{i,g}$ the indicator for whether the ith unit in the gth cluster is treated or not. Let C_g denote the observed baseline covariates for the gth cluster, N_g denote the size of the gth cluster, H_g denote the target fraction of units treated in the gth cluster, and G the number of observed clusters. In addition, define $\mathcal{M}_g$ as the (possibly random) subset of $\{1, ..., N_g\}$ corresponding to the observations within the gth cluster that are sampled by the researcher. Let $M_g = |\mathcal{M}_g|$ denote the number of units in set $\mathcal{M}_g$. In other words, the researcher randomly assigns treatments to all N_g units in the gth cluster but only observes or conducts analysis on a subset of units sampled from the gth cluster (see for example Beuermann et al., 2015; Aramburu et al., 2019; Haushofer and Shapiro, 2016; Haushofer et al., 2019; Hidrobo et al., 2016; Malani et al., 2021; Muralidharan and Sundararaman, 2015; Banerjee et al., 2021). Denote by P_G the distribution of the observed data

$$V^{(G)} := \left(\left(Y_{i,g}, X_{i,g}, Z_{i,g} : i \in \mathcal{M}_g\right), H_g, C_g, N_g : 1 \leq g \leq G\right) \ .$$

This paper considers a setup where units are partitioned into a large number of clusters. In this context, the paper studies a two-stage randomized experiment with binary treatment in both stages. In the first stage, a fraction of π_1 clusters are randomly assigned to the treatment group, while the remaining clusters are assigned to the control group with no treated units. Then, conditional on the assignment in the first stage, a fraction of π_2 individuals from treated clusters are assigned to the treatment group, while the remaining units are assigned to the control group. Such a binary design is widely used in empirical literature (see, e.g., Foos and de Rooij, 2017; Duflo and Saez, 2003; Ichino and Schündeln, 2012; Haushofer and Shapiro, 2016; Haushofer et al., 2019). Moreover, while some experiments have multiple treated fractions, researchers often analyze them as binary designs (see, e.g., Basse and Feller, 2018; Imai et al., 2021; Beuermann et al., 2015).

3.2.1 Potential Outcomes and Interference

In this section, I provide assumptions on the interference structure that assume no interference across clusters and exchangeable/homogeneous interference within clusters. Let $Y_{i,g}(\mathbf{z}, n)$ denote the potential outcome of the ith unit in the gth cluster, where n denotes the cluster size and $\mathbf{z}$ denotes a realized vector of assignment for all units in all clusters, i.e., $\mathbf{z} = ((z_{i,g} : 1 \leq i \leq n) : 1 \leq g \leq G)$, where $z_{i,g} \in \{0, 1\}$ denotes a realized assignment for the ith unit in the gth cluster. Following previous work (see, for example, Hudgens and Halloran, 2008; Basse and Feller, 2018; Basse et al., 2019; Forastiere et al., 2021; Imai et al., 2021), I assume the following about potential outcomes.

Assumption 3.2.1 (Homogeneous partial interference)**.**

$$Y_{i,g}(\mathbf{z}, n) = Y_{i,g}(\mathbf{z}', n) \text{ w.p.1 } \text{ if } z_{i,g} = z'_{i,g} \text{ and } \sum_{1 \leq j \leq n} z_{j,g} = \sum_{1 \leq j \leq n} z'_{j,g}$$

$$\text{for any } 1 \leq i \leq n, 1 \leq g \leq G \ ,$$

where $\mathbf{z}$ and $\mathbf{z}'$ are any realized vectors of assignment, and $z_{i,g}, z'_{i,g}$ are the corresponding individual treatment indicators for i-th unit in g-th cluster.

Under Assumption 3.2.1, potential outcomes can be simplified as $Y_{i,g}(z, n, n_1)$ where n_1

denotes the number of treated units in the cluster. Following this notation, we define

$$Y_{i,g}(z,h) := \sum_{n\geq 1} Y_{i,g}(z,n,\lfloor nh \rfloor) I\{N_g = n\}$$

to be the potential outcome under the individual treatment status $z \in \{0,1\}$ and the cluster target treated fractions $h \in \mathcal{H} \subseteq [0,1]$, where $\mathcal{H}$ is a pre-determined set of treated fractions.[3] As mentioned before, this paper considers binary treatments, i.e. $\mathcal{H} = \{0, \pi_2\}$, throughout the paper.[4] Furthermore, the (observed) outcome and potential outcomes are related to treatment assignment by the relationship $Y_{i,g} = Y_{i,g}(Z_{i,g}, H_g)$. Denote by Q_G the distribution of

$$W^{(G)} := \left(\left(\left(Y_{i,g}(z,h) : z \in \{0,1\}, h \in \mathcal{H}\right), X_{i,g} : 1 \leq i \leq N_g\right), \mathcal{M}_g, C_g, N_g : 1 \leq g \leq G\right) .$$

3.2.2 Distribution and Sampling Procedure

The distribution P_G of observed data and its sampling procedure can be described in three steps. First, $\{(\mathcal{M}_g, C_g, N_g) : 1 \leq g \leq G\}$ are i.i.d samples from a population distribution. Second, potential outcomes and baseline individual covariates are sampled from a conditional distribution $R_G(\mathcal{M}^{(G)}, C^{(G)}, N^{(G)})$, which is defined as follows:

$$\left(\left(\left(Y_{i,g}(z,h) : z \in \{0,1\}, h \in \mathcal{H}\right), X_{i,g} : 1 \leq i \leq N_g\right) : 1 \leq g \leq G\right) \mid \mathcal{M}^{(G)}, C^{(G)}, N^{(G)} .$$

Finally, P_G is jointly determined by the relationship $Y_{i,g} = Y_{i,g}(Z_{i,g}, H_g)$ together with the assignment mechanism, which will be described in Section 3.3 and 3.4, and Q_G, which is

[3] For example, when the cluster size is 3 and the target treated fraction is 0.5, there will be one treated unit in the cluster. Other rounding approaches, like the ceiling function, to handle fractional numbers of treated units can also be easily accommodated.

[4] Extending the designs to accommodate multiple treatment fractions is technically straightforward. Related work can be found in Bugni et al. (2019b).

described in the first two steps. Note that I use $A^{(G)}$ to denote the vector $(A_1, \dots, A_G)$ for any random variable A. The following assumption states my requirements on Q_G using this notation.

Assumption 3.2.2. The distribution Q_G is such that

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

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

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

where $R(\mathcal{M}_g, C_g, N_g)$ denotes the distribution of

$$\big((Y_{i,g}(z, h) : z \in \{0, 1\}, h \in \mathcal{H}) \, , X_{i,g} : 1 \leq i \leq N_g \big)$$

conditional on $\{\mathcal{M}_g, C_g, N_g\}$.

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

(d) For some constant $C < \infty$, $P\Big\{E[Y_{i,g}^2(z, h) \mid N_g, C_g] \leq C \text{ for all } 1 \leq i \leq N_g\Big\} = 1$ for all $z \in \{0, 1\}$ and $h \in \mathcal{H}$ and $1 \leq g \leq G$.

(e) $\mathcal{M}_g \perp \big((Y_{i,g}(z, h) : z \in \{0, 1\}, h \in \mathcal{H}) : 1 \leq i \leq N_g \big) \mid C_g, N_g$ for all $1 \leq g \leq G$.

(f) For all $z \in \{0, 1\}, h \in \mathcal{H}$ and $1 \leq g \leq G$,

$$E\left[\frac{1}{M_g} \sum_{i \in \mathcal{M}_g} Y_{i,g}(z, h) \mid N_g \right] = E\left[\frac{1}{N_g} \sum_{1 \leq i \leq N_g} Y_{i,g}(z, h) \mid N_g \right] \text{w.p.1} \ .$$

The sampling procedure of a cluster randomized experiment used in this paper closely follows that formalized by Bugni et al. (2022b) and Bai et al. (2022a). Assumption 3.2.2 is exactly the same as Assumption 2.2 in Bugni et al. (2022b), which formalizes the sampling procedure of i.i.d. clusters (Assumptions 3.2.2 (a)-(b)) and imposes mild regularity conditions (Assumptions 3.2.2 (c)-(d)). In addition, Assumption 3.2.2 (e) allows the second-stage sampling process in a given cluster to depend on cluster-level covariates and cluster sizes but rules out dependence on the potential outcomes within the cluster. Finally, Assumption 3.2.2 (f) is a high-level assumption that ensures the extrapolation from the observations that are sampled to those that are not sampled.

3.2.3 Parameters of Interest and Estimators

In the context of the sampling framework described above, this paper considers four parameters of interest, including primary and spillover effects that are equally or (cluster) size-weighted. For different choices of (possibly random) weights ω_g, $1 \leq g \leq G$ satisfying $E[\omega_g] = 1$, we define the average *primary effects* and *spillover effects* under general weights as follows.

Definition 3.2.1. *Define the weighted average primary effect under weight w_g as follows:*

$$\theta_\omega^P(Q_G) := E\left[\omega_g\left(\frac{1}{N_g}\sum_{1\leq i\leq N_g} Y_{i,g}(1,\pi_2) - Y_{i,g}(0,0)\right)\right] , \tag{3.1}$$

and the average spillover effect as:

$$\theta_\omega^S(Q_G) := E\left[\omega_g\left(\frac{1}{N_g}\sum_{1\leq i\leq N_g} Y_{i,g}(0,\pi_2) - Y_{i,g}(0,0)\right)\right] . \tag{3.2}$$

Denote by $\theta_1^P(Q_G)$ and $\theta_1^S(Q_G)$ the equally-weighted cluster-level average primary and spillover effects with $\omega_g = 1$, and $\theta_2^P(Q_G)$ and $\theta_2^S(Q_G)$ the size-weighted cluster-level average primary and spillover effects with $\omega_g = N_g/E[N_g]$. The choice of inferential target depends on whether cluster size is meaningful for the parameter of interest. For example, household size may not be important when estimating the effect of an educational program on the average income of households in a city, but if the experiment is clustered by neighborhood, then the cluster size (i.e., the number of households in each neighborhood) may be meaningful and should be taken into account in the analysis. The primary effects $\theta_1^P(Q_G)$ and $\theta_2^P(Q_G)$ are the differences in the averaged potential outcomes of treated units from treated clusters and control units from control clusters. In contrast, the spillover effects $\theta_1^S(Q_G)$ and $\theta_2^S(Q_G)$ are the differences in the averaged potential outcomes of control units from treated clusters and control units from control clusters. In many empirical settings, the estimation and comparison of primary and spillover effects play a crucial role in addressing important research questions (see for example Duflo and Saez, 2003).

In summary, the formulas for the four parameters of interest are listed in Table 3.1. These estimands have been proposed and studied in previous literature (see, e.g., Hudgens and Halloran, 2008; Basse and Feller, 2018; Toulis and Kao, 2013; Imai et al., 2021), but mostly in a finite population framework. This paper adopts the terminology "primary" and "spillover" effects from Basse and Feller (2018), which are respectively referred to as "total" and "indirect" effects in Hudgens and Halloran (2008). Previous works on interference have also studied other estimands, such as direct effects and overall effects (see, e.g., Hudgens and Halloran, 2008; Imai et al., 2021; Hu et al., 2021), but I do not explore these estimands further in this paper.

Parameter of interest	Formula
Equally-weighted primary effect	$\theta_1^P(Q_G) := E\left[\frac{1}{N_g}\sum_{1\le i\le N_g} Y_{i,g}(1,\pi_2) - Y_{i,g}(0,0)\right]$
Equally-weighted spillover effect	$\theta_1^S(Q_G) := E\left[\frac{1}{N_g}\sum_{1\le i\le N_g} Y_{i,g}(0,\pi_2) - Y_{i,g}(0,0)\right]$
Size-weighted primary effect	$\theta_2^P(Q_G) := E\left[\frac{1}{E[N_g]}\sum_{1\le i\le N_g} Y_{i,g}(1,\pi_2) - Y_{i,g}(0,0)\right]$
Size-weighted spillover effect	$\theta_2^S(Q_G) := E\left[\frac{1}{E[N_g]}\sum_{1\le i\le N_g} Y_{i,g}(0,\pi_2) - Y_{i,g}(0,0)\right]$

Table 3.1: Parameters of interest

For estimating the four parameters of interest, I propose the following estimators analogous to the difference-in-"average of averages" estimator in Bugni et al. (2022b):

$$\begin{aligned}
\hat{\theta}_1^P &= \frac{1}{G_1} \sum_{1 \le g \le G} I\{H_g = \pi_2\} \bar{Y}_g^1 - \frac{1}{G_0} \sum_{1 \le g \le G} I\{H_g = 0\} \bar{Y}_g^1 \\
\hat{\theta}_1^S &= \frac{1}{G_1} \sum_{1 \le g \le G} I\{H_g = \pi_2\} \bar{Y}_g^0 - \frac{1}{G_0} \sum_{1 \le g \le G} I\{H_g = 0\} \bar{Y}_g^0 \\
\hat{\theta}_2^P &= \frac{1}{N_1} \sum_{1 \le g \le G} I\{H_g = \pi_2\} N_g \bar{Y}_g^1 - \frac{1}{N_0} \sum_{1 \le g \le G} I\{H_g = 0\} N_g \bar{Y}_g^1 \\
\hat{\theta}_2^S &= \frac{1}{N_1} \sum_{1 \le g \le G} I\{H_g = \pi_2\} N_g \bar{Y}_g^0 - \frac{1}{N_0} \sum_{1 \le g \le G} I\{H_g = 0\} N_g \bar{Y}_g^0 \ ,
\end{aligned}$$

where $G_1 = \sum_{1 \le g \le G} I\{H_g = \pi_2\}$, $G_0 = \sum_{1 \le g \le G} I\{H_g = 0\}$, and $N_1 = \sum_{1 \le g \le G} I\{H_g = \pi_2\} N_g$, $N_0 = \sum_{1 \le g \le G} I\{H_g = 0\} N_g$ and

$$\bar{Y}_g^z = \frac{1}{M_g^z} \sum_{i \in \mathcal{M}_g} Y_{i,g} I\{H_g = \pi_2, Z_{i,g} = z\} + \frac{1}{M_g} \sum_{i \in \mathcal{M}_g} Y_{i,g} I\{H_g = 0\} \ ,$$

where $M_g^z = \sum_{i \in \mathcal{M}_g} I\{Z_{i,g} = z\}$ with $z \in \{0, 1\}$.

By definition, the "first/individual average" $\bar{Y}_g^1$ from the primary effect estimator is taken over all treated units within the g-th cluster if the cluster is treated, and all control units within the g-th cluster if the cluster is assigned to control. When it comes to estimating spillover effects, the "first/individual average" $\bar{Y}_g^0$ is taken over all control units within the g-th cluster if the cluster is treated, and all control units within the g-th cluster if the cluster is assigned to control. Then, the "second/cluster average" is a cluster-level average of $\bar{Y}_g^1$ or $\bar{Y}_g^0$ taken within groups of treated and untreated clusters as featured in a usual difference-in-means estimator. Let $L_{i,g} = I\{H_g = \pi_2\}(1 - Z_{i,g})$ denote the indicator for untreated units within treated clusters. Note that $\hat{\theta}_1^P$ and $\hat{\theta}_1^S$ (or $\hat{\theta}_2^P$ and $\hat{\theta}_2^S$) may be obtained as the estimators of the coefficient of $Z_{i,g}$ and $L_{i,g}$ in a weighted least squares regression of $Y_{i,g}$ on a constant and $Z_{i,g}$ and $L_{i,g}$ with weights equal to $\sqrt{1/M_g}$ (or $\sqrt{N_g/M_g}$) (see Appendix

C.4 for formal derivations).

My estimators are closely related to those studied in previous methodological literature. For example, equally-weighted estimators $\hat{\theta}_1^P$ and $\hat{\theta}_1^S$ are identical to the household-weighted estimators from Basse and Feller (2018), which are closely related to the estimators in Hudgens and Halloran (2008). $\hat{\theta}_1^P$ and $\hat{\theta}_1^S$ may also be obtained through the "household-level regression" proposed in Basse and Feller (2018), which is equivalent to running two separate ordinary least squares regressions of $\bar{Y}_g^1$ on a constant and $I\{H_g = \pi_2\}$, and $\bar{Y}_g^0$ on a constant and $I\{H_g = \pi_2\}$. Size-weighted estimators $\hat{\theta}_2^P$ and $\hat{\theta}_2^S$ are closely related to the individual-weighted estimator proposed by Basse and Feller (2018). In previous studies such as Cruces et al. (2022), Vazquez-Bare (2022), and Basse and Feller (2018), researchers have investigated estimators obtained through a widely used saturated regression in multi-treatment experiments, similar to the form of equation (3.3). These estimators are identical to $\hat{\theta}_2^P$ and $\hat{\theta}_2^S$ when outcomes of all units from each cluster are observed or the number of observed units is proportional to the cluster size. More formally, when $M_g/N_g = c$ for $0 < c \leq 1$, $\hat{\theta}_2^P$ and $\hat{\theta}_2^S$ can be obtained simultaneously as the estimators of the coefficient on $Z_{i,g}$ and $L_{i,g}$ through a regression like the following:

$$Y_{i,g} = \alpha + \beta_1 Z_{i,g} + \beta_2 L_{i,g} + \epsilon_{i,g} \ . \tag{3.3}$$

In empirical literature, various regression estimators are used for estimating primary and spillover effects. One widely used estimator is described in equation (3.3) (see, e.g., Haushofer and Shapiro, 2016; Haushofer et al., 2019). Another estimator that produces the same set of estimators is through the alternative regression $Y_{i,g} = a + b_1 Z_{i,g} + b_2 I\{H_g = \pi_2\} + u_{i,g}$ (see, e.g., Duflo and Saez, 2003; Ichino and Schündeln, 2012), where the estimators are related to those from (3.3) as follows: $\hat{\beta}_1 = \hat{b}_1 + \hat{b}_2$ and $\hat{\beta}_2 = \hat{b}_2$. Some empirical works use either or both of the two separate regressions: $Y_{i,g} = \alpha + \beta_1 Z_{i,g} + \epsilon_{i,g}$ and $Y_{i,g} = \alpha + \beta_2 L_{i,g} + \epsilon_{i,g}$ (see, e.g., Beuermann et al., 2015; Aramburu et al., 2019; Hidrobo et al., 2016). In many cases, estimators obtained from regressions with fixed effects are reported along with those without fixed effects (see, e.g., Ichino and Schündeln, 2012). Section 3.6.2 will examine the validity of statistical tests based on regressions with and without fixed effects.

3.3 Inference for Experiments with Large Strata

In this section, I investigate the asymptotic properties of the estimators presented in Section 3.2.3 in the context of two-stage stratified experiments with a fixed number of large strata in the first stage of the experimental design. Specifically, in the first stage, clusters are partitioned into a fixed number of strata such that the number of clusters within each stratum grows as the total number of clusters increases. Formally, denote by $S^{(G)} = (S_1, \ldots, S_G)$ the vector of strata on clusters, constructed from the observed, baseline covariates C_g and cluster size N_g for gth cluster using a function $S : \text{supp}((C_g, N_g)) \to \mathcal{S}$, where $\mathcal{S}$ is a finite set. Furthermore, I consider a second-stage stratification on units from a given cluster. Denote by $B_g = (B_{i,g} : 1 \leq i \leq N_g)$ the vector of strata on units in the gth cluster, constructed from observed, baseline covariates $X_{i,g}$ for ith unit using a function $B : \text{supp}(X_{i,g}) \to \mathcal{B}_g$.[5]

Example 3.3.1. Section 3.7 presents an illustrative empirical example of such a large-strata experiment conducted by Foos and de Rooij (2017). In the first stage of their experiment, 5,190 two-voter households (i.e., clusters of size 2) were categorized into three strata: "Labour" supporter, "rival party" supporter, and those "unattached" to any party. Within each stratum, households were then randomly allocated to either treatment or control groups. In the subsequent stage, one member from the households in the treatment group was given the treatment. ■

First of all, I provide notations for the quantity of imbalance for each stratum. For $s \in \mathcal{S}$,

[5] Asymptotics are not considered in the second-stage design; thus, the second stage could employ designs with small strata like matched-pair, or those with large strata such as stratified block randomization.

let

$$D_G(s) = \sum_{1 \le g \le G} (I\{H_g = \pi_2\} - \pi_1) I\{S_g = s\}, \tag{3.4}$$

where $\pi_1 \in (0,1)$ is the "target" proportion of clusters to assign to treatment in each stratum. My requirements on the treatment assignment mechanism for the first stage are summarized in the following assumption:

Assumption 3.3.1. The treatment assignment mechanism for the first-stage is such that

(a) $W^{(G)} \perp H^{(G)} \mid S^{(G)}$,

(b) $\left\{ \left\{ \frac{D_G(s)}{\sqrt{G}} \right\}_{s \in \mathcal{S}} \mid S^{(G)} \right\} \xrightarrow{d} N(0, \Sigma_D)$ 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 $s \in \mathcal{S}$, and $p(s) = P\left\{S_g = s\right\}$.

Assumption 3.3.1 (a) simply requires that the treatment assignment mechanism is a function only of the vector of strata and an exogenous randomization device. Assumption 3.3.1 (b) follows Assumption 2.2 (b) of Bugni et al. (2018a). This assumption is commonly satisfied by various experiment designs, such as Bernoulli trials, stratified block randomization, and Efron's biased-coin design, which are widely used in clinical trials and development economics.

The next step is to formalize the assumption of independence between the first and second stage designs. To begin with, I utilize the notation $\{Z_{i,g}(h) : h \in \mathcal{H}\}$, representing the "potential treatment" for various treated fractions $h \in \mathcal{H}$, and relate the (observed) individual treatment indicator and potential individual-level treatment indicator as follows:

$$Z_{i,g} = \sum_{h \in \mathcal{H}} Z_{i,g}(h) I\{H_g = h\} \text{ for } 1 \le i \le N_g \ . \tag{3.5}$$

The underlying motivation for this "potential outcome style" notation becomes evident when considering that in two-stage experiments, the realized treatment assignment in the first stage is almost always correlated with that in the second stage (e.g., $H_g = \frac{1}{N_g} \sum_{1 \le i \le N_g} Z_{i,g}$). Yet, the "potential" individual-level treatment assignment, for any specified target treated fraction, can be independent of the cluster-level assignment of that target treated fraction. This is similar to the classic potential outcome model, where treatment assignment is independent of potential outcomes but likely correlates with observed outcomes.

Then, my requirements on the treatment assignment mechanism for the second stage are summarized in the following assumption:

Assumption 3.3.2. The treatment assignment mechanism for the second-stage is such that

(a) $(((Z_{i,g}(h) : h \in \mathcal{H}) : 1 \le i \le N_g) : 1 \le g \le G) \perp H^{(G)}$,

(b) $W^{(G)} \perp (((Z_{i,g}(h) : h \in \mathcal{H}) : 1 \le i \le N_g) : 1 \le g \le G) \mid (B_g : 1 \le g \le G)$,

(c) For all $1 \le g \le G$, $E[Z_{i,g}(h) \mid B_g] = \frac{1}{M_g} \sum_{i \in \mathcal{M}_g} Z_{i,g}(h)$.

Assumption 3.3.2 (a) rules out any confounders between the first-stage and second-stage treatment assignments, which is typically satisfied in most two-stage experiments. Assumption 3.3.2 (b) is analogous to Assumption 3.3.1 (a). Assumption 3.3.2 (c) requires that the marginal assignment probability is equal to the realized treated fraction on the observed subset of units. An example of this could be (individual-level) stratified block randomization, where the treated fraction remains constant across all strata, with observed units drawn from a random subset of these strata.

The following theorem derives the asymptotic behavior of estimators for equally-weighted effects.[6]

[6] Throughout the paper, $V_1(1), V_2(1), V_3(1)$ and $V_4(1)$ denote the variances of primary effects, while $V_1(0), V_2(0), V_3(0)$ and $V_4(0)$ represent the variances of spillover effects. In other words, the notation $z \in \{0, 1\}$ (as in $V_1(z)$) represents the individual's own treatment status.

Theorem 3.3.1. *Under Assumption 3.2.1-3.2.2 and 3.3.1-3.3.2,*

$$\sqrt{G}\left(\hat{\theta}_1^P - \theta_1^P(Q_G)\right) \to \mathcal{N}(0, V_1(1)) \ , \tag{3.6}$$

and

$$\sqrt{G}\left(\hat{\theta}_1^S - \theta_1^S(Q_G)\right) \to \mathcal{N}(0, V_1(0)) \ , \tag{3.7}$$

where

$$\begin{aligned} V_1(z) = & \frac{1}{\pi_1} \operatorname{Var}\left[\bar{Y}_g(z, \pi_2)\right] + \frac{1}{1-\pi_1} \operatorname{Var}\left[\bar{Y}_g(0,0)\right] \\ & - \pi_1(1-\pi_1) E\left[\left(\frac{1}{\pi_1} m_{z,\pi_2}\left(S_g\right) + \frac{1}{1-\pi_1} m_{0,0}\left(S_g\right)\right)^2\right] \\ & \quad + E\left[\tau\left(S_g\right)\left(\frac{1}{\pi_1} m_{z,\pi_2}\left(S_g\right) + \frac{1}{1-\pi_1} m_{0,0}\left(S_g\right)\right)^2\right] \ , \end{aligned} \tag{3.8}$$

with

$$\bar{Y}_g(1, \pi_2) = \frac{1}{M_g^1} \sum_{i \in \mathcal{M}_g} Y_{i,g}(1, \pi_2) Z_{i,g}(\pi_2) \tag{3.9}$$

$$\bar{Y}_g(0, \pi_2) = \frac{1}{M_g^0} \sum_{i \in \mathcal{M}_g} Y_{i,g}(0, \pi_2)(1 - Z_{i,g}(\pi_2)) \tag{3.10}$$

$$\bar{Y}_g(0,0) = \frac{1}{M_g} \sum_{i \in \mathcal{M}_g} Y_{i,g}(0,0) \tag{3.11}$$

$$m_{z,h}\left(S_g\right) = E[\bar{Y}_g(z,h) \mid S_g] - E[\bar{Y}_g(z,h)] \ . \tag{3.12}$$

Remark 3.3.1. An alternative variance expression, analogous to equation (15) in Bugni et al. (2018a), is:

$$V_1(z) = \frac{1}{\pi_1} \operatorname{Var}\left[\check{Y}_g(z,\pi_2)\right] + \frac{1}{1-\pi_1} \operatorname{Var}\left[\check{Y}_g(0,0)\right] + E\left[\left(m_{z,\pi_2}\left(S_g\right) - m_{0,0}\left(S_g\right)\right)^2\right]$$
$$+ E\left[\tau\left(S_g\right)\left(\frac{1}{\pi_1} m_{z,\pi_2}\left(S_g\right) + \frac{1}{1-\pi_1} m_{0,0}\left(S_g\right)\right)^2\right] \ , \tag{3.13}$$

where $\check{Y}_g(z,h) = \bar{Y}_g(z,h) - E[\bar{Y}_g(z,h) \mid S_g]$. By comparing (3.13) with the variance expression in Bugni et al. (2018a), we conclude that the asymptotic variance in Theorem 3.3.1 corresponds exactly to the asymptotic variance of the difference-in-means estimator for covariate-adaptive experiments with individual-level "one-stage" assignment, as in Bugni et al. (2018a). In fact, when $P(N_g = 1) = 1$ and $\pi_2 = 1$, $V_1(1)$ collapses to their variance expression.

In a special case where covariate information is not used to construct strata and the first stage is a "strong balanced" design with $\mathcal{S} = s$ and $\tau(s) = 0$, the asymptotic variance of the estimated treatment effect can be expressed as follows:

$$V_1(z) = \frac{1}{\pi_1} \operatorname{Var}\left[\bar{Y}_g(z,\pi_2)\right] + \frac{1}{1-\pi_1} \operatorname{Var}\left[\bar{Y}_g(0,0)\right] \ , \tag{3.14}$$

which is equivalent to the identifiable parts of the variance derived in Basse and Feller (2018) under the finite population framework. The asymptotic variance of partial population designs from Cruces et al. (2022) is also closely related to (3.14) under binary settings. Specifically, Cruces et al. (2022) provides an alternative expression of $\operatorname{Var}\left[\bar{Y}_g(z,\pi_2)\right]$ with intra-cluster variances and correlations. Therefore, inference methods based on (3.14), including Basse and Feller (2018) and Cruces et al. (2022), are generally conservative under "strong balanced" covariate-adaptive randomization. ■

Remark 3.3.2. It's worth noting that the setup of the first-stage design has a clear impact on the asymptotic variance $V_1(z)$, as evidenced by the third and fourth term in equation (3.8). Furthermore, the second-stage design also influences the asymptotic variance $V_1(z)$, albeit more implicitly, via the distribution of $Z_{i,g}(\pi_2)$ that the practitioners design. Specifically, the first term in equation (3.8) depends on Var $\left[\check{Y}_g(z,\pi_2)\right]$, which is directly tied to the second-stage design. Thus, the efficacy of designing the first stage versus the second stage can be disentangled into distinct components. This separation could be beneficial for practitioners seeking to assess the relative importance of first-stage design versus second-stage design in optimizing efficiency gains. For instance, a calibrated simulation study using pilot data can be used to estimate the relative efficiency gain obtained at each stage. ■

The following theorem derives the asymptotic behavior of estimators for size-weighted effects.

Theorem 3.3.2. *Under Assumption 3.2.1-3.2.2 and 3.3.1-3.3.2,*

$$\sqrt{G}\left(\hat{\theta}_2^P - \theta_2^P(Q_G)\right) \to \mathcal{N}(0, V_2(1)) \ , \tag{3.15}$$

and

$$\sqrt{G}\left(\hat{\theta}_2^S - \theta_2^S(Q_G)\right) \to \mathcal{N}(0, V_2(0)) \ , \tag{3.16}$$

where

$$\begin{aligned} V_2(z) = \frac{1}{\pi_1}\operatorname{Var}[\tilde{Y}_g(z,\pi_2)] &+ \frac{1}{1-\pi_1}\operatorname{Var}[\tilde{Y}_g(0,0)] \\ &- \pi_1(1-\pi_1)E\left[\left(\frac{1}{\pi_1}E[\tilde{Y}_g(z,\pi_2) \mid S_g] + \frac{1}{1-\pi_1}E[\tilde{Y}_g(0,0) \mid S_g]\right)^2\right] \\ &+ E\left[\tau(S_g)\left(\frac{1}{\pi_1}E[\tilde{Y}_g(z,\pi_2) \mid S_g] + \frac{1}{1-\pi_1}E[\tilde{Y}_g(0,0) \mid S_g]\right)^2\right] \ , \end{aligned} \tag{3.17}$$

with

$$\tilde{Y}_g(z,h) = \frac{N_g}{E[N_g]}\left(\bar{Y}_g(z,h) - \frac{E[\bar{Y}_g(z,h)N_g]}{E[N_g]}\right) \tag{3.18}$$

for $(z,h) \in \{(1,\pi_2),(0,\pi_2),(0,0)\}$.

Remark 3.3.3. Note that the asymptotic variance in Theorem 3.3.2 has the same form as Theorem 3.3.1, with $\tilde{Y}_g(z,h)$ replacing $\bar{Y}_g(z,h)$. Intuitively, $\tilde{Y}_g(z,h)$ is a demeaned and cluster size weighted version of $\bar{Y}_g(z,h)$. Therefore, similar arguments as those made in Remark 3.3.1 and 3.3.2 can be applied here as well. ■

Theorem 3.3.1 and 3.3.2 imply that covariate information is important to establish asymptotically exact inference for the four estimands of interest under covariate-adaptive randomization. In contrast, previous works that do not account for covariate information, such as Basse and Feller (2018) and Cruces et al. (2022), may result in conservative inference. Many empirical studies rely on statistical inference based on the regression in equation (3.3) with HC2 cluster-robust standard errors. While this procedure is also proposed in Basse and Feller (2018) and Cruces et al. (2022), the regression coefficients it produces generally do not provide consistent estimates for the estimands in Table 3.1. Instead, they converge to the primary and spillover effects weighted by the sample sizes of the clusters (see Bugni et al. (2022b)). However, if all units in each cluster are sampled ($N_g = M_g$) or the number of sampled units is proportional to cluster size ($M_g/N_g = c$ for $0 < c < 1$), this procedure yields consistent point estimates for size-weighted effects but may still be conservative (see Appendix C.4). Therefore, I aim to develop asymptotically exact methods based on my theoretical results.

To begin with, I introduce consistent variance estimators for the asymptotic variances from Theorem 3.3.1 and 3.3.2. A natural estimator of $V_1(z)$ may be constructed by replacing population quantities with their sample counterparts. For $z \in \{0,1\}$, Let

$$\begin{aligned}
\bar{Y}_{1,z} &= \frac{1}{G_1} \sum_{1 \le g \le G} \bar{Y}_g^z I\left\{H_g = \pi_2\right\} \\
\bar{Y}_{0,z} &= \frac{1}{G_0} \sum_{1 \le g \le G} \bar{Y}_g^z I\left\{H_g = 0\right\} \ , \\
\hat{\mu}_{1,z}(s) &= \frac{1}{G_1(s)} \sum_{1 \le g \le G} \bar{Y}_g^z I\left\{H_g = \pi_2, S_g = s\right\} \\
\hat{\mu}_{0,z}(s) &= \frac{1}{G_0(s)} \sum_{1 \le g \le G} \bar{Y}_g^z I\left\{H_g = 0, S_g = s\right\} \ ,
\end{aligned}$$

where $G_1(s) = |\{1 \le g \le G : H_g = \pi_2, S_g = s\}|$ and $G_0(s) = |\{1 \le g \le G : H_g = 0, S_g = s\}|$. With this notation, the following estimators can be defined:

$$\begin{aligned}
\hat{V}_1(z) = & \frac{1}{\pi_1} \left(\frac{1}{G_1} \sum_{1 \le g \le G} \left(\bar{Y}_g^z\right)^2 I\{H_g = \pi_2\} - \sum_{s \in \mathcal{S}} \frac{G(s)}{G} \hat{\mu}_{1,z}(s)^2 \right) \\
& + \frac{1}{1-\pi_1} \left(\frac{1}{G_0} \sum_{1 \le g \le G} \left(\bar{Y}_g^z\right)^2 I\{H_g = 0\} - \sum_{s \in \mathcal{S}} \frac{G(s)}{G} \hat{\mu}_{0,0}(s)^2 \right) \\
& + \sum_{s \in \mathcal{S}} \frac{G(s)}{G} \left(\left(\hat{\mu}_{1,z}(s) - \bar{Y}_{1,z}\right) - \left(\hat{\mu}_{0,0}(s) - \bar{Y}_{0,0}\right) \right)^2 \\
& + \sum_{s \in \mathcal{S}} \tau(s) \frac{G(s)}{G} \left(\frac{1}{\pi} \left(\hat{\mu}_{1,z}(s) - \bar{Y}_{1,z}\right) + \frac{1}{1-\pi} \left(\hat{\mu}_{0,0}(s) - \bar{Y}_{0,0}\right) \right)^2 \ .
\end{aligned} \tag{3.19}$$

The estimator for $V_2(z)$ follows the same approach as $\hat{V}_1(z)$, while additionally requires estimation for terms associated with $\tilde{Y}_g(z,h)$. Let $\tilde{Y}_g^z$ denote the observed adjusted outcome.

$$\tilde{Y}_g^z = \frac{N_g}{\frac{1}{G}\sum_{1 \le g \le G} N_g} \left(\bar{Y}_g^z - \frac{\frac{1}{G_g}\sum_{1 \le j \le G} \bar{Y}_j^z I\{H_g = H_j\} N_j}{\frac{1}{G}\sum_{1 \le j \le G} N_j} \right) \ ,$$

where $G_g = \sum_{1\leq j\leq G} I\{H_g = H_j\}$. For $z \in \{0,1\}$, Let

$$\tilde{\mu}_{1,z}(s) = \frac{1}{G_1(s)} \sum_{1\leq g\leq G} \tilde{Y}_g^z I\left\{H_g = \pi_2, S_g = s\right\} \ ,$$
$$\tilde{\mu}_{0,z}(s) = \frac{1}{G_0(s)} \sum_{1\leq g\leq G} \tilde{Y}_g^z I\left\{H_g = 0, S_g = s\right\} \ .$$

To estimate $V_2(z)$, I propose the exact same estimator as $\hat{V}_1(z)$ by simply replacing $\bar{Y}_g^z$ with $\tilde{Y}_g^z$. Thus, the following estimators can be defined:

$$\begin{aligned}\hat{V}_2(z) = & \frac{1}{\pi_1}\left(\frac{1}{G_1}\sum_{1\leq g\leq G}\left(\tilde{Y}_g^z\right)^2 I\{H_g = \pi_2\} - \sum_{s\in\mathcal{S}}\frac{G(s)}{G}\tilde{\mu}_{1,z}(s)^2\right) \\ & + \frac{1}{1-\pi_1}\left(\frac{1}{G_0}\sum_{1\leq g\leq G}\left(\tilde{Y}_g^z\right)^2 I\{H_g = 0\} - \sum_{s\in\mathcal{S}}\frac{G(s)}{G}\tilde{\mu}_{0,0}(s)^2\right) \\ & + \sum_{s\in\mathcal{S}}\frac{G(s)}{G}\left(\tilde{\mu}_{1,z}(s) - \tilde{\mu}_{0,0}(s)\right)^2 + \sum_{s\in\mathcal{S}}\tau(s)\frac{G(s)}{G}\left(\frac{1}{\pi}\tilde{\mu}_{1,z}(s) + \frac{1}{1-\pi}\tilde{\mu}_{0,0}(s)\right)^2 \ .\end{aligned} \tag{3.20}$$

Then, the following consistency result for variance estimators $\hat{V}_1(z)$ and $\hat{V}_2(z)$ can be obtained:

Theorem 3.3.3. *Under Assumption 3.2.1-3.2.2 and 3.3.1-3.3.2, as $n \to \infty$, $\hat{V}_1(z) \xrightarrow{P} V_1(z)$ and $\hat{V}_2(z) \xrightarrow{P} V_2(z)$ for $z \in \{0,1\}$.*

Based on Theorem 3.3.3, I propose the "adjusted" t-test with the aforementioned variance estimators as my method of inference throughout the rest of the paper. As an example, the "adjusted" t-test for equally-weighted primary effect, i.e. $H_0 : \theta_1^P(Q_G) = \theta_0$, is given by

$$\phi_G^{\text{adj}}(V^{(G)}) = I\left\{\left|\sqrt{n}\left(\hat{\theta}_1^P - \theta_0\right)/\hat{V}_1(1)\right| > z_{1-\frac{\alpha}{2}}\right\} \ , \tag{3.21}$$

where $z_{1-\frac{\alpha}{2}}$ represents $1-\frac{\alpha}{2}$ quantile of a standard normal random variable.

Note that the variance estimator $\hat{V}_1(z)$ (or $\hat{V}_2(z)$) depends on the assignment mechanism in the first stage through the strata indicator S_g, but not on the assignment mechanism in the second stage. This means that valid statistical inference based on $\phi_G^{\text{adj}}(V^{(G)})$ does not require knowledge of the assignment mechanism in the second stage. We can see this by observing that the first term in equations (3.8), which is the only term affected by the second-stage design, can be consistently estimated by the first term in equation (3.19).My approach leverages the cluster-level averaged outcomes and benefits from large samples of clusters, without explicitly modeling intra-cluster correlations as done in the previous literature (see, for example, Cruces et al., 2022).

3.4 Inference for Experiments with Small Strata

In this section, I study the asymptotic behavior of the estimators from Section 3.2.3 in two-stage stratified experiments with a large number of small strata. In particular, I consider size of strata to be fixed and assignment mechanism is a completely randomized design (also known as a permuted block design), independently for each stratum. This design is referred to as "matched tuples designs" by Bai et al. (2022b) (also studied as "local randomization" in Cytrynbaum (2023b)), and is commonly-used in the empirical literature (see for example Bold et al., 2018; Brown and Andrabi, 2020; de Mel et al., 2013; Fafchamps et al., 2014). I leverage the result from Bai (2022b) that any stratification is a convex combination of matched tuples designs, so fortunately this section does not lose too much generality by focusing the analysis of "small strata" on matched tuples designs.

To formalize such a design under the settings of two-stage experiments, consider n strata of size k (each stratum consisting of k clusters), formed by matching clusters according to $S_g = S(C_g, N_g)$, whose co-domain could potentially be continuous and multi-dimensional. Within each stratum, l clusters are randomly selected and assigned to the treatment group. Specifically, $G = nk$ and $\pi_1 = l/k$, where $0 < l < k$, and l and k are mutually prime. As an example, a matched-pairs design has $k = 2$, $l = 1$, $\pi_1 = 1/2$ and $G = 2n$.

Example 3.4.1. Duflo and Saez (2003) conducted such a small-strata experiment involving 330 university departments, each averaging 30 staff employees. In the first stage, these departments (clusters with an average size of 30) were grouped into triplets (small strata of size 3) based on their cluster-level covariates. Within each triplet, two departments were randomly chosen to be part of the treated group. In the second stage, individuals from these treated departments were randomly selected to receive treatments. ■

To start with, I impose the following assumption on Q_G in addition to Assumption 3.2.2:

Assumption 3.4.1. The distribution Q_G is such that

(a) $E[\bar{Y}_g^r(z,h)N_g^\ell | S_g = s]$ is Lipschitz in s for $(z,h) \in \{(0,0),(0,\pi_2),(1,\pi_2)\}$ and $r, \ell \in \{0,1,2\}$.

(b) For some $C < \infty$, $P\{E[N_g^2|S_g] \leq C\} = 1$

Assumption 3.4.1(a) is a smoothness requirement analogous to Assumption 3(ii) in Bai (2022b) ensuring that units within clusters which are "close" in terms of their baseline covariates are suitably comparable. Assumption 3.4.1(b) imposes an additional restriction on the distribution of cluster sizes beyond what is stated in Assumption 3.2.2(c).

Next, I describe my assumptions on the treatment assignment mechanism. In words, I consider a treatment assignment mechanism that first stratifies the experimental sample into n blocks of size k using $S^{(G)}$, and then assigns l clusters uniformly at random as treated clusters within each block. Formally, let

$$\lambda_j = \lambda_j(S^{(G)}) \subseteq \{1, \ldots, G\},\ 1 \leq j \leq n$$

denote n sets each consisting of k elements that form a partition of $\{1, \ldots, G\}$.

I assume treatment status is assigned as follows:

Assumption 3.4.2. Treatments are assigned so that $W^{(G)} \perp\!\!\!\perp H^{(G)} | S^{(G)}$ and, conditional on $S^{(G)}$,

$$\{(I\{H_i = \pi_2\}) : i \in \lambda_j) : 1 \leq j \leq n\}$$

are i.i.d. and each uniformly distributed over all permutations of $\left\{ z \in \{0,1\}^k : \sum_{z=i}^{k} z_i = l \right\}$.

Assumption 3.4.2 formally describes the assignment mechanism of a two-stage experiment with matched tuples in the first stage. Additionally, the second-stage design adheres to the specifications outlined in Assumption 3.3.2 from the previous section. Further, units in each pair are required to be "close" in terms of their stratification variable S_g in the following sense:

Assumption 3.4.3. The strata used in determining treatment status satisfy

$$\frac{1}{n} \sum_{1 \leq j \leq n} \max_{i,k \in \lambda_j} |S_i - S_k|^2 \xrightarrow{P} 0 \ .$$

The validity of the variance estimators relies on the following condition that the distances between units in adjacent blocks are considered "close" in relation to their baseline covariates:

Assumption 3.4.4. The strata used in determining treatment status satisfy

$$\frac{1}{n} \sum_{1 \leq j \leq \lfloor n/2 \rfloor} \max_{i \in \lambda_{2j-1}, k \in \lambda_{2j}} |S_i - S_k|^2 \xrightarrow{P} 0 \ .$$

Blocking algorithms which satisfy Assumption 3.4.2-3.4.4 have been thoroughly discussed in recent literature of matched pairs/tuples designs (see for example Bai et al., 2021b; Bai, 2022b; Bai et al., 2022b; Cytrynbaum, 2023b). For instance, when $\dim(S_g) = 1$ and clusters/units are matched into blocks by ordering them according to the values of S_g and grouping the adjacent clusters/units, Theorem 4.1 of Bai et al. (2021b) shows that 3.4.2-3.4.4 are satisfied as long as $E[S_g^2] < \infty$.

When $k = 2$ and $l = 1$, Assumptions 3.4.2-3.4.4 reduce to Assumptions 2.2-2.4 in Bai et al. (2021b) and thus it becomes a matched pairs design. In fact, my framework extends the generalized matched pair designs discussed in Section C.1 of Bai (2022b) to the setting of two-stage experiments. My assumptions on the assignment mechanism follow the matched tuples designs provided in Bai et al. (2022b) for experiments with multiple treatment arms, where the strata size equals the number of treatments. In contrast, this paper focuses on experiments with binary treatments but allow for a general treated fraction. Under these assumptions I obtain the following result:

Theorem 3.4.1. *Suppose Assumption 3.2.1 holds, Q_G satisfies Assumptions 3.2.2 and 3.4.1 and the treatment assignment mechanism satisfies Assumptions 3.3.2, 3.4.2-3.4.3. Then, as $n \to \infty$,*

$$\sqrt{G}\left(\hat{\theta}_1^P - \theta_1^P(Q_G)\right) \to \mathcal{N}(0, V_3(1)) \ , \tag{3.22}$$

$$\sqrt{G}\left(\hat{\theta}_1^S - \theta_1^S(Q_G)\right) \to \mathcal{N}(0, V_3(0)) \ , \tag{3.23}$$

$$\sqrt{G}\left(\hat{\theta}_2^P - \theta_2^P(Q_G)\right) \to \mathcal{N}(0, V_4(1)) \ , \tag{3.24}$$

$$\sqrt{G}\left(\hat{\theta}_2^S - \theta_2^S(Q_G)\right) \to \mathcal{N}(0, V_4(0)) \ , \tag{3.25}$$

where

$$\begin{aligned} V_3(z) = &\frac{1}{\pi_1} \operatorname{Var}\left[\bar{Y}_g(z, \pi_2)\right] + \frac{1}{1-\pi_1} \operatorname{Var}\left[\bar{Y}_g(0,0)\right] \\ &- \pi_1(1-\pi_1) E\left[\left(\frac{1}{\pi_1} m_{z,\pi_2}\left(S_g\right) + \frac{1}{1-\pi_1} m_{0,0}\left(S_g\right)\right)^2\right] \end{aligned} \tag{3.26}$$

and

$$\begin{aligned} V_4(z) = \frac{1}{\pi_1}\operatorname{Var}[\tilde{Y}_g(z,\pi_2)] + \frac{1}{1-\pi_1}\operatorname{Var}[\tilde{Y}_g(0,0)] \\ - \pi_1(1-\pi_1)E\left[\left(\frac{1}{\pi_1}E[\tilde{Y}_g(z,\pi_2) \mid S_g] + \frac{1}{1-\pi_1}E[\tilde{Y}_g(0,0) \mid S_g]\right)^2\right] \end{aligned} \tag{3.27}$$

with $m_{z,h}\left(C_g\right)$ being defined in (3.9), and $\tilde{Y}_g(z,h)$ being defined in (3.18).

Remark 3.4.1. Analogous to Remark 3.3.1, the asymptotic variance in Theorem 3.4.1 corresponds exactly to the asymptotic variance of the difference-in-means estimator for matched-pair experiments with individual-level "one-stage" assignment, as in Bai et al. (2021b) and Bai (2022b). Additionally, $V_4(z)$ has a similar form to the asymptotic variance in a cluster randomized trial with matched pairs, as derived in Bai et al. (2022a). In fact, when $\pi_1 = 1/2$ and $\pi_2 = 1$, my result collapses exactly to theirs. ■

Remark 3.4.2. Note that $V_3(z)$ and $V_4(z)$ have the same formula as the first three terms in $V_1(z)$ and $V_2(z)$. Thus, the way the first and second stage designs show up in the variance expression is similar to Remark 3.3.2. Importantly, the interpretations of S_g in Theorem 3.3.1 and 3.3.2 differ slightly from that in Theorem 3.4.1. In Theorem 3.3.1 and 3.3.2, S_g corresponds to strata indicators on a discrete set, while in Theorem 3.4.1, S_g is a random function with potentially continuous values on which clusters are matched. However, an interesting observation is that a matched tuples design based on a finite-valued S_g leads to the same asymptotic variance as a strong balanced "large strata" design based on the same S_g. Such a matched tuples design can be constructed by first stratifying S_g and then forming tuples within each stratum arbitrarily. In this case, Assumption 3.4.1-3.4.4 are trivially satisfied. ■

The widely used regression method with cluster-robust variance estimator is potentially conservative for matched tuples designs (see Appendix C.4). Therefore, I aim to develop asymptotically exact methods based on my theoretical results. First, I present variance estimators for $V_3(z)$ and $V_4(z)$. Unlike stratified block randomization, there are not enough observations for cluster-level averaged outcomes within each strata to construct variance estimators in a matched tuples design. Thus, I follow the construction of "pairs of pairs" in Bai et al. (2021b) and Bai et al. (2022b), and replace the individual outcomes with the averaged outcomes $\bar{Y}_g^z$ and adjusted averaged outcomes $\tilde{Y}_g$ respectively. Here, I present the construction of variance estimator $\hat{V}_3(z)$ for $V_3(z)$. Similarly, $\hat{V}_4(z)$ can be constructed by simply replacing $\bar{Y}_g^z$ with $\tilde{Y}_g^z$, and thus detalis are omitted. Let $\hat{\Gamma}_n^z(h) = \frac{1}{nk(h)} \sum_{1 \leq g \leq G} \bar{Y}_g^z I\{H_g = h\}$ where $k(h) = \sum_{i \in \lambda_j} I\{H_i = h\}$ denotes the number of units under assignment $H_i = h$ in the j-th strata. In the setup of binary treatment, it becomes that $k(\pi_2) = l$ and $k(0) = k - l$. Finally, my estimator for $V_3(z)$ is then given by

$$\hat{V}_3(z) = \frac{1}{\pi_1}\hat{\mathbb{V}}_{1,n}^z(\pi_2) + \frac{1}{1-\pi_1}\hat{\mathbb{V}}_{1,n}^z(0) + \hat{\mathbb{V}}_{2,n}^z(\pi_2,\pi_2) + \hat{\mathbb{V}}_{2,n}^z(0,0) - 2\hat{\mathbb{V}}_{2,n}^z(\pi_2,0) \tag{3.28}$$

with

$$\begin{aligned}
\hat{\mathbb{V}}_{1,n}^z(h) &= \hat{\mathbb{E}}\left[\operatorname{Var}\left[\bar{Y}_g(z,h) \mid S_g\right]\right] := (\hat{\sigma}_n^z(h))^2 - (\hat{\rho}_n^z(h,h) - (\hat{\Gamma}_n^z(h))^2) \\
\hat{\mathbb{V}}_{2,n}^z(h,h') &= \hat{\operatorname{Cov}}\left[E\left[\bar{Y}_g(z,h) \mid S_g\right], E\left[\bar{Y}_g(z,h') \mid S_g\right]\right] := \hat{\rho}_n^z(h,h') - \hat{\Gamma}_n^z(h)\hat{\Gamma}_n^z(h') ,
\end{aligned}$$

where

$$\begin{aligned}
\hat{\rho}_n^z(h,h) &:= \frac{2}{n}\sum_{1 \leq j \leq \lfloor n/2 \rfloor} \frac{1}{k^2(h)}\Big(\sum_{i \in \lambda_{2j-1}} \bar{Y}_i^z I\{H_i = h\}\Big)\Big(\sum_{i \in \lambda_{2j}} \bar{Y}_i^z I\{H_i = h\}\Big) \\
\hat{\rho}_n^z(\pi_2,0) &:= \frac{1}{n}\sum_{1 \leq j \leq n} \frac{1}{l(k-l)}\Big(\sum_{i \in \lambda_j} \bar{Y}_i^z I\{H_i = \pi_2\}\Big)\Big(\sum_{i \in \lambda_j} \bar{Y}_i^z I\{H_i = 0\}\Big) \\
(\hat{\sigma}_n^z(h))^2 &:= \frac{1}{nk(h)}\sum_{1 \leq g \leq G} (\bar{Y}_g^z - \hat{\Gamma}_n^z(h))^2 I\{H_g = h\} .
\end{aligned}$$

The subsequent analysis yields consistency results for the estimators $\hat{V}_3(z)$ and $\hat{V}_4(z)$:

Theorem 3.4.2. *Suppose Assumption 3.2.1 holds, Q_G satisfies Assumptions 3.2.2, 3.4.1, and the treatment assignment mechanism satisfies Assumption 3.3.2, 3.4.2-3.4.4. Then, as $n \to \infty$, $\hat{V}_3(z) \xrightarrow{P} V_3(z)$ and $\hat{V}_4(z) \xrightarrow{P} V_4(z)$ for $z \in \{0, 1\}$.*

Similarly to the designs of large strata in Section 3.3, the calculation of variance estimators and the validity of statistical inference based on Theorem 3.4.1 and 3.4.2 do not depend on the assignment mechanism in the second-stage. In other words, practitioners can conduct valid statistical inference using only the information from the first-stage design.

In Remark 3.4.2, it is noted that a matched tuples design based on a finite-valued S_g has the same asymptotic variance as a strongly balanced "large strata" design based on S_g. In such cases, $\hat{V}_1(z)$ and $\hat{V}_3(z)$ (or $\hat{V}_2(z)$ and $\hat{V}_4(z)$) are both consistent estimators for the same estimand. However, I recommend using $\hat{V}_1(z)$ in practice for "large strata" experiments, as it is expected to be more efficient than $\hat{V}_3(z)$. As pointed out by Athey and Imbens (2017b) and Bai et al. (2022b), introducing replicates for each treatment arm in a matched tuples design can improve the finite sample performance for the adjusted t-tests based on $\hat{V}_3(z)$.[7] This motivates the use of variance estimators based on "large tuples". To that extent, $\hat{V}_1(z)$, which takes advantage of all observations within a stratum at the same time, is preferable for experiments with large strata. In practice, the choice of variance estimators depends on the sizes of the strata. Specifically, $\hat{V}_1(z)$, whose consistency relies on large numbers of observations within each stratum, is suitable for experiments with large strata, while $\hat{V}_3(z)$ is suitable for experiments with small strata.[8] From this perspective, it is necessary to

[7] When there are duplicates, I no longer need to form "pairs of pairs" for variance estimation. Instead, I could replace $\hat{\rho}_n^z(h,h)$ by

$$\tilde{\rho}_n^z(h,h) = \frac{2}{n} \sum_{1 \le j \le \lfloor n/2 \rfloor} \frac{1}{k^2(h)} \Big(\sum_{i \in \lambda_j} \bar{Y}_i^z I\{H_i = h\} \Big) .$$

[8] In practice, it is often straightforward to distinguish between the two scenarios. Most experimental designs either involve stratification on a small number of categorical variables or matching units into groups with fewer than five units. In cases where it is difficult to decide, it is recommended to choose $\hat{V}_3(z)$ as a safe choice.

divide covariate-adaptive experiments into "large strata" and "small strata" and consider two separate asymptotic regimes. Furthermore, the "large strata" regime provides an analytical framework to study a broader range of experimental designs, including Efron's biased coin design.

3.5 Optimal Stratification for Two-stage Designs

In this section, I introduce two optimality results related to two-stage randomized experiments, as discussed in Sections 3.3 and 3.4. The first result provides insights into the optimal design for the initial stage, while the second addresses the optimal design for the second stage, taking into account additional assumptions about the assignment mechanism and covariance among unit outcomes within clusters. These findings indicate that particular matched tuples designs maximize statistical precision when estimating parameters outlined in Table 3.1.

First, I present a result that identifies the optimal functions for matching in the first-stage matched tuples designs, targeting various parameters of interest.

Theorem 3.5.1. *$V_3(z)$ is minimized when $S_g = E\left[\frac{\bar{Y}_g(z,\pi_2)}{\pi_1} + \frac{\bar{Y}_g(0,0)}{1-\pi_1} \mid C_g, N_g\right]$. Meanwhile, $V_4(z)$ is minimized when $S_g = E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1} + \frac{\tilde{Y}_g(0,0)}{1-\pi_1} \mid C_g, N_g\right]$.*

A direct implication of Theorem 3.5.1 is that it characterizes the optimal functions to match on within the class of matched-tuples designs. Such functions are referred as "index function" in Bai (2022b). With further reasoning, it can also be concluded that the optimal matched tuples design is also asymptotically optimal among all stratified designs described in Section 3.3. To see that, in Section 3.3, strata are constructed from a discrete random variable $S_g = S(C_g, N_g)$, where function S maps to a discrete set $\mathcal{S}$. As noted in Remark 3.4.2, given such a function S, it is possible to construct a matched tuples design that matches on $S(C_g, N_g)$, and thus has a weakly smaller asymptotic variance across all parameters of interest, i.e. $V_1(z) \geq V_3(z)$ and $V_2(z) \geq V_4(z)$ for $z \in \{0,1\}$, where equality holds when designs achieve "strong balance" ($\tau(s) = 0$ for all $s \in \mathcal{S}$). Therefore, any assignment mechanism described in Section 3.3 leads to asymptotic variances that are weakly larger than those of matched tuples designs implied by Theorem 3.5.1.

The subsequent section examines the optimality of matched tuples designs in the second stage of the experiment. The analysis specifically focuses on stratified block randomization, as formalized in the following assumptions.

Assumption 3.5.1. For $1 \leq g \leq G$, units within a given stratum, denoted by $\lambda_b = \{i \in \mathcal{M}_g : B_i = b\}$ for $b \in \mathcal{B}$, are assigned with treatment $(Z_{i,g}(\pi_2) : i \in \lambda_b)$ that is uniformly distributed over $\{z \in \{0,1\}^{|\lambda_b|} : \sum_{j \in \lambda_b} z_j = \lfloor \pi_2 |\lambda_b| \rfloor\}$ and i.i.d across $b \in \mathcal{B}$.

Additionally, I assume that the covariance of outcomes between any pairs of units within a cluster is homogeneous. In other words, the covariance does not depend on the individual-level covariates of units in the same cluster. Formally, the assumption is stated as follows:

Assumption 3.5.2. For $z \in \{0,1\}$, $1 \leq i \neq j \leq N_g$,

$$\text{Cov}\left[Y_{i,g}(z,\pi_2), Y_{j,g}(z,\pi_2) \mid (X_{i,g} : 1 \leq i \leq N_g)\right] = \text{Cov}\left[Y_{i,g}(z,\pi_2), Y_{j,g}(z,\pi_2)\right] \ . \quad (3.29)$$

Assumption 3.5.2 is a weaker assumption than assuming that outcomes of units are independent and identically distributed (i.i.d) within a cluster, as it only requires conditional independence between individual covariates and the covariance of outcomes. It is analogous to the standard homoscedasticity assumption, which assumes constant variance of errors in a regression model, except that it is a statement about covariance instead of variance. Under these two additional assumptions I obtain the following optimality result:

Theorem 3.5.2. *Under Assumption 3.3.2, 3.5.1 and 3.5.2, $V_a(z)$ is minimized when the second-stage design is a matched tuples design that matches on $E\left[Y_{i,g}(z,\pi_2) \mid (X_{i,g} : 1 \leq i \leq N_g)\right]$ for $z \in \{0,1\}$ and $a \in \{1,2,3,4\}$.*

Though practitioners may not have knowledge of the index functions in Theorem 3.5.1 and 3.5.2, optimal stratification can be determined in some special cases. For instance, in experiments where the first-stage design uses only a univariate covariate C_g (see, e.g., Ichino and Schündeln, 2012), and practitioners expect a monotonic relationship between S_g and C_g, the optimal stratification is to order the units by C_g and group adjacent units. Similar results apply to the second-stage design. In more general cases where monotonicity does not hold or the baseline cluster-level covariate is multivariate, Bai (2022b) suggests matching on estimated index functions using pilot data, when available. If optimal stratification is infeasible, and pilot data is unavailable, a suitable matching algorithm (see, e.g., Bai et al., 2021b; Cytrynbaum, 2023b) that directly matches on vectors of covariates can be asymptotically as efficient if the sample size is sufficiently large. In cases where the sample size is not sufficiently large, Bai (2022b) and Bruhn and McKenzie (2009b) suggest matching on the baseline outcome, when available. If none of the aforementioned options is available, matching in a sub-optimal way can still be effective, as both Bai (2022b) and simulation results from Section 3.6 demonstrate that matching units sub-optimally can be more effective than completely randomized designs or some sub-optimal stratification design with large strata.

3.6 Simulations

In this section, I illustrate the results presented in Section 3.3-3.4 with a simulation study. To begin with, potential outcomes are generated according to the equation:

$$Y_{i,g}(z,h) = \mu_{z,h} + \alpha_{z,h} X_{1,i,g}/(X_{2,i,g}+0.1) + \beta_{z,h}\left(C_g - \frac{1}{2}\right) + \gamma\left(N_g - 100\right) + \sigma(C_g, N_g)\epsilon_{i,g}\ ,$$

for $(z,h) \in \{(0,0),(0,\pi_2),(1,\pi_2)\}$, where

- C_g, N_g are i.i.d with $C_g \sim \text{Unif}[0,1]$, and $N_g \sim \text{Unif}\{50,\dots,150\}$, which are mutually independent.

- $X_{1,i,g} = N_g u_{i,g}/100$, where $u_{i,g}$ are i.i.d $N(0, 0.1)$ across i, j. $X_{2,i,g}$ are i.i.d Unif$[0, 1]$ across i, g.

- $\mu_{1,\pi_2} = \mu_{0,\pi_2} + \tau = \mu_{0,0} + \tau + \omega$, i.e. primary and spillover effects are additive and homogeneous.

- $\sigma(C_g, N_g) = C_g(N_g - 100)/100$ and $\epsilon_{i,g} \sim N(0, 10)$, which satisfies Assumption 3.5.2.

All simulations are performed with a sample of 200 clusters, in which all units are sampled, i.e. $N_g = M_g$.

3.6.1 MSE Properties

This section examines the performance of optimal matched tuples designs and several other designs via comparison of their MSEs (Mean Squared Errors). For simplicity, the parameters are given as follows: $\alpha_{z,h} = \beta_{z,h} = 1, \gamma = 1/100$ for all $(z, h) \in \{(0, 0), (0, \pi_2), (1, \pi_2)\}$. This model configuration is referred to as "homogeneous model" since treatment effects are fully captured by $\mu_{z,h}$ and thus are homogeneously additive in this setting. A more complicated "heterogeneous model" will be introduced later. According to Theorem 3.5.1, the optimal matched tuple designs for equally-weighted and size-weighted effects in the first stage are

$$E\left[\frac{\bar{Y}_g(1,\pi_2)}{\pi_1} + \frac{\bar{Y}_g(0,0)}{1-\pi_1} \mid C_g, N_g\right] \propto C_g + N_g/100 \ , \tag{3.30}$$

$$E\left[\frac{\tilde{Y}_g(1,\pi_2)}{\pi_1} + \frac{\tilde{Y}_g(0,0)}{1-\pi_1} \mid C_g, N_g\right] \propto N_g(C_g + N_g/100) - \frac{25}{3}N_g \ . \tag{3.31}$$

In the second stage, the optimal matched tuples design matches on $X_{1,i,g}/(X_{2,i,g} + 0.1)$ according to Theorem 3.5.2. This section considers the following experimental designs for both stages:

First-stage	Parameter	**C**	**S-2**	**S-4**	**S-4O**	**MT-A**	**MT-B**	**MT-C**
C	θ_1^P	1.0000	0.9601	0.9270	0.9235	0.9720	0.9323	**0.9106**
	θ_2^P	1.0000	0.9803	0.9404	**0.9263**	0.9939	0.9573	0.9560
	θ_1^S	1.0000	0.9625	0.9187	0.9197	0.9649	0.9410	**0.9093**
	θ_2^S	1.0000	0.9921	0.9432	**0.9209**	0.9875	0.9709	0.9596
S-2	θ_1^P	0.8437	0.7866	0.7859	**0.7629**	0.8473	0.7981	0.7957
	θ_2^P	0.8227	0.7601	0.7877	**0.7440**	0.8361	0.7880	0.7672
	θ_1^S	0.8396	0.7913	0.7754	**0.7534**	0.8473	0.8052	0.7943
	θ_2^S	0.8244	0.7790	0.7806	**0.7438**	0.8456	0.7904	0.7693
S-4	θ_1^P	0.7772	0.8084	0.7730	0.7835	0.7603	**0.7216**	0.7262
	θ_2^P	0.7759	0.7757	0.7330	0.7473	0.7114	**0.6909**	0.7024
	θ_1^S	0.7711	0.8053	0.7656	0.7749	0.7556	0.7357	**0.7283**
	θ_2^S	0.7773	0.7848	0.7330	0.7482	0.7204	0.7100	**0.7091**
S-4O	θ_1^P	0.2104	0.2102	0.2026	**0.2010**	0.2172	0.2115	0.2035
	θ_2^P	0.2418	0.2428	0.2371	0.2285	0.2339	0.2494	**0.2241**
	θ_1^S	0.2081	0.2136	0.2028	**0.2002**	0.2158	0.2221	0.2004
	θ_2^S	0.2367	0.2489	0.2418	0.2254	0.2396	0.2606	**0.2226**
MT-A	θ_1^P	0.7683	0.8172	0.7573	0.7401	0.7347	0.7744	**0.7097**
	θ_2^P	0.7555	0.7693	0.7202	**0.6726**	0.7159	0.7665	0.6769
	θ_1^S	0.7592	0.8157	0.7573	0.7277	0.7310	0.7882	**0.7035**
	θ_2^S	0.7537	0.7763	0.7221	**0.6644**	0.7123	0.7847	0.6771
MT-B	θ_1^P	0.2935	0.2806	**0.2719**	0.2970	0.2912	0.2847	0.2797
	θ_2^P	0.4175	0.4013	**0.3802**	0.4120	0.4134	0.3953	0.3880
	θ_1^S	0.2866	0.2935	**0.2661**	0.2941	0.2811	0.2810	0.2746
	θ_2^S	0.4143	0.4181	**0.3786**	0.4106	0.4020	0.3934	0.3841
MT-C	θ_1^P	0.1160	0.1140	**0.1047**	0.1125	0.1095	0.1149	0.1069
	θ_2^P	0.0921	0.0873	0.0818	0.0893	0.0842	0.0874	**0.0755**
	θ_1^S	0.1221	0.1183	0.1143	0.1126	0.1076	0.1193	**0.1045**
	θ_2^S	0.0997	0.0930	0.0908	0.0891	0.0829	0.0914	**0.0757**

Table 3.2: Ratio of MSE under all designs against those under complete randomization in both stages

1. **(C)** $(H_g : 1 \leq g \leq G)$ is drawn from a completely randomized design (also known as permuted block design), i.e. uniformly from the assignment space that $\pi_1 G$ (or $\pi_2 N_g$ in the second stage) number of clusters/units get treated.

2. **(S-2)** A stratified design, where the experimental sample is divided into two strata using the midpoint of covariate C_g (or $X_{1,i,g}$ in the second stage) as the cutoff. In each stratum, treatment is assigned as in **C**.

3. **(S-4)** As in **(S-2)**, but with four strata.

4. **(S-4O)** The "optimal" stratification with four strata. Clusters/units are divided into strata using quartiles of (3.30) and (3.31) for equally- and size-weighted estimands respectively (or $X_{1,i,g}/(X_{2,i,g} + 0.1)$ in the second stage).

5. **(MT-A)** Matched tuples design where units are ordered according to C_g (or $X_{1,i,g}$ in the second stage).

6. **(MT1-B)** Matched tuples design where units are ordered according to cluster size N_g (or $X_{2,i,g}$ in the second stage).

7. **(MT-C)** The optimal matched tuples design where units are ordered according to (3.30) and (3.31) for equally- and size-weighted estimands respectively (or $X_{1,i,g}/(X_{2,i,g}+ 0.1)$ in the second stage).

Table 3.2 shows the ratio of the MSE of each design relative to the MSE of the design with completely randomized assignments (**C**) in both stages, computed across 1000 Monte Carlo iterations. The rows indicate first-stage designs, and columns indicate second-stage designs. The lowest values in each row are marked in bold. In all designs, treatment effects are set to zero by assigning $\mu_{z,h} = 0$ for all $(z, h) \in (0, 0), (0, \pi_2), (1, \pi_2)$, and the treated fraction is set to 1/2 in both stages. As expected from Theorem 3.5.1 and 3.5.2, the matched-tuples design with complete matching (**MT-C**) outperforms the other designs in the first stage for all parameters of interest while remaining optimal in the second stage for many cases. However, it is noticeable that the assignment mechanism in the first stage has a greater effect on statistical precision than the second stage.

		$H_0 : \tau = \omega = 0$						$H_1 : \tau = \omega = 0.05$					
First-stage	Parameter	**S-2**	**S-4**	**S-4O**	**MT-A**	**MT-B**	**MT-C**	**S-2**	**S-4**	**S-4O**	**MT-A**	**MT-B**	**MT-C**
S-2	θ_1^P	0.044	0.066	0.063	0.044	0.050	0.050	0.222	0.244	0.244	0.248	0.262	0.258
	θ_2^P	0.045	0.062	0.059	0.049	0.062	0.058	0.224	0.226	0.229	0.239	0.256	0.262
	θ_1^S	0.046	0.061	0.065	0.043	0.052	0.050	0.084	0.100	0.102	0.101	0.095	0.098
	θ_2^S	0.046	0.066	0.066	0.046	0.056	0.061	0.087	0.101	0.091	0.094	0.101	0.094
S-4	θ_1^P	0.050	0.048	0.060	0.058	0.036	0.051	0.241	0.267	0.243	0.275	0.245	0.276
	θ_2^P	0.056	0.055	0.062	0.051	0.037	0.056	0.230	0.261	0.241	0.284	0.250	0.270
	θ_1^S	0.054	0.053	0.062	0.056	0.037	0.048	0.096	0.119	0.105	0.130	0.109	0.112
	θ_2^S	0.058	0.055	0.061	0.054	0.033	0.056	0.087	0.121	0.096	0.127	0.107	0.110
S-4O	θ_1^P	0.048	0.051	0.052	0.058	0.054	0.066	0.692	0.729	0.691	0.708	0.685	0.716
	θ_2^P	0.048	0.059	0.062	0.058	0.055	0.060	0.608	0.629	0.588	0.630	0.582	0.639
	θ_1^S	0.048	0.055	0.047	0.057	0.054	0.057	0.220	0.268	0.247	0.282	0.222	0.241
	θ_2^S	0.052	0.060	0.054	0.055	0.052	0.058	0.220	0.228	0.214	0.246	0.192	0.208
MT-A	θ_1^P	0.060	0.049	0.044	0.050	0.044	0.060	0.270	0.271	0.260	0.252	0.240	0.256
	θ_2^P	0.058	0.049	0.041	0.048	0.050	0.056	0.254	0.260	0.268	0.237	0.236	0.259
	θ_1^S	0.055	0.042	0.040	0.050	0.052	0.058	0.109	0.101	0.100	0.101	0.105	0.106
	θ_2^S	0.055	0.052	0.049	0.046	0.051	0.052	0.115	0.100	0.097	0.096	0.100	0.105
MT-B	θ_1^P	0.044	0.053	0.057	0.031	0.043	0.051	0.565	0.582	0.586	0.553	0.530	0.586
	θ_2^P	0.053	0.047	0.058	0.041	0.045	0.057	0.402	0.419	0.444	0.403	0.378	0.431
	θ_1^S	0.049	0.045	0.052	0.035	0.051	0.052	0.197	0.203	0.216	0.174	0.184	0.198
	θ_2^S	0.053	0.046	0.057	0.038	0.050	0.059	0.148	0.158	0.180	0.131	0.135	0.148
MT-C	θ_1^P	0.058	0.056	0.061	0.044	0.053	0.043	0.920	0.939	0.917	0.917	0.919	0.933
	θ_2^P	0.057	0.045	0.058	0.041	0.052	0.051	0.955	0.975	0.955	0.950	0.941	0.968
	θ_1^S	0.074	0.058	0.059	0.044	0.042	0.044	0.399	0.429	0.427	0.416	0.400	0.411
	θ_2^S	0.058	0.052	0.062	0.034	0.050	0.050	0.430	0.465	0.471	0.472	0.444	0.504

Table 3.3: Rejection probabilities under the null and alternative hypothesis

3.6.2 Inference

In this section, the focus shifts from optimality to studying the finite sample properties of different tests for the following null hypotheses of interest:

$$H_0^{P,1} : \theta_1^P(Q_G) = 0, \quad H_0^{P,2} : \theta_2^P(Q_G) = 0, \quad H_0^{S,1} : \theta_1^S(Q_G) = 0, \quad H_0^{S,2} : \theta_2^S(Q_G) = 0 \ , \tag{3.32}$$

against the alternative hypotheses:

$$H_1^{P,1}: \theta_1^P(Q_G) = \tau+\omega, \quad H_1^{P,2}: \theta_2^P(Q_G) = \tau+\omega, \quad H_1^{S,1}: \theta_1^S(Q_G) = \omega, \quad H_1^{S,2}: \theta_2^S(Q_G) = \omega\,. \tag{3.33}$$

In Table 3.3, the six assignment mechanisms with covariate-adaptive randomization (Design 2-7 in Section 3.6.1) for the first and second stages are considered, resulting in a total of 36 different designs. Hypothesis tests are performed at a significance level of 0.05, and rejection probabilities under the null and alternative hypotheses are computed from 1000 Monte Carlo iterations in each case. Tests are constructed as "adjusted t-tests" using the asymptotic results from Theorem 3.3.1-3.4.2. For stratified designs in the first stage (**S-2**, **S-4** and **S-4O**), tests for equally- and size-weighted effects are performed using my variance estimators $\hat{V}_1(z)$ and $\hat{V}_2(z)$. For matched tuples designs in the first stage (**MT-A**, **MT-B** and **MT-C**), tests for equally- and size-weighted effects are performed using the variance estimators $\hat{V}_3(z)$ and $\hat{V}_4(z)$. The results show that the rejection probabilities are universally around 0.05 under the null hypothesis, which verifies the validity of tests based on my asymptotic results across all the designs. Under the alternative hypotheses, the rejection probabilities vary substantially across the first-stage designs while remaining relatively stable across the second-stage designs. **MT-C** stands out as the most powerful design for the first-stage. These findings are consistent with previous section.

Next, the validity of commonly used regression-based inference methods in the empirical literature is tested. These methods are tested under both the "homogeneous model" from the previous simulation study in Section 3.6.1 and a "heterogeneous model" in which two parameters are modified as follows: $\alpha_{1,\pi_2} = \beta_{1,\pi_2} = 2$, $\alpha_{0,\pi_2} = \beta_{0,\pi_2} = 0.5$, and $\alpha_{0,0} = \beta_{0,0} = 1$. The key difference between the two models is whether the conditional expectations of potential outcomes are identical or different across different exposures (z, h). Four commonly used regression methods are considered in this study:

Model	Inference Method	Effect	**S-4O**	**S-4O**	**S-4O**	**MT-C**	**MT-C**	**MT-C**
			C	**S-4O**	**MT-C**	**C**	**S-4O**	**MT-C**
Homogeneous	OLS robust	Primary	0.184	0.194	0.156	0.062	0.086	0.049
	(standard t-test)	Spillover	0.184	0.167	0.159	0.077	0.048	0.048
	OLS cluster	Primary	0.000	0.000	0.000	0.000	0.000	0.000
	(clustered t-test)	Spillover	0.000	0.000	0.000	0.000	0.000	0.000
	OLS with group	Primary	0.209	0.196	0.179	0.100	0.106	0.077
	fixed effects (robust)	Spillover	0.201	0.184	0.177	0.113	0.100	0.075
	OLS with group	Primary	0.028	0.027	0.029	0.068	0.085	0.071
	fixed effects (clustered)	Spillover	0.036	0.027	0.026	0.064	0.062	0.069
Heterogeneous	OLS robust	Primary	0.118	0.118	0.175	0.061	0.048	0.080
	(standard t-test)	Spillover	0.225	0.213	0.162	0.135	0.144	0.069
	OLS cluster	Primary	0.000	0.001	0.000	0.000	0.000	0.000
	(clustered t-test)	Spillover	0.002	0.000	0.000	0.000	0.000	0.000
	OLS with group	Primary	0.118	0.115	0.172	0.079	0.057	0.125
	fixed effects (robust)	Spillover	0.250	0.253	0.166	0.273	0.265	0.150
	OLS with group	Primary	0.024	0.015	0.023	0.056	0.051	0.047
	fixed effects (clustered)	Spillover	0.027	0.018	0.025	0.045	0.071	0.061

Table 3.4: Rejection probabilities of various inference methods under the null hypothesis

1. OLS robust: regress $Y_{i,g}$ on a constant, individual-level treatment indicator $Z_{i,g}$ and the indicator for untreated units in treated clusters $L_{i,g}$. Tests for primary and spillover effects are performed using standard t-tests under robust standard errors to heteroskedasticity.

2. OLS cluster: run the same regression as "OLS robust" but perform t-tests with clustered standard errors.

3. OLS with group fixed effects (robust): regress $Y_{i,g}$ on a constant, $Z_{i,g}$, $L_{i,g}$ and fixed effects for strata or tuples S_g. Tests are performed using standard t-tests under robust standard errors to heteroskedasticity.

4. OLS with group fixed effects (clustered): run the same regression as "OLS with group fixed effects (robust)" but perform t-tests with clustered standard errors.

Note that due to full sampling, i.e. $N_g = M_g$, regressions without fixed effects ("OLS robust" and "OLS cluster") output the same estimators as the size-weighted estimators $\hat{\theta}_2^P$ and $\hat{\theta}_2^S$. Most of the previous empirical analysis on covariate-adaptive two-stage experiments report cluster-robust standard errors in their main results, which could either be "OLS cluster" (see for example Duflo and Saez, 2003) or "OLS with group fixed effects (clustered)" (see for example Ichino and Schündeln, 2012). For brevity, Table 3.4 includes only six designs: those with either **S-4O** or **MT-C** in the first stage, and **C**, **S-4O**, or **MT-C** in the second stage. The table reveals that test results can be either conservative or invalid across different regression methods and designs. For stratified designs in the first stage, methods based on "robust" standard errors tend to over-reject, while methods based on "clustered" standard errors tend to under-reject. For matched tuples designs, "OLS cluster" is conservative, and the remaining methods could be invalid as they may over-reject the null hypothesis under some model specifications and parameters of interest. Similar results can also be found in the previous literature on covariate-adaptive randomization. For example, Bai et al. (2022b) demonstrated that inferences based on OLS regressions with strata fixed effects could be invalid. On the other hand, de Chaisemartin and Ramirez-Cuellar (2022b) documented that in cluster randomized experiments, t-test based on clustered standard errors tend to over-reject the null hypothesis when strata fixed effects are included, and under-reject otherwise. Therefore, it can be concluded that, with the exception of "OLS cluster" being conservative, the other three inference methods based on regression are generally invalid.

3.7 Empirical Application

In this section, the inference methods introduced in Section 3.3 are illustrated using data collected in Foos and de Rooij (2017). The experiment conducted by Foos and de Rooij (2017) is a randomly assigned spillover experiment in the United Kingdom designed to identify social influence within heterogeneous and homogeneous partisan households. The study first stratified 5190 two-voter households into three blocks based on the latest recorded party preference of the experimental subject[9]: "Labour" supporter,"rival party" supporter and those who were "unattached" to a party. Then experimental subjects or equivalently their households were randomly assigned to three groups: high partisan intensity treatment, low partisan intensity and control. Experimental subjests allocated to treatment groups were called by telephone and encouraged to vote in the PCC election on November 15, 2012. The "high partisan intensity" was formulated in a strongly partisan tone, explicitly mentioning the Labour Party and policies multiple time, while the "low partisan intensity" treatment message avoided all statements about party competition.

In the original analysis of Foos and de Rooij (2017), their main focus was on analyzing treatment effects conditional on a wide range of pre-treatment covariates. That said, in the final column of Table 1 in Foos and de Rooij (2017), they report estimators for (unconditional) primary and spillover effects, which are based on calculations of averages over separate experimental subjects and unassigned subjects. In contrast, my estimators do not distinguish experimental subjects from unassigned subjects and take averages solely based on treatment or spillover status. Another difference in my analysis is that estimators are calculated by pooling the two treatment arms, i.e. high and low partisan intensity, to maintain consistency with the setup of the paper. In contrast, Foos and de Rooij (2017) provide separate estimates for each treatment arm.

Table 3.5 compares point estimates of treatment effect on turnout percentage and confidence intervals obtained from the four regression methods listed in Section 3.6.2 with those based on my theoretical results, namely "adjusted t-test". Since cluster (household) size is fixed, equally-weighted and size-weighted estimators and estimands collapse into one. More-

[9] Before assigning treatments, the researchers randomly selected one individual per household to potentially receive treatments, whom they mark as "experimental subjects".

Table 3.5: Point estimates and confidence intervals for testing the primary and spillover effects

	adjusted t-test	OLS robust	OLS cluster	OLS fe robust	OLS fe cluster
Primary	3.0488	3.0488	3.0488	2.9971	2.9971
	[0.8339, 5.2638]	[0.9962, 5.1014]	[0.8103, 5.2874]	[0.9633, 5.0308]	[0.7812, 5.2129]
Spillover	4.5930	4.5930	4.5930	4.5413	4.5413
	[2.3430, 6.8431]	[2.5046, 6.6815]	[2.3216, 6.8645]	[2.4694, 6.6132]	[2.2904, 6.7922]

Note: The original paper did not mention the target treated fraction π_1. I decide to use the empirical treated fraction, $1/G\sum_{1\leq g\leq G} I\{H_g = \pi_2\}$, to calculate the variance estimators.

over, full sampling ($N_g = M_g = 2$) makes the point estimates of "adjusted t-test" and "OLS robust/cluster" equivalent. In the simulation study, it is found that ‘OLS robust" and "OLS fe robust" tend to over-reject the null hypothesis, which is consistent with the empirical results in Table 3.5 that they both have narrower confidence interval than the "adjusted t-test". Furthermore, "OLS cluster" and "OLS fe cluster" are shown to be conservative in the simulation study, and accordingly, they both have wider confidence intervals than the "adjusted t-test" in Table 3.5. Therefore, the empirical findings are consistent with the simulation study in Table 3.4.

3.8 Recommendations for Empirical Practice

Based on the theoretical results and the supporting simulation study, I conclude with the following recommendations for empirical practice, particularly in conducting inference about the parameters of interest, as listed in Table 3.1. In scenarios where the size of strata is considerably large, such as more than 50 clusters as exemplified in simulation **S-4**, we advise practitioners to utilize $\hat{V}_1(1)$ and $\hat{V}_1(0)$, as defined in (3.19), for estimating the equally-weighted primary effect θ_1^P and the spillover effect θ_1^S. Similarly, $\hat{V}_2(1)$ and $\hat{V}_2(0)$, as detailed in (3.20), should be employed for the size-weighted primary effect θ_2^P and the spillover effect θ_2^S. However, when it is unclear whether the strata size is sufficiently large, or more commonly, when the experimental design involves a matched-tuples design with only one or two observations per treatment arm, we recommend the application of $\hat{V}_3(1), V_3(0)$ and $\hat{V}_4(1), \hat{V}_4(0)$ as indicated in (3.28) for the corresponding equally-weighted and size-weighted effects.

The results of this study have shown that tests based on the regression specified in equation (3.3) with HC2 cluster-robust standard errors are valid but potentially conservative, which would result in a loss of power relative to our proposed test. Further, it's critical to note that regressions using strata fixed effects or heteroskedasticity-robust standard errors have generally been found invalid in the simulation study.

Based on the optimality results and following earlier studies (Bai (2022b), Cytrynbaum (2023b), Bruhn and McKenzie (2009b)), I recommend matching clusters and units on estimated index functions as outlined in Theorem 3.5.1 and 3.5.2 when data from large pilots are available. In cases with limited or no pilot data, alternatives include matching on baseline outcomes and adopting suitable matching algorithms (see, e.g., Bai et al., 2021b; Cytrynbaum, 2023b) when dealing with multiple covariates.

APPENDIX A

APPENDIX FOR CHAPTER 1

A.1 Additional Details

A.1.1 Details for Section 1.5

Proposition A.1.1. *Consider the setting with three treatment statuses* $\{1,2,3\}$*, where* 1 *corresponds to being untreated and* 2 *and* 3 *correspond to two treatments. In a matched quadruplets design where each quadruplet has two untreated units and one unit for each treatment, the test introduced in Section 1.3.1 with* $\mathcal{D}' = \{1,2,3,4\}$ *and*

$$\nu = \begin{pmatrix} -1/2 & -1/2 & 1 & 0 \\ -1/2 & -1/2 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

is valid for testing (1.12)–(1.13) *at level* $\alpha \in (0,1)$*.*

Proof of Proposition A.1.1. Consider a design of matched-quadruplets with two treatments $d = 2,3$ and two controls $d = 1$, i.e a quadruplet consisting of $(1,1,2,3)$. The difference-in-mean estimator for the effect of the first treatment $d = 2$ is

$$\hat{\Delta}_2 = \frac{1}{n}\sum_{i=1}^{4n} I\{D_i = 2\}Y_i - \frac{1}{2n}\sum_{i=1}^{4n} I\{D_i = 1\}Y_i$$

Note that

$$\sqrt{n}\left(\hat{\Delta}_2 - \Delta_2(Q)\right) = A_{n,2} + C_{n,2} - \left(A_{n,1} + C_{n,1}\right) ,$$

where

$$A_{n,2} = \frac{1}{\sqrt{n}}\sum_{1\le i\le 4n} I\{D_i = 2\}(Y_i(2) - E[Y_i(2)|X^{(n)}, D^{(n)}])$$

$$C_{n,2} = \frac{1}{\sqrt{n}}\sum_{1\le i\le 4n} I\{D_i = 2\}(E[Y_i(2)|X^{(n)}, D^{n}] - E[Y_i(2)]) .$$

and

$$\begin{aligned} A_{n,1} &= \frac{1}{2\sqrt{n}} \sum_{1 \le i \le 4n} I\{D_i = 1\}(Y_i(1) - E[Y_i(1)|X^{(n)}, D^{(n)}]) \\ C_{n,1} &= \frac{1}{2\sqrt{n}} \sum_{1 \le i \le 4n} I\{D_i = 1\}(E[Y_i(1)|X^{(n)}, D^n] - E[Y_i(1)]) \ . \end{aligned}$$

Let I_j denote the set of indices for the two untreated units in the j-th tuple. Note

$$\begin{aligned} &\operatorname{Var}[A_{n,1}|X^{(n)}, D^{(n)}] \\ &= \frac{1}{2 \cdot 2n} \sum_{1 \le i \le 4n} I\{D_i = 1\} \operatorname{Var}[Y_i(1)|X_i] \\ &= \frac{1}{2 \cdot 4n} \sum_{1 \le i \le 4n} \operatorname{Var}[Y_i(1)|X_i] - \frac{1}{8n} \sum_{1 \le j \le n} \frac{1}{2} \sum_{i_j \in I_j} \sum_{k \in \lambda_j : k \notin I_j} (\operatorname{Var}[Y_k(1)|X_k] - \operatorname{Var}[Y_{i_j}|X_{i_j}]) \end{aligned}$$

It follows from similar arguments as in the proof of Theorem 1.3.1 that the second term goes to zero. Therefore,

$$\operatorname{Var}[A_{n,1}|X^{(n)}, D^{(n)}] \xrightarrow{P} \frac{1}{2} E[\operatorname{Var}[Y_i(1)|X_i]] \ .$$

It therefore follows from Lemma S.1.2 of Bai et al. (2021a) that

$$\gamma\left(\left((A_{n,2}, A_{n,1})' \, |X^{(n)}, D^{(n)}\right), N\left(0, \begin{bmatrix} E[\operatorname{Var}[Y_i(2)|X_i]] & 0 \\ 0 & \frac{1}{2} E[\operatorname{Var}[Y_i(1)|X_i]] \end{bmatrix}\right)\right) \xrightarrow{P} 0 \ ,$$

where γ is any metric that metrizes weak convergence.

Next, note

$$\begin{aligned} E[C_{n,2}|X^{(n)}] &= \frac{1}{\sqrt{n}} \sum_{1 \le i \le 4n} \frac{1}{4}(E[Y_i(2)|X^{(n)}] - E[Y_i(2)]) = \frac{1}{4\sqrt{n}} \sum_{1 \le i \le 4n} (\Gamma_2(X_i) - \Gamma_2) \\ E[C_{n,1}|X^{(n)}] &= \frac{1}{2\sqrt{n}} \sum_{1 \le i \le 4n} \frac{1}{2}(E[Y_i(1)|X^{(n)}] - E[Y_i(1)]) = \frac{1}{4\sqrt{n}} \sum_{1 \le i \le 4n} (\Gamma_1(X_i) - \Gamma_1) \ . \end{aligned}$$

Therefore,

$$(C_{n,2}, C_{n,1})' \xrightarrow{d} N\left(0, \frac{1}{4}\begin{bmatrix} \operatorname{Var}(\Gamma_2(X_i)) & \operatorname{Cov}(\Gamma_2(X_i), \Gamma_1(X_i)) \\ \operatorname{Cov}(\Gamma_2(X_i), \Gamma_1(X_i)) & \operatorname{Var}(\Gamma_1(X_i))] \end{bmatrix}\right) \ .$$

It then follows from Lemma S.1.2 of Bai et al. (2021a) that

$$\sqrt{n}\left(\hat{\Delta}_2 - \Delta_2(Q)\right) \xrightarrow{d} N\left(0, \mathbb{V}_2\right) \ ,$$

where

$$\mathbb{V}_2 = E[\mathrm{Var}[Y_i(2)|X_i]] + \frac{1}{2}E[\mathrm{Var}[Y_i(1)|X_i]] + \frac{1}{4}\left(\mathrm{Var}(\Gamma_2(X_i)) + \mathrm{Var}(\Gamma_1(X_i)) - 2\,\mathrm{Cov}(\Gamma_2(X_i), \Gamma_1(X_i))\right) \ .$$

Now, suppose we pretend the two untreated units are assigned to two distinct treatment levels and denote the two untreated levels and two treated levels by $d \in \{1, 2, 3, 4\}$, where $d = 1, 2$ actually corresponds to the untreated units. Our estimand can then be defined as

$$\tilde{\Delta}_2(Q) = \Gamma_3(Q) - \frac{1}{2}\left(\Gamma_1(Q) + \Gamma_2(Q)\right)$$

Applying the existing results in Theorem 1.3.1 with $\nu = (-1/2, -1/2, 1, 0)$. It follows that

$$\sqrt{n}\left(\hat{\Delta}_2 - \tilde{\Delta}_2(Q)\right) \xrightarrow{d} N\left(0, \tilde{\mathbb{V}}_2\right) \ ,$$

where

$$\begin{aligned}\tilde{\mathbb{V}}_2 &= E[\mathrm{Var}[Y_i(3)|X_i]] + \frac{1}{4}\left(E[\mathrm{Var}[Y_i(1)|X_i]] + E[\mathrm{Var}[Y_i(2)|X_i]]\right) \\ &+ \frac{1}{4}\Bigg(\mathrm{Var}(\Gamma_3(X_i)) + \frac{1}{4}\mathrm{Var}(\Gamma_1(X_i)) + \frac{1}{4}\mathrm{Var}(\Gamma_2(X_i)) + \frac{1}{2}\mathrm{Cov}(\Gamma_2(X_i), \Gamma_1(X_i)) \\ &- \mathrm{Cov}(\Gamma_3(X_i), \Gamma_1(X_i)) - \mathrm{Cov}(\Gamma_3(X_i), \Gamma_2(X_i))\Bigg) \\ &= \mathbb{V}_2 \ ,\end{aligned}$$

where the last equality follows by setting $d = 1, 2, 3$ to $d = 1, 1, 2$. The same argument holds for $v = (-1/2, -1/2, 0, 1)$. As for $\nu = (0, 0, -1, 1)$, the estimation and inference of the third and fourth arms is not affected by treatment status in the first two arms. Therefore, pretending two controls are two different treatment levels yields the true asymptotic variance, meaning that the inference is still valid. ■

A.2 Proofs of Main Results

A.2.1 Proof of Theorem 1.3.1

We derive the limiting distribution of $\sqrt{n}(\hat{\Gamma}_n(d) - \Gamma_d(Q) : d \in \mathcal{D})$, from which the conclusion of the theorem follows by an application of the continuous mapping theorem. Note that

$$\sqrt{n}(\hat{\Gamma}_n(d) - \Gamma_d(Q) : d \in \mathcal{D})' = A_n + C_n \ ,$$

where $A_n = (A_{n,d} : d \in \mathcal{D})'$, $C_n = (C_{n,d} : d \in \mathcal{D})'$, and

$$\begin{aligned} A_{n,d} &= \frac{1}{\sqrt{n}} \sum_{1 \le i \le |\mathcal{D}| n} I\{D_i = d\}(Y_i(d) - E[Y_i(d)|X^{(n)}, D^{(n)}]) \\ C_{n,d} &= \frac{1}{\sqrt{n}} \sum_{1 \le i \le |\mathcal{D}| n} I\{D_i = d\}(E[Y_i(d)|X^{(n)}, D^n] - E[Y_i(d)]) \ . \end{aligned}$$

Note that conditional on $X^{(n)}, D^{(n)}$, $C_{n,d}$'s are constants, and $A_{n,d}$'s are independent. By Assumption 1.2.2, for $d \in \mathcal{D}$, $E[Y_i(d)|X^{(n)}, D^{(n)}] = E[Y_i(d)|X_i]$. Fix $d \in \mathcal{D}$. Let $i_j \in \lambda_j$ be such that $D_{i_j} = d$. Note

$$\begin{aligned} &\operatorname{Var}[A_{n,d}|X^{(n)}, D^{(n)}] \\ &= \frac{1}{n} \sum_{1 \le i \le |\mathcal{D}| n} I\{D_i = d\} \operatorname{Var}[Y_i(d)|X_i] \\ &= \frac{1}{|\mathcal{D}| n} \sum_{1 \le i \le |\mathcal{D}| n} \operatorname{Var}[Y_i(d)|X_i] - \frac{1}{|\mathcal{D}| n} \sum_{1 \le j \le n} \sum_{k \in \lambda_j : k \ne i_j} (\operatorname{Var}[Y_k(d)|X_k] - \operatorname{Var}[Y_{i_j}(d)|X_{i_j}]) \end{aligned}$$

where the first equality follows from Assumption 1.2.2. By Assumption 1.2.1(b) and the weak law of large numbers,

$$\frac{1}{|\mathcal{D}| n} \sum_{1 \le i \le |\mathcal{D}| n} \operatorname{Var}[Y_i(d)|X_i] \xrightarrow{P} E[\operatorname{Var}[Y_i(d)|X_i]] \ .$$

By Assumptions 1.2.1(c) and 3.4.3, we have

$$\begin{aligned}
&\Big|\frac{1}{|\mathcal{D}|n}\sum_{1\le j\le n}\sum_{k\in\lambda_j:k\neq i_j}(\operatorname{Var}[Y_k(d)|X_k]-\operatorname{Var}[Y_{i_j}(d)|X_{i_j}])\Big| \\
&\le \frac{1}{|\mathcal{D}|n}\sum_{1\le j\le n}\sum_{k\in\lambda_j:k\neq i_j}|\operatorname{Var}[Y_k(d)|X_k]-\operatorname{Var}[Y_{i_j}(d)|X_{i_j}]| \\
&\lesssim \frac{1}{n}\sum_{1\le j\le n}\sum_{k\in\lambda_j:k\neq i_j}\|X_k-X_{i_j}\| \\
&\le \frac{|\mathcal{D}|-1}{n}\sum_{1\le j\le n}\max_{i,k\in\lambda_j}\|X_i-X_k\|\xrightarrow{P}0\ .
\end{aligned}$$

Therefore, $\operatorname{Var}[A_{n,d}|X^{(n)},D^{(n)}]\xrightarrow{P}E[\operatorname{Var}[Y_i(d)|X_i]]$. We can then verify Lindeberg's condition as in the proof of Lemma S.1.4 of Bai et al. (2021a). It follows that

$$\gamma(((A_{n,d}:d\in\mathcal{D})'|X^{(n)},D^{(n)})),N(0,\mathbb{V}_1))\xrightarrow{P}0\ ,$$

where $\mathbb{V}_1=\operatorname{diag}(E[\operatorname{Var}[Y_i(d)|X_i]]:d\in\mathcal{D})$ and γ is any metric that metrizes weak convergence.

Next,

$$E[C_{n,d}|X^{(n)}]=\frac{1}{|\mathcal{D}|\sqrt{n}}\sum_{1\le i\le|\mathcal{D}|n}(\Gamma_d(X_i)-\Gamma_d)$$

and

$$\begin{aligned}
\operatorname{Var}[C_{n,d}|X^{(n)}]&=\frac{1}{n}\sum_{1\le j\le n}\sum_{i\in\lambda_j}\frac{1}{|\mathcal{D}|}\Big(\Gamma_d(X_i)-\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}\Gamma_d(X_k)\Big)^2 \\
&\lesssim \frac{1}{n}\sum_{1\le j\le n}\max_{i,k\in\lambda_j}\|X_i-X_k\|^2\xrightarrow{P}0
\end{aligned}$$

by Assumption 1.2.1(c) and 3.4.3. Therefore, by repeating the argument which establishes (S.24) in the proof of Lemma S.1.4 of Bai et al. (2021a), it follows that

$$C_{n,d}=\frac{1}{|\mathcal{D}|\sqrt{n}}\sum_{1\le i\le|\mathcal{D}|n}(\Gamma_d(X_i)-\Gamma_d)+o_P(1)\ .$$

Therefore,

$$(C_{n,d}:d\in\mathcal{D})\xrightarrow{d}N(0,\mathbb{V}_2)\ ,$$

where $(\mathbb{V}_2)_{d,d'} = \frac{1}{|\mathcal{D}|} \operatorname{Cov}(\Gamma_d(X_i), \Gamma_{d'}(X_i))$. It then follows from Lemma S.1.2 of Bai et al. (2021a) that

$$\sqrt{n}(\hat{\Gamma}_n(d) - \Gamma_d : d \in \mathcal{D})' \xrightarrow{d} N(0, \mathbb{V}_1 + \mathbb{V}_2) \ .$$

The conclusion now follows. ■

A.2.2 Proof of Theorem 1.3.2

The conclusion follows from Lemmas A.3.1–A.3.3 together with the continuous mapping theorem. ■

A.2.3 Proof of Theorem 1.3.3

Define

$$C_i = (I\{D_i = 2\}, \dots, I\{D_i = |\mathcal{D}|\})' \ .$$

To begin, note it follows from the Frisch-Waugh-Lovell theorem and Assumption 1.2.2 that

$$\begin{pmatrix} \hat{\beta}_n(2) \\ \vdots \\ \hat{\beta}_n(|\mathcal{D}|) \end{pmatrix} = \left(\sum_{1 \le i \le |\mathcal{D}|n} \tilde{C}_i \tilde{C}_i' \right)^{-1} \sum_{1 \le i \le |\mathcal{D}|n} \tilde{C}_i Y_i \ ,$$

where

$$\tilde{C}_i = \left(I\{D_i = 2\} - \frac{1}{|\mathcal{D}|}, \dots, I\{D_i = |\mathcal{D}|\} - \frac{1}{|\mathcal{D}|} \right)' \ .$$

Next, note for

$$\sum_{1 \le i \le |\mathcal{D}|n} \tilde{C}_i \tilde{C}_i' \ ,$$

the diagonal entries are $\frac{|\mathcal{D}|-1}{|\mathcal{D}|} n$ and the off-diagonal entries are $-\frac{1}{|\mathcal{D}|} n$. It follows from element calculation that the diagonal entries of $\left(\sum_{1 \le i \le |\mathcal{D}|n} \tilde{C}_i \tilde{C}_i' \right)^{-1}$ are $\frac{2}{n}$ and the off-diagonal entries are $\frac{1}{n}$. Furthermore,

$$\sum_{1 \le i \le |\mathcal{D}|n} \tilde{C}_i Y_i = \begin{pmatrix} n\hat{\Gamma}_n(2) - \frac{1}{|\mathcal{D}|} \sum_{1 \le i \le |\mathcal{D}|n} Y_i \\ \vdots \\ n\hat{\Gamma}_n(|\mathcal{D}|) - \frac{1}{|\mathcal{D}|} \sum_{1 \le i \le |\mathcal{D}|n} Y_i \end{pmatrix} \ .$$

Therefore, for $d \in \mathcal{D}\backslash\{1\}$,

$$\begin{aligned}\hat{\beta}_n(d) &= \frac{2}{n}\Big(n\hat{\Gamma}_n(d) - \frac{1}{|\mathcal{D}|}\sum_{1 \le i \le |\mathcal{D}|n} Y_i\Big) + \frac{1}{|\mathcal{D}|}\sum_{d' \in \mathcal{D}\backslash\{1,d\}} \hat{\Gamma}_n(d') - \frac{|\mathcal{D}|-2}{|\mathcal{D}|n}\sum_{1 \le i \le |\mathcal{D}|n} Y_i \\ &= \hat{\Gamma}_n(d) - \hat{\Gamma}_n(1) \ .\end{aligned}$$

The first conclusion of the theorem then follows.

Next, note by the properties of the OLS estimator that

$$\begin{aligned}\hat{\delta}_{j,n} &= \left(\sum_{1 \le i \le |\mathcal{D}|n} I\{i \in \lambda_j\}\right)^{-1} \sum_{1 \le i \le |\mathcal{D}|n} I\{i \in \lambda_j\}\left(Y_i - \sum_{d \in \mathcal{D}\backslash\{1\}} \hat{\beta}_n(d) I\{D_i = d\}\right) \\ &= \frac{1}{|\mathcal{D}|}\sum_{i \in \lambda_j} Y_i - \frac{1}{|\mathcal{D}|}\sum_{d \in \mathcal{D}\backslash\{1\}} \hat{\beta}_n(d) \ .\end{aligned}$$

Therefore,

$$\hat{\epsilon}_i = \begin{cases} Y_i - \sum_{1 \le j \le n} I\{i \in \lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k \in \lambda_j} Y_k + \frac{1}{|\mathcal{D}|}\sum_{d' \in \mathcal{D}\backslash\{1\}} \hat{\beta}_n(d') \ , \text{ if } D_i = 1 \\ Y_i - \hat{\beta}_n(d) - \sum_{1 \le j \le n} I\{i \in \lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k \in \lambda_j} Y_k + \frac{1}{|\mathcal{D}|}\sum_{d' \in \mathcal{D}\backslash\{1\}} \hat{\beta}_n(d') \ , \text{ if } D_i = d \neq 1 \ . \end{cases}$$

It follows from Lemma A.3.4 that the heteroskedasticity-robust variance estimator of $(\hat{\beta}_n(2), \dots, \hat{\beta}_n(|\mathcal{D}|))'$ equals

$$\left(\sum_{1 \le i \le |\mathcal{D}|n} \tilde{C}_i \tilde{C}_i'\right)^{-1}\left(\sum_{1 \le i \le |\mathcal{D}|n} \hat{\epsilon}_i^2 \tilde{C}_i \tilde{C}_i'\right)\left(\sum_{1 \le i \le |\mathcal{D}|n} \tilde{C}_i \tilde{C}_i'\right)^{-1} \ .$$

For $d \in \mathcal{D}\backslash\{1\}$, the corresponding $(d-1)$-th diagonal term of $\mathbb{A} = \sum_{1 \le i \le |\mathcal{D}|n} \hat{\epsilon}_i^2 \tilde{C}_i \tilde{C}_i'$ equals

$$\mathbb{A}_d = \sum_{1 \le i \le |\mathcal{D}|n} I\{D_i = 1\}\frac{1}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 + \sum_{1 \le i \le |\mathcal{D}|n} I\{D_i = d\}\frac{(|\mathcal{D}|-1)^2}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 + \sum_{\tilde{d} \in \mathcal{D}\backslash\{1,d\}}\sum_{1 \le i \le |\mathcal{D}|n} I\{D_i = \tilde{d}\}\frac{1}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 \ .$$

For $\tilde{d} \neq \check{d} \in \mathcal{D}\backslash\{1\}$, the correponding $(\tilde{d}-1, \check{d}-1)$-th term of $\sum_{1 \le i \le |\mathcal{D}|n} \hat{\epsilon}_i^2 \tilde{C}_i \tilde{C}_i'$ equals

$$\begin{aligned}\mathbb{A}_{\tilde{d},\check{d}} = &\sum_{1 \le i \le |\mathcal{D}|n} I\{D_i = 1\}\frac{1}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 + \sum_{1 \le i \le |\mathcal{D}|n} I\{D_i = \tilde{d}\}\frac{-(|\mathcal{D}|-1)}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 \\ &+ \sum_{1 \le i \le |\mathcal{D}|n} I\{D_i = \check{d}\}\frac{-(|\mathcal{D}|-1)}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 + \sum_{d' \in \mathcal{D}\backslash\{1,\tilde{d},\check{d}\}}\sum_{1 \le i \le |\mathcal{D}|n} I\{D_i = d'\}\frac{1}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 \ .\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\hat{\mathbb{V}}_n^{\mathrm{sfe}}(d,1) \\
&= \frac{4}{n^2}\mathbb{A}_d + \frac{1}{n^2}\sum_{\tilde{d}\in\mathcal{D}\setminus\{1,d\}}\mathbb{A}_{\tilde{d}} + \frac{4}{n^2}\sum_{\tilde{d}\in\mathcal{D}\setminus\{1,d\}}\mathbb{A}_{d,\tilde{d}} + \frac{2}{n^2}\sum_{\tilde{d}<\check{d}\in\mathcal{D}\setminus\{1,d\}}\mathbb{A}_{\tilde{d},\check{d}} \\
&= \frac{4+|\mathcal{D}|-2+4(|\mathcal{D}|-2)+2(|\mathcal{D}|-2)(|\mathcal{D}|-3)/2}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=1\}\frac{1}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 \\
&\qquad + \frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=d\}\frac{4(|\mathcal{D}|-1)^2+|\mathcal{D}|-2-4(|\mathcal{D}|-1)(|\mathcal{D}|-2)+2(|\mathcal{D}|-2)(|\mathcal{D}|-3)/2}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 \\
&\qquad + \frac{1}{n^2}\sum_{\tilde{d}\in\mathcal{D}\setminus\{1,d\}}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=\tilde{d}\} \\
&\qquad \times \frac{4+(|\mathcal{D}|-1)^2+|\mathcal{D}|-3-4(|\mathcal{D}|-1)+4(|\mathcal{D}|-3)-2(|\mathcal{D}|-1)(|\mathcal{D}|-3)+2(|\mathcal{D}|-3)(|\mathcal{D}|-4)/2}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 \\
&= \frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=1\}\hat{\epsilon}_i^2 + \frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=d\}\hat{\epsilon}_i^2 \\
&= \frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=1\}\left(Y_i - \sum_{1\le j\le n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k + \frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}\setminus\{1\}}\hat{\beta}_n(d')\right)^2 \\
&\qquad + \frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=d\}\left(Y_i - \hat{\beta}_n(d) - \sum_{1\le j\le n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k + \frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}\setminus\{1\}}\hat{\beta}_n(d')\right)^2 \\
&= \frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=1\}\left(Y_i - \hat{\Gamma}_n(1) - \sum_{1\le j\le n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k + \frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat{\Gamma}_n(d')\right)^2 \\
&\qquad + \frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=d\}\left(Y_i - \hat{\Gamma}_n(d) - \sum_{1\le j\le n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k + \frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat{\Gamma}_n(d')\right)^2 \\
&= \frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=1\}\left(Y_i - \hat{\Gamma}_n(1) - \sum_{1\le j\le n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k + \frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat{\Gamma}_n(d')\right)^2 \\
&\qquad + \frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=d\}\left(Y_i - \hat{\Gamma}_n(d) - \sum_{1\le j\le n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k + \frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat{\Gamma}_n(d')\right)^2 \\
&= \frac{1}{n^2}\sum_{1\le j\le n}\left(\sum_{i\in\lambda_j}\left(I\{D_i=1\}-\frac{1}{|\mathcal{D}|}\right)Y_i\right)^2 - \frac{1}{n}\left(\hat{\Gamma}_n(1)-\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat{\Gamma}_n(d')\right)^2 \\
&\qquad + \frac{1}{n^2}\sum_{1\le j\le n}\left(\sum_{i\in\lambda_j}\left(I\{D_i=d\}-\frac{1}{|\mathcal{D}|}\right)Y_i\right)^2 - \frac{1}{n}\left(\hat{\Gamma}_n(d)-\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat{\Gamma}_n(d')\right)^2 ,
\end{aligned}
$$

where in the last equality we used the fact that for $d \in \mathcal{D}$,

$$\begin{aligned}
&\sum_{1 \leq i \leq |\mathcal{D}| n} I\{D_i = d\} \left(Y_i - \sum_{1 \leq j \leq n} I\{i \in \lambda_j\} \frac{1}{|\mathcal{D}|} \sum_{k \in \lambda_j} Y_k \right) \left(\hat{\Gamma}_n(d) - \frac{1}{|\mathcal{D}|} \sum_{d' \in \mathcal{D}} \hat{\Gamma}_n(d') \right) \\
&= n \left(\hat{\Gamma}_n(d) - \frac{1}{|\mathcal{D}|} \sum_{d' \in \mathcal{D}} \hat{\Gamma}_n(d') \right)^2 ~.
\end{aligned}$$

It follows from Assumptions 1.2.1 and 3.4.3 as well as Lemmas A.3.1–A.3.3 that as $n \to \infty$,

$$\begin{aligned}
&\hat{\Gamma}_n(d) \xrightarrow{P} E[Y_i(d)] \text{ for all } d \in \mathcal{D} \\
&\frac{1}{n} \sum_{1 \leq j \leq n} \sum_{i \in \lambda_j} I\{D_i = d\} Y_i^2 \xrightarrow{P} E[Y_i^2(d)] \\
&\frac{1}{n} \sum_{1 \leq j \leq n} \left(\sum_{i \in \lambda_j} I\{D_i = d\} Y_i \right) \left(\sum_{i \in \lambda_j} I\{D_i = d'\} Y_i \right) \xrightarrow{P} E[\Gamma_d(X_i) \Gamma_{d'}(X_i)] \text{ for all } d \neq d' \in \mathcal{D} ~.
\end{aligned}$$

Therefore,

$$\begin{aligned}
n \hat{\mathbb{V}}_n^{\text{sfe}}(d, 1) \xrightarrow{P} \text{Var} &\left[\Gamma_1(X_i) - \frac{1}{|\mathcal{D}|} \sum_{d' \in \mathcal{D}} \Gamma_{d'}(X_i) \right] + \left(1 - \frac{1}{|\mathcal{D}|} \right)^2 E[\text{Var}[Y_i(1) | X_i]] \\
&+ \frac{1}{|\mathcal{D}|^2} \sum_{d' \in \mathcal{D} \setminus \{1\}} E[\text{Var}[Y_i(d') | X_i]] + \text{Var} \left[\Gamma_d(X_i) - \frac{1}{|\mathcal{D}|} \sum_{d' \in \mathcal{D}} \Gamma_{d'}(X_i) \right] \\
&+ \left(1 - \frac{1}{|\mathcal{D}|} \right)^2 E[\text{Var}[Y_i(d) | X_i]] + \frac{1}{|\mathcal{D}|^2} \sum_{d' \in \mathcal{D} \setminus \{d\}} E[\text{Var}[Y_i(d') | X_i]] ~.
\end{aligned}$$

Finally, note by Theorem 1.3.1 that the actual limiting variance for $\hat{\Gamma}_n(d) - \hat{\Gamma}_n(1)$ is

$$E\left[\text{Var}[Y_i(d) | X_i] \right] + E[\text{Var}[Y_i(1) | X_i]] + \frac{1}{|\mathcal{D}|} E\left[((\Gamma_d(X_i) - \Gamma_d) - (\Gamma_1(X_i) - \Gamma_1))^2 \right] ~.$$

Consider the special case where $E[\text{Var}[Y_i(d') | X_i]]$ are identical across $d' \in \mathcal{D}$ and

$$\Gamma_1(X_i) = \Gamma_d(X_i) = \frac{1}{|\mathcal{D}|} \sum_{d' \in \mathcal{D}} \Gamma_{d'}(X_i) \text{ with probability one } .$$

Then, the probability limit of $n \hat{\mathbb{V}}_n^{\text{sfe}}(d, 1)$ is clearly strictly smaller than the actual limiting variance for $\hat{\Gamma}_n(d) - \hat{\Gamma}_n(1)$.

For variance estimator HC 1, consider the special case where $\Gamma_d(X_i)$ are identical across $d \in \mathcal{D}$,

$E[\text{Var}[Y_i(d)|X_i]] > 0$, $E[\text{Var}[Y_i(1)|X_i]] > 0$, and $E[\text{Var}[Y_i(d')|X_i]]$ is zero for all $d' \in \mathcal{D}\backslash\{1, d\}$. Then,

$$
\begin{aligned}
n\hat{\mathbb{V}}_n^{\text{sfe}}(d,1) \times \frac{|\mathcal{D}|n}{|\mathcal{D}|n - (|\mathcal{D}| - 1 + n)} &\xrightarrow{P} \frac{|\mathcal{D}|}{|\mathcal{D}| - 1}\left(\left(1 - \frac{1}{|\mathcal{D}|}\right)^2 + \frac{1}{|\mathcal{D}|^2}\right)(E[\text{Var}[Y_i(d)|X_i]] + E[\text{Var}[Y_i(1)|X_i]]) \\
&= \frac{|\mathcal{D}|^2 - 2|\mathcal{D}| + 2}{|\mathcal{D}|^2 - |\mathcal{D}|}(E[\text{Var}[Y_i(d)|X_i]] + E[\text{Var}[Y_i(1)|X_i]]) \ .
\end{aligned}
$$

Note that

$$
\frac{|\mathcal{D}|^2 - 2|\mathcal{D}| + 2}{|\mathcal{D}|^2 - |\mathcal{D}|} < 1
$$

if and only if $|\mathcal{D}| > 2$. By a continuity argument, the result then follows for the case where $E[\text{Var}[Y_i(d')|X_i]]$ is sufficiently close to zero for all $d' \in \mathcal{D}\backslash\{1, d\}$. ■

A.2.4 Proof of Theorem 1.3.4

First, note that

$$
\left(\frac{1}{n}\sum_{1 \le j \le n}\sum_{i \in \lambda_j} C_i C_i'\right)^{-1} = \begin{pmatrix} |\mathcal{D}| & 1 & 1 & \dots & 1 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -1 & -1 & \dots & -1 \\ -1 & 2 & 1 & \dots & 1 \\ -1 & 1 & 2 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 1 & 1 & \dots & 2 \end{pmatrix} ,
$$

and note that

$$
\sum_{i \in \lambda_j} \hat{\epsilon}_i C_i = \begin{pmatrix} \sum_{i \in \lambda_j}\sum_{d \in \mathcal{D}\backslash\{1\}}(Y_i - \hat{\gamma}_n(d))I\{D_i = d\} + Y_i I\{D_i = 1\} - \hat{\gamma}_n(1) \\ \sum_{i \in \lambda_j}(Y_i - \hat{\gamma}_n(2) - \hat{\gamma}_n(1))I\{D_i = 2\} \\ \sum_{i \in \lambda_j}(Y_i - \hat{\gamma}_n(3) - \hat{\gamma}_n(1))I\{D_i = 3\} \\ \vdots \\ \sum_{i \in \lambda_j}(Y_i - \hat{\gamma}_n(|\mathcal{D}|) - \hat{\gamma}_n(1))I\{D_i = |\mathcal{D}|\} \end{pmatrix} .
$$

Combining these expressions, it follows that the d-th diagonal element of $n \cdot \hat{\mathbb{V}}_n^{\mathrm{bcve}}$ is equal to

$$
\begin{aligned}
n \cdot \hat{\mathbb{V}}_n^{\mathrm{bcve}}(d) &= \frac{1}{n} \sum_{1 \leq j \leq n} \left(\sum_{i \in \lambda_j} (Y_i - \hat{\gamma}_n(d) - \hat{\gamma}_n(1)) I\{D_i = d\} - \sum_{i \in \lambda_j} (Y_i - \hat{\gamma}_n(1)) I\{D_i = 1\} \right)^2 \\
&= \frac{1}{n} \sum_{1 \leq j \leq n} \left(\sum_{i \in \lambda_j} Y_i I\{D_i = d\} - \sum_{i \in \lambda_j} Y_i I\{D_i = 1\} \right)^2 - (\hat{\Gamma}_n(d) - \hat{\Gamma}_n(1))^2 \ .
\end{aligned}
$$

Where the second equality exploits the fact that $\hat{\gamma}_n(d) = \hat{\Gamma}_n(d) - \hat{\Gamma}_n(1)$ for $d \in \mathcal{D} \backslash \{1\}$ and $\hat{\gamma}_n(1) = \hat{\Gamma}_n(1)$. It thus follows from Lemmas A.3.1–A.3.2 and the continuous mapping theorem that

$$
n \cdot \hat{\mathbb{V}}_n^{\mathrm{bcve}}(d) \xrightarrow{p} E[\mathrm{Var}[Y_i(d)|X_i]] + E[\mathrm{Var}[Y_i(1)|X_i]] + E\left[((\Gamma_d(X_i) - \Gamma_d) - (\Gamma_1(X_i) - \Gamma_1))^2 \right] \ .
$$

Next, note that by Theorem 1.3.1, the actual limiting variance of $\hat{\Gamma}_n(d) - \hat{\Gamma}_n(1)$ is given by

$$
E\left[\mathrm{Var}[Y_i(d)|X_i]\right] + E[\mathrm{Var}[Y_i(1)|X_i]] + \frac{1}{|\mathcal{D}|} E\left[((\Gamma_d(X_i) - \Gamma_d) - (\Gamma_1(X_i) - \Gamma_1))^2 \right] \ .
$$

Therefore, the test defined in (1.8) is conservative unless

$$
E\left[((\Gamma_d(X_i) - \Gamma_d) - (\Gamma_1(X_i) - \Gamma_1))^2 \right] = 0 \ ,
$$

as desired. ■

A.2.5 Proof of Theorem 1.3.5

The proof is similar to the proof of Theorem 1.3.1, with the difference being that two units are assigned to each treatment status in each block. The necessary modification follows from arguing similarly as in Lemma B.3 of Bai (2022a) and is omitted. ■

A.2.6 Proof of Theorem 1.3.6

First note

$$\begin{aligned}
&E[\tilde{\rho}_n(d,d)|X^{(n)}] \\
&= \frac{2}{n} \sum_{1 \le j \le \lfloor n/2 \rfloor} \frac{1}{\binom{2|\mathcal{D}|}{2}} \sum_{i<l, i,l \in \lambda_j} E[Y_i(d)Y_l(d)|X^{(n)}] \\
&= \frac{2}{n} \sum_{1 \le j \le \lfloor n/2 \rfloor} \frac{1}{\binom{2|\mathcal{D}|}{2}} \sum_{i<l, i,l \in \lambda_j} E[Y_i(d)|X_i]E[Y_l(d)|X_l] ~,
\end{aligned}$$

where the first equality follows from the conditional independence assumption in Assumption 1.2.2 and the fact that in each block, there are $\binom{2|\mathcal{D}|}{2}$ ways to choose 2 units out of $2|\mathcal{D}|$ units and assign them to treatment arm d, and the second equality follows from the fact that conditional on $X^{(n)}$, $Y^{(n)}(d)$ are i.i.d. across units. (1.11) then follows by arguing similarly as in the proof of Lemma A.3.3 below (see also Section 4.7 of Bai (2022a)). ■

A.2.7 Proof of Theorem 1.3.7

First we show that

$$\sqrt{n}(\hat{\Delta}_{\nu,n} - \Delta_\nu(Q)) \xrightarrow{d} N(0, \sigma^2_{h,\nu}) ~,$$

under the stratified factorial design defined by $h(\cdot)$. To show this, we derive the limiting distribution of $\sqrt{n}(\hat{\Gamma}_n(d) - \Gamma_d(Q) : d \in \mathcal{D})$. To that end, note that

$$\sqrt{n}(\hat{\Gamma}_n(d) - \Gamma_d(Q) : d \in \mathcal{D})' = A_n + B_n + C_n + o_P(1) ~,$$

where $A_n = (A_{n,d} : d \in \mathcal{D})'$, $B_n = (B_{n,d} : d \in \mathcal{D})'$, $C_n = (C_{n,d} : d \in \mathcal{D})'$, with

$$\begin{aligned}
A_{n,d} &= \sqrt{|\mathcal{D}|}\frac{1}{\sqrt{J_n}} \sum_{1 \le i \le J_n} (Y_i(d) - E[Y_i(d)|h(X_i)])I\{D_i = d\} \\
B_{n,d} &= \sqrt{|\mathcal{D}|}\frac{1}{\sqrt{J_n}} \sum_{1 \le i \le J_n} (I\{D_i = d\} - \pi)(E[Y_i(d)|h(X_i)] - E[Y_i(d)]) \\
C_{n,d} &= \sqrt{|\mathcal{D}|}\frac{1}{\sqrt{J_n}} \sum_{1 \le i \le J_n} \pi(E[Y_i(d)|h(X_i)] - E[Y_i(d)]) ~,
\end{aligned}$$

where $\pi := \frac{1}{|\mathcal{D}|}$. Re-writing each of these terms using the fact that

$$E[Y_i(d)|h(X_i)] = \sum_{1 \le s \le S} E[Y_i(d)|h(X_i)] I\{h(X_i) = s\} = \sum_{1 \le s \le S} E[Y_i(d)|h(X_i) = s] I\{h(X_i) = s\} \ ,$$

we obtain

$$\begin{aligned}
A_{n,d} &= \sqrt{|\mathcal{D}|} \sum_{1 \le s \le S} \frac{1}{\sqrt{J_n}} \sum_{1 \le i \le J_n} (E[Y_i(d)|h(X_i)] - E[Y_i(d)]) I\{D_i = d, h(X_i) = s\} \\
B_{n,d} &= \sqrt{|\mathcal{D}|} \sum_{1 \le s \le S} (E[Y_i(d)|h(X_i) = s] - E[Y_i(d)]) \frac{J_n(s)}{J_n} \sqrt{J_n} \left(\frac{J_{n,d}(s)}{J_n(s)} - \pi \right) \\
C_{n,d} &= \sqrt{|\mathcal{D}|} \sum_{1 \le s \le S} \pi (E[Y_i(d)|h(X_i) = s] - E[Y_i(d)]) \sqrt{J_n} \left(\frac{J_n(s)}{J_n} - p(s) \right) \ ,
\end{aligned}$$

where $J_n(s) = \sum_{1 \le i \le J_n} I\{h(X_i) = s\}$, $J_{n,d}(s) = \sum_{1 \le i \le J_n} I\{h(X_i) = s, D_i = d\}$, $p(s) = P(h(X_i) = s)$, and importantly for $C_{n,d}$ we have used the fact that

$$\sum_{1 \le s \le S} (E[Y_i(d)|h(X_i) = s] - E[Y_i(d)]) p(s) = 0 \ ,$$

which follows by the law of iterated expectations. By the law of large numbers, $J_n(s)/J_n \xrightarrow{p} p(s)$, and by the properties of stratified block randomization (see Example 3.4 in Bugni et al. (2018a)),

$$\sqrt{J_n} \left(\frac{J_{n,d}(s)}{J_n(s)} - \pi \right) \xrightarrow{p} 0 \ ,$$

and hence we can conclude that $B_{n,d} \xrightarrow{p} 0$ for every $d \in \mathcal{D}$. Using Lemma C.1. in Bugni et al. (2019a), it can then be shown that

$$\begin{pmatrix} A_n \\ C_n \end{pmatrix} \xrightarrow{d} N \left(0, \begin{bmatrix} \mathbb{V}_{h,1} & 0 \\ 0 & \mathbb{V}_{h,2} \end{bmatrix} \right) \ ,$$

and hence the first result follows. Next, let ν be a $1 \times |\mathcal{D}|$ vector of constants, then it can be shown that

$$\nu \mathbb{V} \nu' = \sum_{d \in \mathcal{D}} \nu_d^2 \text{Var}[Y_i(d)] - \sum_{d \neq d' \in \mathcal{D}} \frac{1}{|\mathcal{D}|} \text{Var}[\nu_d E[Y_i(d)|X_i] - \nu_{d'} E[Y_i(d')|X_i]] \ ,$$

and

$$\nu \mathbb{V}_h \nu' = \sum_{d \in \mathcal{D}} \nu_d^2 \text{Var}[Y_i(d)] - \sum_{d \neq d' \in \mathcal{D}} \frac{1}{|\mathcal{D}|} \text{Var}[\nu_d E[Y_i(d)|h(X_i)] - \nu_{d'} E[Y_i(d')|h(X_i)]] \ .$$

It then follows from similar arguments to those used in the proof of Theorem C.2 of Bai (2022a) that $\nu \mathbb{V} \nu' \leq \nu \mathbb{V}_h \nu'$. In particular, note that

$$\begin{aligned}
&\mathrm{Var}[\nu_d E[Y_i(d)|X_i] - \nu_{d'} E[Y_i(d')|X_i]] \\
&= E[(\nu_d E[Y_i(d)|X_i] - \nu_{d'} E[Y_i(d')|X_i] - (\nu_d E[Y_i(d)] - \nu_{d'} E[Y_i(d')]))^2] \\
&= E[(\nu_d E[Y_i(d)|X_i] - \nu_{d'} E[Y_i(d')|X_i] - (\nu_d E[Y_i(d)|h(X_i)] - \nu_{d'} E[Y_i(d')|h(X_i)]) \\
&\qquad + (\nu_d E[Y_i(d)|h(X_i)] - \nu_{d'} E[Y_i(d')|h(X_i)]) - (\nu_d E[Y_i(d)] - \nu_{d'} E[Y_i(d')]))^2] \\
&= E[(\nu_d E[Y_i(d)|X_i] - \nu_{d'} E[Y_i(d')|X_i] - (\nu_d E[Y_i(d)|h(X_i)] - \nu_{d'} E[Y_i(d')|h(X_i)]))^2] \\
&\qquad + E[(\nu_d E[Y_i(d)|h(X_i)] - \nu_{d'} E[Y_i(d')|h(X_i)]) - (\nu_d E[Y_i(d)] - \nu_{d'} E[Y_i(d')]))^2] ~,
\end{aligned}$$

where the last equality follows because

$$\begin{aligned}
&E[(\nu_d E[Y_i(d)|X_i] - \nu_{d'} E[Y_i(d')|X_i] - (\nu_d E[Y_i(d)|h(X_i)] - \nu_{d'} E[Y_i(d')|h(X_i)])) \\
&\qquad ((\nu_d E[Y_i(d)|h(X_i)] - \nu_{d'} E[Y_i(d')|h(X_i)]) - (\nu_d E[Y_i(d)] - \nu_{d'} E[Y_i(d')]))] \\
&= E[E[(\nu_d E[Y_i(d)|X_i] - \nu_{d'} E[Y_i(d')|X_i] - (\nu_d E[Y_i(d)|h(X_i)] - \nu_{d'} E[Y_i(d')|h(X_i)])) \\
&\qquad ((\nu_d E[Y_i(d)|h(X_i)] - \nu_{d'} E[Y_i(d')|h(X_i)]) - (\nu_d E[Y_i(d)] - \nu_{d'} E[Y_i(d')]))|h(X_i)]] \\
&= E[E[(\nu_d E[Y_i(d)|X_i] - \nu_{d'} E[Y_i(d')|X_i] - (\nu_d E[Y_i(d)|h(X_i)] - \nu_{d'} E[Y_i(d')|h(X_i)]))|h(X_i)] \\
&\qquad ((\nu_d E[Y_i(d)|h(X_i)] - \nu_{d'} E[Y_i(d')|h(X_i)]) - (\nu_d E[Y_i(d)] - \nu_{d'} E[Y_i(d')]))] \\
&= 0 ~,
\end{aligned}$$

where the last equality follows from the law of iterated expectations. We can thus conclude that the matched tuples design is asymptotically more efficient than the large stratum design, in the sense that the difference in variances between the large stratum and matched tuples designs, $\mathbb{V}_h - \mathbb{V}$, is positive semidefinite. ■

A.2.8 Proof of Theorem 1.3.8

To begin, note that

$$\hat{\Delta}_{\nu_k,n} = \frac{1}{n} \sum_{1 \leq i \leq J_n} \sum_{d \in \mathcal{D}} I\{\iota_k(d) = +1\} I\{D_i = d\} Y_i(d) - \frac{1}{n} \sum_{1 \leq i \leq J_n} \sum_{d \in \mathcal{D}} I\{\iota_k(d) = -1\} I\{D_i = d\} Y_i(d) ~.$$

Let $A_i, 1 \leq i \leq J_n$ denote a sequence of i.i.d. random vectors, each of which is a $K-1$ vector of i.i.d. Rademacher random variables. Further assume they are independent of $Y^{(n)}(d), d \in \mathcal{D}$, $D^{(n)}$, and $X^{(n)}$. Define $\iota_{-k}(d)$ as the vector of all entries of $\iota(d)$ except the kth entry. Then, we consider the following "averaged" potential outcomes over these $K-1$ factors defined as follows:

$$\tilde{Y}_i(+1) := \sum_{d \in \mathcal{D}} I\{\iota_k(d) = +1\} I\{\iota_{-k}(d) = A_i\} Y_i(d)$$

$$\tilde{Y}_i(-1) := \sum_{d \in \mathcal{D}} I\{\iota_k(d) = -1\} I\{\iota_{-k}(d) = A_i\} Y_i(d) \ .$$

With this notation, define

$$\tilde{\Delta}_{\nu_k,n} = \frac{1}{n} \sum_{1 \leq i \leq J_n} I\{\iota_k(D_i) = +1\} \tilde{Y}_i(+1) - \frac{1}{n} \sum_{1 \leq i \leq J_n} I\{\iota_k(D_i) = -1\} \tilde{Y}_i(-1) \ .$$

It then follows from the definition of the factor k-specific design that $\tilde{\Delta}_{\nu_k,n}$ has the same distribution as $\hat{\Delta}_{\nu_k,n}$. To see it, note

$$\begin{aligned} &\frac{1}{n} \sum_{1 \leq i \leq J_n} \sum_{d \in \mathcal{D}} I\{\iota_k(d) = +1\} I\{D_i = d\} Y_i(d) \\ &= \frac{1}{n} \sum_{1 \leq i \leq J_n} \sum_{d \in \mathcal{D}} I\{\iota_k(D_i) = +1\} I\{\iota_k(d) = +1\} I\{\iota_{-k}(D_i) = \iota_{-k}(d)\} Y_i(d) \end{aligned}$$

and

$$\frac{1}{n} \sum_{1 \leq i \leq J_n} I\{\iota_k(D_i) = +1\} \tilde{Y}_i(+1) = \frac{1}{n} \sum_{1 \leq i \leq J_n} \sum_{d \in \mathcal{D}} I\{\iota_k(D_i) = +1\} I\{\iota_k(d) = +1\} I\{\iota_{-k}(d) = A_i\} Y_i(d)$$

and $\iota_{-k}(D_i)$ and A_i follow the same distribution independently of everything else.

Note $\tilde{\Delta}_{\nu_k,n}/2^{K-1}$ can be thought of as the difference-in-means estimator where the treatment has two levels $+1$ and -1 and the potential outcomes are $\tilde{Y}_i(+1)$ and $\tilde{Y}_i(-1)$. The conditions in Lemma S.1.4 in Bai et al. (2021a) can be verified straightforwardly and therefore we have

$$\sqrt{2^{K-1} n} \left(\frac{\hat{\Delta}_{\nu_k,n}}{2^{K-1}} - \frac{\Delta_{\nu_k}(Q)}{2^{K-1}} \right) \xrightarrow{d} N(0, \mathbb{V}_{\nu_k,mp}) \ ,$$

where

$$\begin{aligned}\mathbb{V}_{\nu_k,mp} &:= E[\mathrm{Var}[\tilde{Y}_i(+1)|X_i]] + E[\mathrm{Var}[\tilde{Y}_i(-1)|X_i]] \\ &\quad + \frac{1}{2}E[(E[\tilde{Y}_i(+1)|X_i] - E[\tilde{Y}_i(+1)] - (E[\tilde{Y}_i(-1)|X_i] - E[\tilde{Y}_i(-1)]))^2] \ .\end{aligned}$$

Note that by Assumption 1.2.2,

$$\begin{aligned}E[\tilde{Y}_i(+1)|X_i] &= E\left[\sum_{d\in\mathcal{D}} I\{\iota_k(d) = +1\}I\{\iota_{-k}(d) = A_i\}Y_i(d)\middle| X_i\right] \\ &= \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}} I\{\iota_k(d) = +1\}\Gamma_d(X_i) \ .\end{aligned}$$

Therefore,

$$\begin{aligned}&\frac{1}{2}E[(E[\tilde{Y}_i(+1)|X_i] - E[\tilde{Y}_i(+1)] - (E[\tilde{Y}_i(-1)|X_i] - E[\tilde{Y}_i(0)]))^2] \\ &= \frac{1}{2}\cdot\frac{1}{2^{2(K-1)}}E\left[\left(\sum_{d\in\mathcal{D}} I\{\iota_k(d) = +1\}(\Gamma_d(X_i) - \Gamma_d) - \sum_{d\in\mathcal{D}} I\{\iota_k(d) = -1\}(\Gamma_d(X_i) - \Gamma_d)\right)^2\right] \\ &= \frac{1}{2}\cdot\frac{1}{2^{2(K-1)}}E\left[(\nu_k'(\Gamma_d(X_i) - \Gamma_d : d\in\mathcal{D})))^2\right] \\ &= \frac{2^{K-1}}{2^{2(K-1)}}\nu_k' E\left[\frac{1}{2^K}\,\mathrm{Cov}[\Gamma_d(X_i), \Gamma_{d'}(X_i)]\right]_{d,d'\in\mathcal{D}}\nu_k \\ &= \frac{1}{2^{(K-1)}}\nu_k'\mathbb{V}_2\nu_k \ .\end{aligned}$$

Moreover,

$$
\begin{aligned}
&\operatorname{Var}[\tilde{Y}_i(+1)|X_i] \\
&= \operatorname{Var}\left[\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}I\{\iota_{-k}(d)=A_i\}Y_i(d)\middle|X_i\right] \\
&= E\left[\operatorname{Var}\left[\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}I\{\iota_{-k}(d)=A_i\}Y_i(d)\middle|X_i,A_i\right]\middle|X_i\right] \\
&\qquad + \operatorname{Var}\left[E\left[\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}I\{\iota_{-k}(d)=A_i\}Y_i(d)\middle|X_i,A_i\right]\middle|X_i\right] \\
&= E\left[\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}I\{\iota_{-k}(d)=A_i\}\operatorname{Var}[Y_i(d)|X_i]\middle|X_i\right] \\
&\qquad + \operatorname{Var}\left[\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}I\{\iota_{-k}(d)=A_i\}\Gamma_d(X_i)\middle|X_i\right] \\
&= \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}:\iota_k(d)=+1}\operatorname{Var}[Y_i(d)|X_i] + \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}:\iota_k(d)=+1}\left(\Gamma_d(X_i)-\frac{1}{2^{K-1}}\sum_{d'\in\mathcal{D}:\iota_k(d')=+1}\Gamma_{d'}(X_i)\right)^2 \\
&= \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}:\iota_k(d)=+1}\left(\operatorname{Var}[Y_i(d)|X_i] + \left(\Gamma_d(X_i)-\frac{1}{2^{K-1}}\sum_{d'\in\mathcal{D}:\iota_k(d')=+1}\Gamma_{d'}(X_i)\right)^2\right)~.
\end{aligned}
$$

A similar calculation applies to $\operatorname{Var}[\tilde{Y}_i(-1)|X_i]$. Finally,

$$
\begin{aligned}
\mathbb{V}_{\nu_k,mp} = \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}} E[\operatorname{Var}[Y_i(d)|X_i]] &+ \frac{1}{2^{K-1}}\nu_k'\mathbb{V}_2\nu_k \\
&+ \frac{1}{2^{K-1}}E\left[\sum_{d\in\mathcal{D}:\iota_k(d)=+1}\left(\Gamma_d(X_i)-\frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}:\iota_k(d)=+1}\Gamma_d(X_i)\right)^2\right] \\
&+ \frac{1}{2^{K-1}}E\left[\sum_{d\in\mathcal{D}:\iota_k(d')=-1}\left(\Gamma_{d'}(X_i)-\frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}::\iota_k(d)=-1}\Gamma_d(X_i)\right)^2\right]~.
\end{aligned}
$$

The conclusion therefore follows. ■

A.3 Auxiliary Lemmas

Lemma A.3.1. *Suppose Assumptions 1.2.1–3.4.3 hold. Then, for* $r = 1, 2$,

$$
\frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n} Y_i^r(d)I\{D_i=d\} \xrightarrow{P} E[Y_i^r(d)]~.
$$

Proof of Lemma A.3.1. We prove the conclusion for $r = 1$ only and the proof for $r = 2$ follows similarly. To this end, write

$$\frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n} Y_i(d)I\{D_i = d\} = \frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n}(Y_i(d)I\{D_i = d\} - E[Y_i(d)I\{D_i = d\}|X^{(n)}, D^{(n)}]) \\ + \frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n} E[Y_i(d)I\{D_i = d\}|X^{(n)}, D^{(n)}] \ .$$

Note

$$\frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n} E[Y_i(d)I\{D_i = d\}|X^{(n)}, D^{(n)}] = \frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i = d\}E[Y_i(d)|X_i] \xrightarrow{P} E[E[Y_i(d)|X_i]] = E[Y_i(d)] \ ,$$

where the equality follows from Assumption 1.2.2 and the convergence in probability follows from Assumption 3.4.3 and similar arguments to those used in the proof of Theorem 1.3.1. To complete the proof, we argue

$$\frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n}(Y_i(d)I\{D_i = d\} - E[Y_i(d)I\{D_i = d\}|X^{(n)}, D^{(n)}]) \xrightarrow{P} 0 \ .$$

For this purpose, we proceed by verifying the uniform integrability condition in Lemma S.1.3 of Bai et al. (2021a) conditional on $X^{(n)}$ and $D^{(n)}$. Note for any $m > 0$ that

$$\begin{aligned}
&\frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n} E[|Y_i(d)I\{D_i = d\} - E[Y_i(d)I\{D_i = d\}|X^{(n)}, D^{(n)}]) \\
&\qquad\qquad \times I\{|Y_i(d)I\{D_i = d\} - E[Y_i(d)I\{D_i = d\}|X^{(n)}, D^{(n)}]| > m\}|X^{(n)}, D^{(n)}] \\
&= \frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n} E[|Y_i(d)I\{D_i = d\} - E[Y_i(d)|X_i]I\{D_i = d\}| \\
&\qquad\qquad \times I\{|Y_i(d)I\{D_i = d\} - E[Y_i(d)|X_i]I\{D_i = d\}| > m\}|X^{(n)}, D^{(n)}] \\
&\le \frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n} E[|Y_i(d) - E[Y_i(d)|X_i]|I\{|Y_i(d) - E[Y_i(d)|X_i]| > m\}|X^{(n)}, D^{(n)}] \\
&= \frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n} E[|Y_i(d) - E[Y_i(d)|X_i]|I\{|Y_i(d) - E[Y_i(d)|X_i]| > m\}|X_i] \\
&\xrightarrow{P} E[|Y_i(d) - E[Y_i(d)|X_i]|I\{|Y_i(d) - E[Y_i(d)|X_i]| > m\}] \ ,
\end{aligned}$$

where the first equality holds because of Assumption 1.2.2, the inequality holds because $0 \le I\{D_i = d\} \le 1$, the second equality holds because of Assumption 1.2.2 again, and the convergence in probability follows from the weak law of large numbers because

$$\begin{aligned} &E[|Y_i(d) - E[Y_i(d)|X_i]|I\{|Y_i(d) - E[Y_i(d)|X_i]| > m\}] \leq E[|Y_i(d) - E[Y_i(d)|X_i]|] \\ &\qquad \leq E[|Y_i(d)|] + E[|E[Y_i(d)|X_i]|] \leq E[|Y_i(d)|] + E[E[|Y_i(d)||X_i]] = 2E[|Y_i(d)|] \ . \end{aligned}$$

The proof could then be completed using the subsequencing argument as in (S.29) of the proof of Lemma S.1.5 of Bai et al. (2021a). ■

Lemma A.3.2. *Suppose Assumptions 1.2.1–3.4.3 hold. Then,* $\hat{\rho}_n(d,d') \xrightarrow{P} E[\Gamma_d(X_i)\Gamma_{d'}(X_i)]$ *as* $n \to \infty$.

Proof of Lemma A.3.2. To begin with, note

$$\begin{aligned} &E[\hat{\rho}_n(d,d')|X^{(n)}] \\ &= \frac{1}{n}\sum_{1\leq j\leq n}\frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)}\sum_{\{i,k\}\subset\lambda_j}(\Gamma_d(X_i)\Gamma_{d'}(X_k) + \Gamma_d(X_k)\Gamma_{d'}(X_i)) \\ &= \frac{1}{n}\sum_{1\leq j\leq n}\frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)}\sum_{\{i,k\}\subset\lambda_j}(\Gamma_d(X_i)\Gamma_{d'}(X_i) + \Gamma_d(X_k)\Gamma_{d'}(X_k) \\ &\qquad - \frac{1}{n}\sum_{1\leq j\leq n}\frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)}\sum_{\{i,k\}\subset\lambda_j}(\Gamma_d(X_i) - \Gamma_d(X_k))(\Gamma_{d'}(X_i) - \Gamma_{d'}(X_k))) \\ &= \frac{1}{|\mathcal{D}|n}\sum_{1\leq i\leq|\mathcal{D}|n}\Gamma_d(X_i)\Gamma_{d'}(X_i) - \frac{1}{n}\sum_{1\leq j\leq n}\frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)}\sum_{\{i,k\}\subset\lambda_j}(\Gamma_d(X_i) - \Gamma_d(X_k))(\Gamma_{d'}(X_i) - \Gamma_{d'}(X_k)) \\ &\xrightarrow{P} E[\Gamma_d(X_i)\Gamma_{d'}(X_i)] \ , \end{aligned}$$

where the convergence in probability follows from Assumptions 1.2.1(c) and 3.4.3. To conclude the proof, we show

$$\hat{\rho}_n(d,d') - E[\hat{\rho}_n(d,d')|X^{(n)}] \xrightarrow{P} 0 \ . \tag{A.1}$$

In order for this, we proceed to verify the uniform integrability condition in Lemma S.1.3 of Bai et al. (2021a) conditional on $X^{(n)}$. Define

$$\hat{\rho}_{n,j}(d,d') = \Big(\sum_{i\in\lambda_j}Y_iI\{D_i = d\}\Big)\Big(\sum_{i\in\lambda_j}Y_iI\{D_i = d'\}\Big) \ .$$

In what follows, we repeatedly use the following inequalities:

$$\begin{aligned}
I\Big\{\Big|\sum_{1\le j\le k} a_j\Big| > \lambda\Big\} &\le \sum_{1\le j\le k} I\Big\{|a_j| > \frac{\lambda}{k}\Big\} \\
\Big|\sum_{1\le j\le k} a_j\Big| I\Big\{\Big|\sum_{1\le j\le k} a_j\Big| > \lambda\Big\} &\le \sum_{1\le j\le k} k|a_j| I\Big\{|a_j| > \frac{\lambda}{k}\Big\} \\
|ab| I\{|ab| > \lambda\} &\le a^2 I\{|a| > \sqrt{\lambda}\} + b^2 I\{|b| > \sqrt{\lambda}\}\ .
\end{aligned}$$

We will also repeatedly use the facts that $0 \le I\{D_i = d\} \le 1$ and $I\{D_i = d\}I\{D_k = d\} = 0$ for $i \ne k$ in the same stratum. Note

$$\begin{aligned}
&E[|\hat{\rho}_{n,j}(d,d') - E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]|I\{|\hat{\rho}_{n,j}(d,d') - E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]| > \lambda\}|X^{(n)}] \\
&\le E\Big[|\hat{\rho}_{n,j}(d,d')|I\Big\{|\hat{\rho}_{n,j}(d,d')| > \frac{\lambda}{2}\Big\}\Big|X^{(n)}\Big] + E\Big[|E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]|I\Big\{|E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]| > \frac{\lambda}{2}\Big\}\Big|X^{(n)}\Big] \\
&= E\Big[|\hat{\rho}_{n,j}(d,d')|I\Big\{|\hat{\rho}_{n,j}(d,d')| > \frac{\lambda}{2}\Big\}\Big|X^{(n)}\Big] + |E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]|I\Big\{|E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]| > \frac{\lambda}{2}\Big\} \\
&\le E\Big[\Big|\sum_{i\in\lambda_j} Y_i(d)I\{D_i = d\}\sum_{i\in\lambda_j} Y_i(d')I\{D_i = d'\}\Big|I\Big\{\Big|\sum_{i\in\lambda_j} Y_i(d)I\{D_i = d\}\sum_{i\in\lambda_j} Y_i(d')I\{D_i = d'\}\Big| > \frac{\lambda}{2}\Big\}\Big|X^{(n)}\Big] \\
&\quad + \Big|\frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)}\sum_{\{i,k\}\subset\lambda_j}(\Gamma_d(X_i)\Gamma_{d'}(X_k) + \Gamma_d(X_k)\Gamma_{d'}(X_i)) \\
&\quad \times \Big|I\Big\{\Big|\frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)}\sum_{\{i,k\}\subset\lambda_j}(\Gamma_d(X_i)\Gamma_{d'}(X_k) + \Gamma_d(X_k)\Gamma_{d'}(X_i))\Big| > \frac{\lambda}{2}\Big\} \\
&\lesssim E\Big[\sum_{i\in\lambda_j} Y_i^2(d)I\{D_i = d\}I\Big\{\Big|\sum_{i\in\lambda_j} Y_i(d)I\{D_i = d\}\Big| > \sqrt{\frac{\lambda}{2}}\Big\}\Big|X^{(n)}\Big] \\
&\quad + E\Big[\sum_{i\in\lambda_j} Y_i^2(d')I\{D_i = d'\}I\Big\{\Big|\sum_{i\in\lambda_j} Y_i(d')I\{D_i = d'\}\Big| > \sqrt{\frac{\lambda}{2}}\Big\}\Big|X^{(n)}\Big] \\
&\quad + \sum_{\{i,k\}\subset\lambda_j}\Big(|\Gamma_d(X_i)\Gamma_{d'}(X_k)|I\Big\{|\Gamma_d(X_i)\Gamma_{d'}(X_k)| > \frac{\lambda}{2}\Big\} + |\Gamma_d(X_k)\Gamma_{d'}(X_i)|I\Big\{|\Gamma_d(X_k)\Gamma_{d'}(X_i)| > \frac{\lambda}{2}\Big\}\Big) \\
&\le \sum_{i\in\lambda_j} E\Big[Y_i^2(d)I\Big\{|Y_i(d)| > \sqrt{\frac{\lambda}{4}}\Big\}\Big|X_i\Big] + \sum_{i\in\lambda_j} E\Big[Y_i^2(d')I\Big\{|Y_i(d')| > \sqrt{\frac{\lambda}{4}}\Big\}\Big|X_i\Big] \\
&\quad + \sum_{i\in\lambda_j}\Gamma_d^2(X_i)I\Big\{|\Gamma_d(X_i)| > \sqrt{\frac{\lambda}{4}}\Big\} + \sum_{i\in\lambda_j}\Gamma_{d'}^2(X_i)I\Big\{|\Gamma_{d'}(X_i)| > \sqrt{\frac{\lambda}{4}}\Big\}\ .
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\frac{1}{|\mathcal{D}|n}\sum_{1\le j\le n} E[|\hat{\rho}_{n,j}(d,d') - E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]|I\{|\hat{\rho}_{n,j}(d,d') - E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]| > \lambda\}|X^{(n)}] \\
&\lesssim \frac{1}{|\mathcal{D}|n}\sum_{1\le i\le |\mathcal{D}|n} E\Big[Y_i^2(d)I\Big\{|Y_i(d)| > \sqrt{\frac{\lambda}{4}}\Big\}\Big|X_i\Big] + \frac{1}{|\mathcal{D}|n}\sum_{1\le i\le |\mathcal{D}|n} E\Big[Y_i^2(d')I\Big\{|Y_i(d')| > \sqrt{\frac{\lambda}{4}}\Big\}\Big|X_i\Big] \\
&\quad + \frac{1}{|\mathcal{D}|n}\sum_{1\le i\le |\mathcal{D}|n} \Gamma_d^2(X_i)I\Big\{|\Gamma_d(X_i)| > \sqrt{\frac{\lambda}{4}}\Big\} + \frac{1}{|\mathcal{D}|n}\sum_{1\le i\le |\mathcal{D}|n} \Gamma_{d'}^2(X_i)I\Big\{|\Gamma_d(X_i)| > \sqrt{\frac{\lambda}{4}}\Big\} \\
&\xrightarrow{P} E\Big[Y_i^2(d)I\Big\{|Y_i(d)| > \sqrt{\frac{\lambda}{4}}\Big\}\Big] + E\Big[Y_i^2(d')I\Big\{|Y_i(d')| > \sqrt{\frac{\lambda}{4}}\Big\}\Big] \\
&\quad + E\Big[\Gamma_d^2(X_i)I\Big\{|\Gamma_d(X_i)| > \sqrt{\frac{\lambda}{4}}\Big\}\Big] + E\Big[\Gamma_{d'}^2(X_i)I\Big\{|\Gamma_d(X_i)| > \sqrt{\frac{\lambda}{4}}\Big\}\Big] \ ,
\end{aligned}
$$

where the convergence in probability follows from the weak law or large numbers. Because $E[Y_i^2(d)] < \infty$, $E[Y_i^2(d')] < \infty$, $E[\Gamma_d^2(X_i)] \le E[Y_i^2(d)] < \infty$, and $E[\Gamma_{d'}^2(X_i)] \le E[Y_i^2(d')] < \infty$, we have

$$
\begin{aligned}
\lim_{\lambda\to\infty} E\Big[Y_i^2(d)I\Big\{|Y_i(d)| > \sqrt{\frac{\lambda}{4}}\Big\}\Big] &= 0 \\
\lim_{\lambda\to\infty} E\Big[Y_i^2(d')I\Big\{|Y_i(d')| > \sqrt{\frac{\lambda}{4}}\Big\}\Big] &= 0 \\
\lim_{\lambda\to\infty} E\Big[\Gamma_d^2(X_i)I\Big\{|\Gamma_d(X_i)| > \sqrt{\frac{\lambda}{4}}\Big\}\Big] &= 0 \\
\lim_{\lambda\to\infty} E\Big[\Gamma_{d'}^2(X_i)I\Big\{|\Gamma_{d'}(X_i)| > \sqrt{\frac{\lambda}{4}}\Big\}\Big] &= 0 \ .
\end{aligned}
$$

It follows from a subsequencing argument as in (S.29) of the proof of Lemma S.1.5 of Bai et al. (2021a) that (A.1) holds. The conclusion therefore follows. ■

Lemma A.3.3. *Suppose Assumptions 1.2.1–1.2.4 hold. Then,* $\hat{\rho}_n(d,d) \xrightarrow{P} E[\Gamma_d^2(X_i)]$ *as* $n\to\infty$.

Proof of Lemma A.3.3. For $1 \le j \le \frac{n}{2}$, define

$$
\hat{\rho}_{n,j}(d,d) = \sum_{i\in\lambda_{2j-1}} Y_i I\{D_i = d\} \sum_{i\in\lambda_{2j}} Y_i I\{D_i = d\} \ .
$$

By defintition, $\hat{\rho}_n(d,d) = \frac{2}{n}\sum_{1\le j\le \frac{n}{2}} \hat{\rho}_{n,j}$. Note by Assumption 1.2.2,

$$
E[\hat{\rho}_{n,j}(d,d)|X^{(n)}] = \frac{1}{|\mathcal{D}|^2}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}} \Gamma_d(X_i)\Gamma_d(X_k) \ .
$$

Further note

$$\Gamma_d(X_i)\Gamma_d(X_k) = \frac{1}{2}\Gamma_d^2(X_i) + \frac{1}{2}\Gamma_d^2(X_k) - \frac{1}{2}(\Gamma_d(X_i) - \Gamma_d(X_k))^2 \ .$$

Therefore,

$$\begin{aligned}
&E[\hat{\rho}_n(d,d)|X^{(n)}] = \frac{2}{n}\sum_{1\le j\le \frac{n}{2}} E[\hat{\rho}_{n,j}(d,d)|X^{(n)}] \\
&= \frac{2}{n}\sum_{1\le j\le \frac{n}{2}} \frac{1}{|\mathcal{D}|^2}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}} \Big(\frac{1}{2}\Gamma_d^2(X_i) + \frac{1}{2}\Gamma_d^2(X_k) - \frac{1}{2}(\Gamma_d(X_i) - \Gamma_d(X_k))^2\Big) \\
&= \frac{1}{|\mathcal{D}|n}\sum_{1\le i\le |\mathcal{D}|n} \Gamma_d^2(X_i) - \frac{1}{n|\mathcal{D}|^2}\sum_{1\le j\le \frac{n}{2}}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}} (\Gamma_d(X_i) - \Gamma_d(X_k))^2 \\
&\xrightarrow{P} E[\Gamma_d^2(X_i)] \ ,
\end{aligned}$$

where the convergence in probability follows from Assumptions 1.2.1(c) and 1.2.4 as well as the weak law of large numbers. To conclude the proof, we show

$$\hat{\rho}_n(d,d) - E[\hat{\rho}_n(d,d)|X^{(n)}] \xrightarrow{P} 0 \ . \tag{A.2}$$

In order for this, we proceed to verify the uniform integrability condition in Lemma S.1.3 of Bai et al. (2021a) conditional on $X^{(n)}$. In what follows, we repeatedly use the following inequalities:

$$\begin{aligned}
I\Big\{\Big|\sum_{1\le j\le k} a_j\Big| > \lambda\Big\} &\le \sum_{1\le j\le k} I\Big\{|a_j| > \frac{\lambda}{k}\Big\} \\
\Big|\sum_{1\le j\le k} a_j\Big| I\Big\{\Big|\sum_{1\le j\le k} a_j\Big| > \lambda\Big\} &\le \sum_{1\le j\le k} k|a_j| I\Big\{|a_j| > \frac{\lambda}{k}\Big\} \\
|ab|I\{|ab| > \lambda\} &\le a^2 I\{|a| > \sqrt{\lambda}\} + b^2 I\{|b| > \sqrt{\lambda}\} \ .
\end{aligned}$$

We will also repeatedly use the facts that $0 \le I\{D_i = d\} \le 1$ and $I\{D_i = d\}I\{D_k = d\} = 0$ for $i \ne k$ in the same stratum. Note

$$\begin{aligned}
&E[|\hat{\rho}_{n,j}(d,d) - E[\hat{\rho}_{n,j}(d,d)|X^{(n)}]|I\{|\hat{\rho}_{n,j}(d,d) - E[\hat{\rho}_{n,j}(d,d)|X^{(n)}]| > \lambda\}|X^{(n)}] \\
&\leq E\Big[|\hat{\rho}_{n,j}(d,d)|I\Big\{|\hat{\rho}_{n,j}(d,d)| > \frac{\lambda}{2}\Big\}\Big|X^{(n)}\Big] + E\Big[|E[\hat{\rho}_{n,j}(d,d)|X^{(n)}]|I\Big\{|E[\hat{\rho}_{n,j}(d,d)|X^{(n)}]| > \frac{\lambda}{2}\Big\}\Big|X^{(n)}\Big] \\
&= E\Big[|\hat{\rho}_{n,j}(d,d)|I\Big\{|\hat{\rho}_{n,j}(d,d)| > \frac{\lambda}{2}\Big\}\Big|X^{(n)}\Big] + |E[\hat{\rho}_{n,j}(d,d)|X^{(n)}]|I\Big\{|E[\hat{\rho}_{n,j}(d,d)|X^{(n)}]| > \frac{\lambda}{2}\Big\} \\
&\leq E\Big[\Big|\sum_{i\in\lambda_{2j-1}} Y_i(d)I\{D_i = d\}\sum_{i\in\lambda_{2j}} Y_i(d)I\{D_i = d\}\times \\
&\qquad\qquad \Big|I\Big\{\Big|\sum_{i\in\lambda_{2j-1}} Y_i(d)I\{D_i = d\}\sum_{i\in\lambda_{2j}} Y_i(d)I\{D_i = d\}\Big| > \frac{\lambda}{2}\Big\}\Big|X^{(n)}\Big] \\
&\quad + \Big|\frac{1}{|\mathcal{D}|^2}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}\Gamma_d(X_i)\Gamma_d(X_k)\Big|I\Big\{\Big|\frac{1}{|\mathcal{D}|^2}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}\Gamma_d(X_i)\Gamma_d(X_k)\Big| > \frac{\lambda}{2}\Big\} \\
&\lesssim E\Big[\sum_{i\in\lambda_{2j-1}} Y_i^2(d)I\{D_i = d\}I\Big\{\Big|\sum_{i\in\lambda_{2j-1}} Y_i(d)I\{D_i = d\}\Big| > \sqrt{\frac{\lambda}{2}}\Big\}\Big|X^{(n)}\Big] \\
&\quad + E\Big[\sum_{i\in\lambda_{2j}} Y_i^2(d)I\{D_i = d\}I\Big\{\Big|\sum_{i\in\lambda_{2j}} Y_i(d)I\{D_i = d\}\Big| > \sqrt{\frac{\lambda}{2}}\Big\}\Big|X^{(n)}\Big] \\
&\quad + \sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}|\Gamma_d(X_i)\Gamma_d(X_k)|I\Big\{|\Gamma_d(X_i)\Gamma_d(X_k)| > \frac{\lambda}{2}\Big\} \\
&\leq \sum_{i\in\lambda_{2j-1}} E\Big[Y_i^2(d)I\Big\{|Y_i(d)| > \sqrt{\frac{\lambda}{2}}\Big\}\Big|X_i\Big] + \sum_{i\in\lambda_{2j}} E\Big[Y_i^2(d)I\Big\{|Y_i(d)| > \sqrt{\frac{\lambda}{2}}\Big\}\Big|X_i\Big] \\
&\quad + \sum_{i\in\lambda_{2j-1}}\Gamma_d^2(X_i)I\Big\{|\Gamma_d(X_i)| > \sqrt{\frac{\lambda}{2}}\Big\} + \sum_{i\in\lambda_{2j}}\Gamma_d^2(X_i)I\Big\{|\Gamma_d(X_i)| > \sqrt{\frac{\lambda}{2}}\Big\}\ .
\end{aligned}$$

Therefore

$$\begin{aligned}
&\frac{1}{|\mathcal{D}|n}\sum_{1\leq j\leq\frac{n}{2}} E[|\hat{\rho}_{n,j}(d,d) - E[\hat{\rho}_{n,j}(d,d)|X^{(n)}]|I\{|\hat{\rho}_{n,j}(d,d) - E[\hat{\rho}_{n,j}(d,d)|X^{(n)}]| > \lambda\}|X^{(n)}] \\
&\lesssim \frac{1}{|\mathcal{D}|n}\sum_{1\leq i\leq|\mathcal{D}|n} E\Big[Y_i^2(d)I\Big\{|Y_i(d)| > \sqrt{\frac{\lambda}{2}}\Big\}\Big|X_i\Big] + \frac{1}{|\mathcal{D}|n}\sum_{1\leq i\leq|\mathcal{D}|n}\Gamma_d^2(X_i)I\Big\{|\Gamma_d(X_i)| > \sqrt{\frac{\lambda}{2}}\Big\} \\
&\xrightarrow{P} E\Big[Y_i^2(d)I\Big\{|Y_i(d)| > \sqrt{\frac{\lambda}{2}}\Big\}\Big] + E\Big[\Gamma_d^2(X_i)I\Big\{|\Gamma_d(X_i)| > \sqrt{\frac{\lambda}{2}}\Big\}\Big]\ ,
\end{aligned}$$

where the convergence in probability follows from the weak law or large numbers. Because $E[Y_i^2(d)] < \infty$

and $E[\Gamma_d^2(X_i)] \leq E[Y_i^2(d)] < \infty$, we have

$$\lim_{\lambda \to \infty} E\Big[Y_i^2(d) I\Big\{|Y_i(d)| > \sqrt{\frac{\lambda}{4}}\Big\}\Big] = 0$$
$$\lim_{\lambda \to \infty} E\Big[\Gamma_d^2(X_i) I\Big\{|\Gamma_d(X_i)| > \sqrt{\frac{\lambda}{4}}\Big\}\Big] = 0 \ .$$

It follows from a subsequencing argument as in the proof of Lemma S.1.5 of Bai et al. (2021a) that (A.2) holds. The conclusion therefore follows. ■

Lemma A.3.4. *Suppose $(Y_i, X_{1,i}', X_{2,i}')'$, $1 \leq i \leq n$ is an i.i.d. sequence of random vectors, where Y_i takes values in $\mathbf{R}$, $X_{1,i}$ takes values in $\mathbf{R}^{k_1}$, and $X_{2,i}$ takes values in $\mathbf{R}^{k_2}$. Consider the linear regression*

$$Y_i = X_{1,i}'\beta_1 + X_{2,i}'\beta_2 + \epsilon_i \ .$$

Define $\mathbb{X} = (X_1, \ldots, X_n)'$, $\mathbb{X}_1 = (X_{1,1}, \ldots, X_{1,n})'$, and $\mathbb{X}_2 = (X_{2,1}, \ldots, X_{2,n})'$. Define $\mathbb{P}_2 = \mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1}\mathbb{X}_2'$ and $\mathbb{M}_2 = \mathbb{I} - \mathbb{P}_2$. Let $\hat{\beta}_{1,n}$ and $\hat{\beta}_{2,n}$ denote the OLS estimator of β_1 and β_2. Define $\hat{\epsilon}_i = Y_i - X_{1,i}'\hat{\beta}_{1,n} - X_{2,i}'\hat{\beta}_{2,n}$. Define

$$\tilde{\mathbb{X}}_1 = \mathbb{M}_2\mathbb{X}_1 \ .$$

Let

$$\hat{\Omega}_n = (\mathbb{X}'\mathbb{X})^{-1}(\mathbb{X}'\text{diag}(\hat{\epsilon}_i^2 : 1 \leq i \leq n)\mathbb{X})(\mathbb{X}'\mathbb{X})^{-1}$$

denote the heteroskedasticity-robust variance estimator of $(\hat{\beta}_{1,n}, \hat{\beta}_{2,n})$. Then, the upper-left $k_1 \times k_1$ block of $\hat{\Omega}_n$ equals

$$(\tilde{\mathbb{X}}_1'\tilde{\mathbb{X}}_1)^{-1}(\tilde{\mathbb{X}}_1'\text{diag}(\hat{\epsilon}_i^2 : 1 \leq i \leq n)\tilde{\mathbb{X}}_1)(\tilde{\mathbb{X}}_1'\tilde{\mathbb{X}}_1)^{-1} \ .$$

Proof of Lemma A.3.4. By the formula for the inverse of a partitioned matrix, the first k_1 rows of $(\mathbb{X}'\mathbb{X})^{-1}$ equal

$$\begin{pmatrix}(\mathbb{X}_1'\mathbb{M}_2\mathbb{X}_1)^{-1} & -(\mathbb{X}_1'\mathbb{M}_2\mathbb{X}_1)^{-1}\mathbb{X}_1'\mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1}\end{pmatrix} \ .$$

Furthermore,

$$\mathbb{X}'\text{diag}(\hat{\epsilon}_i^2 : 1 \leq i \leq n)\mathbb{X} = \begin{pmatrix}\mathbb{X}_1'\text{diag}(\hat{\epsilon}_i^2 : 1 \leq i \leq n)\mathbb{X}_1 & \mathbb{X}_1'\text{diag}(\hat{\epsilon}_i^2 : 1 \leq i \leq n)\mathbb{X}_2 \\ \mathbb{X}_2'\text{diag}(\hat{\epsilon}_i^2 : 1 \leq i \leq n)\mathbb{X}_1 & \mathbb{X}_2'\text{diag}(\hat{\epsilon}_i^2 : 1 \leq i \leq n)\mathbb{X}_2\end{pmatrix} \ .$$

The conclusion then follows from elementary calculations. ■

A.4 Additional Tables and Figures

A.4.1 Power Plots

In Section 1.4.3, we presented truncated power plots for the first and third configurations in order to make the horizontal axes the same as that of the second power plot. Here we present plots showing the entire "S" shape of the power curves for **MT** and **MT2** under all three configurations.

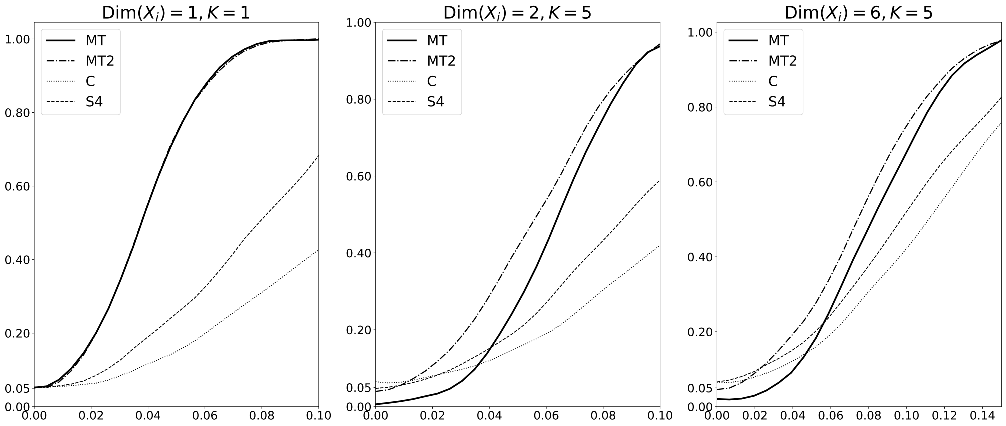

Figure A.1: Reject probability under various τs for the alternative hypothesis

A.4.2 Comparing Super Population and Finite Population Inference

In this section, we compare the coverage properties of confidence intervals constructed using our proposed variance estimator versus two other well-known estimators, under both the super and finite population approaches to inference. First, we revisit the setting introduced in Section 1.4.2, but now we consider only the matched-tuples design **(MT)**, and construct confidence intervals for the parameter $\Delta_{\nu^1_{-1}}$ using one of three variance estimators:

1. the variance estimator $\hat{\mathbb{V}}_{\nu,n}$ introduced in Section 1.3.1,
2. a standard heteroskedasticity-robust variance estimator obtained from the regression in (1.4), and
3. the block-cluster variance estimator considered in Theorem 1.3.4.

For the super population simulations, we generate the data as in Section 1.4.2. For the finite population simulations, we simply use each DGP to generate the covariates and outcomes *once*, and then fix these in repeated samples.

Table A.1 presents coverage probabilities and average confidence interval lengths (in parentheses) with varying sample sizes, based on $2,000$ Monte Carlo replications. As expected given our theoretical results, $\hat{\mathbb{V}}_{\nu,n}$ delivers exact coverage in large samples under the super-population framework in all cases, whereas the robust variance estimator and BCVE are both generally conservative. In the finite population framework, we find that both $\hat{\mathbb{V}}_{\nu,n}$ and BCVE deliver exact coverage for some model specifications in large populations, but all three methods are generally conservative. $\hat{\mathbb{V}}_{\nu,n}$ displays some under-coverage in small populations relative to BCVE, but as the population size increases, $\hat{\mathbb{V}}_{\nu,n}$ generally produces narrower confidence intervals.

Next, we repeat the above exercise using a calibrated simulation design analogous to that used in Section 1.4.3, but utilizing the wave 6 data from Fafchamps et al. (2014). To construct our data generating process, we run an OLS regression of Y_i on a constant and the seven covariates X_i employed for matching, obtaining $\hat{\beta}$ and residuals $\hat{\epsilon}$. Subsequently, for $d \in \{0, 1, 2\}$ we compute $Y_i(d)$ based on the following model:

$$Y_i(d) = X_i'\hat{\beta} + (X_i - \bar{X}_i)'\hat{\beta} \cdot \gamma \cdot d + \epsilon_i \ ,$$

with X_i drawn from the empirical distribution of the data and $\epsilon_i \sim N(0, \text{var}(\hat{\epsilon}))$. Note that when $\gamma = 0$ we obtain a model with a constant treatment effect of zero, but that as γ increases so does the amount of treatment effect heterogeneity. For the super-population simulations, the data is re-generated for each of the Monte Carlo replications. For the finite population simulations, the data is generated only *once* and then fixed in repeated samples. In each experimental assignment we match the units into triplets and assign one unit to each of $d \in \{0, 1, 2\}$.

Table A.2 presents coverage probabilities and average confidence interval lengths (in parentheses) for the parameter $\Delta_\nu = E[Y_i(1) - Y_i(0)]$, based on 2,000 Monte Carlo replications. Our first observation is that given the results for $\gamma = 0$, it is clear that the covariates X_i explain little of the variation in experimental outcomes in our simulation design since all three variance estimators obtain exact coverage. However, as we artificially increase the amount of treatment effect heterogeneity by increasing the parameter γ, we find that, in line with our theoretical results, both the robust variance estimator and BCVE become slightly conservative. Moreover, in the finite population framework, $\hat{\mathbb{V}}_{\nu,n}$ starts to become conservative as well.

A.4.3 Calibrated Simulation Design Details

In this section we provide details for the calibrated simulation study considered in Section 1.4.3. Following Branson et al. (2016), we consider data obtained from the New York Department of Education, who were

		Super Population					Finite Population				
Model	Method	$4n$=40	$4n$=80	$4n$=160	$4n$=480	$4n$=1000	$4n$=40	$4n$=80	$4n$=160	$4n$=480	$4n$=1000
1	$\hat{\mathbb{V}}_{\nu,n}$	0.9340	0.9445	0.9435	0.9460	0.9470	0.9620	0.9550	0.9335	0.9445	0.9535
		(1.810)	(1.253)	(0.881)	(0.508)	(0.351)	(2.002)	(1.547)	(0.923)	(0.480)	(0.354)
	Robust	0.9855	0.9910	0.9930	0.9890	0.9920	0.9905	0.9895	0.9860	0.9950	0.9970
		(2.375)	(1.727)	(1.226)	(0.714)	(0.495)	(2.373)	(1.891)	(1.208)	(0.702)	(0.506)
	BCVE	0.9350	0.9470	0.9400	0.9455	0.9455	0.9185	0.9390	0.9405	0.9470	0.9525
		(1.821)	(1.262)	(0.885)	(0.509)	(0.351)	(1.822)	(1.475)	(0.938)	(0.483)	(0.354)
2	$\hat{\mathbb{V}}_{\nu,n}$	0.9295	0.9395	0.9400	0.9525	0.9505	0.9495	0.9375	0.9405	0.9370	0.9520
		(1.897)	(1.299)	(0.896)	(0.509)	(0.352)	(1.829)	(1.309)	(0.848)	(0.505)	(0.354)
	Robust	0.9850	0.9905	0.9955	0.9965	0.9955	0.9870	0.9820	0.9970	0.9945	0.9980
		(2.489)	(1.809)	(1.290)	(0.751)	(0.522)	(2.337)	(1.560)	(1.354)	(0.749)	(0.540)
	BCVE	0.9185	0.9395	0.9415	0.9545	0.9515	0.9340	0.9395	0.9425	0.9415	0.9530
		(1.858)	(1.282)	(0.893)	(0.508)	(0.352)	(1.789)	(1.311)	(0.852)	(0.518)	(0.356)
3	$\hat{\mathbb{V}}_{\nu,n}$	0.9445	0.9545	0.9600	0.9435	0.9450	0.9970	0.9790	0.9975	0.9890	0.9945
		(2.499)	(1.702)	(1.193)	(0.679)	(0.469)	(2.439)	(1.710)	(1.144)	(0.686)	(0.468)
	Robust	0.9800	0.9915	0.9920	0.9905	0.9910	1.0000	0.9985	1.0000	0.9995	1.0000
		(3.080)	(2.222)	(1.593)	(0.922)	(0.640)	(3.112)	(2.228)	(1.485)	(0.916)	(0.654)
	BCVE	0.9915	0.9940	0.9980	0.9960	0.9965	0.9995	0.9995	1.0000	1.0000	1.0000
		(3.748)	(2.578)	(1.811)	(1.032)	(0.714)	(3.766)	(2.628)	(1.729)	(1.015)	(0.709)
4	$\hat{\mathbb{V}}_{\nu,n}$	0.9355	0.9480	0.9375	0.9445	0.9470	0.9310	0.9345	0.9540	0.9535	0.9640
		(1.889)	(1.319)	(0.927)	(0.534)	(0.371)	(1.674)	(1.292)	(1.015)	(0.562)	(0.373)
	Robust	0.9470	0.9680	0.9580	0.9635	0.9655	0.9435	0.9560	0.9695	0.9685	0.9770
		(1.931)	(1.406)	(1.005)	(0.584)	(0.406)	(1.751)	(1.410)	(1.085)	(0.599)	(0.407)
	BCVE	0.9550	0.9740	0.9700	0.9710	0.9750	0.9730	0.9760	0.9750	0.9760	0.9815
		(2.208)	(1.543)	(1.077)	(0.617)	(0.428)	(2.190)	(1.572)	(1.149)	(0.655)	(0.432)
5	$\hat{\mathbb{V}}_{\nu,n}$	0.9315	0.9435	0.9495	0.9465	0.9530	0.9620	0.9615	0.9735	0.9625	0.9680
		(2.012)	(1.386)	(0.962)	(0.550)	(0.381)	(2.244)	(1.153)	(0.975)	(0.554)	(0.377)
	Robust	0.9530	0.9660	0.9790	0.9770	0.9850	0.9805	0.9870	0.9950	0.9870	0.9875
		(2.152)	(1.570)	(1.117)	(0.650)	(0.452)	(2.472)	(1.415)	(1.162)	(0.655)	(0.448)
	BCVE	0.9615	0.9730	0.9790	0.9785	0.9845	0.9610	0.9915	0.9930	0.9880	0.9870
		(2.419)	(1.667)	(1.155)	(0.662)	(0.458)	(2.506)	(1.530)	(1.151)	(0.656)	(0.453)
6	$\hat{\mathbb{V}}_{\nu,n}$	0.9065	0.9290	0.9305	0.9425	0.9505	0.9105	0.9675	0.9655	0.9715	0.9665
		(4.730)	(3.361)	(2.388)	(1.388)	(0.961)	(4.846)	(3.244)	(2.233)	(1.425)	(1.025)
	Robust	0.9425	0.9600	0.9615	0.9660	0.9670	0.9625	0.9835	0.9855	0.9835	0.9765
		(5.001)	(3.624)	(2.606)	(1.521)	(1.055)	(5.392)	(3.449)	(2.437)	(1.549)	(1.090)
	BCVE	0.9560	0.9675	0.9660	0.9725	0.9735	0.9670	0.9875	0.9865	0.9865	0.9860
		(5.623)	(3.930)	(2.767)	(1.595)	(1.101)	(5.886)	(3.812)	(2.537)	(1.611)	(1.166)

Table A.1: Coverage rate and average CI length (parentheses) under the super and finite population approaches to inference

considering implementing a 2^5 factorial experiment to study five new intervention programs: a quality review, a periodic assessment, inquiry teams, a school-wide performance bonus program and an online resource program; details about each of these programs can be found in Dasgupta et al. (2015). The data-set contains covariate information for 1,376 schools. As in Branson et al. (2016), we consider experimental designs

		Super Population					Finite Population				
Model	Method	$3n$=60	$3n$=120	$3n$=360	$3n$=750	$3n$=1200	$3n$=60	$3n$=120	$3n$=360	$3n$=750	$3n$=1200
$\gamma = 0$	$\hat{\mathbb{V}}_{\nu,n}$	0.949	0.943	0.946	0.946	0.952	0.950	0.940	0.955	0.946	0.953
		(225.457)	(160.525)	(92.715)	(64.226)	(50.706)	(225.896)	(159.946)	(92.607)	(64.235)	(50.771)
	Robust	0.950	0.943	0.950	0.947	0.952	0.947	0.943	0.955	0.951	0.955
		(223.224)	(160.560)	(93.791)	(65.160)	(51.503)	(224.081)	(160.511)	(93.731)	(65.128)	(51.553)
	BCVE	0.948	0.938	0.943	0.940	0.946	0.953	0.944	0.954	0.943	0.950
		(229.461)	(162.261)	(92.762)	(64.198)	(50.674)	(230.041)	(161.019)	(92.765)	(64.089)	(50.685)
$\gamma = 1$	$\hat{\mathbb{V}}_{\nu,n}$	0.940	0.946	0.953	0.960	0.959	0.946	0.941	0.947	0.948	0.953
		(229.287)	(164.518)	(94.925)	(65.239)	(51.591)	(233.870)	(165.423)	(94.580)	(65.390)	(51.554)
	Robust	0.936	0.955	0.961	0.970	0.963	0.945	0.950	0.954	0.958	0.960
		(230.262)	(166.659)	(97.449)	(67.499)	(53.449)	(232.131)	(167.113)	(97.281)	(67.482)	(53.420)
	BCVE	0.936	0.945	0.957	0.961	0.959	0.949	0.946	0.950	0.950	0.956
		(232.063)	(165.622)	(95.388)	(65.468)	(51.662)	(237.561)	(166.805)	(94.836)	(65.553)	(51.658)
$\gamma = 3$	$\hat{\mathbb{V}}_{\nu,n}$	0.947	0.949	0.963	0.966	0.957	0.948	0.952	0.953	0.947	0.952
		(251.942)	(180.451)	(101.057)	(70.280)	(55.300)	(253.653)	(177.162)	(102.184)	(70.042)	(55.324)
	Robust	0.961	0.962	0.978	0.977	0.975	0.951	0.961	0.962	0.968	0.968
		(255.377)	(188.130)	(108.362)	(76.242)	(60.466)	(257.964)	(185.413)	(109.376)	(75.993)	(60.422)
	BCVE	0.947	0.955	0.969	0.971	0.963	0.958	0.957	0.954	0.959	0.961
		(256.837)	(185.391)	(103.913)	(72.470)	(57.259)	(260.735)	(181.843)	(105.186)	(72.325)	(57.091)
$\gamma = 5$	$\hat{\mathbb{V}}_{\nu,n}$	0.945	0.947	0.966	0.964	0.957	0.940	0.959	0.978	0.968	0.966
		(285.897)	(199.748)	(111.957)	(78.191)	(60.960)	(284.327)	(200.163)	(113.900)	(77.267)	(60.890)
	Robust	0.959	0.965	0.986	0.981	0.977	0.955	0.970	0.986	0.983	0.982
		(295.771)	(215.171)	(125.135)	(88.824)	(70.149)	(293.489)	(215.318)	(127.164)	(88.177)	(70.040)
	BCVE	0.949	0.958	0.975	0.976	0.970	0.949	0.962	0.981	0.975	0.975
		(296.164)	(209.731)	(119.286)	(83.916)	(65.873)	(293.557)	(209.593)	(121.447)	(83.287)	(65.842)

Table A.2: Coverage rate and average CI length (parentheses) under the super and finite population approaches to inference

constructed using nine covariates which were deemed likely to be correlated with schools' performance scores: total number of students, proportion of male students, enrollment rate, poverty rate, and five additional variables recording the proportion of students of various races.

Since the NYDE has yet to run such an experiment, and given the limitations of the available dataset, we select one covariate (“number of teachers") from the original dataset to use as the potential outcome under control, and then construct the potential outcomes under the various treatment combinations using the model described in Section 1.4.3. Specifically, we first demean and standardize all 9 covariates (denoted $\tilde{X}_i$), and then estimate a parameter vector β by ordinary least squares in the following linear model specification for $Y_i(-1,-1,\dots,-1)$:

$$Y_i(-1,-1,\dots,-1) = \gamma_{(-1,-1,\dots,-1)}\tilde{X}_i'\beta + \epsilon_i \ , \tag{A.3}$$

where $\gamma_{(-1,-1,\dots,-1)} = -1$ as defined in Section 1.4.3. Table A.3 presents the regression results. For each

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
constant	2.824e-06	0.007	0.000	1.000	-0.014	0.014
Total	-0.9808	0.016	-60.609	0.000	-1.012	-0.949
nativeAmerican	0.0374	0.054	0.699	0.485	-0.068	0.143
black	2.9378	3.175	0.925	0.355	-3.285	9.160
latino	2.6158	2.836	0.922	0.356	-2.942	8.174
asian	1.6866	1.822	0.926	0.355	-1.884	5.258
white	1.9064	2.150	0.887	0.375	-2.308	6.121
male	-0.0379	0.007	-5.355	0.000	-0.052	-0.024
stability	0.0045	0.007	0.636	0.525	-0.009	0.018
povertyRate	-0.1818	0.011	-16.350	0.000	-0.204	-0.160

Table A.3: Model (A.3) OLS Regression Results

treatment combination d, we then compute $Y_i(d)$ using the model from Section 1.4.3 given by

$$Y_i(d) = \tau \cdot \left(d^{(1)} + \frac{\sum_{k=2}^{K} d^{(k)}}{K-1} \right) + \gamma_d \tilde{X}_i' \beta + \epsilon_i \ ,$$

where $\tilde{X}_i$ is drawn from the empirical distribution of the data and $\epsilon_i \sim N(0, 0.1)$, where we note that 0.1 is approximately equal to the sample variance of the residuals of the regression in (A.3).

A.4.4 More Results for the Empirical Application

In this section we repeat our analysis for the data on long-term effects obtained through the final round (wave 7) of surveys from the original paper. For the analysis of long-term effects, we follow the same procedure as in the original paper, except we additionally drop the four groups with sizes ranging from 5 to 8. Note that the estimated effects are different for the fixed-effect regression. This is because, as in the analysis in the original paper, we do *not* drop entire quadruplets from our dataset whenever one member of the quadruplet was missing due to non-response in the final survey round.

		All			High initial	Low initial
		firms	Males	Females	Profit women	Profit women
		(1)	(2)	(3)	(4)	(5)
OLS without group fixed effects	Cash treatment	18.02	56.17	-8.43	-15.32	-3.84
		(29.66)	(67.95)	(18.25)	(38.99)	(17.14)
	In-kind treatment	31.59	62.02	4.63	42.10	-13.40
		(21.63)	(40.60)	(20.97)	(48.82)	(16.08)
	Cash=in-kind (*p*-val)	0.680	0.938	0.484	0.171	0.554
Matched-Tuples	Cash treatment	18.02	56.17	-8.43	-15.32	-3.84
		(26.07)	(60.09)	(17.25)	(42.10)	(16.60)
	In-kind treatment	31.59	62.02	4.63	42.10	-13.40
		(19.47)	(39.02)	(18.57)	(45.30)	(14.32)
	Cash=in-kind (*p*-val)	0.641	0.931	0.456	0.147	0.556

Table A.4: Point estimates and standard errors for testing the treatment effects of cash and in-kind grants using different methods (wave 7)

APPENDIX B

APPENDIX FOR CHAPTER 2

B.1 Proofs of Main Results

B.1.1 Proof of Proposition 2.3.1

Proof. By the Cauchy-Schwarz inequality

$$\frac{1}{G}\sum_{g=1}^{G} N_{\pi(2g)}^{\ell}|W_{\pi(2g)} - W_{\pi(2g-1)}|^{r} \le \left[\left(\frac{1}{G}\sum_{g=1}^{G} N_{\pi(2g)}^{2\ell}\right)\left(\frac{1}{G}\sum_{g=1}^{G}|W_{\pi(2g)} - W_{\pi(2g-1)}|^{2r}\right)\right]^{1/2} ,$$

$\frac{1}{G}\sum_{g=1}^{G} N_{\pi(2g)}^{2\ell} \le \frac{1}{G}\sum_{g=1}^{2G} N_g^{2\ell} = O_P(1)$ by the law of large numbers, $\frac{1}{G}\sum_g |W_{\pi(2g)} - W_{\pi(2g-1)}|^{2r} \xrightarrow{p} 0$ by assumption, hence the result follows. ■

B.1.2 Proof of Theorem 2.3.1

Proof. We have that

$$\hat{\Delta}_G = \frac{\frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_g(1)N_g D_g}{\frac{1}{G}\sum_{1\le g\le 2G} N_g D_g} - \frac{\frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_g(0)N_g(1-D_g)}{\frac{1}{G}\sum_{1\le g\le 2G} N_g(1-D_g)} .$$

In particular, for $h(x,y,z,w) = \frac{x}{y} - \frac{z}{w}$, observe that

$$\hat{\Delta}_G = h\left(\frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_g(1)N_g D_g, \frac{1}{G}\sum_{1\le g\le 2G} N_g D_g, \frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_g(0)N_g(1-D_g), \frac{1}{G}\sum_{1\le g\le 2G} N_g(1-D_g)\right)$$

and the Jacobian is

$$D_h(x,y,z,w) = \left(\frac{1}{y}, -\frac{x}{y^2}, -\frac{1}{w}, \frac{z}{w^2}\right) .$$

By Assumption 2.3.1,

$$\sqrt{G}\Big(\frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_g N_g D_g - E[\bar{Y}_g(1)N_g]\Big) = \frac{1}{\sqrt{G}}\sum_{1\le g\le 2G}(\bar{Y}_g(1)N_g D_g - E[\bar{Y}_g(1)N_g]D_g)$$

and similarly for the other three terms. The desired conclusion then follows from Lemma B.1.1 together with an application of the delta method. To see this, note by the laws of total variance and total covariance that $\mathbb{V}$ in Lemma B.1.1 is symmetric with entries

$$
\begin{aligned}
\mathbb{V}_{11} &= \operatorname{Var}[\bar{Y}_g(1)N_g] - \frac{1}{2}\operatorname{Var}[E[\bar{Y}_g(1)N_g|X_g]] \\
\mathbb{V}_{12} &= \operatorname{Cov}[\bar{Y}_g(1)N_g, N_g] - \frac{1}{2}\operatorname{Cov}[E[\bar{Y}_g(1)N_g|X_g], E[N_g|X_g]] \\
\mathbb{V}_{13} &= \frac{1}{2}\operatorname{Cov}[E[\bar{Y}_g(1)N_g|X_g], E[\bar{Y}_g(0)N_g|X_g]] \\
\mathbb{V}_{14} &= \frac{1}{2}\operatorname{Cov}[E[\bar{Y}_g(1)N_g|X_g], E[N_g|X_g]] \\
\mathbb{V}_{22} &= \operatorname{Var}[N_g] - \frac{1}{2}\operatorname{Var}[E[N_g|X_g]] \\
\mathbb{V}_{23} &= \frac{1}{2}\operatorname{Cov}[E[N_g|X_g], E[\bar{Y}_g(0)N_g|X_g]] \\
\mathbb{V}_{24} &= \frac{1}{2}\operatorname{Cov}[E[N_g|X_g], E[N_g|X_g]] \\
\mathbb{V}_{33} &= \operatorname{Var}[\bar{Y}_g(0)N_g] - \frac{1}{2}\operatorname{Var}[E[\bar{Y}_g(0)N_g|X_g]] \\
\mathbb{V}_{34} &= \operatorname{Cov}[\bar{Y}_g(0)N_g, N_g] - \frac{1}{2}\operatorname{Cov}[E[\bar{Y}_g(0)N_g|X_g], E[N_g|X_g]] \\
\mathbb{V}_{44} &= \operatorname{Var}[N_g] - \frac{1}{2}\operatorname{Var}[E[N_g|X_g]] \ .
\end{aligned}
$$

We separately calculate the variance terms involving conditional expectations and those that don't. The terms not involving conditional expectations are

$$
\begin{aligned}
&\frac{\operatorname{Var}[\bar{Y}_g(1)N_g]}{E[N_g]^2} + \frac{\operatorname{Var}[N_g]E[\bar{Y}_g(1)N_g]^2}{E[N_g]^4} + \frac{\operatorname{Var}[\bar{Y}_g(0)N_g]}{E[N_g]^2} + \frac{\operatorname{Var}[N_g]E[\bar{Y}_g(0)N_g]^2}{E[N_g]^4} \\
&\qquad - \frac{2\operatorname{Cov}[\bar{Y}_g(1)N_g, N_g]E[\bar{Y}_g(1)N_g]}{E[N_g]^3} - \frac{2\operatorname{Cov}[\bar{Y}_g(0)N_g, N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^3} \\
&= \frac{E[\bar{Y}_g^2(1)N_g^2] - E[\bar{Y}_g(1)N_g]^2}{E[N_g]^2} + \frac{E[N_g^2]E[\bar{Y}_g(1)N_g]^2 - E[N_g]^2E[\bar{Y}_g(1)N_g]^2}{E[N_g]^4} \\
&\qquad + \frac{E[\bar{Y}_g^2(0)N_g^2] - E[\bar{Y}_g(0)N_g]^2}{E[N_g]^2} + \frac{E[N_g^2]E[\bar{Y}_g(0)N_g]^2 - E[N_g]^2E[\bar{Y}_g(0)N_g]^2}{E[N_g]^4} \\
&\qquad - \frac{2E[\bar{Y}_g(1)N_g^2]E[\bar{Y}_g(1)N_g]}{E[N_g]^3} + \frac{2E[\bar{Y}_g(1)N_g]E[N_g]E[\bar{Y}_g(1)N_g]}{E[N_g]^3} \\
&\qquad - \frac{2E[\bar{Y}_g(0)N_g^2]E[\bar{Y}_g(0)N_g]}{E[N_g]^3} + \frac{2E[\bar{Y}_g(0)N_g]E[N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^3} \\
&= \frac{E[\bar{Y}_g^2(1)N_g^2]}{E[N_g]^2} + \frac{E[\bar{Y}_g^2(0)N_g^2]}{E[N_g]^2} + \frac{E[N_g^2]E[\bar{Y}_g(1)N_g]^2}{E[N_g]^4} + \frac{E[N_g^2]E[\bar{Y}_g(0)N_g]^2}{E[N_g]^4} \\
&\qquad - \frac{2E[\bar{Y}_g(1)N_g^2]E[\bar{Y}_g(1)N_g]}{E[N_g]^3} - \frac{2E[\bar{Y}_g(0)N_g^2]E[\bar{Y}_g(0)N_g]}{E[N_g]^3} \\
&= E[\tilde{Y}_g^2(1)] + E[\tilde{Y}_g^2(0)] \ ,
\end{aligned}
$$

where

$$\tilde{Y}_g(d) = \frac{N_g}{E[N_g]} \left(\bar{Y}_g(d) - \frac{E[\bar{Y}_g(d) N_g]}{E[N_g]} \right)$$

for $d \in \{0, 1\}$.

Next, the terms involving conditional expectations are

$$
\begin{aligned}
&-\frac{\mathrm{Var}[E[\bar{Y}_g(1)N_g|X_g]]}{2E[N_g]^2}-\frac{\mathrm{Var}[E[N_g|X_g]]E[\bar{Y}_g(1)N_g]^2}{2E[N_g]^4}\\
&\quad-\frac{\mathrm{Var}[E[\bar{Y}_g(0)N_g|X_g]]}{2E[N_g]^2}-\frac{\mathrm{Var}[E[N_g|X_g]]E[\bar{Y}_g(0)N_g]^2}{2E[N_g]^4}\\
&\quad+\frac{\mathrm{Cov}[E[\bar{Y}_g(1)N_g|X_g],E[N_g|X_g]]E[\bar{Y}_g(1)N_g]}{E[N_g]^3}+\frac{\mathrm{Cov}[E[\bar{Y}_g(0)N_g|X_g],E[N_g|X_g]]E[\bar{Y}_g(0)N_g]}{E[N_g]^3}\\
&\quad-\frac{\mathrm{Cov}[E[\bar{Y}_g(1)N_g|X_g],E[\bar{Y}_g(0)N_g|X_g]]}{E[N_g]^2}+\frac{\mathrm{Cov}[E[\bar{Y}_g(1)N_g|X_g],E[N_g|X_g]]E[\bar{Y}_g(0)N_g]}{E[N_g]E[N_g]^2}\\
&\quad+\frac{\mathrm{Cov}[E[N_g|X_g],E[\bar{Y}_g(0)N_g|X_g]]E[\bar{Y}_g(1)N_g]}{E[N_g]^2E[N_g]}\\
&\quad-\frac{\mathrm{Cov}[E[N_g|X_g],E[N_g|X_g]]E[\bar{Y}_g(1)N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^2E[N_g]^2}\\
&=-\frac{E[E[\bar{Y}_g(1)N_g|X_g]^2]-E[\bar{Y}_g(1)N_g]^2}{2E[N_g]^2}-\frac{(E[E[N_g|X_g]^2]-E[N_g]^2)E[\bar{Y}_g(1)N_g]^2}{2E[N_g]^4}\\
&\quad-\frac{E[E[\bar{Y}_g(0)N_g|X_g]^2]-E[\bar{Y}_g(0)N_g]^2}{2E[N_g]^2}-\frac{(E[E[N_g|X_g]^2]-E[N_g]^2)E[\bar{Y}_g(0)N_g]^2}{2E[N_g]^4}\\
&\quad+\frac{(E[E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]]-E[\bar{Y}_g(1)N_g]E[N_g])E[\bar{Y}_g(1)N_g]}{E[N_g]^3}\\
&\quad+\frac{(E[E[\bar{Y}_g(0)N_g|X_g]E[N_g|X_g]]-E[\bar{Y}_g(0)N_g]E[N_g])E[\bar{Y}_g(0)N_g]}{E[N_g]^3}\\
&\quad-\frac{E[E[\bar{Y}_g(1)N_g|X_g]E[\bar{Y}_g(0)N_g|X_g]]-E[\bar{Y}_g(1)N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]E[N_g]}\\
&\quad+\frac{(E[E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]]-E[\bar{Y}_g(1)N_g]E[N_g])E[\bar{Y}_g(0)N_g]}{E[N_g]E[N_g]^2}\\
&\quad+\frac{(E[E[\bar{Y}_g(0)N_g|X_g]E[N_g|X_g]]-E[\bar{Y}_g(0)N_g]E[N_g])E[\bar{Y}_g(1)N_g]}{E[N_g]^2E[N_g]}\\
&\quad-\frac{(E[E[N_g|X_g]E[N_g|X_g]]-E[N_g]E[N_g])E[\bar{Y}_g(1)N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^2E[N_g]^2}\\
&=-\frac{E[E[\bar{Y}_g(1)N_g|X_g]^2]}{2E[N_g]^2}-\frac{E[E[N_g|X_g]^2]E[\bar{Y}_g(1)N_g]^2}{2E[N_g]^4}-\frac{E[E[\bar{Y}_g(0)N_g|X_g]^2]}{2E[N_g]^2}-\frac{E[E[N_g|X_g]^2]E[\bar{Y}_g(0)N_g]^2}{2E[N_g]^4}\\
&\quad+\frac{E[E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]]E[\bar{Y}_g(1)N_g]}{E[N_g]^3}+\frac{E[E[\bar{Y}_g(0)N_g|X_g]E[N_g|X_g]]E[\bar{Y}_g(0)N_g]}{E[N_g]^3}\\
&\quad-\frac{E[E[\bar{Y}_g(1)N_g|X_g]E[\bar{Y}_g(0)N_g|X_g]]}{E[N_g]^2}+\frac{E[E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]]E[\bar{Y}_g(0)N_g]}{E[N_g]^3}\\
&\quad+\frac{E[E[\bar{Y}_g(0)N_g|X_g]E[N_g|X_g]]E[\bar{Y}_g(1)N_g]}{E[N_g]^3}-\frac{E[E[N_g|X_g]^2]E[\bar{Y}_g(1)N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^4}\\
&=-\frac{1}{2}E[E[\tilde{Y}_g(1)|X_g]^2]-\frac{1}{2}E[E[\tilde{Y}_g(0)|X_g]^2]-E[E[\tilde{Y}_g(1)|X_g]E[\tilde{Y}_g(0)|X_g]]\\
&=-\frac{1}{2}E[(E[\tilde{Y}_g(1)+\tilde{Y}_g(0)|X_g])^2]\ .
\end{aligned}
$$

■

Lemma B.1.1. *Suppose Q satisfies Assumptions 2.2.1 and 2.3.3 and the treatment assignment mechanism satisfies Assumptions 2.3.1–2.3.2. Define*

$$\begin{aligned}
\mathbb{L}_G^{\mathrm{YN1}} &= \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (\bar{Y}_g(1) N_g D_g - E[\bar{Y}_g(1) N_g] D_g) \\
\mathbb{L}_G^{\mathrm{N1}} &= \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (N_g D_g - E[N_g] D_g) \\
\mathbb{L}_G^{\mathrm{YN0}} &= \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (\bar{Y}_g(0) N_g (1 - D_g) - E[\bar{Y}_g(0) N_g](1 - D_g)) \\
\mathbb{L}_G^{\mathrm{N0}} &= \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (N_g (1 - D_g) - E[N_g](1 - D_g)) \ .
\end{aligned}$$

Then, as $G \to \infty$,

$$(\mathbb{L}_G^{\mathrm{YN1}}, \mathbb{L}_G^{\mathrm{N1}}, \mathbb{L}_G^{\mathrm{YN0}}, \mathbb{L}_G^{\mathrm{N0}})' \xrightarrow{d} N(0, \mathbb{V}) \ ,$$

where

$$\mathbb{V} = \mathbb{V}_1 + \mathbb{V}_2$$

for

$$\mathbb{V}_1 = \begin{pmatrix} \mathbb{V}_1^1 & 0 \\ 0 & \mathbb{V}_1^0 \end{pmatrix}$$

$$\begin{aligned}
\mathbb{V}_1^1 &= \begin{pmatrix} E[\mathrm{Var}[\bar{Y}_g(1) N_g | X_g]] & E[\mathrm{Cov}[\bar{Y}_g(1) N_g, N_g | X_g]] \\ E[\mathrm{Cov}[\bar{Y}_g(1) N_g, N_g | X_g]] & E[\mathrm{Var}[N_g | X_g]] \end{pmatrix} \\
\mathbb{V}_1^0 &= \begin{pmatrix} E[\mathrm{Var}[\bar{Y}_g(0) N_g | X_g]] & E[\mathrm{Cov}[\bar{Y}_g(0) N_g, N_g | X_g]] \\ E[\mathrm{Cov}[\bar{Y}_g(0) N_g, N_g | X_g]] & E[\mathrm{Var}[N_g | X_g]] \end{pmatrix}
\end{aligned}$$

$$\mathbb{V}_2 = \frac{1}{2} \mathrm{Var}[(E[\bar{Y}_g(1) N_g | X_g], E[N_g | X_g], E[\bar{Y}_g(0) N_g | X_g], E[N_g | X_g])'] \ .$$

Proof of Lemma B.1.1. Note

$$(\mathbb{L}_G^{\mathrm{YN1}}, \mathbb{L}_G^{\mathrm{N1}}, \mathbb{L}_G^{\mathrm{YN0}}, \mathbb{L}_G^{\mathrm{N0}}) = (\mathbb{L}_{1,G}^{\mathrm{YN1}}, \mathbb{L}_{1,G}^{\mathrm{N1}}, \mathbb{L}_{1,G}^{\mathrm{YN0}}, \mathbb{L}_{1,G}^{\mathrm{N0}}) + (\mathbb{L}_{2,G}^{\mathrm{YN1}}, \mathbb{L}_{2,G}^{\mathrm{N1}}, \mathbb{L}_{2,G}^{\mathrm{YN0}}, \mathbb{L}_{2,G}^{\mathrm{N0}}) \ ,$$

where

$$\mathbb{L}^{\text{YN1}}_{1,G} = \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (\bar{Y}_g(1) N_g D_g - E[\bar{Y}_g(1) N_g D_g | X^{(G)}, D^{(G)}])$$
$$\mathbb{L}^{\text{YN1}}_{2,G} = \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (E[\bar{Y}_g(1) N_g D_g | X^{(G)}, D^{(G)}] - E[\bar{Y}_g(1) N_g] D_g)$$

and similarly for the rest. Next, note $(\mathbb{L}^{\text{YN1}}_{1,G}, \mathbb{L}^{\text{N1}}_{1,G}, \mathbb{L}^{\text{YN0}}_{1,G}, \mathbb{L}^{\text{N0}}_{1,G}), G \ge 1$ is a triangular array of normalized sums of random vectors. Conditional on $X^{(G)}, D^{(G)}$, $(\mathbb{L}^{\text{YN1}}_{1,G}, \mathbb{L}^{\text{N1}}_{1,G}) \perp\!\!\!\perp (\mathbb{L}^{\text{YN0}}_{1,G}, \mathbb{L}^{\text{N0}}_{1,G})$. Moreover, it follows from $Q_G = Q^{2G}$ and Assumption 2.3.1 that

$$\text{Var}\left[\begin{pmatrix} \mathbb{L}^{\text{YN1}}_{1,G} \\ \mathbb{L}^{\text{N1}}_{1,G} \end{pmatrix} \middle| X^{(G)}, D^{(G)} \right] =$$
$$\begin{pmatrix} \frac{1}{G} \sum_{1 \le g \le 2G} \text{Var}[\bar{Y}_g(1) N_g | X_g] D_g & \frac{1}{G} \sum_{1 \le g \le 2G} \text{Cov}[\bar{Y}_g(1) N_g, N_g | X_g] D_g \\ \frac{1}{G} \sum_{1 \le g \le 2G} \text{Cov}[\bar{Y}_g(1) N_g, N_g | X_g] D_g & \frac{1}{G} \sum_{1 \le g \le 2G} \text{Var}[N_g | X_g] D_g \end{pmatrix} .$$

For the upper left component, we have

$$\frac{1}{G} \sum_{1 \le g \le 2G} \text{Var}[\bar{Y}_g(1) N_g | X_g] D_g = \frac{1}{G} \sum_{1 \le g \le 2G} E[\bar{Y}_g^2(1) N_g^2 | X_g] D_g - \frac{1}{G} \sum_{1 \le g \le 2G} E[\bar{Y}_g(1) N_g | X_g]^2 D_g \ . \tag{B.1}$$

Note

$$\frac{1}{G} \sum_{1 \le g \le 2G} E[\bar{Y}_g^2(1) N_g^2 | X_g] D_g$$
$$= \frac{1}{2G} \sum_{1 \le g \le 2G} E[\bar{Y}_g^2(1) N_g^2 | X_g] + \frac{1}{2} \Big(\frac{1}{G} \sum_{1 \le g \le 2G : D_g = 1} E[\bar{Y}_g^2(1) N_g^2 | X_g] - \frac{1}{G} \sum_{1 \le g \le 2G : D_g = 0} E[\bar{Y}_g^2(1) N_g^2 | X_g] \Big) \ .$$

It follows from the weak law of large numbers, the application of which is permitted by Lemma C.2.3, that

$$\frac{1}{2G} \sum_{1 \le g \le 2G} E[\bar{Y}_g^2(1) N_g^2 | X_g] \xrightarrow{P} E[\bar{Y}_g^2(1) N_g^2] \ .$$

On the other hand, it follows from Assumptions 2.3.2 and 2.3.3(a) that

$$
\begin{aligned}
&\Big|\frac{1}{G}\sum_{1\le g\le 2G:D_g=1} E[\bar{Y}_g^2(1)N_g^2|X_g] - \frac{1}{G}\sum_{1\le g\le 2G:D_g=0} E[\bar{Y}_g^2(1)N_g^2|X_g]\Big| \\
&\le \frac{1}{G}\sum_{1\le j\le G} |E[\bar{Y}^2_{\pi(2j-1)}(1)N^2_{\pi(2j-1)}|X_{\pi(2j-1)}] - E[\bar{Y}^2_{\pi(2j)}(1)N^2_{\pi(2j)}|X_{\pi(2j)}]| \\
&\lesssim \frac{1}{G}\sum_{1\le j\le G} |X_{\pi(2j-1)} - X_{\pi(2j)}| \xrightarrow{P} 0 ~.
\end{aligned}
$$

Therefore,

$$
\frac{1}{G}\sum_{1\le g\le 2G} E[\bar{Y}_g^2(1)N_g^2|X_g]D_g \xrightarrow{P} E[\bar{Y}_g^2(1)N_g^2] ~.
$$

Meanwhile,

$$
\begin{aligned}
&\frac{1}{G}\sum_{1\le g\le 2G} E[\bar{Y}_g(1)N_g|X_g]^2 D_g \\
&= \frac{1}{2G}\sum_{1\le g\le 2G} E[\bar{Y}_g(1)N_g|X_g]^2 + \frac{1}{2}\Big(\frac{1}{G}\sum_{1\le g\le 2G:D_g=1} E[\bar{Y}_g(1)N_g|X_g]^2 - \frac{1}{G}\sum_{1\le g\le 2G:D_g=0} E[\bar{Y}_g(1)N_g|X_g]^2\Big) ~.
\end{aligned}
$$

It follows from the weak law of large numbers (the application of which is permitted by Lemma C.2.3) that

$$
\frac{1}{2G}\sum_{1\le g\le 2G} E[\bar{Y}_g(1)N_g|X_g]^2 \xrightarrow{P} E[E[\bar{Y}_g(1)N_g|X_g]^2] ~.
$$

Next,

$$\begin{aligned}
&\Big|\frac{1}{G}\sum_{1\le g\le 2G:D_g=1} E[\bar{Y}_g(1)N_g|X_g]^2-\frac{1}{G}\sum_{1\le g\le 2G:D_g=0} E[\bar{Y}_g(1)N_g|X_g]^2\Big| \\
&\le \frac{1}{G}\sum_{1\le j\le G}|E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|X_{\pi(2j-1)}]-E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|X_{\pi(2j)}]| \\
&\qquad\qquad\qquad\qquad \times |E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|X_{\pi(2j-1)}]+E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|X_{\pi(2j)}]| \\
&\lesssim \Big(\frac{1}{G}\sum_{1\le j\le G}|X_{\pi(2j-1)}-X_{\pi(2j)}|^2\Big)^{1/2}\times \\
&\qquad \Big(\frac{1}{G}\sum_{1\le j\le G}(|E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|X_{\pi(2j-1)}]+E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|X_{\pi(2j)}]|)^2\Big)^{1/2} \\
&\lesssim \Big(\frac{1}{G}\sum_{1\le j\le G}|X_{\pi(2j-1)}-X_{\pi(2j)}|^2\Big)^{1/2}\times \\
&\qquad \Big(\frac{1}{G}\sum_{1\le j\le G}(|E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|X_{\pi(2j-1)}]|^2+|E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|X_{\pi(2j)}]|^2)\Big)^{1/2} \\
&\le \Big(\frac{1}{G}\sum_{1\le j\le G}|X_{\pi(2j-1)}-X_{\pi(2j)}|^2\Big)^{1/2}\Big(\frac{1}{G}\sum_{1\le g\le 2G}E[\bar{Y}_g(1)N_g|X_g]^2\Big)^{1/2}\xrightarrow{P}0\ ,
\end{aligned}$$

where the first inequality follows by inspection, the second follows from Assumption 2.3.3(a) and the Cauchy-Schwarz inequality, the third follows from $(a+b)^2\le 2a^2+2b^2$, the last follows by inspection again and the convergence in probability follows from Assumption 2.3.2 and the law of large numbers. Therefore,

$$\frac{1}{G}\sum_{1\le g\le 2G}E[\bar{Y}_g(1)N_g|X_g]^2D_g\xrightarrow{P}E\left[E[\bar{Y}_g(1)N_g|X_g]^2\right]\ ,$$

and hence it follows from (C.4) that

$$\frac{1}{G}\sum_{1\le g\le 2G}\mathrm{Var}[\bar{Y}_g(1)N_g|X_g]D_g\xrightarrow{P}E[\mathrm{Var}[\bar{Y}_g(1)N_g|X_g]]\ .$$

An identical argument establishes that

$$\frac{1}{G}\sum_{1\le g\le 2G}\mathrm{Var}[N_g|X_g]D_g\xrightarrow{P}E[\mathrm{Var}[N_g|X_g]]\ .$$

To study the off-diagonal components, note that

$$\frac{1}{G}\sum_{1\le g\le 2G}\text{Cov}[\bar{Y}_g(1)N_g, N_g|X_g]D_g = \frac{1}{G}\sum_{1\le g\le 2G}E[\bar{Y}_g(1)N_g^2|X_g]D_g - \frac{1}{G}\sum_{1\le g\le 2G}E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]D_g \ . \tag{B.2}$$

By a similar argument to that used above, it can be shown that

$$\frac{1}{G}\sum_{1\le g\le 2G}E[\bar{Y}_g(1)N_g^2|X_g]D_g \xrightarrow{P} E[\bar{Y}_g(1)N_g^2] \ .$$

Meanwhile,

$$\begin{aligned}&\frac{1}{G}\sum_{1\le g\le 2G}E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]D_g\\ &= \frac{1}{2G}\sum_{1\le g\le 2G}E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]\\ &+ \frac{1}{2}\Big(\frac{1}{G}\sum_{1\le g\le 2G:D_g=1}E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g] - \frac{1}{G}\sum_{1\le g\le 2G:D_g=0}E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]\Big) \ .\end{aligned}$$

Note that

$$E[E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]] = E[[N_gE[\bar{Y}_g(1)|W_g]|X_g]E[N_g|X_g]] \lesssim E[N_g^2] < \infty \ ,$$

where the equality follows by the law of iterated expectations and the inequality by Lemma C.2.3 and Jensen's inequality, and the law of iterated expectations. Thus by the weak law of large numbers,

$$\frac{1}{2G}\sum_{1\le g\le 2G}E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g] \xrightarrow{P} E[E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]] \ .$$

Next, by the triangle inequality

$$\begin{aligned}&\Big|\frac{1}{G}\sum_{1\le g\le 2G:D_g=1}E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g] - \frac{1}{G}\sum_{1\le g\le 2G:D_g=0}E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]\Big|\\ &\le \frac{1}{G}\sum_{1\le j\le G}\Big|E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|X_{\pi(2j-1)}]E[N_{\pi(2j-1)}|X_{\pi(2j-1)}]\\ &\qquad - E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|X_{\pi(2j)}]E[N_{\pi(2j)}|X_{\pi(2j)}]\Big| \ ,\end{aligned}$$

and for each j,

$$\begin{aligned}
&\big|E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|X_{\pi(2j-1)}]E[N_{\pi(2j-1)}|X_{\pi(2j-1)}] - E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|X_{\pi(2j)}]E[N_{\pi(2j)}|X_{\pi(2j)}]\big| \\
&= \Big|(E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|X_{\pi(2j-1)}] - E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|X_{\pi(2j)}])E[N_{\pi(2j)}|X_{\pi(2j)}] \\
&+ (E[N_{\pi(2j-1)}|X_{\pi(2j-1)}] - E[N_{\pi(2j)}|X_{\pi(2j)}])E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|X_{\pi(2j-1)}]\Big| \\
&\lesssim \big|E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|X_{\pi(2j-1)}] - E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|X_{\pi(2j)}]\big| \\
&\qquad + \big|E[N_{\pi(2j-1)}|X_{\pi(2j-1)}] - E[N_{\pi(2j)}|X_{\pi(2j)}]\big| \ ,
\end{aligned}$$

where the final inequality follows from the triangle inequality, Assumption 2.3.3(b) and Lemma C.2.3. Thus we have that

$$\begin{aligned}
&\Big|\frac{1}{G}\sum_{1\le g\le 2G:D_g=1} E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g] - \frac{1}{G}\sum_{1\le g\le 2G:D_g=0} E[\bar{Y}_g(1)N_g|X_g]E[N_g|X_g]\Big| \\
&\lesssim \frac{1}{G}\sum_{1\le j\le G} \big|E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|X_{\pi(2j-1)}] - E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|X_{\pi(2j)}]\big| \\
&\qquad + \big|E[N_{\pi(2j-1)}|X_{\pi(2j-1)}] - E[N_{\pi(2j)}|X_{\pi(2j)}]\big| \\
&\lesssim \frac{1}{G}\sum_{1\le j\le G} |X_{\pi(2j-1)} - X_{\pi(2j)}| \xrightarrow{P} 0 \ ,
\end{aligned}$$

where the final inequality follows from Assumptions 2.3.3 and the convergence in probability follows from Assumption 2.3.1. Proceeding as in the case of the upper left component, we obtain that

$$\frac{1}{G}\sum_{1\le g\le 2G} \mathrm{Cov}[\bar{Y}_g(1)N_g, N_g|X_g]D_g \xrightarrow{P} E[\mathrm{Cov}[\bar{Y}_g(1)N_g, N_g|X_g]] \ .$$

Thus we have established that

$$\mathrm{Var}\left[\begin{pmatrix}\mathbb{L}^{\mathrm{YN1}}_{1,G} \\ \mathbb{L}^{\mathrm{N1}}_{1,G}\end{pmatrix}\middle| X^{(G)}, D^{(G)}\right] \xrightarrow{P} \mathbb{V}^1_1 \ .$$

Similarly,

$$\mathrm{Var}\left[\begin{pmatrix}\mathbb{L}^{\mathrm{YN0}}_{1,G} \\ \mathbb{L}^{\mathrm{N0}}_{1,G}\end{pmatrix}\middle| X^{(G)}, D^{(G)}\right] \xrightarrow{P} \mathbb{V}^0_1 \ .$$

It thus follows from similar arguments to those used in Lemma B.1.2 that

$$\rho(\mathcal{L}((\mathbb{L}^{\mathrm{YN1}}_{1,G}, \mathbb{L}^{\mathrm{N1}}_{1,G}, \mathbb{L}^{\mathrm{YN0}}_{1,G}, \mathbb{L}^{\mathrm{N0}}_{1,G})'|X^{(G)}, D^{(G)}), N(0, \mathbb{V}_1)) \xrightarrow{P} 0 \ , \tag{B.3}$$

where $\mathcal{L}(\cdot)$ denotes the law of a random variable and ρ is any metric that metrizes weak convergence.

Next, we study $(\mathbb{L}_{2,G}^{\text{YN1}}, \mathbb{L}_{2,G}^{\text{N1}}, \mathbb{L}_{2,G}^{\text{YN0}}, \mathbb{L}_{2,G}^{\text{N0}})$. It follows from $Q_G = Q^{2G}$ and Assumption 2.3.1 that

$$\begin{pmatrix} \mathbb{L}_{2,G}^{\text{YN1}} \\ \mathbb{L}_{2,G}^{\text{N1}} \\ \mathbb{L}_{2,G}^{\text{YN0}} \\ \mathbb{L}_{2,G}^{\text{N0}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} D_g(E[\bar{Y}_g(1) N_g | X_g] - E[\bar{Y}_g(1) N_g]) \\ \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} D_g(E[N_g | X_g] - E[N_g]) \\ \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (1 - D_g)(E[\bar{Y}_g(0) N_g | X_g] - E[\bar{Y}_g(0) N_g]) \\ \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (1 - D_g)(E[N_g | X_g] - E[N_g]) \end{pmatrix} .$$

For $\mathbb{L}_{2,G}^{\text{YN1}}$, note it follows from Assumption 2.3.1 that

$$\begin{aligned} \text{Var}[\mathbb{L}_{2,G}^{\text{YN1}} | X^{(G)}] &= \frac{1}{4G} \sum_{1 \le j \le G} (E[\bar{Y}_{\pi(2j-1)}(1) N_{\pi(2j-1)} | X_{\pi(2j-1)}] - E[\bar{Y}_{\pi(2j)}(1) N_{\pi(2j)} | X_{\pi(2j)}])^2 \\ &\lesssim \frac{1}{G} \sum_{1 \le j \le G} |X_{\pi(2j-1)} - X_{\pi(2j)}|^2 \xrightarrow{P} 0 . \end{aligned}$$

Therefore, it follows from Markov's inequality conditional on $X^{(G)}$ and $D^{(G)}$, and the fact that probabilities are bounded and hence uniformly integrable, that

$$\mathbb{L}_{2,G}^{\text{YN1}} = E[\mathbb{L}_{2,G}^{\text{YN1}} | X^{(G)}] + o_P(1) .$$

Applying a similar argument to each of $L_{2,G}^{\text{N1}}$, $L_{2,G}^{\text{YN0}}$, $L_{2,G}^{\text{N0}}$ allows us to conclude that

$$\begin{pmatrix} \mathbb{L}_{2,G}^{\text{YN1}} \\ \mathbb{L}_{2,G}^{\text{N1}} \\ \mathbb{L}_{2,G}^{\text{YN0}} \\ \mathbb{L}_{2,G}^{\text{N0}} \end{pmatrix} = \begin{pmatrix} \frac{1}{2\sqrt{G}} \sum_{1 \le g \le 2G} (E[\bar{Y}_g(1) N_g | X_g] - E[\bar{Y}_g(1) N_g]) \\ \frac{1}{2\sqrt{G}} \sum_{1 \le g \le 2G} (E[N_g | X_g] - E[N_g]) \\ \frac{1}{2\sqrt{G}} \sum_{1 \le g \le 2G} (E[\bar{Y}_g(0) N_g | X_g] - E[\bar{Y}_g(0) N_g]) \\ \frac{1}{2\sqrt{G}} \sum_{1 \le g \le 2G} (E[N_g | X_g] - E[N_g]) \end{pmatrix} + o_P(1) .$$

It thus follows from the central limit theorem (the application of which is justified by Jensen's inequality combined with Assumption 2.2.1(b), and Lemma C.2.3) that

$$(\mathbb{L}_{2,G}^{\text{YN1}}, \mathbb{L}_{2,G}^{\text{N1}}, \mathbb{L}_{2,G}^{\text{YN0}}, \mathbb{L}_{2,G}^{\text{N0}})' \xrightarrow{d} N(0, \mathbb{V}_2) .$$

Because (C.5) holds and $(\mathbb{L}_{2,G}^{\text{YN1}}, \mathbb{L}_{2,G}^{\text{N1}}, \mathbb{L}_{2,G}^{\text{YN0}}, \mathbb{L}_{2,G}^{\text{N0}})$ is deterministic conditional on $X^{(G)}, D^{(G)}$, the conclusion of the theorem follows from Lemma S.1.3 in Bai et al. (2022c). ■

B.1.3 Proof of Theorem 2.3.2

Proof. We have that

$$\hat{\Delta}_G = \frac{\frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_g(1)N_gD_g}{\frac{1}{G}\sum_{1\le g\le 2G}N_gD_g} - \frac{\frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_g(0)N_g(1-D_g)}{\frac{1}{G}\sum_{1\le g\le 2G}N_g(1-D_g)} \ .$$

In particular, for $h(x,y,z,w) = \frac{x}{y} - \frac{z}{w}$, observe that

$$\hat{\Delta}_G = h\left(\frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_g(1)N_gD_g, \frac{1}{G}\sum_{1\le g\le 2G}N_gD_g, \frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_g(0)N_g(1-D_g), \frac{1}{G}\sum_{1\le g\le 2G}N_g(1-D_g)\right)$$

and the Jacobian is

$$D_h(x,y,z,w) = \left(\frac{1}{y}, -\frac{x}{y^2}, -\frac{1}{w}, \frac{z}{w^2}\right) \ .$$

By Assumption 2.3.4,

$$\sqrt{G}\Big(\frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_gN_gD_g - E[\bar{Y}_g(1)N_g]\Big) = \frac{1}{\sqrt{G}}\sum_{1\le g\le 2G}(\bar{Y}_g(1)N_gD_g - E[\bar{Y}_g(1)N_g]D_g)$$

and similarly for the other three terms. The desired conclusion then follows from Lemma B.1.2 together with an application of the Delta method. To see this, note by the laws of total variance and total covariance that $\mathbb{V}$ in Lemma B.1.2 is symmetric with entries

$$\begin{aligned}
\mathbb{V}_{11} &= \text{Var}[\bar{Y}_g(1)N_g] - \frac{1}{2}\text{Var}[E[\bar{Y}_g(1)N_g|W_g]] \\
\mathbb{V}_{12} &= \text{Cov}[E[\bar{Y}_g(1)N_g|W_g], N_g] - \frac{1}{2}\text{Cov}[E[\bar{Y}_g(1)N_g|W_g], N_g] \\
\mathbb{V}_{13} &= \frac{1}{2}\text{Cov}[E[\bar{Y}_g(1)N_g|W_g], E[\bar{Y}_g(0)N_g|W_g]] \\
\mathbb{V}_{14} &= \frac{1}{2}\text{Cov}[E[\bar{Y}_g(1)N_g|W_g], N_g] \\
\mathbb{V}_{22} &= \text{Var}[N_g] - \frac{1}{2}\text{Var}[N_g] \\
\mathbb{V}_{23} &= \frac{1}{2}\text{Cov}[N_g, E[\bar{Y}_g(0)N_g|X_g]] \\
\mathbb{V}_{24} &= \frac{1}{2}\text{Var}[N_g] \\
\mathbb{V}_{33} &= \text{Var}[\bar{Y}_g(0)N_g] - \frac{1}{2}\text{Var}[E[\bar{Y}_g(0)N_g|W_g]] \\
\mathbb{V}_{34} &= \text{Cov}[E[\bar{Y}_g(0)N_g|W_g], N_g] - \frac{1}{2}\text{Cov}[E[\bar{Y}_g(0)N_g|W_g], N_g] \\
\mathbb{V}_{44} &= \text{Var}[N_g] - \frac{1}{2}\text{Var}[N_g] \ .
\end{aligned}$$

We proceed by mirroring the algebra in Theorem 2.3.1. Expanding and simplifying the first half of the expression:

$$
\begin{aligned}
&\frac{\mathrm{Var}[\bar{Y}_g(1)N_g]}{E[N_g]^2} + \frac{\mathrm{Var}[N_g]E[\bar{Y}_g(1)N_g]^2}{E[N_g]^4} + \frac{\mathrm{Var}[\bar{Y}_g(0)N_g]}{E[N_g]^2} + \frac{\mathrm{Var}[N_g]E[\bar{Y}_g(0)N_g]^2}{E[N_g]^4} \\
&\quad - \frac{2\,\mathrm{Cov}[E[\bar{Y}_g(1)N_g|W_g], N_g]E[\bar{Y}_g(1)N_g]}{E[N_g]^3} - \frac{2\,\mathrm{Cov}[E[\bar{Y}_g(0)N_g|W_g], N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^3} \\
&= \frac{E[\bar{Y}_g^2(1)N_g^2] - E[\bar{Y}_g(1)N_g]^2}{E[N_g]^2} + \frac{E[N_g^2]E[\bar{Y}_g(1)N_g]^2 - E[N_g]^2E[\bar{Y}_g(1)N_g]^2}{E[N_g]^4} \\
&\quad + \frac{E[\bar{Y}_g^2(0)N_g^2] - E[\bar{Y}_g(0)N_g]^2}{E[N_g]^2} + \frac{E[N_g^2]E[\bar{Y}_g(0)N_g]^2 - E[N_g]^2E[\bar{Y}_g(0)N_g]^2}{E[N_g]^4} \\
&\quad - \frac{2E[\bar{Y}_g(1)N_g^2]E[\bar{Y}_g(1)N_g]}{E[N_g]^3} + \frac{2E[\bar{Y}_g(1)N_g]E[N_g]E[\bar{Y}_g(1)N_g]}{E[N_g]^3} \\
&\quad - \frac{2E[\bar{Y}_g(0)N_g^2]E[\bar{Y}_g(0)N_g]}{E[N_g]^3} + \frac{2E[\bar{Y}_g(0)N_g]E[N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^3} \\
&= \frac{E[\bar{Y}_g^2(1)N_g^2]}{E[N_g]^2} + \frac{E[\bar{Y}_g^2(0)N_g^2]}{E[N_g]^2} + \frac{E[N_g^2]E[\bar{Y}_g(1)N_g]^2}{E[N_g]^4} + \frac{E[N_g^2]E[\bar{Y}_g(0)N_g]^2}{E[N_g]^4} \\
&\quad - \frac{2E[\bar{Y}_g(1)N_g^2]E[\bar{Y}_g(1)N_g]}{E[N_g]^3} - \frac{2E[\bar{Y}_g(0)N_g^2]E[\bar{Y}_g(0)N_g]}{E[N_g]^3} \\
&= E[\tilde{Y}_g^2(1)] + E[\tilde{Y}_g^2(0)]\ ,
\end{aligned}
$$

where

$$\tilde{Y}_g(d) = \frac{N_g}{E[N_g]}\left(\bar{Y}_g(d) - \frac{E[\bar{Y}_g(d)N_g]}{E[N_g]}\right)$$

for $d \in \{0, 1\}$.

Expanding the second half of the expression:

$$\begin{aligned}
&-\frac{\text{Var}[E[\bar{Y}_g(1)N_g|W_g]]}{2E[N_g]^2}-\frac{\text{Var}[N_g]E[\bar{Y}_g(1)N_g]^2}{2E[N_g]^4}\\
&\quad-\frac{\text{Var}[E[\bar{Y}_g(0)N_g|W_g]]}{2E[N_g]^2}-\frac{\text{Var}[N_g]E[\bar{Y}_g(0)N_g]^2}{2E[N_g]^4}\\
&\quad+\frac{\text{Cov}[E[\bar{Y}_g(1)N_g|W_g],N_g]E[\bar{Y}_g(1)N_g]}{E[N_g]^3}+\frac{\text{Cov}[E[\bar{Y}_g(0)N_g|W_g],N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^3}\\
&\quad-\frac{\text{Cov}[E[\bar{Y}_g(1)N_g|W_g],E[\bar{Y}_g(0)N_g|W_g]]}{E[N_g]^2}+\frac{\text{Cov}[E[\bar{Y}_g(1)N_g|W_g],N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]E[N_g]^2}\\
&\quad+\frac{\text{Cov}[N_g,E[\bar{Y}_g(0)N_g|W_g]]E[\bar{Y}_g(1)N_g]}{E[N_g]^2E[N_g]}\\
&\quad-\frac{\text{Cov}[N_g,N_g]E[\bar{Y}_g(1)N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^2E[N_g]^2}\\
&=-\frac{E[E[\bar{Y}_g(1)N_g|W_g]^2]-E[\bar{Y}_g(1)N_g]^2}{2E[N_g]^2}-\frac{(E[N_g^2]-E[N_g]^2)E[\bar{Y}_g(1)N_g]^2}{2E[N_g]^4}\\
&\quad-\frac{E[E[\bar{Y}_g(0)N_g|W_g]^2]-E[\bar{Y}_g(0)N_g]^2}{2E[N_g]^2}-\frac{(E[N_g^2]-E[N_g]^2)E[\bar{Y}_g(0)N_g]^2}{2E[N_g]^4}\\
&\quad+\frac{(E[E[\bar{Y}_g(1)N_g|W_g]N_g]-E[\bar{Y}_g(1)N_g]E[N_g])E[\bar{Y}_g(1)N_g]}{E[N_g]^3}\\
&\quad+\frac{(E[E[\bar{Y}_g(0)N_g|W_g]N_g]-E[\bar{Y}_g(0)N_g]E[N_g])E[\bar{Y}_g(0)N_g]}{E[N_g]^3}\\
&\quad-\frac{E[E[\bar{Y}_g(1)N_g|W_g]E[\bar{Y}_g(0)N_g|W_g]]-E[\bar{Y}_g(1)N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]E[N_g]}\\
&\quad+\frac{(E[E[\bar{Y}_g(1)N_g|W_g]N_g]-E[\bar{Y}_g(1)N_g]E[N_g])E[\bar{Y}_g(0)N_g]}{E[N_g]E[N_g]^2}\\
&\quad+\frac{(E[E[\bar{Y}_g(0)N_g|W_g]N_g]-E[\bar{Y}_g(0)N_g]E[N_g])E[\bar{Y}_g(1)N_g]}{E[N_g]^2E[N_g]}\\
&\quad-\frac{(E[N_g^2]-E[N_g]^2)E[\bar{Y}_g(1)N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^2E[N_g]^2}\\
&=-\frac{E[E[\bar{Y}_g(1)N_g|W_g]^2]}{2E[N_g]^2}-\frac{E[N_g^2]E[\bar{Y}_g(1)N_g]^2}{2E[N_g]^4}-\frac{E[E[\bar{Y}_g(0)N_g|W_g]^2]}{2E[N_g]^2}-\frac{E[N_g^2]E[\bar{Y}_g(0)N_g]^2}{2E[N_g]^4}\\
&\quad+\frac{E[E[\bar{Y}_g(1)N_g|W_g]N_g]E[\bar{Y}_g(1)N_g]}{E[N_g]^3}+\frac{E[E[\bar{Y}_g(0)N_g|W_g]N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^3}\\
&\quad-\frac{E[E[\bar{Y}_g(1)N_g|W_g]E[\bar{Y}_g(0)N_g|W_g]]}{E[N_g]^2}+\frac{E[E[\bar{Y}_g(1)N_g|W_g]N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^3}\\
&\quad+\frac{E[E[\bar{Y}_g(0)N_g|W_g]N_g]E[\bar{Y}_g(1)N_g]}{E[N_g]^3}-\frac{E[N_g^2]E[\bar{Y}_g(1)N_g]E[\bar{Y}_g(0)N_g]}{E[N_g]^4}\\
&=-\frac{1}{2}E[E[\tilde{Y}_g(1)|W_g]^2]-\frac{1}{2}E[E[\tilde{Y}_g(0)|W_g]^2]-E[E[\tilde{Y}_g(1)|W_g]E[\tilde{Y}_g(0)|W_g]]\\
&=-\frac{1}{2}E[(E[\tilde{Y}_g(1)+\tilde{Y}_g(0)|W_g])^2]\ .
\end{aligned}$$

■

Lemma B.1.2. *Suppose Q satisfies Assumptions 2.2.1 and 2.3.6 and the treatment assignment mechanism satisfies Assumptions 2.3.4–2.3.5. Define*

$$\begin{aligned}
\mathbb{L}_G^{\mathrm{YN1}} &= \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (\bar{Y}_g(1) N_g D_g - E[\bar{Y}_g(1) N_g] D_g) \\
\mathbb{L}_G^{\mathrm{N1}} &= \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (N_g D_g - E[N_g] D_g) \\
\mathbb{L}_G^{\mathrm{YN0}} &= \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (\bar{Y}_g(0) N_g (1 - D_g) - E[\bar{Y}_g(0) N_g](1 - D_g)) \\
\mathbb{L}_G^{\mathrm{N0}} &= \frac{1}{\sqrt{G}} \sum_{1 \le g \le 2G} (N_g (1 - D_g) - E[N_g](1 - D_g)) \ .
\end{aligned}$$

Then, as $G \to \infty$,

$$(\mathbb{L}_G^{\mathrm{YN1}}, \mathbb{L}_G^{\mathrm{N1}}, \mathbb{L}_G^{\mathrm{YN0}}, \mathbb{L}_G^{\mathrm{N0}})' \xrightarrow{d} N(0, \mathbb{V}) \ ,$$

where

$$\mathbb{V} = \mathbb{V}_1 + \mathbb{V}_2$$

for

$$\mathbb{V}_1 = \begin{pmatrix} \mathbb{V}_1^1 & 0 \\ 0 & \mathbb{V}_1^0 \end{pmatrix}$$

$$\begin{aligned}
\mathbb{V}_1^1 &= \begin{pmatrix} E[\mathrm{Var}[\bar{Y}_g(1) N_g | W_g]] & 0 \\ 0 & 0 \end{pmatrix} \\
\mathbb{V}_1^0 &= \begin{pmatrix} E[\mathrm{Var}[\bar{Y}_g(0) N_g | W_g]] & 0 \\ 0 & 0 \end{pmatrix}
\end{aligned}$$

$$\mathbb{V}_2 = \frac{1}{2} \mathrm{Var}[(E[\bar{Y}_g(1) N_g | W_g], N_g, E[\bar{Y}_g(0) N_g | W_g], N_g)'] \ .$$

Proof of Lemma B.1.2. Note

$$(\mathbb{L}_G^{\mathrm{YN1}}, \mathbb{L}_G^{\mathrm{N1}}, \mathbb{L}_G^{\mathrm{YN0}}, \mathbb{L}_G^{\mathrm{N0}}) = (\mathbb{L}_{1,G}^{\mathrm{YN1}}, 0, \mathbb{L}_{1,G}^{\mathrm{YN0}}, 0) + (\mathbb{L}_{2,G}^{\mathrm{YN1}}, \mathbb{L}_G^{\mathrm{N1}}, \mathbb{L}_{2,G}^{\mathrm{YN0}}, \mathbb{L}_G^{\mathrm{N0}}) \ ,$$

where

$$\begin{aligned}\mathbb{L}^{\mathrm{YN1}}_{1,G} &= \frac{1}{\sqrt{G}}\sum_{1\le g\le 2G}(\bar{Y}_g(1)N_gD_g - E[\bar{Y}_g(1)N_gD_g|N^{(G)},X^{(G)},D^{(G)}])\\ \mathbb{L}^{\mathrm{YN1}}_{2,G} &= \frac{1}{\sqrt{G}}\sum_{1\le g\le 2G}(E[\bar{Y}_g(1)N_gD_g|N^{(G)},X^{(G)},D^{(G)}] - E[\bar{Y}_g(1)N_g]D_g)\end{aligned}$$

and similarly for $\mathbb{L}^{\mathrm{YN0}}_G$. Next, note $(\mathbb{L}^{\mathrm{YN1}}_{1,G}, 0, \mathbb{L}^{\mathrm{YN0}}_{1,G}, 0), G \ge 1$ is a triangular array of normalized sums of random vectors. Conditional on $N^{(G)}, X^{(G)}, D^{(G)}$, $\mathbb{L}^{\mathrm{YN1}}_{1,G} \perp\!\!\!\perp \mathbb{L}^{\mathrm{YN0}}_{1,G}$. Moreover, it follows from $Q_G = Q^{2G}$ and Assumption 2.3.4 that

$$\mathrm{Var}\left[\mathbb{L}^{\mathrm{YN1}}_{1,G}\middle| N^{(G)}, X^{(G)}, D^{(G)}\right] = \mathrm{Var}[\bar{Y}_g(1)N_g|W_g]D_g \ .$$

We have

$$\frac{1}{G}\sum_{1\le g\le 2G}\mathrm{Var}[\bar{Y}_g(1)N_g|W_g]D_g = \frac{1}{G}\sum_{1\le g\le 2G}E[\bar{Y}^2_g(1)N^2_g|W_g]D_g - \frac{1}{G}\sum_{1\le g\le 2G}E[\bar{Y}_g(1)N_g|W_g]^2D_g \ . \tag{B.4}$$

Note

$$\begin{aligned}&\frac{1}{G}\sum_{1\le g\le 2G}E[\bar{Y}^2_g(1)N^2_g|W_g]D_g\\ &= \frac{1}{2G}\sum_{1\le g\le 2G}E[\bar{Y}^2_g(1)N^2_g|W_g] + \frac{1}{2}\Big(\frac{1}{G}\sum_{1\le g\le 2G:D_g=1}E[\bar{Y}^2_g(1)N^2_g|W_g] - \frac{1}{G}\sum_{1\le g\le 2G:D_g=0}E[\bar{Y}^2_g(1)N^2_g|W_g]\Big) \ .\end{aligned}$$

It follows from the weak law of large numbers, the application of which is permitted by Lemma C.2.3,

$$\frac{1}{2G}\sum_{1\le g\le 2G}E[\bar{Y}^2_g(1)N^2_g|W_g] \xrightarrow{P} E[\bar{Y}^2_g(1)N^2_g] \ .$$

On the other hand,

$$\begin{aligned}
&\Big|\frac{1}{G}\sum_{1\le g\le 2G:D_g=1} E[\bar{Y}_g^2(1)N_g^2|W_g] - \frac{1}{G}\sum_{1\le g\le 2G:D_g=0} E[\bar{Y}_g^2(1)N_g^2|W_g]\Big| \\
&\le \frac{1}{G}\sum_{1\le j\le G} |N_{\pi(2j-1)}^2 E[\bar{Y}_{\pi(2j-1)}^2(1)|W_{\pi(2j-1)}] - N_{\pi(2j)}^2 E[\bar{Y}_{\pi(2j)}^2(1)|W_{\pi(2j)}]| \\
&\le \frac{1}{G}\sum_{1\le j\le G} N_{\pi(2j)}^2 |E[\bar{Y}_{\pi(2j-1)}^2(1)|W_{\pi(2j-1)}] - E[\bar{Y}_{\pi(2j)}^2(1)|W_{\pi(2j)}]| \\
&\qquad + \frac{1}{G}\sum_{1\le j\le G} |N_{\pi(2j)}^2 - N_{\pi(2j-1)}^2||E[\bar{Y}_{\pi(2j-1)}^2(1)|W_{\pi(2j-1)}]| \\
&\lesssim \frac{1}{G}\sum_{1\le j\le G} N_{\pi(2j)}^2 |W_{\pi(2j-1)} - W_{\pi(2j)}| + \frac{1}{G}\sum_{1\le j\le G} |N_{\pi(2j)}^2 - N_{\pi(2j-1)}^2| \xrightarrow{P} 0 ,
\end{aligned}$$

where the first inequality follows from Assumption 2.3.4 and the triangle inequality, the second inequality by some algebraic manipulations, the final inequality by Assumption 2.3.6 and Lemma C.2.3, and the convergence in probability follows from Assumption 2.3.5 and Lemma B.2.2. Therefore,

$$\frac{1}{G}\sum_{1\le g\le 2G} E[\bar{Y}_g^2(1)N_g^2|W_g]D_g \xrightarrow{P} E[\bar{Y}_g^2(1)N_g^2] .$$

Meanwhile,

$$\begin{aligned}
&\frac{1}{G}\sum_{1\le g\le 2G} E[\bar{Y}_g(1)N_g|W_g]^2 D_g \\
&= \frac{1}{2G}\sum_{1\le g\le 2G} E[\bar{Y}_g(1)N_g|W_g]^2 + \frac{1}{2}\Big(\frac{1}{G}\sum_{1\le g\le 2G:D_g=1} E[\bar{Y}_g(1)N_g|W_g]^2 - \frac{1}{G}\sum_{1\le g\le 2G:D_g=0} E[\bar{Y}_g(1)N_g|W_g]^2\Big) .
\end{aligned}$$

It follows from the weak law of large numbers, the application of which is permitted by Lemma C.2.3 and Assumption 2.2.1(c) that

$$\frac{1}{2G}\sum_{1\le g\le 2G} E[\bar{Y}_g(1)N_g|W_g]^2 \xrightarrow{P} E[E[\bar{Y}_g(1)N_g|W_g]^2] .$$

Next,

$$
\begin{aligned}
&\Big|\frac{1}{G}\sum_{1\le g\le 2G: D_g=1} E[\bar{Y}_g(1)N_g|W_g]^2 - \frac{1}{G}\sum_{1\le g\le 2G: D_g=0} E[\bar{Y}_g(1)N_g|W_g]^2\Big| \\
&\le \frac{1}{G}\sum_{1\le j\le G} |E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|W_{\pi(2j-1)}] - E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|W_{\pi(2j)}]| \\
&\qquad\qquad\qquad\qquad \times |E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|W_{\pi(2j-1)}] + E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|W_{\pi(2j)}]| \\
&\le \Big(\frac{1}{G}\sum_{1\le j\le G} |E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|W_{\pi(2j-1)}] - E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|W_{\pi(2j)}]|^2\Big)^{1/2} \\
&\quad\cdot \Big(\frac{1}{G}\sum_{1\le j\le G} |E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|W_{\pi(2j-1)}] + E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|W_{\pi(2j)}]|^2\Big)^{1/2} \\
&\lesssim \Big(\frac{1}{G}\sum_{1\le j\le G} |E[\bar{Y}_{\pi(2j-1)}(1)N_{\pi(2j-1)}|W_{\pi(2j-1)}] - E[\bar{Y}_{\pi(2j)}(1)N_{\pi(2j)}|W_{\pi(2j)}]|^2\Big)^{1/2} \times \\
&\qquad \Big(\frac{1}{G}\sum_{1\le g\le 2G} E[\bar{Y}_g(1)N_g|W_g]^2\Big)^{1/2} \\
&\xrightarrow{P} 0 \ ,
\end{aligned}
$$

where the first inequality follows by inspection, the second follows from Cauchy-Schwarz, the third follows from $(a+b)^2 \le 2a^2 + 2b^2$, and the convergence in probability follows from Assumptions 2.3.6, 2.3.5 and the law of large numbers. Therefore,

$$
\frac{1}{G}\sum_{1\le g\le 2G} E[\bar{Y}_g(1)N_g|W_g]^2 D_g \xrightarrow{P} E\left[E[\bar{Y}_g(1)N_g|W_g]^2\right] \ ,
$$

and hence it follows from (B.4) that

$$
\frac{1}{G}\sum_{1\le g\le 2G} \text{Var}[\bar{Y}_g(1)N_g|W_g] D_g \xrightarrow{P} E[\text{Var}[\bar{Y}_g(1)N_g|W_g]] \ .
$$

Similarly,

$$
\frac{1}{G}\sum_{1\le g\le 2G} \text{Var}[\bar{Y}_g(0)N_g|W_g] D_g \xrightarrow{P} E[\text{Var}[\bar{Y}_g(0)N_g|W_g]] \ .
$$

We now establish

$$
\rho(\mathcal{L}((\mathbb{L}_{1,G}^{\text{YN1}}, 0, \mathbb{L}_{1,G}^{\text{YN0}}, 0)|W^{(G)}, D^{(G)}), N(0, \mathbb{V}_1)) \xrightarrow{P} 0 \ , \qquad \text{(B.5)}
$$

where $\mathcal{L}(\cdot)$ is used to denote the law of a random variable and ρ is any metric that metrizes weak convergence. For that purpose note that we only need to show that for any subsequence $\{G_k\}$ there exists a further subsequence $\{G_{k_l}\}$ along which

$$\rho(\mathcal{L}((\mathbb{L}^{\mathrm{YN1}}_{1,G_{k_l}}, 0, \mathbb{L}^{\mathrm{YN0}}_{1,G_{k_l}}, 0)|W^{(G_{k_l})}, D^{(G^{k_l})}, N(0, \mathbb{V}_1)) \to 0 \text{ with probability one .} \tag{B.6}$$

In order to extract such a subsequence, we verify the conditions in the Lindeberg central limit theorem in Proposition 2.27 of van der Vaart (1998). First note that by the results proved so far,

$$\mathrm{Var}[(\mathbb{L}^{\mathrm{YN1}}_{1,G}, 0, \mathbb{L}^{\mathrm{YN0}}_{1,G}, 0)'|W^{(G)}, D^{(G)}] \xrightarrow{P} \mathbb{V}_1 \ .$$

Next, We will use the inequality

$$\left|\sum_{1\le j\le k} a_j\right| I\left\{\left|\sum_{1\le j\le k} a_j\right| > \epsilon\right\} \le \sum_{1\le j\le k} k|a_j| I\left\{|a_j| > \frac{\epsilon}{k}\right\} \ . \tag{B.7}$$

It follows from (B.7) that

$$\begin{aligned}
&\frac{1}{G}\sum_{1\le g\le 2G} E[(D_g(\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g]))^2 + ((1-D_g)(\bar{Y}_g(0)N_g - E[\bar{Y}_g(0)N_g|W_g]))^2 \\
&\quad \times I\{(D_g(\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g]))^2 + ((1-D_g)(\bar{Y}_g(0)N_g - E[\bar{Y}_g(0)N_g|W_g]))^2 > \epsilon^2 G\}|W^{(G)}, D^{(G)}] \\
&\lesssim \frac{1}{G}\sum_{1\le g\le 2G} E[D_g(\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g])^2 I\{D_g(\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g])^2 > \epsilon^2 G/2\}|W^{(G)}, D^{(G)}] \\
&\qquad + \frac{1}{G}\sum_{1\le g\le 2G} E[(1-D_g)(\bar{Y}_g(0)N_g - E[\bar{Y}_g(0)N_g|W_g])^2 \\
&\qquad\qquad \times I\{(1-D_g)(\bar{Y}_g(0)N_g - E[\bar{Y}_g(0)N_g|W_g])^2 > \epsilon^2 G/2\}|W^{(G)}, D^{(G)}] \\
&\le \frac{1}{G}\sum_{1\le g\le 2G} E[(\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g])^2 I\{|\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g]| > \epsilon\sqrt{G}/\sqrt{2}\}|W_g] \\
&\qquad + \frac{1}{G}\sum_{1\le g\le 2G} E[(\bar{Y}_g(0)N_g - E[\bar{Y}_g(0)N_g|W_g])^2 I\{|\bar{Y}_g(0)N_g - E[\bar{Y}_g(0)N_g|W_g]| > \epsilon\sqrt{G}/\sqrt{2}\}|W_g] \ .
\end{aligned}$$

Fix any $m > 0$. For G large enough, the previous line

$$\begin{aligned}
&\le \frac{1}{G}\sum_{1\le g\le 2G} E[(\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g])^2 I\{|\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g]| > m\}|W_g] \\
&\qquad + \frac{1}{G}\sum_{1\le g\le 2G} E[(\bar{Y}_g(0)N_g - E[\bar{Y}_g(0)N_g|W_g])^2 I\{|\bar{Y}_g(0)N_g - E[\bar{Y}_g(0)N_g|W_g]| > m|W_g] \\
&\xrightarrow{P} 2E[(\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g])^2 I\{|\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g]| > m\}] \\
&\qquad + E[(\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g])^2 I\{|\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g]| > m\}] \ .
\end{aligned}$$

because $E[(\bar{Y}_g(1)N_g - E[\bar{Y}_g(1)N_g|W_g])^2] < \infty$ and $E[(\bar{Y}_g(0)N_g - E[\bar{Y}_g(0)N_g|W_g])^2] < \infty$. As $m \to \infty$, the last expression goes to 0. Therefore, it follows from similar arguments to those in the proof of Lemma B.3 of Bai (2022a) that both conditions in Proposition 2.27 of van der Vaart (1998) hold in probability, and therefore there must be a subsequence along which they hold almost surely, so (B.6) and hence (B.5) holds.

Next, we study $(\mathbb{L}_{2,G}^{\mathrm{YN1}}, \mathbb{L}_G^{\mathrm{N1}}, \mathbb{L}_{2,G}^{\mathrm{YN0}}, \mathbb{L}_G^{\mathrm{N0}})$. It follows from $Q_G = Q^{2G}$ and Assumption 2.3.4 that

$$\begin{pmatrix} \mathbb{L}_{2,G}^{\mathrm{YN1}} \\ \mathbb{L}_{2,G}^{\mathrm{N1}} \\ \mathbb{L}_{2,G}^{\mathrm{YN0}} \\ \mathbb{L}_{2,G}^{\mathrm{N0}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{G}} \sum_{1\le g\le 2G} D_g(E[\bar{Y}_g(1)N_g|W_g] - E[\bar{Y}_g(1)N_g]) \\ \frac{1}{\sqrt{G}} \sum_{1\le g\le 2G} D_g(N_g - E[N_g]) \\ \frac{1}{\sqrt{G}} \sum_{1\le g\le 2G} (1-D_g)(E[\bar{Y}_g(0)N_g|W_g] - E[\bar{Y}_g(0)N_g]) \\ \frac{1}{\sqrt{G}} \sum_{1\le g\le 2G} (1-D_g)(N_g - E[N_g]) \end{pmatrix} .$$

For $\mathbb{L}_{2,G}^{\mathrm{YN1}}$, it follows from similar arguments to those used above that $\mathrm{Var}[\mathbb{L}_{2,G}^{\mathrm{YN1}}|W^{(G)}] \xrightarrow{P} 0$. Therefore, it follows from Markov's inequality conditional on $W^{(G)}$ and $D^{(G)}$, and the fact that probabilities are bounded and hence uniformly integrable, that

$$\mathbb{L}_{2,G}^{\mathrm{YN1}} = E[\mathbb{L}_{2,G}^{\mathrm{YN1}}|W^{(G)}] + o_P(1) \ .$$

Applying a similar argument to each of L_G^{N1}, L_{2G}^{YN0} and L_G^{N0} allows us to conclude that

$$\begin{pmatrix} \mathbb{L}_{2,G}^{\mathrm{YN1}} \\ \mathbb{L}_{G}^{\mathrm{N1}} \\ \mathbb{L}_{2,G}^{\mathrm{YN0}} \\ \mathbb{L}_{G}^{\mathrm{N0}} \end{pmatrix} = \begin{pmatrix} \frac{1}{2\sqrt{G}} \sum_{1\le g\le 2G} (E[\bar{Y}_g(1)N_g|W_g] - E[\bar{Y}_g(1)N_g]) \\ \frac{1}{2\sqrt{G}} \sum_{1\le g\le 2G} (N_g - E[N_g]) \\ \frac{1}{2\sqrt{G}} \sum_{1\le g\le 2G} (E[\bar{Y}_g(0)N_g|W_g] - E[\bar{Y}_g(0)N_g]) \\ \frac{1}{2\sqrt{G}} \sum_{1\le g\le 2G} (N_g - E[N_g]) \end{pmatrix} + o_P(1) \ .$$

It thus follows from the central limit theorem (the application of which is justified by Assumption 2.2.1(c) and Lemma C.2.3) that

$$(\mathbb{L}_{2,G}^{\mathrm{YN1}}, \mathbb{L}_{G}^{\mathrm{N1}}, \mathbb{L}_{2,G}^{\mathrm{YN0}}, \mathbb{L}_{G}^{\mathrm{N0}})' \xrightarrow{d} N(0, \mathbb{V}_2) \ .$$

Because (C.5) holds and $(\mathbb{L}_{2,G}^{\mathrm{YN1}}, \mathbb{L}_{G}^{\mathrm{N1}}, \mathbb{L}_{2,G}^{\mathrm{YN0}}, \mathbb{L}_{G}^{\mathrm{N0}})$ is deterministic conditional on $N^{(G)}, X^{(G)}, D^{(G)}$, the conclusion of the theorem follows from Lemma S.1.3 in Bai et al. (2022c). ■

B.1.4 Proof of Theorem 3.4.2

The desired conclusion follows immediately from Lemmas B.2.4-B.2.6. ■

B.1.5 Proof of Theorem 2.4.2

By the derivation in Theorem 3.6 in Bugni et al. (2022a),

$$\hat{\omega}^2_{\mathrm{CR,G}} = \frac{1}{2}\left(\hat{\omega}^2_{\mathrm{CR,G}}(1) + \hat{\omega}^2_{\mathrm{CR,G}}(0)\right) \ , \tag{B.8}$$

(where we note that the factor of 1/2 appears since we are normalizing by the number of *pairs*), and

$$\hat{\omega}^2_{\mathrm{CR,G}}(d) := \frac{1}{\left(\frac{1}{2G}\sum_{1\le g\le 2G} N_g I\{D_g = d\}\right)^2} \frac{1}{2G} \sum_{1\le g\le 2G} \left[\left(\frac{N_g}{|\mathcal{M}_g|}\right)^2 I\{D_g = d\} \left(\sum_{i\in\mathcal{M}_g} \hat{\epsilon}_{i,g}(d)\right)^2\right] \ ,$$

with

$$\hat{\epsilon}_{i,g}(d) := Y_{i,g} - \frac{1}{\sum_{1\le g\le 2G} N_g I\{D_g = d\}} \sum_{1\le g\le 2G} N_g \bar{Y}_g I\{D_g = d\} \ .$$

Fix $d \in \{0,1\}$, $r \in \{0,1,2\}$, $\ell \in \{1,2\}$ arbitrarily. Then by Lemma S.1.5 in Bai et al. (2022c) applied to the observations $(N_g^\ell \bar{Y}_g^r(d) : 1 \le g \le 2G)$,

$$\frac{1}{2G} \sum_{1\le g\le 2G} N_g^\ell \bar{Y}_g^r(d) I\{D_g = d\} \xrightarrow{P} \frac{E[N^l \bar{Y}_g^r(d)]}{2} \ .$$

The result then follows by an identical derivation to that of Theorem 3.6 in Bugni et al. (2022a). ■

B.1.6 Proof of Theorem 2.4.3

Let $\mathbf{1}_K$ denote a column of ones of length K. Then consider the following cluster-robust variance estimator where clusters are defined at the level of the *pair*:

$$\left(\frac{1}{G}\sum_{1\le j\le G}\sum_{g\in\lambda_j} X_g' X_g\right)^{-1}\left(\frac{1}{G}\sum_{1\le j\le G}\left(\sum_{g\in\lambda_j} X_g'\hat{\epsilon}_g\right)\left(\sum_{g\in\lambda_j} X_g'\hat{\epsilon}_g\right)'\right)\left(\frac{1}{G}\sum_{1\le g\le G}\sum_{g\in\lambda_j} X_g' X_g\right)^{-1}, \qquad \text{(B.9)}$$

where $\lambda_j := \{\pi(2j-1), \pi(2j)\}$, and

$$\begin{aligned} X_g &:= \left(\begin{array}{cc} \mathbf{1}_{|\mathcal{M}_g|}\cdot\sqrt{\frac{N_g}{|\mathcal{M}_g|}}, & \mathbf{1}_{|\mathcal{M}_g|}\cdot\sqrt{\frac{N_g}{|\mathcal{M}_g|}}D_g \end{array}\right) \\ \hat{\epsilon}_g &:= \sqrt{\frac{N_g}{|\mathcal{M}_g|}}\left(Y_{i,g} - (\hat{\mu}_G(1) - \hat{\mu}_G(0))D_g - \hat{\mu}_G(0)\ :\ i\in\mathcal{M}_g\right)'\ . \end{aligned}$$

Imposing the condition that $N_g = k$ are equal and fixed and $|\mathcal{M}_g| = N_g$, and then following the algebra in, for instance, the proof of Theorem 3.4 in Bai et al. (2023c), it can be shown that

$$\hat{\omega}^2_{\text{PCVE,G}} = \frac{1}{G}\sum_{1\le j\le G}\left(\sum_{g\in\lambda_j}\bar{Y}_g I\{D_g = 1\} - \sum_{g\in\lambda_j}\bar{Y}_g I\{D_g = 0\}\right)^2 - (\hat{\mu}_G(1) - \hat{\mu}_G(0))^2\ .$$

By Lemmas S.1.5 and S.1.6 of Bai et al. (2022c) applied to the observations $(\bar{Y}_g(d) : 1\le g\le 2G)$, and the continuous mapping theorem, we thus obtain that

$$\begin{aligned} \hat{\omega}^2_{\text{PCVE,G}} \xrightarrow{P}\ & E[\text{Var}[\bar{Y}_g(1)|X_g]] + E[\text{Var}[\bar{Y}_g(1)|X_g]] \\ & + E[\left((E[\bar{Y}_g(1)|X_g] - E[\bar{Y}_g(1)]) - (E[\bar{Y}_g(0)|X_g] - E[\bar{Y}_g(0)])\right)^2]\ . \end{aligned}$$

Simplifying using the law of total variance and the fact that $\tilde{Y}_g(d) = \bar{Y}_g(d) - E[\bar{Y}_g(d)]$ once we impose that $N_g = k$, we then obtain

$$\hat{\omega}^2_{\text{PCVE,G}} \xrightarrow{P} E[\tilde{Y}_g^2(1)] + E[\tilde{Y}_g^2(0)] - \frac{1}{2}E[(E[\tilde{Y}_g(1) + \tilde{Y}_g(0)|X_g])^2] + \frac{1}{2}E\left[(E[\tilde{Y}_g(1) - \tilde{Y}_g(0)|X_g])^2\right]\ .$$

The conclusion then follows. ■

B.1.7 Proof of Theorem 2.5.1

Proof. Note that the null hypothesis (2.8) combined with Assumption 2.2.1(e) implies that

$$\bar{Y}_g(1)|(X_g, N_g) \stackrel{d}{=} \bar{Y}_g(0)|(X_g, N_g) \ . \tag{B.10}$$

If the assignment mechanism satisfies Assumption 2.3.4, the result then follows by applying Theorem 3.4 in Bai et al. (2022c) to the cluster-level outcomes $\{(\bar{Y}_g, D_g, X_g, N_g) : 1 \leq g \leq 2G\}$. If instead the assignment mechanism satisfies Assumption 2.3.1, then note that (B.10) is in fact equivalent to the statement

$$(\bar{Y}_g(1), N_g)|X_g \stackrel{d}{=} (\bar{Y}_g(0), N_g)|X_g \ . \tag{B.11}$$

The result then follows by applying Theorem 3.4 in Bai et al. (2022c) using (B.11) as the null hypothesis. To establish this equivalence, we first begin with (B.10) and verify that for any Borel sets A and B,

$$P\{\bar{Y}_g(1) \in A, N_g \in B|X_g\} = P\{\bar{Y}_g(0) \in A, N_g \in B|X_g\} \text{ a.s.}$$

By the definition of a conditional expectation, note we only need to verify for all Borel sets C,

$$E[P\{\bar{Y}_g(1) \in A, N_g \in B|X_g\}I\{X_g \in C\}] = P\{\bar{Y}_g(0) \in A, N_g \in B, X_g \in C\} \ .$$

We have

$$\begin{aligned}
&E[P\{\bar{Y}_g(1) \in A, N_g \in B|X_g\}I\{X_g \in C\}] \\
&= P\{\bar{Y}_g(1) \in A, N_g \in B, X_g \in C\} \\
&= E[P\{\bar{Y}_g(1) \in A|X_g, N_g\}I\{N_g \in B\}I\{X_g \in C\}] \\
&= E[P\{\bar{Y}_g(0) \in A|X_g, N_g\}I\{N_g \in B\}I\{X_g \in C\}] \\
&= P\{\bar{Y}_g(0) \in A, N_g \in B, X_g \in C\} \ ,
\end{aligned}$$

where the first and second equalities follow from the definition of conditional expectations, the the third follows from (B.10), and the last follows again from the definition of a conditional expectation. The opposite implication follows from a similar argument and is thus omitted. ■

B.1.8 Proof of Theorem 2.5.2

Note that

$$\begin{aligned}
\sqrt{G}\hat{\Delta}_G &= \sqrt{G}\left(\frac{1}{N(1)}\sum_{1\le g\le 2G} D_g N_g \bar{Y}_g - \frac{1}{N(0)}\sum_{1\le g\le 2G}(1-D_g)N_g\bar{Y}_g\right) \\
&= \frac{1}{N(1)}\sqrt{G}\sum_{1\le g\le 2G}\big(D_g N_g\bar{Y}_g - (1-D_g)N_g\bar{Y}_g\big) + \left(\frac{1}{N(1)} - \frac{1}{N(0)}\right)\sqrt{G}\sum_{1\le g\le 2G}(1-D_g)N_g\bar{Y}_g \\
&= \frac{1}{N(1)/G}\frac{1}{\sqrt{G}}\sum_{1\le j\le G}\left(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)}\right)\left(D_{\pi(2j)} - D_{\pi(2j-1)}\right) \\
&\qquad + \frac{\frac{1}{\sqrt{G}}(N(0)-N(1))}{\frac{N(1)}{G}\frac{N(0)}{G}}\frac{1}{G}\sum_{1\le g\le 2G}(1-D_g)N_g\bar{Y}_g \\
&= \frac{1}{N(1)/G}\frac{1}{\sqrt{G}}\sum_{1\le j\le G}\left(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)}\right)\left(D_{\pi(2j)} - D_{\pi(2j-1)}\right) \\
&\qquad - \frac{\frac{1}{\sqrt{G}}\sum_{1\le j\le G}(N_{\pi(2j)} - N_{\pi(2j-1)})(D_{\pi(2j)} - D_{\pi(2j-1)})}{\frac{N(1)}{G}\frac{N(0)}{G}}\frac{1}{G}\sum_{1\le g\le 2G}(1-D_g)N_g\bar{Y}_g\ .
\end{aligned}$$

Hence the randomization distribution of $\sqrt{G}\hat{\Delta}_G$ is given by

$$\tilde{R}_G(t) := P\left\{\sqrt{G}\check{\Delta}(\epsilon_1,\ldots,\epsilon_G) \le t \,\middle|\, Z^{(G)}\right\}\ , \tag{B.12}$$

where

$$\begin{aligned}
\sqrt{G}\check{\Delta}(\epsilon_1,\ldots,\epsilon_G) &= \frac{1}{\tilde{N}(1)/G}\frac{1}{\sqrt{G}}\sum_{1\le j\le G}\epsilon_j\left(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)}\right)\left(D_{\pi(2j)} - D_{\pi(2j-1)}\right) \\
&\qquad - \frac{\frac{1}{\sqrt{G}}\sum_{1\le j\le G}\epsilon_j(N_{\pi(2j)} - N_{\pi(2j-1)})(D_{\pi(2j)} - D_{\pi(2j-1)})}{\frac{\tilde{N}(1)}{G}\frac{\tilde{N}(0)}{G}}\frac{1}{G}\sum_{1\le g\le 2G}(1-\tilde{D}_g)N_g\bar{Y}_g\ ,
\end{aligned}$$

ϵ_j, $j = 1,\ldots,G$ are i.i.d. Rademacher random variables generated independently of $Z^{(G)}$, $\{\tilde{D}_g : 1 \le g \le 2G\}$ denotes the assignment of cluster g after applying the transformation implied by $\{\epsilon_j : 1 \le j \le G\}$, and

$$\tilde{N}(d) = \sum_{1\le g\le 2G} N_g I\{\tilde{D}_g = d\}\ .$$

By construction, $\hat{v}_G^2$ evaluated at the transformation of the data implied by $\{\epsilon_j : 1 \leq j \leq G\}$ is given by

$$\check{v}_G^2(\epsilon_1, \ldots, \epsilon_G) = \hat{\tau}_G^2 - \frac{1}{2}\check{\lambda}_G^2\left(\epsilon_1, \ldots, \epsilon_G\right) \tag{B.13}$$

where $\hat{\tau}_G^2$ is defined in (2.5), and

$$\begin{aligned}&\check{\lambda}_G^2\left(\epsilon_1, \ldots, \epsilon_G\right) = \\ &\qquad \frac{2}{G}\sum_{1 \leq j \leq \lfloor G/2 \rfloor} \epsilon_{2j-1}\epsilon_{2j}\Big(\left(\hat{Y}_{\pi(4j-3)} - \hat{Y}_{\pi(4j-2)}\right)\left(\hat{Y}_{\pi(4j-1)} - \hat{Y}_{\pi(4j)}\right) \times \\ &\qquad\qquad \left(D_{\pi(4j-3)} - D_{\pi(4j-2)}\right)\left(D_{\pi(4j-1)} - D_{\pi(4j)}\right)\Big) \ .\end{aligned}$$

The desired conclusion then follows from Lemmas B.2.7 and B.2.9, along with Theorem 5.2 in Chung and Romano (2013). ■

B.1.9 Proof of Theorem 2.6.1

We first show that $\hat{\beta}_G \xrightarrow{P} \beta^*$. The proof follows almost verbatim Theorem 4.2 in Bai et al. (2023a) with a few minor differences because we match on N_g, which could all be resolved as in the proof of Lemma C.3.1. To establish the limiting distribution, first define

$$\bar{\psi}_{d,G} = \frac{1}{G}\sum_{1 \leq g \leq 2G} \psi_g I\{D_g = d\}$$

for $d \in \{0, 1\}$. Note that

$$\begin{aligned}&\frac{1}{G}\sum_{1 \leq g \leq 2G}(\bar{Y}_g(1)N_g - (\psi_g - \bar{\psi}_G)'\hat{\beta}_G)D_g \\ &= \frac{1}{G}\sum_{1 \leq g \leq 2G}(\bar{Y}_g(1)N_g - (\psi_g - \bar{\psi}_G)'\beta^*)D_g - \frac{1}{G}\sum_{1 \leq g \leq 2G}(\psi_g - \bar{\psi}_{1,G})'(\hat{\beta}_G - \beta^*)D_g - (\bar{\psi}_{1,G} - \bar{\psi}_G)'(\hat{\beta}_G - \beta^*) \\ &= \frac{1}{G}\sum_{1 \leq g \leq 2G}(\bar{Y}_g(1)N_g - (\psi_g - \bar{\psi}_G)'\beta^*)D_g - O_P(G^{-1/2})o_P(1) \\ &= \frac{1}{G}\sum_{1 \leq g \leq 2G}(\bar{Y}_g(1)N_g - (\psi_g - \bar{\psi}_G)'\beta^*)D_g + o_P(G^{-1/2}) \\ &= \frac{1}{G}\sum_{1 \leq g \leq 2G}(\bar{Y}_g(1)N_g - (\psi_g - E[\psi_g])'\beta^*)D_g - (\bar{\psi}_G - E[\psi_g])'\beta^* + o_P(G^{-1/2}) \ .\end{aligned}$$

where the second equality follows because $\hat{\beta}_G - \beta^* = o_P(1)$,

$$\frac{1}{G}\sum_{1\le g\le 2G}(\psi_g - \bar{\psi}_{1,G})D_g = 0 \ ,$$

and

$$\sqrt{G}(\bar{\psi}_{1,G} - \bar{\psi}_G) = O_P(1) \ .$$

The last equality follows from the arguments that establish (50) in Bai et al. (2023a). Define

$$\tilde{\Delta}_G^{\text{adj}} = \frac{\frac{1}{G}\sum_{1\le g\le 2G}(\bar{Y}_g(1)N_g - (\psi_g - E[\psi_g])'\beta^*)D_g}{\frac{1}{G}\sum_{1\le g\le 2G}N_gD_g} - \frac{\frac{1}{G}\sum_{1\le g\le 2G}(\bar{Y}_g(0)N_g - (\psi_g - E[\psi_g])'\beta^*)(1-D_g)}{\frac{1}{G}\sum_{1\le g\le 2G}N_g(1-D_g)} \ .$$

It follows from previous arguments that

$$\begin{aligned}
&\sqrt{G}(\hat{\Delta}_G^{\text{adj}} - \Delta) - \sqrt{G}(\tilde{\Delta}_G^{\text{adj}} - \Delta) \\
&= \sqrt{G}(\bar{\psi}_G - E[\psi_g])'\beta^*\left(\frac{1}{\frac{1}{G}\sum_{1\le g\le 2G}N_gD_g} - \frac{1}{\frac{1}{G}\sum_{1\le g\le 2G}N_g(1-D_g)}\right) + o_P(1) \\
&= o_P(1) \ .
\end{aligned}$$

Recall that

$$\nu^2 = E[\text{Var}[\tilde{Y}_g(1)|X_g,N_g]] + E[\text{Var}[\tilde{Y}_g(0)|X_g,N_g]] + \frac{1}{2}E[(E[\tilde{Y}_g(1) - \tilde{Y}_g(0)|X_g,N_g] - \Delta)^2] \ .$$

It then follows from the proof of Theorem 2.3.2 that $\sqrt{G}(\hat{\Delta}_G^{\text{adj}} - \Delta) \xrightarrow{d} N(0,\varsigma^2)$, where

$$\varsigma^2 = E[\text{Var}[Y_g^*(1)|X_g,N_g]] + E[\text{Var}[Y_g^*(0)|X_g,N_g]] + \frac{1}{2}E[(E[Y_g^*(1) - Y_g^*(0)|X_g,N_g] - \Delta)^2] \ ,$$

where

$$Y_g^*(d) = \frac{\bar{Y}_g(d)N_g - (\psi_g - E[\psi_g])'\beta^*}{E[N_g]} - \frac{N_g}{E[N_g]}\frac{E[\bar{Y}_g(d)N_g - (\psi_g - E[\psi_g])'\beta^*]}{E[N_g]} = \tilde{Y}_g(d) - \frac{(\psi_g - E[\psi_g])'\beta^*}{E[N_g]}$$

for $d \in \{0,1\}$. All relevant assumptions for Theorem 2.3.2 have their counterparts stated in Theorem 2.6.1.

Next we show that $\varsigma^2 \le \nu^2$. First note that by definition it follows immediately that

$$E[(E[\tilde{Y}_g(1) - \tilde{Y}_g(0)|X_g,N_g] - \Delta)^2] = E[(E[Y_g^*(1) - Y_g^*(0)|X_g,N_g] - \Delta)^2] \ .$$

It thus remains to show that

$$E[\text{Var}[Y_g^*(1)|X_g,N_g]]+E[\text{Var}[Y_g^*(0)|X_g,N_g]]\le E[\text{Var}[\tilde{Y}_g(1)|X_g,N_g]]+E[\text{Var}[\tilde{Y}_g(0)|X_g,N_g]]\ .$$

To that end,

$$\begin{aligned}
&E[\text{Var}[Y_g^*(1)|X_g,N_g]]+E[\text{Var}[Y_g^*(0)|X_g,N_g]]\\
&=E\left[\text{Var}\left[\tilde{Y}_g(1)-\frac{(\psi_g-E[\psi_g])'\beta^*}{E[N_g]}\Big|X_g,N_g\right]\right]+E\left[\text{Var}\left[\tilde{Y}_g(0)-\frac{(\psi_g-E[\psi_g])'\beta^*}{E[N_g]}\Big|X_g,N_g\right]\right]\\
&=E[\text{Var}[\tilde{Y}_g(1)|X_g,N_g]]+E[\text{Var}[\tilde{Y}_g(0)|X_g,N_g]]-\frac{2E[((\psi_g-E[\psi_g|X_g,N_g])'\beta^*)^2]}{E[N_g]^2}\\
&\qquad-2E[\text{Cov}[N_g,\psi_g'\beta^*|X_g,N_g]]\frac{E[\bar{Y}_g(1)N_g]+E[\bar{Y}_g(0)N_g]}{E[N_g]^3}\ ,
\end{aligned}$$

where the first equality follows by definition, the second equality by noting that β^* is the projection coefficient of $\frac{1}{2}(\bar{Y}_g(1)N_g+\bar{Y}_g(0)N_g-E[\bar{Y}_g(1)N_g+\bar{Y}_g(0)N_g|X_g,N_g])$ on $\psi_g-E[\psi_g|X_g,N_g]$,

$$\begin{aligned}
&E[(\bar{Y}_g(1)N_g+\bar{Y}_g(0)N_g-E[\bar{Y}_g(1)N_g+\bar{Y}_g(0)N_g|X_g,N_g])(\psi_g-E[\psi_g|X_g,N_g])'\beta^*]\\
&=2E[((\psi_g-E[\psi_g|X_g,N_g])'\beta^*)^2]\ ,
\end{aligned}$$

or equivalently,

$$E[\text{Cov}[\bar{Y}_g(1)N_g+\bar{Y}_g(0)N_g,\psi_g'\beta^*|X_g,N_g]]=2E[\text{Var}[\psi_g'\beta^*|X_g,N_g]]\ . \tag{B.14}$$

We thus obtain

$$\varsigma^2=\nu^2-\kappa^2$$

once we notice that $\text{Cov}[N_g,\psi_g'\beta^*|X_g,N_g]=0$, as desired. Finally, note that if we do not match on N_g, then we have that

$$\begin{aligned}
&E[\text{Var}[Y_g^*(1)|X_g]]+E[\text{Var}[Y_g^*(0)|X_g]]\\
&=E[\text{Var}[\tilde{Y}_g(1)|X_g]]+E[\text{Var}[\tilde{Y}_g(0)|X_g]]-\frac{2E[((\psi_g-E[\psi_g|X_g])'\beta^*)^2]}{E[N_g]^2}\\
&\qquad-2E[\text{Cov}[N_g,\psi_g'\beta^*|X_g]]\frac{E[\bar{Y}_g(1)N_g]+E[\bar{Y}_g(0)N_g]}{E[N_g]^3}\ ,
\end{aligned}$$

but the last term no longer necessarily evaluates to zero.

B.1.10 Proof of Theorem 2.6.2

The theorem follows from combining the arguments used to establish Theorem 3.4.2 and those used to establish Theorem 3.2 in Bai et al. (2023a). ■

B.2 Auxiliary Lemmas

Lemma B.2.1. *If Assumption 2.2.1 holds,*

$$\left|E[\bar{Y}_g^r(d)|X_g, N_g]\right| \leq C \quad a.s. \ ,$$

for $r \in \{1, 2\}$ for some constant $C > 0$,

$$E\left[\bar{Y}_g^r(d) N_g^\ell\right] < \infty \ ,$$

for $r \in \{1, 2\}, \ell \in \{0, 1, 2\}$, and

$$E\left[E[\bar{Y}_g(d) N_g | X_g]^2\right] < \infty \ .$$

Proof. We show the first statement for $r = 2$, since the case $r = 1$ follows similarly. By the Cauchy-Schwarz inequality,

$$\bar{Y}_g(d)^2 = \left(\frac{1}{|\mathcal{M}_g|} \sum_{i \in \mathcal{M}_g} Y_{i,g}(d)\right)^2 \leq \frac{1}{|\mathcal{M}_g|} \sum_{i \in \mathcal{M}_g} Y_{i,g}(d)^2 \ ,$$

and hence

$$\left|E[\bar{Y}_g(d)^2|X_g, N_g]\right| \leq E\left[\frac{1}{|\mathcal{M}_g|} \sum_{i \in \mathcal{M}_g} E[Y_{i,g}(d)^2|X_g, N_g]\middle| X_g, N_g\right] \leq C \ ,$$

where the first inequality follows from the above derivation, Assumption 2.2.1(e) and the law of iterated expectations, and final inequality follows from Assumption 2.2.1(d). We show the next statement for $r = \ell = 2$, since the other cases follow similarly. By the law of iterated expectations,

$$\begin{aligned} E\left[\bar{Y}_g^2(d) N_g^2\right] &= E\left[N_g^2 E[\bar{Y}_g^2(d)|X_g, N_g]\right] \\ &\lesssim E\left[N_g^2\right] < \infty \ , \end{aligned}$$

where the final line follows by Assumption 2.2.1 (c). Finally,

$$\begin{aligned} E\left[E[\bar{Y}_g(d)N_g|X_g]^2\right] &= E\left[E[N_g E[\bar{Y}_g(d)|X_g, N_g]|X_g]^2\right] \\ &\lesssim E\left[E[N_g|X_g]^2\right] < \infty \ , \end{aligned}$$

where the final line follows from Jensen's inequality and Assumption 2.2.1(c). ■

Lemma B.2.2. *If Assumptions 2.2.1 and 2.3.5 hold,*

$$\frac{1}{G}\sum_{g=1}^{G}\left|N_{\pi(2g)}^2 - N_{\pi(2g-1)}^2\right| \xrightarrow{p} 0 \ .$$

Proof.

$$\begin{aligned} \frac{1}{G}\sum_{g=1}^{G}\left|N_{\pi(2g)}^2 - N_{\pi(2g-1)}^2\right| &= \frac{1}{G}\sum_{g=1}^{G}\left|N_{\pi(2g)} - N_{\pi(2g-1)}\right|\left|N_{\pi(2g)} + N_{\pi(2g-1)}\right| \\ &\leq \left[\left(\frac{1}{G}\sum_{g=1}^{G}\left|N_{\pi(2g)} - N_{\pi(2g-1)}\right|^2\right)\left(\frac{1}{G}\sum_{g=1}^{G}\left|N_{\pi(2g)} + N_{\pi(2g-1)}\right|^2\right)\right]^{1/2} \ , \end{aligned}$$

where the inequality follows by Cauchy-Schwarz. It follows from an argument similar to the proof of Proposition 2.3.1 that $\frac{1}{G}\sum_{g=1}^{G}\left|N_{\pi(2g)} + N_{\pi(2g-1)}\right|^2 = O_P(1)$. By Assumption 2.3.5, $\frac{1}{G}\sum_{g=1}^{G}\left|N_{\pi(2g)} - N_{\pi(2g-1)}\right|^2 \xrightarrow{p} 0$. Hence the result follows. ■

Lemma B.2.3. *If Assumptions 2.2.1 holds, and additionally Assumptions 2.3.2-2.3.3, 2.4.1 (or Assumptions 2.3.5-2.3.6, 2.4.2) hold, then*

1. $E\left[\tilde{Y}_g^2(d)\right] < \infty$ *for* $d \in \{0,1\}$.
2. $((\tilde{Y}_g(1), \tilde{Y}_g(0)) : 1 \leq g \leq 2G) \perp D^{(G)} \mid X^{(G)}$ *(or* $((\tilde{Y}_g(1), \tilde{Y}_g(0)) : 1 \leq g \leq 2G) \perp D^{(G)} \mid W^{(G)}$*)*
3. $\frac{1}{G}\sum_{j=1}^{G}\left|\mu_d(X_{\pi(2j)}) - \mu_d(X_{\pi(2j-1)})\right| \xrightarrow{P} 0$, *where we use* $\mu_d(X_g)$ *to denote* $E[\tilde{Y}_g(d) \mid X_g]$ *for* $d \in \{0,1\}$.

   (or $\frac{1}{G}\sum_{j=1}^{G}\left|\mu_d(W_{\pi(2j)}) - \mu_d(W_{\pi(2j-1)})\right| \xrightarrow{P} 0$*)*
4. $\frac{1}{G}\sum_{j=1}^{G}\left|\left(\mu_1(X_{\pi(2j)}) - \mu_1(X_{\pi(2j-1)})\right)\left(\mu_0(X_{\pi(2j)}) - \mu_0(X_{\pi(2j-1)})\right)\right| \xrightarrow{P} 0$.

   (or $\frac{1}{G}\sum_{j=1}^{G}\left|\left(\mu_1(W_{\pi(2j)}) - \mu_1(W_{\pi(2j-1)})\right)\left(\mu_0(W_{\pi(2j)}) - \mu_0(W_{\pi(2j-1)})\right)\right| \xrightarrow{P} 0$*)*
5. $\frac{1}{4G}\sum_{k\in\{2,3\},\ell\in\{0,1\}}\sum_{1\leq j\leq\frac{G}{2}}\left(\mu_d\left(X_{\pi(4j-\ell)}\right) - \mu_d\left(X_{\pi(4j-k)}\right)\right)^2 \xrightarrow{P} 0$.

   (or $\frac{1}{4G}\sum_{k\in\{2,3\},\ell\in\{0,1\}}\sum_{1\leq j\leq\frac{G}{2}}\left(\mu_d\left(W_{\pi(4j-\ell)}\right) - \mu_d\left(W_{\pi(4j-k)}\right)\right)^2 \xrightarrow{P} 0$*)*

Proof. Note that

$$
\begin{aligned}
E\left[\tilde{Y}_g^2(d)\right] &\le E\left[N_g^2\left(\bar{Y}_g(d)-\frac{E\left[\bar{Y}_g(d)N_g\right]}{E\left[N_g\right]}\right)^2\right] \\
&\lesssim E\left[N_g^2\bar{Y}_g^2(d)\right]+\left(\frac{E\left[\bar{Y}_g(d)N_g\right]}{E\left[N_g\right]}\right)^2 E[N_g^2]<\infty
\end{aligned}
$$

where the inequality follows by Lemma C.2.3. The second result follows directly by inspection and Assumption 2.3.4 (or Assumption 2.3.1). In terms of the third result, by Assumption 2.3.3 and 2.3.2,

$$
\frac{1}{G}\sum_{j=1}^{G}\left|\mu_1(X_{\pi(2j)})-\mu_1(X_{\pi(2j-1)})\right| \lesssim \frac{1}{G}\sum_{j=1}^{G}\left|X_{\pi(2j)}-X_{\pi(2j-1)}\right| \xrightarrow{P} 0\ .
$$

Meanwhile,

$$
\begin{aligned}
&\frac{1}{G}\sum_{j=1}^{G}\left|\mu_1(W_{\pi(2j)})-\mu_1(W_{\pi(2j-1)})\right| \\
&\lesssim \frac{1}{G}\sum_{j=1}^{G}\left|E[N_{\pi(2j)}\bar{Y}_{\pi(2j)}(d)\mid W_{\pi(2j)}]-E[N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)}(d)\mid W_{\pi(2j-1)}]\right| \\
&\qquad+\frac{1}{G}\sum_{j=1}^{G}\left|E[N_{\pi(2j)}\mid W_{\pi(2j)}]-E[N_{\pi(2j-1)}\mid W_{\pi(2j-1)}]\right| \\
&\lesssim \frac{1}{G}\sum_{j=1}^{G}\left|N_{\pi(2j)}\left(E[\bar{Y}_{\pi(2j)}(d)\mid W_{\pi(2j)}]-E[\bar{Y}_{\pi(2j-1)}(d)\mid W_{\pi(2j-1)}]\right)\right|+\frac{1}{G}\sum_{j=1}^{G}\left|N_{\pi(2j)}-N_{\pi(2j-1)}\right| \\
&\qquad+\frac{1}{G}\sum_{j=1}^{G}\left|(N_{\pi(2j)}-N_{\pi(2j-1)})E[\bar{Y}_{\pi(2j-1)}(d)\mid W_{\pi(2j-1)}]\right| \\
&\lesssim \frac{1}{G}\sum_{j=1}^{G}N_{\pi(2j)}\left|W_{\pi(2j)}-W_{\pi(2j-1)}\right|\ ,
\end{aligned}
$$

which converges to zero in probability by Assumption 2.3.5. To prove the fourth result, by Assumption 2.3.3 and 2.3.2,

$$
\frac{1}{G}\sum_{j=1}^{G}\left|\left(\mu_1(X_{\pi(2j)})-\mu_1(X_{\pi(2j-1)})\right)\left(\mu_0(X_{\pi(2j)})-\mu_0(X_{\pi(2j-1)})\right)\right| \lesssim \frac{1}{G}\sum_{j=1}^{G}\left|X_{\pi(2j)}-X_{\pi(2j-1)}\right|^2 \xrightarrow{P} 0\ .
$$

Similarly,

$$
\begin{aligned}
&\frac{1}{G}\sum_{j=1}^{G}\left|\left(\mu_1(W_{\pi(2j)})-\mu_1(W_{\pi(2j-1)})\right)\left(\mu_0(W_{\pi(2j)})-\mu_0(W_{\pi(2j-1)})\right)\right| \\
&\qquad\le \frac{1}{G}\sum_{j=1}^{G}\left|\mu_1(W_{\pi(2j)})-\mu_1(W_{\pi(2j-1)})\right|\left|\mu_0(W_{\pi(2j)})-\mu_0(W_{\pi(2j-1)})\right| \\
&\qquad\lesssim \frac{1}{G}\sum_{j=1}^{G}N_{\pi(2j)}^2\left|W_{\pi(2j)}-W_{\pi(2j-1)}\right|^2 \xrightarrow{P} 0 \ ,
\end{aligned}
$$

where the last step follows by Assumption 2.3.5. Finally, fifth result follows the same argument by Assumption 2.4.2 (or Assumption 2.4.1). ■

Lemma B.2.4. *Consider the following adjusted potential outcomes:*

$$\hat{Y}_g(d)=\frac{N_g}{\frac{1}{2G}\sum_{1\le j\le 2G}N_j}\left(\bar{Y}_g(d)-\frac{\frac{1}{G}\sum_{1\le j\le 2G}\bar{Y}_j(d)I\{D_j=d\}N_j}{\frac{1}{G}\sum_{1\le j\le 2G}I\{D_j=d\}N_j}\right) \ .$$

Note the usual relationship still holds for adjusted outcomes, i.e. $\hat{Y}_g=D_g\hat{Y}_g(1)+(1-D_g)\hat{Y}_g(0)$*. If Assumptions 2.2.1 holds, and additionally Assumptions 2.3.2–2.3.3 (or Assumptions 2.3.5–2.3.6) hold, then*

$$
\begin{aligned}
\hat{\mu}_G(d)&=\frac{1}{G}\sum_{1\le g\le 2G}\hat{Y}_g(d)I\left\{D_g=d\right\}\xrightarrow{P}0 \\
\hat{\sigma}_G^2(d)&=\frac{1}{G}\sum_{1<g<2G}\left(\hat{Y}_g-\hat{\mu}_G(d)\right)^2 I\left\{D_g=d\right\}\xrightarrow{P}\mathrm{Var}\left[\tilde{Y}_g(d)\right] \ .
\end{aligned}
$$

Proof. It suffices to show that

$$\frac{1}{G}\sum_{1\le g\le 2G}\hat{Y}_g^r(d)I\left\{D_g=d\right\}\xrightarrow{P}E\left[\tilde{Y}_g^r(d)\right]$$

for $r\in\{1,2\}$. We prove this result only for $r=1$ and $d=1$; the other cases can be proven similarly. To this end, write

$$\frac{1}{G}\sum_{1\le g\le 2G}\hat{Y}_g(1)I\left\{D_g=1\right\}=\frac{1}{G}\sum_{1\le g\le 2G}\hat{Y}_g(1)D_g=\frac{1}{G}\sum_{1\le g\le 2G}\tilde{Y}_g(1)D_g+\frac{1}{G}\sum_{1\le g\le 2G}\left(\hat{Y}_g(1)-\tilde{Y}_g(1)\right)D_g \ .$$

Note that

$$\frac{1}{G}\sum_{1\le g\le 2G}\left(\hat{Y}_g(1)-\tilde{Y}_g(1)\right)D_g=\left(\frac{1}{\frac{1}{2G}\sum_{1\le g\le 2G}N_g}-\frac{1}{E[N_g]}\right)\left(\frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_g(1)N_gD_g\right)$$
$$-\left(\frac{\frac{1}{G}\sum_{1\le g\le 2G}\bar{Y}_g(d)I\{D_g=d\}N_g}{\left(\frac{1}{2G}\sum_{1\le g\le 2G}N_g\right)^2}-\frac{E[\bar{Y}_g(d)N_g]}{E[N_g]^2}\right)\left(\frac{1}{G}\sum_{1\le g\le 2G}N_gD_g\right)$$

By weak law of large number, Lemma B.1.2 (or Lemma B.1.1) and Slutsky's theorem, we have

$$\frac{1}{G}\sum_{1\le g\le 2G}\left(\hat{Y}_g(1)-\tilde{Y}_g(1)\right)D_g\xrightarrow{P}0\ .$$

By applying Lemma S.1.5 from Bai et al. (2022c) and Lemma C.3.1, we have

$$\frac{1}{G}\sum_{1\le g\le 2G}\tilde{Y}_g(d)D_g\xrightarrow{P}E\left[\tilde{Y}_g(d)\right]=0\ .$$

Thus, the result follows. ■

Lemma B.2.5. *If Assumptions 2.2.1 holds, and Assumptions 2.3.2-2.3.3 hold, then*

$$\hat{\tau}_G^2\xrightarrow{P}E\left[\text{Var}\left[\tilde{Y}_g(1)\mid X_g\right]\right]+E\left[\text{Var}\left[\tilde{Y}_g(0)\mid X_g\right]\right]+E\left[\left(E\left[\tilde{Y}_g(1)\mid X_g\right]-E\left[\tilde{Y}_g(0)\mid X_g\right]\right)^2\right]$$

in the case where we match on cluster size. Instead, if Assumptions 2.2.1 and 2.3.5-2.3.6 hold, then

$$\hat{\tau}_G^2\xrightarrow{P}E\left[\text{Var}\left[\tilde{Y}_g(1)\mid W_g\right]\right]+E\left[\text{Var}\left[\tilde{Y}_g(0)\mid W_g\right]\right]+E\left[\left(E\left[\tilde{Y}_g(1)\mid W_g\right]-E\left[\tilde{Y}_g(0)\mid W_g\right]\right)^2\right]$$

in the case where we do not match on cluster size.

Proof. Note that

$$\hat{\tau}_G^2=\frac{1}{G}\sum_{1\le j\le G}\left(\hat{Y}_{\pi(2j)}-\hat{Y}_{\pi(2j-1)}\right)^2=\frac{1}{G}\sum_{1\le g\le 2G}\hat{Y}_g^2-\frac{2}{G}\sum_{1\le j\le G}\hat{Y}_{\pi(2j)}\hat{Y}_{\pi(2j-1)}.$$

Since

$$\frac{1}{G}\sum_{1\le g\le 2G}\hat{Y}_g^2=\hat{\sigma}_G^2(1)-\hat{\mu}_G^2(1)+\hat{\sigma}_G^2(0)-\hat{\mu}_G^2(0)$$

It follows from Lemma B.2.4 that

$$\frac{1}{G}\sum_{1\le g\le 2G}\hat{Y}_g^2 \xrightarrow{P} E[\tilde{Y}_g^2(1)] + E[\tilde{Y}_g^2(0)]$$

Next, we argue that

$$\frac{2}{G}\sum_{1\le j\le G}\hat{Y}_{\pi(2j)}\hat{Y}_{\pi(2j-1)} \xrightarrow{P} 2E[\mu_1(W_g)\mu_0(W_g)]~,$$

where we use the notation $\mu_d(W_g)$ to denote $E[\tilde{Y}_g(d) \mid W_g]$. To this end, first note that

$$\frac{2}{G}\sum_{1\le j\le G}\hat{Y}_{\pi(2j)}\hat{Y}_{\pi(2j-1)} = \frac{2}{G}\sum_{1\le j\le G}\tilde{Y}_{\pi(2j)}\tilde{Y}_{\pi(2j-1)} + \frac{2}{G}\sum_{1\le j\le G}\hat{Y}_{\pi(2j)}\hat{Y}_{\pi(2j-1)} - \tilde{Y}_{\pi(2j)}\tilde{Y}_{\pi(2j-1)}~.$$

Note that

$$\begin{aligned}
&\frac{2}{G}\sum_{1\le j\le G}\left(\hat{Y}_{\pi(2j)}(1)\hat{Y}_{\pi(2j-1)}(0) - \tilde{Y}_{\pi(2j)}(1)\tilde{Y}_{\pi(2j-1)}(0)\right)D_{\pi(2j)} \\
&= \frac{2}{G}\sum_{1\le j\le G}\left(\hat{Y}_{\pi(2j)}(1) - \tilde{Y}_{\pi(2j)}(1)\right)\hat{Y}_{\pi(2j-1)}(0)D_{\pi(2j)} + \left(\hat{Y}_{\pi(2j-1)}(0) - \tilde{Y}_{\pi(2j-1)}(0)\right)\tilde{Y}_{\pi(2j)}(1)D_{\pi(2j)} \\
&= \frac{2}{G}\sum_{1\le j\le G}\left(\hat{Y}_{\pi(2j)}(1) - \tilde{Y}_{\pi(2j)}(1)\right)\tilde{Y}_{\pi(2j-1)}(0)D_{\pi(2j)} \\
&\qquad + \left(\hat{Y}_{\pi(2j)}(1) - \tilde{Y}_{\pi(2j)}(1)\right)\left(\hat{Y}_{\pi(2j-1)}(0) - \tilde{Y}_{\pi(2j-1)}(0)\right)D_{\pi(2j)} \\
&\qquad\qquad + \left(\hat{Y}_{\pi(2j-1)}(0) - \tilde{Y}_{\pi(2j-1)}(0)\right)\tilde{Y}_{\pi(2j)}(1)D_{\pi(2j)}~,
\end{aligned}$$

for which the first term is given as follows:

$$\begin{aligned}
&\frac{2}{G}\sum_{1\le j\le G}\left(\hat{Y}_{\pi(2j)}(1) - \tilde{Y}_{\pi(2j)}(1)\right)\tilde{Y}_{\pi(2j-1)}(0)D_{\pi(2j)} \\
&= \left(\frac{1}{\frac{1}{2G}\sum_{1\le g\le 2G}N_g} - \frac{1}{E[N_g]}\right)\left(\frac{2}{G}\sum_{1\le j\le G}N_{\pi(2j)}\bar{Y}_{\pi(2j)}(1)\tilde{Y}_{\pi(2j-1)}(0)D_{\pi(2j)}\right) \\
&\quad - \left(\frac{\frac{1}{2G}\sum_{1\le g\le 2G}\bar{Y}_g(1)I\{D_g=1\}N_g}{\left(\frac{1}{2G}\sum_{1\le g\le 2G}N_g\right)^2} - \frac{E[\bar{Y}_g(1)N_g]}{E[N_g]^2}\right)\left(\frac{2}{G}\sum_{1\le j\le G}N_{\pi(2j)}\tilde{Y}_{\pi(2j-1)}(0)D_{\pi(2j)}\right)~.
\end{aligned}$$

By following the same argument in Lemma S.1.6 from Bai et al. (2022c) and Lemma C.3.1, we have

$$\frac{2}{G}\sum_{1\leq j\leq G} N_{\pi(2j)}\bar{Y}_{\pi(2j)}(1)\tilde{Y}_{\pi(2j-1)}(0)D_{\pi(2j)} \xrightarrow{P} E[E[N_g\bar{Y}_g(1)\mid X_g]E[\bar{Y}_g(0)\mid X_g]]$$

$$\frac{2}{G}\sum_{1\leq j\leq G} N_{\pi(2j)}\tilde{Y}_{\pi(2j-1)}(0)D_{\pi(2j)} \xrightarrow{P} E[E[N_g\mid X_g]E[\bar{Y}_g(0)\mid X_g]]$$

for the case of not matching on cluster sizes. As for the case where we match on cluster sizes,

$$\frac{2}{G}\sum_{1\leq j\leq G} N_{\pi(2j)}\bar{Y}_{\pi(2j)}(1)\tilde{Y}_{\pi(2j-1)}(0)D_{\pi(2j)} \xrightarrow{P} E[N_gE[\bar{Y}_g(1)\mid W_g]E[\bar{Y}_g(0)\mid W_g]]$$

$$\frac{2}{G}\sum_{1\leq j\leq G} N_{\pi(2j)}\tilde{Y}_{\pi(2j-1)}(0)D_{\pi(2j)} \xrightarrow{P} E[N_gE[\bar{Y}_g(0)\mid W_g]]$$

Then, by weak law of large number, Lemma B.1.2 (or Lemma B.1.1) and Slutsky's theorem, we have

$$\frac{2}{G}\sum_{1\leq j\leq G}\left(\hat{Y}_{\pi(2j)}(1)-\tilde{Y}_{\pi(2j)}(1)\right)\tilde{Y}_{\pi(2j-1)}(0)D_{\pi(2j)} \xrightarrow{P} 0 \ .$$

By repeating the same arguments for the other two terms, we conclude that

$$\frac{2}{G}\sum_{1\leq j\leq G}\left(\hat{Y}_{\pi(2j)}(1)\hat{Y}_{\pi(2j-1)}(0)-\tilde{Y}_{\pi(2j)}(1)\tilde{Y}_{\pi(2j-1)}(0)\right)D_{\pi(2j)} \xrightarrow{P} 0 \ ,$$

which immediately implies

$$\frac{2}{G}\sum_{1\leq j\leq G}\hat{Y}_{\pi(2j)}\hat{Y}_{\pi(2j-1)}-\tilde{Y}_{\pi(2j)}\tilde{Y}_{\pi(2j-1)} \xrightarrow{P} 0 \ .$$

Thus, it is left to show that

$$\frac{2}{G}\sum_{1\leq j\leq G}\tilde{Y}_{\pi(2j)}\tilde{Y}_{\pi(2j-1)} \xrightarrow{P} 2E[\mu_1(W_g)\mu_0(W_g)] \ ,$$

for the case of matching on cluster sizes, and for the case of not matching on cluster size,

$$\frac{2}{G}\sum_{1\leq j\leq G}\tilde{Y}_{\pi(2j)}\tilde{Y}_{\pi(2j-1)} \xrightarrow{P} 2E[\mu_1(X_g)\mu_0(X_g)] \ ,$$

both of which can be proved by applying Lemma S.1.6 from Bai et al. (2022c) and Lemma C.3.1. Hence, in the case where we match on cluster size,

$$\begin{aligned}
\hat{\tau}_n^2 &\xrightarrow{P} E\left[\tilde{Y}_g^2(1)\right] + E\left[\tilde{Y}_g^2(0)\right] - 2E\left[\mu_1\left(W_g\right)\mu_0\left(W_g\right)\right] \\
&= E\left[\operatorname{Var}\left[\tilde{Y}_g(1) \mid W_g\right]\right] + E\left[\operatorname{Var}\left[\tilde{Y}_g(0) \mid W_g\right]\right] + E\left[\left(\mu_1\left(W_g\right) - \mu_0\left(W_g\right)\right)^2\right] \\
&= E\left[\operatorname{Var}\left[\tilde{Y}_g(1) \mid W_g\right]\right] + E\left[\operatorname{Var}\left[\tilde{Y}_g(0) \mid W_g\right]\right] + E\left[\left(E\left[\tilde{Y}_g(1) \mid X_i\right] - E\left[\tilde{Y}_g(0) \mid W_g\right]\right)^2\right] .
\end{aligned}$$

And corresponding result holds in the case where we do not match on cluster size. ■

Lemma B.2.6. *If Assumptions 2.2.1 holds, and Assumptions 2.2.1 and 2.3.2-2.3.3, 2.4.1 hold, then*

$$\hat{\lambda}_G^2 \xrightarrow{P} E\left[\left(E\left[\tilde{Y}_g(1) \mid X_g\right] - E\left[\tilde{Y}_g(0) \mid X_g\right]\right)^2\right]$$

in the case where we match on cluster size. Instead, if Assumptions 2.3.5-2.3.6, 2.4.2 hold, then

$$\hat{\lambda}_G^2 \xrightarrow{P} E\left[\left(E\left[\tilde{Y}_g(1) \mid W_g\right] - E\left[\tilde{Y}_g(0) \mid W_g\right]\right)^2\right]$$

in the case where we do not match on cluster size.

Proof. Note that

$$\begin{aligned}
\hat{\lambda}_G^2 &= \frac{2}{G} \sum_{1 \le j \le \lfloor G/2 \rfloor} \left(\left(\hat{Y}_{\pi(4j-3)} - \hat{Y}_{\pi(4j-2)}\right)\left(\hat{Y}_{\pi(4j-1)} - \hat{Y}_{\pi(4j)}\right)\left(D_{\pi(4j-3)} - D_{\pi(4j-2)}\right)\left(D_{\pi(4j-1)} - D_{\pi(4j)}\right)\right) \\
&= \underbrace{\frac{2}{G} \sum_{1 \le j \le \lfloor G/2 \rfloor} \left(\left(\tilde{Y}_{\pi(4j-3)} - \tilde{Y}_{\pi(4j-2)}\right)\left(\tilde{Y}_{\pi(4j-1)} - \tilde{Y}_{\pi(4j)}\right)\left(D_{\pi(4j-3)} - D_{\pi(4j-2)}\right)\left(D_{\pi(4j-1)} - D_{\pi(4j)}\right)\right)}_{:=\tilde{\lambda}_G^2} \\
&+ \frac{2}{G} \sum_{1 \le j \le \lfloor G/2 \rfloor} \left(\left(\left(\hat{Y}_{\pi(4j-3)} - \hat{Y}_{\pi(4j-2)}\right)\left(\hat{Y}_{\pi(4j-1)} - \hat{Y}_{\pi(4j)}\right) - \left(\tilde{Y}_{\pi(4j-3)} - \tilde{Y}_{\pi(4j-2)}\right)\left(\tilde{Y}_{\pi(4j-1)} - \tilde{Y}_{\pi(4j)}\right)\right)\right. \\
&\times \left(D_{\pi(4j-3)} - D_{\pi(4j-2)}\right)\left(D_{\pi(4j-1)} - D_{\pi(4j)}\right)\Big)
\end{aligned}$$

Note that

$$
\begin{aligned}
&\left(\hat{Y}_{\pi(4j-3)}(1)-\hat{Y}_{\pi(4j-2)}(0)\right)\left(\hat{Y}_{\pi(4j-1)}(1)-\hat{Y}_{\pi(4j)}(0)\right)D_{\pi(4j-3)}D_{\pi(4j-1)}\\
&\quad-\left(\tilde{Y}_{\pi(4j-3)}(1)-\tilde{Y}_{\pi(4j-2)}(0)\right)\left(\tilde{Y}_{\pi(4j-1)}(1)-\tilde{Y}_{\pi(4j)}(0)\right)D_{\pi(4j-3)}D_{\pi(4j-1)}\\
&=\left(\hat{Y}_{\pi(4j-3)}(1)-\hat{Y}_{\pi(4j-2)}(0)-\left(\tilde{Y}_{\pi(4j-3)}(1)-\tilde{Y}_{\pi(4j-2)}(0)\right)\right)\left(\tilde{Y}_{\pi(4j-1)}(1)-\tilde{Y}_{\pi(4j)}(0)\right)D_{\pi(4j-3)}D_{\pi(4j-1)}\\
&+\left(\hat{Y}_{\pi(4j-3)}(1)-\hat{Y}_{\pi(4j-2)}(0)-\left(\tilde{Y}_{\pi(4j-3)}(1)-\tilde{Y}_{\pi(4j-2)}(0)\right)\right)\\
&\qquad\times\left(\hat{Y}_{\pi(4j-1)}(1)-\hat{Y}_{\pi(4j)}(0)-\left(\tilde{Y}_{\pi(4j-1)}(1)-\tilde{Y}_{\pi(4j)}(0)\right)\right)D_{\pi(4j-3)}D_{\pi(4j-1)}\\
&+\left(\hat{Y}_{\pi(4j-1)}(1)-\hat{Y}_{\pi(4j)}(0)-\left(\tilde{Y}_{\pi(4j-1)}(1)-\tilde{Y}_{\pi(4j)}(0)\right)\right)\left(\tilde{Y}_{\pi(4j-3)}(1)-\tilde{Y}_{\pi(4j-2)}(0)\right)D_{\pi(4j-3)}D_{\pi(4j-1)}\ .
\end{aligned}
$$

Then we can show that each term converges to zero in probability by repeating the arguments in Lemma B.2.5. The results should hold for any treatment combination, which implies $\hat{\lambda}_G^2-\tilde{\lambda}_G^2\xrightarrow{P}0$. Finally, by Lemma S.1.7 of Bai et al. (2022c) and Lemma C.3.1, we have

$$
\hat{\lambda}_G^2=\tilde{\lambda}_G^2+o_P(1)\xrightarrow{P}E\left[\left(E\left[\tilde{Y}_g(1)\mid W_g\right]-E\left[\tilde{Y}_g(0)\mid W_g\right]\right)^2\right]
$$

in the case where we match on cluster size, and

$$
\hat{\lambda}_G^2=\tilde{\lambda}_G^2+o_P(1)\xrightarrow{P}E\left[\left(E\left[\tilde{Y}_g(1)\mid X_g\right]-E\left[\tilde{Y}_g(0)\mid X_g\right]\right)^2\right]
$$

in the case where we do not match on cluster size. ■

Lemma B.2.7. *Let $\tilde{R}_G(t)$ denote the randomization distribution of $\sqrt{G}\hat{\Delta}_G$ (see equation* (B.12)*). Then under the null hypothesis* (2.9)*, we have that*

$$
\sup_{t\in\mathbf{R}}|\tilde{R}_G(t)-\Phi(t/\tau)|\xrightarrow{P}0\ ,
$$

where, in the case where we match on cluster size,

$$
\tau^2=E[\mathrm{Var}[\tilde{Y}_g(1)|W_g]]+E[\mathrm{Var}[\tilde{Y}_g(0)|W_g]]+E\left[(E[\tilde{Y}_g(1)|W_g]-E[\tilde{Y}_g(0)|X_g])^2\right]\ ,
$$

and in the case where we do not match on cluster size,

$$
\tau^2=E[\mathrm{Var}[\tilde{Y}_g(1)|X_g]]+E[\mathrm{Var}[\tilde{Y}_g(0)|X_g]]+E\left[(E[\tilde{Y}_g(1)|X_g]-E[\tilde{Y}_g(0)|X_g])^2\right]\ ,
$$

with (in both cases)

$$\tilde{Y}_g(d) = \frac{N_g}{E[N_g]} \left(\bar{Y}_g(d) - \frac{E[\bar{Y}_g(d) N_g]}{E[N_g]} \right) \ .$$

Proof. For a random transformation of the data, it follows as a consequence of Lemmas B.1.1 and B.1.2 that

$$\frac{1}{G} \sum_{1 \leq g \leq 2G} I\{\tilde{D}_g = d\} N_g \xrightarrow{P} E[N_g] \ ,$$

$$\frac{1}{G} \sum_{1 \leq g \leq 2G} (1 - \tilde{D}_g) N_g \bar{Y}_g \xrightarrow{p} E[N_g \bar{Y}_g(0)] \ .$$

Combining this with Lemma B.2.8 and a straightforward modification of Lemma A.3. in Chung and Romano (2013) to two dimensional distributions, we obtain that

$$\sup_{t \in \mathbf{R}} |\tilde{R}_G(t) - \Phi(t/\tau)| \xrightarrow{P} 0 \ ,$$

where when we match on cluster size

$$\tau^2 = \frac{1}{E[N_g]^2} \left(E[\text{Var}(N_g \bar{Y}_g(1) | W_g)] + E[\text{Var}(N_g \bar{Y}_g(0) | W_g)] + E\left[(E[N_g \bar{Y}_g(1) | W_g] - E[N_g \bar{Y}_g(0) | W_g])^2 \right] \right) \ ,$$

and when we do *not* match on cluster size

$$\begin{aligned} \tau^2 = \frac{1}{E[N_g]^2} \Big(& E[\text{Var}(N_g \bar{Y}_g(1) | X_g)] + E[\text{Var}(N_g \bar{Y}_g(0) | X_g)] + E\left[(E[N_g \bar{Y}_g(1) | X_g] - E[N_g \bar{Y}_g(0) | X_g])^2 \right] + \\ & - 2 \frac{E[N_g \bar{Y}_g(0)]}{E[N_g]} \left(E[N_g^2 \bar{Y}_g(1)] + E[N_g^2 \bar{Y}_g(0)] \right. \\ & \quad - \left(E\left[E[N_g \bar{Y}_g(1) | X_g] E[N_g | X_g] \right] + E\left[E[N_g \bar{Y}_g(0) | X_g] E[N_g | X_g] \right] \right) \Big) \\ & \quad + \left(\frac{E[N_g \bar{Y}_g(0)]}{E[N_g]} \right)^2 2 E[\text{Var}(N_g | X_g)] \Big) \ . \end{aligned}$$

The result then follows from further algebraic manipulations to simplify τ in each case (see for instance Lemma B.2.10). ■

Lemma B.2.8.

$$\rho\left(\mathcal{L}\left((\mathbb{K}_G^{YN}, \mathbb{K}_G^N)' | Z^{(G)} \right), N\left(0, \mathbb{V}_R\right) \right) \xrightarrow{P} 0 \ ,$$

where

$$\begin{pmatrix} \mathbb{K}_G^{YN} \\ \mathbb{K}_G^N \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{G}} \sum_{1 \leq j \leq G} \epsilon_j \left(N_{\pi(2j)} \bar{Y}_{\pi(2j)} - N_{\pi(2j-1)} \bar{Y}_{\pi(2j-1)} \right) \left(D_{\pi(2j)} - D_{\pi(2j-1)} \right) \\ \frac{1}{\sqrt{G}} \sum_{1 \leq j \leq G} \epsilon_j (N_{\pi(2j)} - N_{\pi(2j-1)})(D_{\pi(2j)} - D_{\pi(2j-1)}) \end{pmatrix} \ ,$$

and where, in the case where we match on cluster size,

$$\mathbb{V}_R = \begin{pmatrix} \mathbb{V}_R^1 & 0 \\ 0 & 0 \end{pmatrix} ~,$$

with

$$\mathbb{V}_R^1 = E[\text{Var}(N_g \bar{Y}_g(1)|W_g)] + E[\text{Var}(N_g \bar{Y}_g(0)|W_g)] + E\left[(E[N_g \bar{Y}_g(1)|W_g] - E[N_g \bar{Y}_g(0)|W_g])^2\right] ~,$$

and when we do not match on cluster size,

$$\mathbb{V}_R = \begin{pmatrix} \mathbb{V}_R^{1,1} & \mathbb{V}_R^{1,2} \\ \mathbb{V}_R^{1,2} & \mathbb{V}_R^{2,2} \end{pmatrix} ~,$$

with

$$\begin{aligned} \mathbb{V}_R^{1,1} &= E[\text{Var}(N_g \bar{Y}_g(1)|X_g)] + E[\text{Var}(N_g \bar{Y}_g(0)|X_g)] + E\left[(E[N_g \bar{Y}_g(1)|X_g] - E[N_g \bar{Y}_g(0)|X_g])^2\right] \\ \mathbb{V}_R^{1,2} &= E[N_g^2 \bar{Y}_g(1)] + E[N_g^2 \bar{Y}_g(0)] - \left(E\left[E[N_g \bar{Y}_g(1)|X_g] E[N_g|X_g]\right] + E\left[E[N_g \bar{Y}_g(0)|X_g] E[N_g|X_g]\right]\right) \\ \mathbb{V}_R^{2,2} &= 2E[\text{Var}(N_g|X_g)] ~. \end{aligned}$$

Proof. Using the fact that ϵ_j, $j = 1, \dots, G$ and $\epsilon_j(D_{\pi(2j)} - D_{\pi(2j-1)})$, $j = 1, \dots, G$ have the same distribution conditional on $Z^{(G)}$, it suffices to study the limiting distribution of $(\tilde{\mathbb{K}}_G^{YN}, \tilde{\mathbb{K}}_G^N)'$ conditional on $Z^{(G)}$, where

$$\begin{aligned} \tilde{\mathbb{K}}_G^{YN} &:= \frac{1}{\sqrt{G}} \sum_{1 \le j \le G} \epsilon_j \left(N_{\pi(2j)} \bar{Y}_{\pi(2j)} - N_{\pi(2j-1)} \bar{Y}_{\pi(2j-1)}\right) ~, \\ \tilde{\mathbb{K}}_G^N &:= \frac{1}{\sqrt{G}} \sum_{1 \le j \le G} \epsilon_j \left(N_{\pi(2j)} - N_{\pi(2j-1)}\right) ~. \end{aligned}$$

We will show

$$\rho\left(\mathcal{L}\left((\tilde{\mathbb{K}}_G^{YN}, \tilde{\mathbb{K}}_G^N)' | Z^{(G)}\right), N(0, \mathbb{V}_R)\right) \xrightarrow{P} 0 ~, \tag{B.15}$$

where $\mathcal{L}(\cdot)$ denote the law and ρ is any metric that metrizes weak convergence. To that end, we will employ the Lindeberg central limit theorem in Proposition 2.27 of van der Vaart (1998) and a subsequencing argument. Indeed, to verify (B.15), note we need only show that for any subsequence $\{G_k\}$ there exists a further subsequence $\{G_{k_l}\}$ such that

$$\rho\left(\mathcal{L}\left((\tilde{\mathbb{K}}^{YN}_{G_{k_l}}, \tilde{\mathbb{K}}^{N}_{G_{k_l}})'|Z^{(G_{k_l})}\right), N(0, \mathbb{V}_R)\right) \to 0 \text{ with probability one} . \tag{B.16}$$

To that end, define

$$\mathbb{V}_{R,n} = \begin{pmatrix} \mathbb{V}^{1,1}_{R,n} & \mathbb{V}^{1,2}_{R,n} \\ \mathbb{V}^{1,2}_{R,n} & \mathbb{V}^{2,2}_{R,n} \end{pmatrix} = \text{Var}[(\tilde{\mathbb{K}}^{YN}_G, \tilde{\mathbb{K}}^N_G)'|Z^{(G)}] ,$$

where

$$\begin{aligned}
\mathbb{V}^{1,1}_{R,n} &= \frac{1}{G}\sum_{1\le j\le G}(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)})^2 \\
\mathbb{V}^{1,2}_{R,n} &= \frac{1}{G}\sum_{1\le j\le G}(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)})(N_{\pi(2j)} - N_{\pi(2j-1)}) \\
\mathbb{V}^{2,2}_{R,n} &= \frac{1}{G}\sum_{1\le j\le G}(N_{\pi(2j)} - N_{\pi(2j-1)})^2 .
\end{aligned}$$

First consider the case where we match on cluster size. By arguing as in Lemma S.1.6 of Bai et al. (2022c), it can be shown that

$$\mathbb{V}^{1,1}_{R,n} \xrightarrow{P} E[\text{Var}[N_g\bar{Y}_g(1)]|W_g] + E[\text{Var}[N_g\bar{Y}_g(0)]|W_g] + E\left[(E[N_g\bar{Y}_g(1)|W_g] - E[N_g\bar{Y}_g(0)|W_g])^2\right] .$$

Next, we show that in this case $\mathbb{V}^{1,2}_{R,n}$ and $\mathbb{V}^{2,2}_{R,n}$ are $o_P(1)$. For $\mathbb{V}^{2,2}_{R,n}$ this follows immediately from Assumption 2.3.5. For $\mathbb{V}^{1,2}_{R,n}$ note that by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\frac{1}{G}\sum_{1\le j\le G}\left(\left(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)}\right)\left(N_{\pi(2j)} - N_{\pi(2j-1)}\right)\right) \\
&\le \left(\left(\frac{1}{G}\sum_{1\le j\le G}\left(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)}\right)^2\right)\left(\frac{1}{G}\sum_{1\le j\le G}\left(N_{\pi(2j)} - N_{\pi(2j-1)}\right)^2\right)\right)^{1/2} .
\end{aligned}$$

The second term of the product on the RHS is $o_P(1)$ by Assumption 2.3.5. The first term is $O_P(1)$ since

$$\frac{1}{G}\sum_{1\le j\le G}\left(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)}\right)^2 \lesssim \frac{1}{G}\sum_{1\le g\le 2G}N_g^2\bar{Y}_g(1)^2 + \frac{1}{G}\sum_{1\le g\le 2G}N_g^2\bar{Y}_g(0)^2 = O_P(1) ,$$

where the first inequality follows from exploiting the fact that $|a-b|^2 \le 2(a^2+b^2)$ and the definition of $\bar{Y}_g$, and the final equality follows from Lemma C.2.3 and the law of large numbers. We can thus conclude that

$\mathbb{V}^{1,2}_{R,n} = o_P(1)$ when matching on cluster size.

$$\mathbb{V}_{R,n} \xrightarrow{P} \mathbb{V}_R~. \tag{B.17}$$

In the case where we do *not* match on cluster size, again by arguing as in Lemma S.1.6 of Bai et al. (2022c), it can be shown that (B.17) holds. Next, we verify the Lindeberg condition in Proposition 2.27 of van der Vaart (1998). Note that

$$\begin{aligned}
&\frac{1}{G}\sum_{1\le j\le G} E[((\epsilon_j(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)}))^2 + (\epsilon_j(N_{\pi(2j)} - N_{\pi(2j-1)}))^2) \\
&\qquad \times I\{((\epsilon_j(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)}))^2 + (\epsilon_j(N_{\pi(2j)} - N_{\pi(2j-1)}))^2) > \epsilon^2 G\}|Z^{(G)}] \\
&= \frac{1}{G}\sum_{1\le j\le G} E[((N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)})^2 + (N_{\pi(2j)} - N_{\pi(2j-1)})^2) \\
&\qquad \times I\{((N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)})^2 + (N_{\pi(2j)} - N_{\pi(2j-1)})^2) > \epsilon^2 G\}|Z^{(G)}] \\
&\lesssim \frac{1}{G}\sum_{1\le j\le G} (N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)})^2 I\{(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)})^2 > \epsilon^2 G/2\} \\
&\qquad + \frac{1}{G}\sum_{1\le j\le G} (N_{\pi(2j)} - N_{\pi(2j-1)})^2 I\{(N_{\pi(2j)} - N_{\pi(2j-1)})^2 > \epsilon^2 G/2\}~.
\end{aligned}$$

where the inequality follows from (B.7) and the fact that $(N_g, \bar{Y}_g), 1 \le g \le 2G$ are all constants conditional on $Z^{(G)}$. The last line converges in probability to zero as long as we can show

$$\begin{aligned}
&\frac{1}{G}\max_{1\le j\le G}(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)})^2 \xrightarrow{P} 0 \\
&\frac{1}{G}\max_{1\le j\le G}(N_{\pi(2j)} - N_{\pi(2j-1)})^2 \xrightarrow{P} 0~.
\end{aligned}$$

Note

$$\begin{aligned}
\frac{1}{G}\max_{1\le j\le G}(N_{\pi(2j)}\bar{Y}_{\pi(2j)} - N_{\pi(2j-1)}\bar{Y}_{\pi(2j-1)})^2 &\lesssim \frac{1}{G}\max_{1\le j\le G}\left(N^2_{\pi_{(2j-1)}}\bar{Y}^2_{\pi(2j-1)} + N^2_{\pi_{(2j)}}\bar{Y}^2_{\pi(2j)}\right) \\
&\lesssim \frac{1}{G}\max_{1\le g\le 2G}\left(N_g^2\bar{Y}_g^2(1) + N_g^2\bar{Y}_g^2(0)\right) \xrightarrow{P} 0
\end{aligned}$$

Where the first inequality follows from the fact that $|a-b|^2 \le 2(a^2+b^2)$, the second by inspection, and the convergence by Lemma S.1.1 in Bai et al. (2022c) along with Assumption 2.2.1(c) and Lemma C.2.3. The second statement follows similarly. Therefore, we have verified both conditions in Proposition 2.27 of van der Vaart (1998) hold in probability, and therefore for each subsequence there must exists a further subsequence along which both conditions hold with probability one, so (B.16) holds, and the conclusion of the lemma follows. ■

Lemma B.2.9. *Let* $\check{v}_G^2(\epsilon_1, \ldots, \epsilon_G)$ *be defined as in equation* (B.13). *If Assumption 2.2.1 holds, and Assumptions 2.3.6-2.3.5 (or Assumptions 2.3.3-2.3.2) hold,*

$$\check{v}_G^2(\epsilon_1, \ldots, \epsilon_G) \xrightarrow{P} \tau^2 \ ,$$

where τ^2 *is defined in (B.2.7).*

Proof. From Lemma B.2.5, we see that $\hat{\tau}_G^2 \xrightarrow{P} \tau^2$. It therefore suffices to show that $\check{\lambda}_G^2\left(\epsilon_1, \ldots, \epsilon_G\right) \xrightarrow{P} 0$. In order to do so, note that $\check{\lambda}_G^2\left(\epsilon_1, \ldots, \epsilon_G\right)$ may be decomposed into sums of the form

$$\frac{2}{G} \sum_{1 \leq j \leq \left\lfloor \frac{G}{2} \right\rfloor} \epsilon_{2j-1}\epsilon_{2j} \hat{Y}_{\pi(4j-k)} \hat{Y}_{\pi(4j-\ell)} D_{\pi(4j-k')} D_{\pi(4j-\ell')} \ ,$$

where $(k, k') \in \{2, 3\}^2$ and $(l, l') \in \{0, 1\}^2$. Note that

$$\begin{aligned} &\frac{2}{G} \sum_{1 \leq j \leq \left\lfloor \frac{G}{2} \right\rfloor} \epsilon_{2j-1}\epsilon_{2j} \hat{Y}_{\pi(4j-k)} \hat{Y}_{\pi(4j-\ell)} D_{\pi(4j-k')} D_{\pi(4j-\ell')} \\ &= \frac{2}{G} \sum_{1 \leq j \leq \left\lfloor \frac{G}{2} \right\rfloor} \epsilon_{2j-1}\epsilon_{2j} \tilde{Y}_{\pi(4j-k)} \tilde{Y}_{\pi(4j-\ell)} D_{\pi(4j-k')} D_{\pi(4j-\ell')} \\ &\quad + \frac{G}{n} \sum_{1 \leq j \leq \left\lfloor \frac{G}{2} \right\rfloor} \epsilon_{2j-1}\epsilon_{2j} \left(\hat{Y}_{\pi(4j-k)} \hat{Y}_{\pi(4j-\ell)} - \tilde{Y}_{\pi(4j-k)} \tilde{Y}_{\pi(4j-\ell)} \right) D_{\pi(4j-k')} D_{\pi(4j-\ell')} \ . \end{aligned}$$

By following the arguments in Lemma S.1.9 of Bai et al. (2022c) and Lemma C.3.1, we have that

$$\frac{2}{G} \sum_{1 \leq j \leq \left\lfloor \frac{G}{2} \right\rfloor} \epsilon_{2j-1}\epsilon_{2j} \tilde{Y}_{\pi(4j-k)} \tilde{Y}_{\pi(4j-\ell)} D_{\pi(4j-k')} D_{\pi(4j-\ell')} \xrightarrow{P} 0 \ .$$

As for the second term, we show that it convergences to zero in probability in the case where $k = k' = 3$

and $\ell = \ell' = 1$. And the other cases should hold by repeating the same arguments.

$$
\begin{aligned}
&\frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} \left(\hat{Y}_{\pi(4j-3)} \hat{Y}_{\pi(4j-1)} - \tilde{Y}_{\pi(4j-3)} \tilde{Y}_{\pi(4j-1)} \right) D_{\pi(4j-3)} D_{\pi(4j-1')} \\
&= \frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} \left(\hat{Y}_{\pi(4j-3)}(1) \hat{Y}_{\pi(4j-1)}(1) - \tilde{Y}_{\pi(4j-3)}(1) \tilde{Y}_{\pi(4j-1)}(1) \right) D_{\pi(4j-3)} D_{\pi(4j-1')} \\
&= \frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} \left(\hat{Y}_{\pi(4j-3)}(1) - \tilde{Y}_{\pi(4j-3)}(1) \right) \tilde{Y}_{\pi(4j-1)}(1) D_{\pi(4j-3)} D_{\pi(4j-1')} \\
&\quad + \frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} \left(\hat{Y}_{\pi(4j-3)}(1) - \tilde{Y}_{\pi(4j-3)}(1) \right) \left(\hat{Y}_{\pi(4j-1)}(1) - \tilde{Y}_{\pi(4j-1)}(1) \right) D_{\pi(4j-3)} D_{\pi(4j-1')} \\
&\quad + \frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} \left(\hat{Y}_{\pi(4j-1)}(1) - \tilde{Y}_{\pi(4j-1)}(1) \right) \tilde{Y}_{\pi(4j-3)}(1) D_{\pi(4j-3)} D_{\pi(4j-1')} ~,
\end{aligned}
$$

for which the first term is given as follows:

$$
\begin{aligned}
&\frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} \left(\hat{Y}_{\pi(4j-3)}(1) - \tilde{Y}_{\pi(4j-3)}(1) \right) \tilde{Y}_{\pi(4j-1)}(1) D_{\pi(4j-3)} D_{\pi(4j-1')} \\
&= \left(\frac{1}{\frac{1}{2G} \sum_{1 \leq g \leq 2G} N_g} - \frac{1}{E[N_g]} \right) \left(\frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} N_{\pi(4j-3)} \bar{Y}_{\pi(4j-3)}(1) \tilde{Y}_{\pi(4j-1)}(1) D_{\pi(4j-3)} D_{\pi(4j-1')} \right) \\
&\quad - \left(\frac{\frac{1}{2G} \sum_{1 \leq g \leq 2G} \bar{Y}_g(1) I\{D_g = 1\} N_g}{\left(\frac{1}{2G} \sum_{1 \leq g \leq 2G} N_g \right)^2} - \frac{E[\bar{Y}_g(1) N_g]}{E[N_g]^2} \right) \left(\frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} N_{\pi(4j-3)} \tilde{Y}_{\pi(4j-1)}(1) D_{\pi(4j-3)} D_{\pi(4j-1')} \right) ~.
\end{aligned}
$$

by following the same argument in Lemma S.1.6 from Bai et al. (2022c) and Lemma C.3.1, we have

$$
\begin{aligned}
&\frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} N_{\pi(4j-3)} \bar{Y}_{\pi(4j-3)}(1) \tilde{Y}_{\pi(4j-1)}(1) D_{\pi(4j-3)} D_{\pi(4j-1')} \xrightarrow{P} 0 \\
&\frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} N_{\pi(4j-3)} \tilde{Y}_{\pi(4j-1)}(1) D_{\pi(4j-3)} D_{\pi(4j-1')} \xrightarrow{P} 0 ~.
\end{aligned}
$$

Then, by weak law of large number, Lemma B.1.2 (or Lemma B.1.1) and Slutsky's theorem, we have

$$
\frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} \left(\hat{Y}_{\pi(4j-3)}(1) - \tilde{Y}_{\pi(4j-3)}(1) \right) \tilde{Y}_{\pi(4j-1)}(1) D_{\pi(4j-3)} D_{\pi(4j-1')} \xrightarrow{P} 0 ~.
$$

By repeating the same arguments for the other two terms, we conclude that

$$
\frac{2}{G} \sum_{1 \leq j \leq \lfloor \frac{G}{2} \rfloor} \epsilon_{2j-1} \epsilon_{2j} \left(\hat{Y}_{\pi(4j-3)} \hat{Y}_{\pi(4j-1)} - \tilde{Y}_{\pi(4j-3)} \tilde{Y}_{\pi(4j-1)} \right) D_{\pi(4j-3)} D_{\pi(4j-1')} \xrightarrow{P} 0 ~.
$$

Therefore, for $(k, k') \in \{2, 3\}^2$ and $(l, l') \in \{0, 1\}^2$,

$$\frac{2}{G} \sum_{1 \leq j \leq \left\lfloor \frac{G}{2} \right\rfloor} \epsilon_{2j-1}\epsilon_{2j} \hat{Y}_{\pi(4j-k)} \hat{Y}_{\pi(4j-\ell)} D_{\pi(4j-k')} D_{\pi(4j-\ell')} \xrightarrow{P} 0 \ ,$$

which implies $\check{\lambda}^2_G\,(\epsilon_1, \ldots, \epsilon_G) \xrightarrow{P} 0$, and thus $\check{\nu}^2_G(\epsilon_1, \ldots, \epsilon_G) \xrightarrow{P} \tau^2$. ■

Lemma B.2.10. *If $E[N_g \bar{Y}_g(1)] = E[N_g \bar{Y}_g(0)]$, then for τ defined in Lemma B.2.7 (when not matching on cluster size),*

$$\tau^2 = E[\mathrm{Var}[\tilde{Y}_g(1)|X_g]] + E[\mathrm{Var}[\tilde{Y}_g(0)|X_g]] + E[(E[\tilde{Y}_g(1)|X_g] - E[\tilde{Y}_g(0)|X_g])^2] \ .$$

Proof. Note if $E[N_g \bar{Y}_g(1)] = E[N_g \bar{Y}_g(0)]$, then

$$\begin{aligned}
&E[\mathrm{Var}[\tilde{Y}_g(1)|X_g]] + E[\mathrm{Var}[\tilde{Y}_g(0)|X_g]] + E[(E[\tilde{Y}_g(1)|X_g] - E[\tilde{Y}_g(0)|X_g])^2] \\
&= \frac{E[\mathrm{Var}[N_g \bar{Y}_g(1)|X_g]]}{E[N_g]^2} + \frac{E[\mathrm{Var}[N_g \bar{Y}_g(0)|X_g]]}{E[N_g]^2} + \frac{2E[\mathrm{Var}[N_g|X_g]]E[N_g \bar{Y}_g(d)]^2}{E[N_g]^4} \\
&\quad + \frac{E[(E[N_g \bar{Y}_g(1)|X_g] - E[N_g \bar{Y}_g(0)|X_g])^2}{E[N_g]^2} \\
&\quad - 2\frac{E[N_g \bar{Y}_g(1)](E[N_g^2 \bar{Y}_g(1)] - E[E[N_g \bar{Y}_g(1)|X_g]E[N_g|X_g]])}{E[N_g]^3} \\
&\quad - 2\frac{E[N_g \bar{Y}_g(0)](E[N_g^2 \bar{Y}_g(0)] - E[E[N_g \bar{Y}_g(0)|X_g]E[N_g|X_g]])}{E[N_g]^3} \ .
\end{aligned}$$

The result then follows immediately. ■

B.3 Analysis of Matched Tuples designs

In this section we state generalizations of the results presented in Sections 2.3 and 2.4 to settings with more than two treatments. We focus on the case when not matching on cluster size; similar results should follow for the case of matching on cluster size analogously.

B.3.1 Setup and Main Results

We follow the general setup of Bai et al. (2023c) generalized to a setting with clustered assignment. Let $D_g \in \mathcal{D}$ denote treatment status for the gth cluster, where $\mathcal{D} = \{1, \ldots, |\mathcal{D}|\}$ denotes a finite set of values of the treatment. For $d \in \mathcal{D}$, let $Y_{i,g}(d)$ denote the potential outcome for the ith unit in the gth cluster if its treatment status were d. The observed outcome and potential outcomes are related to treatment status by the expression

$$Y_{i,g} = \sum_{d \in \mathcal{D}} Y_{i,g}(d) I\{D_g = d\} \ .$$

We suppose our sample consists of $J_G := (|\mathcal{D}|)G$ i.i.d. clusters. Now we have

$$Z^{(G)} := (((Y_{i,g} : i \in \mathcal{M}_g), D_g, X_g, N_g) : 1 \leq g \leq J_G)$$

and

$$((((Y_{i,g}(d) : d \in \mathcal{D}) : 1 \leq i \leq N_g), \mathcal{M}_g, X_g, N_g) : 1 \leq g \leq J_G) \ .$$

Our object of interest will generically be defined as a vector of linear contrasts over the collection of size-weighted cluster-level expected potential outcomes across treatments. Formally, let

$$\Gamma(Q_G) := (\Gamma_1(Q_G), \ldots, \Gamma_{|\mathcal{D}|}(Q_G))',$$

where

$$\Gamma_d(Q_G) := \frac{1}{E[N_g]} E\left[\sum_{i=1}^{N_g} Y_{ig}(d)\right]$$

for $d \in \mathcal{D}$. Let ν be a real-valued $m \times |\mathcal{D}|$ matrix. Define

$$\Delta_\nu(Q_G) := \nu\Gamma(Q_G) \in \mathbf{R}^m \ ,$$

as our generic parameter of interest. We maintain the following generalization of Assumptions 2.2.1 and 2.3.3.

Assumption B.3.1. The distribution Q_G is such that

(a) $\{(\mathcal{M}_g, X_g, N_g), 1 \leq g \leq J_G\}$ is an i.i.d. sequence of random variables.

(b) For some family of distributions $\{R(s, x, n) : (s, x, n) \in \text{supp}(\mathcal{M}_g, X_g, N_g)\}$,

$$R_G(\mathcal{M}_g^{(G)}, X^{(G)}, N^{(G)}) = \prod_{1 \leq g \leq J_G} R(\mathcal{M}_g, X_g, N_g) \ .$$

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

(d) For some $C < \infty$, $P\{E[Y_{i,g}^2(d)|X_g, N_g] \leq C$ for all $1 \leq i \leq N_g\} = 1$ for all $d \in \mathcal{D}$ and $1 \leq g \leq J_G$.

(e) $\mathcal{M}_g \perp\!\!\!\perp ((Y_{i,g}(d) : d \in \mathcal{D}) : 1 \leq i \leq N_g) \bigm| X_g, N_g$ for all $1 \leq g \leq J_G$.

(f) For $d \in \mathcal{D}$ and $1 \leq g \leq J_G$,

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

(g) For some $C < \infty$, $P\{E[N_g|X_g] \leq C\} = 1$

(h) $E[\bar{Y}_g^r(d)N_g^\ell|X_g = x]$, are Lipschitz for $d \in \mathcal{D}$, $r, \ell \in \{0, 1, 2\}$.

Following Bai et al. (2023c), the G blocks in a matched tuples design may then be represented by the sets

$$\lambda_j = \lambda_j(X^{(G)}) \subseteq \{1, 2, \ldots, J_G\} ,$$

for $1 \leq j \leq G$. We then maintain the following two assumptions on the treatment assignment mechanism which generalize Assumptions 2.3.1, 2.3.2, and 2.4.1:

Assumption B.3.2. Treatment is assigned so that $\left\{\left((Y_{ig}(d) : d \in \mathcal{D})_{1\leq i\leq N_g}, N_g\right)\right\}_{g=1}^{G} \perp\!\!\!\perp D^{(G)}|X^{(G)}$, and, conditional on $X^{(G)}$,

$$\{(D_g : g \in \lambda_j) : 1 \leq j \leq G\} ,$$

are i.i.d. and each uniformly distributed over all permutations of $(1, 2, \ldots, |\mathcal{D}|)$.

Assumption B.3.3. The blocks satisfy

$$\frac{1}{G}\sum_{1\leq j\leq G}\max_{i,k\in\lambda_j}|X_i - X_k|^2 \xrightarrow{P} 0 .$$

Assumption B.3.4. The blocks satisfy

$$\frac{1}{G}\sum_{1\leq j\leq\left\lfloor\frac{G}{2}\right\rfloor}\max_{i\in\lambda_{2j-1},k\in\lambda_{2j}}|X_i - X_k|^2 \xrightarrow{P} 0 .$$

The estimator for $\Delta_\nu(Q_G)$ is given by

$$\hat{\Delta}_{\nu,G} := \nu\hat{\Gamma}_G ,$$

where for $d \in \mathcal{D}$ we define

$$\hat{\Gamma}_G(d) := \frac{1}{N(d)} \sum_{1 \le g \le J_G} I\{D_g = d\} \frac{N_g}{|\mathcal{M}_g|} \sum_{i \in \mathcal{M}_g} Y_{ig} \ ,$$

with

$$N(d) = \sum_{1 \le g \le J_G} N_g I\{D_g = d\} \ .$$

and let $\hat{\Gamma}_G = (\hat{\Gamma}_G(1), \ldots, \hat{\Gamma}_G(|\mathcal{D}|))'$.

Our first result derives the limiting distribution of $\hat{\Delta}_{\nu,G}$ under our maintained assumptions.

Theorem B.3.1. *Suppose Assumptions B.3.1-B.3.3 holds. Then,*

$$\sqrt{G}(\hat{\Delta}_{\nu,G} - \Delta_\nu(Q)) \xrightarrow{d} N(0, \mathbf{V}_\nu) \ ,$$

where $\mathbf{V}_\nu := \nu \mathbf{V} \nu'$, *with*

$$\mathbf{V} := \mathbf{V}_1 + \mathbf{V}_2 \ , \tag{B.18}$$

$$\mathbf{V}_1 := \operatorname{diag}(E[\operatorname{Var}[\tilde{Y}_g(d)|X_i]] : d \in \mathcal{D}) \ ,$$

$$\mathbf{V}_2 := \left[\frac{1}{|\mathcal{D}|} \operatorname{Cov}[E[\tilde{Y}_g(d)|X_i], E[\tilde{Y}_g(d'))|X_i]]\right]_{d,d' \in \mathcal{D}} \ .$$

Proof. We show that $\sqrt{G}(\hat{\Gamma}_G(d) - \Gamma_G(Q) : d \in \mathcal{D}) \xrightarrow{d} N(0, \mathbf{V})$, from which the conclusion of the theorem follows by an application of the continuous mapping theorem. To show this we repeat the arguments from the proof of Theorem 2.3.1 while using the Delta method for vector-valued functions with $h(x_1, y_1, \ldots, x_{|\mathcal{D}|}, y_{|\mathcal{D}|}) = (x_d/y_d : d \in \mathcal{D})$ and using the fact that

$$(\hat{\Gamma}_G(d) : d \in \mathcal{D}) = \left(\frac{\frac{1}{\sqrt{G}} \sum_{1 \le g \le J_G} \bar{Y}_g(d) N_g I\{D_g = d\}}{\frac{1}{\sqrt{G}} \sum_{1 \le g \le J_G} N_g I\{D_g = d\}} : d \in \mathcal{D} \right) \ .$$

The Jacobian is given by

$$D_h(x_1, y_1, \ldots, x_{|\mathcal{D}|}, y_{|\mathcal{D}|}) = \begin{bmatrix} \frac{1}{y_1} & 0 & \ldots & 0 \\ -\frac{x_1}{y_1^2} & 0 & \ldots & 0 \\ 0 & \frac{1}{y_2} & \ldots & 0 \\ 0 & -\frac{x_2}{y_2^2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \frac{1}{y_{|\mathcal{D}|}} \\ 0 & 0 & \ldots & -\frac{x_{|\mathcal{D}|}}{y_{|\mathcal{D}|}^2} \ . \end{bmatrix} \ .$$

Repeating the algebra in proof of binary case, we obtain

$$D_h((E[\bar{Y}_g(d)N_g], E[N_g]) : d \in \mathcal{D})' \mathbb{V} D_h((E[\bar{Y}_g(d)N_g], E[N_g]) : d \in \mathcal{D}) = \mathbf{V} \ ,$$

where $\mathbb{V}$ is defined in the statement of Lemma B.3.1. ■

Following Bai et al. (2023c), our estimator for $\mathbf{V}_\nu$ is then given by $\hat{\mathbf{V}}_{\nu,G} := \nu \hat{\mathbf{V}}_G \nu'$, where

$$\begin{aligned} \hat{\mathbf{V}}_G &:= \hat{\mathbf{V}}_{1,G} + \hat{\mathbf{V}}_{2,G} \\ \hat{\mathbf{V}}_{1,G} &:= \operatorname{diag}\left(\hat{\mathbf{V}}_{1,G}(d) : d \in \mathcal{D}\right) \\ \hat{\mathbf{V}}_{2,G} &:= \left[\hat{\mathbf{V}}_{2,G}(d, d')\right]_{d,d' \in \mathcal{D}} \ , \end{aligned}$$

with

$$\begin{aligned} \hat{\mathbf{V}}_{1,G}(d) &:= \hat{\sigma}_G^2(d) - \hat{\rho}_G(d, d) \\ \hat{\mathbf{V}}_{2,G}(d, d') &:= \frac{1}{|\mathcal{D}|} \hat{\rho}_G(d, d') \ , \end{aligned}$$

where

$$\begin{aligned} \hat{\rho}_G(d, d) &:= \frac{2}{G} \sum_{1 \le j \le \lfloor G/2 \rfloor} \Big(\sum_{g \in \lambda_{2j-1}} \hat{Y}_g I\{D_g = d\} \Big) \Big(\sum_{g \in \lambda_{2j}} \hat{Y}_g I\{D_g = d\} \Big) \\ \hat{\rho}_G(d, d') &:= \frac{1}{G} \sum_{1 \le j \le G} \Big(\sum_{g \in \lambda_j} \hat{Y}_g I\{D_g = d\} \Big) \Big(\sum_{g \in \lambda_j} \hat{Y}_g I\{D_g = d'\} \Big) \text{ if } d \neq d' \\ \hat{\sigma}_G^2(d) &:= \frac{1}{G} \sum_{1 \le g \le J_G} \hat{Y}_g^2 I\{D_g = d\} \ . \end{aligned}$$

Suppose Assumptions B.3.1–B.3.4 hold, then consistency of our variance estimator follows by adapting the arguments from Bai et al. (2023c) to the proof of Theorem 3.4.2.

Lemma B.3.1. *Suppose Assumptions B.3.1–B.3.3 holds. Define*

$$\begin{aligned}\mathbb{L}_G^{\mathrm{YN}}(d) &= \frac{1}{\sqrt{G}} \sum_{1 \leq g \leq J_G} (\bar{Y}_g(d) N_g I\{D_g = d\} - E[\bar{Y}_g(d) N_g] I\{D_g = d\}) \\ \mathbb{L}_G^{\mathrm{N}}(d) &= \frac{1}{\sqrt{G}} \sum_{1 \leq g \leq J_G} (N_g I\{D_g = d\} - E[N_g] I\{D_g = d\}) \ .\end{aligned}$$

Then, as $G \to \infty$,

$$((\mathbb{L}_G^{\mathrm{YN}}(d), \mathbb{L}_G^{\mathrm{N}}(d)) : d \in \mathcal{D})' \xrightarrow{d} N(0, \mathbb{V}) \ ,$$

where

$$\mathbb{V} = \mathbb{V}_1 + \mathbb{V}_2$$

for

$$\mathbb{V}_1 = \mathrm{diag}(\mathbb{V}_1^d : d \in \mathcal{D})$$

$$\mathbb{V}_1^d = \begin{pmatrix} E[\mathrm{Var}[\bar{Y}_g(d) N_g | X_g]] & E[\mathrm{Cov}[\bar{Y}_g(d) N_g, N_g | X_g]] \\ E[\mathrm{Cov}[\bar{Y}_g(d) N_g, N_g | X_g]] & E[\mathrm{Var}[N_g | X_g]] \end{pmatrix}$$

$$\mathbb{V}_2 = \frac{1}{|\mathcal{D}|} \mathrm{Var}[((E[\bar{Y}_g(d) N_g | X_g], E[N_g | X_g]) : d \in \mathcal{D})'] \ .$$

Proof. The proof is omitted, but follows similarly to previous results using arguments from the proofs of Theorem 3.1 in Bai et al. (2023c) and Lemma B.1.2. ■

APPENDIX C

APPENDIX FOR CHAPTER 3

C.1 Proofs of Main Results

C.1.1 Proof of Theorem 3.3.1

To begin with, both estimators can be written as follows.

$$\hat{\theta}_1^P = \frac{1}{G_1} \sum_{1 \le g \le G} I\{H_g = \pi_2\} \bar{Y}_g(1, \pi_2) - \frac{1}{G_0} \sum_{1 \le g \le G} I\{H_g = 0\} \bar{Y}_g(0, 0) \ ,$$
$$\hat{\theta}_1^S = \frac{1}{G_1} \sum_{1 \le g \le G} I\{H_g = \pi_2\} \bar{Y}_g(0, \pi_2) - \frac{1}{G_0} \sum_{1 \le g \le G} I\{H_g = 0\} \bar{Y}_g(0, 0) \ .$$

By Lemma 5.1 of Bugni et al. (2022b) and Assumption 3.2.2 (a)-(b), we have $((\bar{Y}_g(1, \pi_2), \bar{Y}_g(0, \pi_2), \bar{Y}_g(0, 0)) : 1 \le g \le G)$ being an i.i.d sequence of random variables. Then, by the law of iterated expectation and Assumption 3.3.2 and 3.2.2 (f),

$$\begin{aligned}
E\left[\bar{Y}_g(1, \pi_2)\right] &= E\left[E\left[\frac{1}{M_g^1} \sum_{i \in \mathcal{M}_g} Y_{i,g}(1, \pi_2) Z_{i,g}(\pi_2) \mid B_g, \mathcal{M}_g\right]\right] \\
&= E\left[\frac{1}{M_g^1} \sum_{i \in \mathcal{M}_g} E\left[Y_{i,g}(1, \pi_2) Z_{i,g}(\pi_2) \mid B_g, \mathcal{M}_g\right]\right] \\
&= E\left[\frac{1}{M_g^1} \sum_{i \in \mathcal{M}_g} E\left[Y_{i,g}(1, \pi_2) \mid B_g, \mathcal{M}_g\right] E\left[Z_{i,g}(\pi_2) \mid B_g\right]\right] \\
&= E\left[\frac{1}{M_g} \sum_{i \in \mathcal{M}_g} E\left[Y_{i,g}(1, \pi_2) \mid B_g, \mathcal{M}_g\right]\right] = E\left[\frac{1}{N_g} \sum_{1 \le i \le N_g} Y_{i,g}(1, \pi_2)\right] \ .
\end{aligned}$$

Similarly,

$$E[\bar{Y}_g(0, \pi_2)] = E\left[\frac{1}{N_g} \sum_{1 \le i \le N_g} Y_{i,g}(0, \pi_2)\right] , \quad E[\bar{Y}_g(0, 0)] = E\left[\frac{1}{N_g} \sum_{1 \le i \le N_g} Y_{i,g}(0, 0)\right] \ .$$

Thus, $\theta_1^P = E\left[\bar{Y}_g(1, \pi_2)\right] - E[\bar{Y}_g(0, 0)]$ and $\theta_1^S = E\left[\bar{Y}_g(0, \pi_2)\right] - E[\bar{Y}_g(0, 0)]$. By Assumption 3.3.1-3.3.2, we have

$$H^{(G)} \perp\!\!\!\perp ((\bar{Y}_g(1, \pi_2), \bar{Y}_g(0, \pi_2), \bar{Y}_g(0, 0)) : 1 \le g \le G) \mid S^{(G)} \ .$$

By Assumption 3.2.2 (c)-(d),

$$E\left[\bar{Y}_g^2(1,\pi_2)\right] = E\left[\left(\frac{1}{M_g^1}\sum_{i\in\mathcal{M}_g} Y_{i,g}(1,\pi_2)Z_{i,g}(\pi_2)\right)^2\right] \leq E\left[\left(\max_{1\leq i\leq N_g} Y_{i,g}(1,\pi_2)\right)^2\right] < \infty \ .$$

Same conclusions hold for $\bar{Y}_g^2(0,\pi_2)$ and $\bar{Y}_g^2(0,0)$. Then, the result follows directly by Theorem 4.1 of Bugni et al. (2018a) and Lemma C.2.3 and Assumption 3.3.1-3.2.2. ■

C.1.2 Proof of Theorem 3.3.2

To preserve space, I only present proof for primary effect as the proof for spillover effect follows the same argument. Define $\mathbf{L}_G = \left(\mathbf{L}_G^{\mathrm{YN1}}, \mathbf{L}_G^{\mathrm{N1}}, \mathbf{L}_G^{\mathrm{YN0}}, \mathbf{L}_G^{\mathrm{N0}}\right)$ as follows.

$$\begin{aligned}
\mathbf{L}_G^{\mathrm{YN1}} &:= \frac{1}{G_1}\sum_{1\leq g\leq G}\left(\bar{Y}_g(1,\pi_2)N_g - E\left[\bar{Y}_g(1,\pi_2)N_g\right]\right) I\{H_g=\pi_2\} \ ,\\
\mathbf{L}_G^{\mathrm{N1}} &:= \frac{1}{G_1}\sum_{1\leq g\leq G}\left(N_g - E\left[N_g\right]\right) I\{H_g=\pi_2\} \ ,\\
\mathbf{L}_G^{\mathrm{YN0}} &:= \frac{1}{G_0}\sum_{1\leq g\leq G}\left(\bar{Y}_g(0,0)N_g - E\left[\bar{Y}_g(0,0)N_g\right]\right) I\{H_g=0\} \ ,\\
\mathbf{L}_G^{\mathrm{N0}} &:= \frac{1}{G_0}\sum_{1\leq g\leq G}\left(N_g - E\left[N_g\right]\right) I\{H_g=0\} \ .
\end{aligned}$$

By the law of iterated expectation and Assumption 3.2.2 (f),

$$\begin{aligned}
E\left[\bar{Y}_g(1,\pi_2)N_g\right] &= E\left[N_g E\left[\bar{Y}_g(1,\pi_2)\mid N_g\right]\right] = E\left[N_g E\left[\frac{1}{M_g}\sum_{i\in\mathcal{M}_g} Y_{i,g}(1,\pi_2)\mid N_g\right]\right]\\
&= E\left[N_g E\left[\frac{1}{N_g}\sum_{1\leq i\leq N_g} Y_{i,g}(1,\pi_2)\mid N_g\right]\right] = E\left[\sum_{1\leq i\leq N_g} Y_{i,g}(1,\pi_2)\right] .
\end{aligned}$$

Thus,

$$\theta_2^P = \frac{E\left[\bar{Y}_g(1,\pi_2)N_g\right]}{E[N_g]} - \frac{E\left[\bar{Y}_g(0,0)N_g\right]}{E[N_g]} \text{ and } \theta_2^S = \frac{E\left[\bar{Y}_g(0,\pi_2)N_g\right]}{E[N_g]} - \frac{E\left[\bar{Y}_g(0,0)N_g\right]}{E[N_g]} \ .$$

Note that $\frac{G_1}{G} = \frac{D_G}{G} + \pi_1$. Thus,

$$\sqrt{G}\mathbf{L}_G^{\mathrm{YN1}} = \left(\frac{D_G}{G}+\pi_1\right)^{-1}\left(1-\pi_1-\frac{D_G}{G}\right)^{-1}\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\left(\left(1-\pi-\frac{D_G}{G}\right)\left(\bar{Y}_g(1,\pi_2)N_g-\mu_1\right) I\{H_g=\pi_2\}\right) \ ,$$

where $E\left[\bar{Y}_g(1,\pi_2)N_g\right]=\mu_1$. By Lemma B.1 and B.3 of Bugni et al. (2018a), Lemma C.2.3 and Assumption 3.3.1-3.2.2, we have

$$\sqrt{G}\mathbf{L}_G^{\mathrm{YN1}}=(\pi_1(1-\pi_1))^{-1}\underbrace{\frac{1}{\sqrt{G}}\sum_{1\le g\le G}\left((1-\pi_1)\left(\bar{Y}_g(1,\pi_2)N_g-E\left[\bar{Y}_g(1,\pi_2)N_g\right]\right)I\{H_g=\pi_2\}\right)}_{:=L_G^{\mathrm{YN1}}}+o_P(1)\ .$$

Similarly,

$$\sqrt{G}\mathbf{L}_G^{\mathrm{N1}}=(\pi_1(1-\pi_1))^{-1}\underbrace{\frac{1}{\sqrt{G}}\sum_{1\le g\le G}\left((1-\pi_1)\left(N_g-E\left[N_g\right]\right)I\{H_g=\pi_2\}\right)}_{:=L_G^{\mathrm{N1}}}+o_P(1)\ ,$$

$$\sqrt{G}\mathbf{L}_G^{\mathrm{YN0}}=(\pi_1(1-\pi_1))^{-1}\underbrace{\frac{1}{\sqrt{G}}\sum_{1\le g\le G}\left(\pi_1\left(\bar{Y}_g(0,0)N_g-E\left[\bar{Y}_g(0,0)N_g\right]\right)I\{H_g=0\}\right)}_{:=L_G^{\mathrm{YN0}}}+o_P(1)\ ,$$

$$\sqrt{G}\mathbf{L}_G^{\mathrm{N0}}=(\pi_1(1-\pi_1))^{-1}\underbrace{\frac{1}{\sqrt{2n}}\sum_{1\le i\le 2n}\left(\pi_1\left(N_g-E\left[N_g\right]\right)I\{H_g=0\}\right)}_{:=L_G^{\mathrm{N0}}}+o_P(1)\ .$$

Define

$$\begin{aligned}
&\tilde{Y}_g^{\mathrm{N}}(z,h)=\bar{Y}_g(z,h)N_g-E\left[\bar{Y}_g(z,h)N_g\mid S_g\right]\ ,\\
&\tilde{N}_g=N_g-E\left[N_g\mid S_g\right]\ ,\\
&m_{z,h}^{\mathrm{YN}}(S_g)=E\left[Y_g(z,h)N_g\mid S_g\right]-E\left[Y_g(z,h)N_g\right]\ ,\\
&m^{\mathrm{N}}(S_g)=E\left[N_g\mid S_g\right]-E\left[N_g\right]\ ,
\end{aligned}$$

and consider the following decomposition for L_G^{YN1}:

$$\begin{aligned}
L_G^{\mathrm{YN1}}&=R_{n,1}+R_{n,2}+R_{n,3}\\
&=\frac{\pi_1(1-\pi_1)}{\sqrt{G}}\sum_{1\le g\le G}\frac{1}{\pi_1}\tilde{Y}_g^{\mathrm{N}}(1,\pi_2)I\{H_g=\pi_2\}+\pi_1(1-\pi_1)\sum_{s\in\mathcal{S}}\frac{D_G(s)}{\sqrt{G}}\frac{1}{\pi_1}m_{1,\pi_2}^{\mathrm{YN}}(S_g)\\
&\quad+\pi_1(1-\pi_1)\sum_{s\in\mathcal{S}}\sqrt{G}\left(\frac{G(s)}{G}-p(s)\right)m_{1,\pi_2}^{\mathrm{YN}}(S_g)\ .
\end{aligned}$$

Similarly, we have the same decomposition for $L_G^{\mathrm{YN}0}, L_G^{\mathrm{N}1}, L_G^{\mathrm{N}0}$. Define

$$\begin{aligned}
\mathbf{d} &:= \left(\frac{D_G(s)}{\sqrt{G}} : s \in \mathcal{S}\right)' \\
\mathbf{n} &:= \left(\sqrt{G}\left(\frac{G(s)}{G} - p(s)\right) : s \in \mathcal{S}\right)' \\
\mathbf{m}_{z,h}^{\mathrm{YN}} &:= \left(E\left[m_{z,h}^{\mathrm{YN}}(C_g) \mid S_g = s\right] : s \in \mathcal{S}\right)' \\
\mathbf{m}^{\mathrm{N}} &:= \left(E\left[m^{\mathrm{N}}(C_g) \mid S_g = s\right] : s \in \mathcal{S}\right)' \ .
\end{aligned}$$

Then, we can write

$$(\pi_1(1-\pi_1))^{-1}\begin{pmatrix} L_G^{\mathrm{YN}1} \\ L_G^{\mathrm{N}1} \\ L_G^{\mathrm{YN}0} \\ L_G^{\mathrm{N}0} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & \frac{1}{\pi_1}\left(\mathbf{m}_{1,\pi_2}^{\mathrm{YN}}\right)' & \left(\mathbf{m}_{1,\pi_2}^{\mathrm{YN}}\right)' \\ 0 & 1 & 0 & 0 & \frac{1}{\pi_1}\left(\mathbf{m}^{\mathrm{N}}\right)' & \left(\mathbf{m}^{\mathrm{N}}\right)' \\ 0 & 0 & 1 & 0 & -\frac{1}{1-\pi_1}\left(\mathbf{m}_{0,0}^{\mathrm{YN}}\right)' & \left(\mathbf{m}_{0,0}^{\mathrm{YN}}\right)' \\ 0 & 0 & 0 & 1 & -\frac{1}{1-\pi_1}\left(\mathbf{m}^{\mathrm{N}}\right)' & \left(\mathbf{m}^{\mathrm{N}}\right)' \end{pmatrix}}_{:=M'} \underbrace{\begin{pmatrix} \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\frac{1}{\pi_1}\tilde{Y}_g^{\mathrm{YN}}(1,\pi_2)I\{H_g=\pi_2\} \\ \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\frac{1}{\pi_1}\tilde{N}_g I\{H_g=\pi_2\} \\ \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\frac{1}{1-\pi_1}\tilde{Y}_g^{\mathrm{YN}}(0,0)I\{H_g=0\} \\ \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\frac{1}{1-\pi_1}\tilde{N}_g I\{H_g=0\} \\ \mathbf{d} \\ \mathbf{n} \end{pmatrix}}_{:=\mathbf{y}_n}$$

Following Lemma B.2 from Bugni et al. (2018a), we have

$$\mathbf{y}_n \xrightarrow{d} \mathcal{N}(0,\Sigma) \ ,$$

where

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

for

$$\Sigma_1 = \begin{pmatrix} \frac{\mathrm{Var}\left[\tilde{Y}_g^{\mathrm{YN}}(1,\pi_2)\right]}{\pi_1} & \frac{E\left[\tilde{Y}_g^{\mathrm{YN}}(1,\pi_2)N_g\right]}{\pi_1} \\ \frac{E\left[\tilde{Y}_g^{\mathrm{YN}}(1,\pi_2)N_g\right]}{\pi_1} & \frac{\mathrm{Var}[N_g]}{\pi_1} \end{pmatrix}, \qquad \Sigma_0 = \begin{pmatrix} \frac{\mathrm{Var}\left[\tilde{Y}_g^{\mathrm{YN}}(0,0)\right]}{1-\pi_1} & \frac{E\left[\tilde{Y}_g^{\mathrm{YN}}(0,0)N_g\right]}{1-\pi_1} \\ \frac{E\left[\tilde{Y}_g^{\mathrm{YN}}(0,0)N_g\right]}{1-\pi_1} & \frac{\mathrm{Var}[N_g]}{1-\pi_1} \end{pmatrix} \ ,$$

$$\Sigma_D = \mathrm{diag}\left(p(s)\tau(s) : s \in \mathcal{S}\right), \qquad \Sigma_N = \mathrm{diag}\left(p(s) : s \in \mathcal{S}\right) - \left(p(s) : s \in \mathcal{S}\right)\left(p(s) : s \in \mathcal{S}\right)' \ .$$

Let $\mathbf{m}(S_g) = \left(m_{1,\pi_2}^{\mathrm{YN}}(S_g), m_0^{\mathrm{N}}(S_g), m_{0,0}^{\mathrm{YN}}(S_g), m_0^{\mathrm{N}}(S_g)\right)'$. We have

$$\mathbb{V} = M'\Sigma M = \mathbb{V}_1 + \mathbb{V}_2 + \mathbb{V}_3,$$

where

$$\begin{aligned}
\mathbb{V}_1 &= \begin{pmatrix} \frac{1}{\pi_1}\mathrm{Var}\left[\tilde{Y}_g^{\mathrm{YN}}(1,\pi_2)\right] & \frac{1}{\pi_1}E\left[\tilde{Y}_g^{\mathrm{YN}}(1,\pi_2)N_g\right] & 0 & 0 \\ \frac{1}{\pi_1}E\left[\tilde{Y}_g^{\mathrm{YN}}(1,\pi_2)N_g\right] & \frac{1}{\pi_1}\mathrm{Var}\left[N_g\right] & 0 & 0 \\ 0 & 0 & \frac{1}{1-\pi_1}\mathrm{Var}\left[\tilde{Y}_g^{\mathrm{YN}}(0,0)\right] & \frac{1}{1-\pi_1}E\left[\tilde{Y}_g^{\mathrm{YN}}(0,0)N_g\right] \\ 0 & 0 & \frac{1}{1-\pi_1}E\left[\tilde{Y}_g^{\mathrm{YN}}(0,0)N_g\right] & \frac{1}{1-\pi_1}\mathrm{Var}\left[N_g\right] \end{pmatrix}, \\
\mathbb{V}_2 &= \mathrm{Var}\left[\mathbf{m}(S_g)\right] \ , \\
\mathbb{V}_3 &= E\left[\tau(S_g)\left(\Lambda\mathbf{m}(S_g)\mathbf{m}(S_g)'\Lambda\right)\right] \text{ with } \Lambda = \mathrm{diag}\left(\frac{1}{\pi_1}, \frac{1}{\pi_1}, -\frac{1}{1-\pi_1}, -\frac{1}{1-\pi_1}\right) \ .
\end{aligned}$$

Alternatively,

$$
\begin{aligned}
\mathbb{V}_{11} &= \frac{1}{\pi_1}\operatorname{Var}\left[\bar{Y}_g(1,\pi_2)N_g\right] - \frac{1-\pi_1}{\pi_1}\operatorname{Var}\left[E\left[\bar{Y}_g(1,\pi_2)N_g \mid S_g\right]\right] \\
&\quad + E\left[\frac{\tau(S_g)}{\pi_1^2}\left(E[\bar{Y}_g(1,\pi_2)N_g \mid S_g] - E[\bar{Y}_g(1,\pi_2)N_g]\right)^2\right] \\
\mathbb{V}_{12} &= \frac{1}{\pi_1}\operatorname{Cov}[\bar{Y}_g(1,\pi_2)N_g, N_g] - \frac{1-\pi_1}{\pi_1}\operatorname{Cov}[E[\bar{Y}_g(1,\pi_2)N_g|S_g], E[N_g|S_g]] \\
&\quad + E\left[\frac{\tau(S_g)}{\pi_1^2}\left(E[\bar{Y}_g(1,\pi_2)N_g \mid S_g] - E[\bar{Y}_g(1,\pi_2)N_g]\right)\left(E[N_g \mid S_g] - E[N_g]\right)\right] \\
\mathbb{V}_{13} &= \operatorname{Cov}[E[\bar{Y}_g(1,\pi_2)N_g|S_g], E[\bar{Y}_g(0,0)N_g|S_g]] \\
&\quad - E\left[\frac{\tau(S_g)}{\pi_1(1-\pi_1)}\left(E[\bar{Y}_g(1,\pi_2)N_g \mid S_g] - E[\bar{Y}_g(1,\pi_2)N_g]\right)\left(E[\bar{Y}_g(0,0)N_g \mid S_g] - E[\bar{Y}_g(0,0)N_g]\right)\right] \\
\mathbb{V}_{14} &= \operatorname{Cov}[E[\bar{Y}_g(1,\pi_2)N_g|S_g], E[N_g|S_g]] \\
&\quad - E\left[\frac{\tau(S_g)}{\pi_1(1-\pi_1)}\left(E[\bar{Y}_g(1,\pi_2)N_g \mid S_g] - E[\bar{Y}_g(1,\pi_2)N_g]\right)\left(E[N_g \mid S_g] - E[N_g]\right)\right] \\
\mathbb{V}_{22} &= \frac{1}{\pi_1}\operatorname{Var}[N_g] - \frac{1-\pi_1}{\pi_1}\operatorname{Var}[E[N_g|S_g]] \\
&\quad + E\left[\frac{\tau(S_g)}{\pi_1^2}\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(0,0)N_g|S_g]] \\
&\quad - E\left[\frac{\tau(S_g)}{\pi_1(1-\pi_1)}\left(E[N_g \mid S_g] - E[N_g]\right)\left(E[\bar{Y}_g(0,0)N_g \mid S_g] - E[\bar{Y}_g(0,0)N_g]\right)\right] \\
\mathbb{V}_{24} &= \operatorname{Cov}[E[N_g|S_g], E[N_g|S_g]] \\
&\quad - E\left[\frac{\tau(S_g)}{\pi_1(1-\pi_1)}\left(E[N_g \mid S_g] - E[N_g]\right)^2\right] \\
\mathbb{V}_{33} &= \frac{1}{1-\pi_1}\operatorname{Var}[\bar{Y}_g(0,0)N_g] - \frac{\pi_1}{1-\pi_1}\operatorname{Var}[E[\bar{Y}_g(0,0)N_g|S_g]] \\
&\quad + E\left[\frac{\tau(S_g)}{(1-\pi_1)^2}\left(E[\bar{Y}_g(0,0)N_g \mid S_g] - E[\bar{Y}_g(0,0)N_g]\right)^2\right] \\
\mathbb{V}_{34} &= \frac{1}{1-\pi_1}\operatorname{Cov}[\bar{Y}_g(0,0)N_g, N_g] - \frac{\pi_1}{1-\pi_1}\operatorname{Cov}[E[\bar{Y}_g(0,0)N_g|S_g], E[N_g|S_g]] \\
&\quad + E\left[\frac{\tau(S_g)}{(1-\pi_1)^2}\left(E[\bar{Y}_g(0,0)N_g \mid S_g] - E[\bar{Y}_g(0,0)N_g]\right)\left(E[N_g \mid S_g] - E[N_g]\right)\right] \\
\mathbb{V}_{44} &= \frac{1}{1-\pi_1}\operatorname{Var}[N_g] - \frac{\pi_1}{1-\pi_1}\operatorname{Var}[E[N_g|S_g]] \\
&\quad + E\left[\frac{\tau(S_g)}{(1-\pi_1)^2}\left(E[N_g \mid S_g] - E[N_g]\right)^2\right] .
\end{aligned}
$$

Therefore,

$$\sqrt{G}(\hat{\beta}-\beta) := \sqrt{G}\left(\mathbf{L}_G^{\mathrm{YN1}}, \mathbf{L}_G^{\mathrm{N1}}, \mathbf{L}_G^{\mathrm{YN0}}, \mathbf{L}_G^{\mathrm{N0}}\right)' = (\pi(1-\pi))^{-1} \cdot \left(L_G^{\mathrm{YN1}}, L_G^{\mathrm{N1}}, L_G^{\mathrm{YN0}}, L_G^{\mathrm{N0}}\right)' + o_P(1) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}) .$$

Let $g(x,y,z,w) = \frac{x}{y} - \frac{z}{w}$. Note that the Jacobian is

$$D_g(x,y,z,w) = \left(\frac{1}{y}, -\frac{x}{y^2}, -\frac{1}{w}, \frac{z}{w^2}\right) .$$

By delta method,

$$\sqrt{2n}(\hat{\theta}_2^P - \theta_2^P) = \sqrt{2n}(g(\hat{\beta}) - g(\beta)) \xrightarrow{d} \mathcal{N}(0, V_2(1)) ,$$

where

$$V_2(1) = D_g' \left(\mathbb{V}_1 + \mathbb{V}_2 + \mathbb{V}_3\right) D_g$$

for

$$D_g = \left(\frac{1}{\pi_1 E[N_g]}, -\frac{E[\bar{Y}_g(1,\pi_2)N_g]}{\pi_1 E[N_g]^2}, -\frac{1}{(1-\pi_1)E[N_g]}, \frac{E[\bar{Y}_g(0,0)N_g]}{(1-\pi_1)E[N_g]^2}\right)' .$$

By simple calculation,

$$\begin{aligned} D_g' \left(\mathbb{V}_1 + \mathbb{V}_2\right) D_g &= \frac{1}{\pi_1} \operatorname{Var}[\tilde{Y}_g(z,\pi_2)] + \frac{1}{1-\pi_1} \operatorname{Var}[\tilde{Y}_g(0,0)] \\ &\quad - E\left[E\left[\sqrt{\frac{1-\pi_1}{\pi_1}}\tilde{Y}_g(z,\pi_2) + \sqrt{\frac{\pi_1}{1-\pi_1}}\tilde{Y}_i(0,0)\middle| S_g\right]^2\right] \\ D_g' \mathbb{V}_3 D_g &= E\left[\tau(S_g)\left(\frac{m_{1,\pi_2}^{\mathrm{YN}}(S_g)}{\pi_1 E[N_g]} - \frac{E[Y_i(1)N_g]m^{\mathrm{N}}(S_g)}{\pi_1 E[N_g]^2} + \frac{m_{0,0}^{\mathrm{YN}}(S_g)}{(1-\pi_1)E[N_g]} - \frac{E[Y_i(0)N_g]m^{\mathrm{N}}(S_g)}{(1-\pi_1)E[N_g]^2}\right)^2\right] \\ &= E\left[\tau(S_g)\left(\frac{1}{\pi_1}E[\tilde{Y}_g(z,\pi_2) \mid S_g] + \frac{1}{1-\pi_1}E[\tilde{Y}_g(0,0) \mid S_g]\right)^2\right] . \end{aligned}$$

Thus, the result follows. ■

C.1.3 Proof of Theorem 3.3.3

The conclusion follows by continuous mapping theorem and by showing the following results:

(a) $\frac{G(s)}{G} \xrightarrow{P} p(s)$.

(b) $\frac{1}{G_a} \sum_{1 \leq g \leq G} \left(\bar{Y}_g^z\right)^r I\{H_g = h\} \xrightarrow{P} E[\bar{Y}_g(z,h)^r]$ for $r, z \in \{0,1\}$ and $(a,h) \in \{(1,\pi_2),(0,0)\}$.

(c) $\frac{1}{G_a(s)} \sum_{1 \leq g \leq G} \bar{Y}_g^z I\left\{H_g = h, S_g = s\right\} \xrightarrow{P} E[\bar{Y}_g(z,h) \mid S_g]$ for $z \in \{0,1\}$ and $(a,h) \in \{(1,\pi_2),(0,0)\}$.

(d) $\frac{1}{G_a} \sum_{1 \leq g \leq G} \left(\tilde{Y}_g^z\right)^r I\{H_g = h\} \xrightarrow{P} E[\tilde{Y}_g(z,h)^r]$ for $r, z \in \{0,1\}$ and $(a,h) \in \{(1,\pi_2),(0,0)\}$.

(e) $\frac{1}{G_a(s)} \sum_{1 \leq g \leq G} \tilde{Y}_g^z I\left\{H_g = h, S_g = s\right\} \xrightarrow{P} E[\tilde{Y}_g(z,h) \mid S_g]$ for $z \in \{0,1\}$ and $(a,h) \in \{(1,\pi_2),(0,0)\}$

By following the arguments in Appendix A.2 of Bugni et al. (2018a), Lemma C.2.3 and Assumption 3.3.1-3.2.2, we conclude that (a), (b) and (c) hold. Next, I first show the results hold for $\tilde{Y}_g(z,h)$ and then analyze the difference between $\tilde{Y}_g(z,h)$ and adjusted version $\hat{Y}_g^z(h)$ defined as follows:

$$\begin{aligned}\hat{Y}_g^z(\pi_2) &= \frac{N_g}{\frac{1}{G}\sum_{1\leq g\leq G} N_g}\left(\bar{Y}_g(z,\pi_2) - \frac{\frac{1}{G_1}\sum_{1\leq j\leq G}\bar{Y}_j(z,\pi_2)I\{H_j=\pi_2\}N_j}{\frac{1}{G}\sum_{1\leq j\leq G}N_j}\right)\\ \hat{Y}_g^z(0) &= \frac{N_g}{\frac{1}{G}\sum_{1\leq g\leq G} N_g}\left(\bar{Y}_g(0,0) - \frac{\frac{1}{G_0}\sum_{1\leq j\leq G}\bar{Y}_j(0,0)I\{H_j=0\}N_j}{\frac{1}{G}\sum_{1\leq j\leq G}N_j}\right)\ ,\end{aligned} \tag{C.1}$$

for which the usual relationship still holds for adjusted outcomes, i.e. $\tilde{Y}_g^z = \sum_{h\in\{0,\pi_2\}} I\{H_g=h\}\hat{Y}_g^z(h)$. Note that

$$\begin{aligned}E[\tilde{Y}_g(z,h)^2] &= E\left[\frac{N_g^2}{E[N_g]^2}\left(\bar{Y}_g(z,h) - \frac{E[\bar{Y}_g(z,h)N_g]}{E[N_g]}\right)^2\right] \leq 2E\left[\frac{N_g^2}{E[N_g]^2}\left(\bar{Y}_g(z,h)^2 + \frac{E[\bar{Y}_g(z,h)N_g]^2}{E[N_g]^2}\right)\right]\\ &\leq 2E\left[N_g^2\bar{Y}_g(z,h)^2\right] + 2E[\bar{Y}_g(z,h)N_g]^2E[N_g^2] < \infty\ .\end{aligned}$$

where the first inequality holds by the fact $(a-b)^2 \leq 2a^2+2b^2$, the second inequality follows by the fact that $E[N_g]\geq 1$, and the last inequality follows by Lemma C.2.3. Therefore, again by following the arguments in Appendix A.2 of Bugni et al. (2018a), we conclude that for $r, z \in \{0,1\}$ and $(a,h)\in\{(1,\pi_2),(0,0)\}$,

$$\begin{aligned}&\frac{1}{G_a}\sum_{1\leq g\leq G}\tilde{Y}_g(z,h)^r I\{H_g=h\} \xrightarrow{P} E[\tilde{Y}_g(z,h)^r]\\ &\frac{1}{G_a(s)}\sum_{1\leq g\leq G}\tilde{Y}_g(z,h)I\left\{H_g=h, S_g=s\right\} \xrightarrow{P} E[\tilde{Y}_g(z,h)\mid S_g]\ ,\end{aligned}$$

Finally, I show the difference between the above equations with $\tilde{Y}_g(z,h)$ and $\tilde{Y}_g^z$ go to zero. Here, I prove this for the following case,

$$\frac{1}{G_1}\sum_{1\leq g\leq G}\left(\tilde{Y}_g(1,\pi_2)^2 - \left(\tilde{Y}_g^1\right)^2\right)I\{H_g=\pi_2\} \xrightarrow{P} 0\ ; \tag{C.2}$$

an analogous argument establishes the rest. Note that

$$
\begin{aligned}
&\frac{1}{G_1}\sum_{1\leq g\leq G}\left(\tilde{Y}_g(1,\pi_2)^2-\left(\hat{Y}_g^1\right)^2\right)I\{H_g=\pi_2\}\\
&=\frac{1}{G_1}\sum_{1\leq g\leq G}\left(\tilde{Y}_g(1,\pi_2)-\hat{Y}_g^1(\pi_2)\right)\left(\tilde{Y}_g(1,\pi_2)+\hat{Y}_g^1(\pi_2)\right)I\{H_g=\pi_2\}\\
&=\frac{1}{G_1}\sum_{1\leq g\leq G}\left(\frac{1}{E[N_g]}-\frac{1}{\frac{1}{G}\sum_{1\leq g\leq G}N_g}\right)\bar{Y}_g(1,\pi_2)N_g\left(\tilde{Y}_g(1,\pi_2)+\hat{Y}_g^1(\pi_2)\right)I\{H_g=\pi_2\}\\
&\quad-\frac{1}{G_1}\sum_{1\leq g\leq G}\left(\frac{\frac{1}{G}\sum_{1\leq g\leq G}\bar{Y}_g(1,\pi_2)I\{H_g=\pi_2\}N_g}{\left(\frac{1}{G}\sum_{1\leq g\leq G}N_g\right)^2}-\frac{E[\bar{Y}_g(1,\pi_2)N_g]}{E[N_g]^2}\right)\\
&\qquad\times N_g\left(\tilde{Y}_g(1,\pi_2)+\hat{Y}_g^1(\pi_2)\right)I\{H_g=\pi_2\}\ .
\end{aligned}
$$

I then proceed to prove the following statement

$$
\frac{1}{G_1}\sum_{1\leq g\leq G}\bar{Y}_g(1,\pi_2)N_g\left(\tilde{Y}_g(1,\pi_2)+\hat{Y}_g^1(\pi_2)\right)I\{H_g=\pi_2\}\xrightarrow{P}2E[\tilde{Y}_g(1,\pi_2)\bar{Y}_g(1,\pi_2)N_g]\ , \tag{C.3}
$$

and similar arguments would prove the following statement

$$
\frac{1}{G_1}\sum_{1\leq g\leq G}N_g\left(\tilde{Y}_g(1,\pi_2)+\hat{Y}_g^1(\pi_2)\right)I\{H_g=\pi_2\}\xrightarrow{P}2E[N_g\tilde{Y}_g(1,\pi_2)]\ .
$$

Note that

$$
\begin{aligned}
&\frac{1}{G_1}\sum_{1\leq g\leq G}\bar{Y}_g(1,\pi_2)N_g\left(\tilde{Y}_g(1,\pi_2)+\hat{Y}_g^1(\pi_2)\right)I\{H_g=\pi_2\}\\
&=\frac{1}{G_1}\sum_{1\leq g\leq G}2\bar{Y}_g(1,\pi_2)N_g\tilde{Y}_g(1,\pi_2)I\{H_g=\pi_2\}+\frac{1}{G_1}\sum_{1\leq g\leq G}\bar{Y}_g(1,\pi_2)N_g\left(\hat{Y}_g^1(\pi_2)-\tilde{Y}_g(1,\pi_2)\right)I\{H_g=\pi_2\}\ .
\end{aligned}
$$

By weak law of large number, Slutsky's theorem and arguments in the proof of Theorem 3.3.2, we have

$$
\begin{aligned}
\frac{1}{\frac{1}{G}\sum_{1\leq g\leq G}N_g}&\xrightarrow{P}\frac{1}{E[N_g]}\\
\frac{\frac{1}{G}\sum_{1\leq g\leq G}\bar{Y}_g(1,\pi_2)I\{H_g=\pi_2\}N_g}{\left(\frac{1}{G}\sum_{1\leq g\leq G}N_g\right)^2}&\xrightarrow{P}\frac{E[\bar{Y}_g(1,\pi_2)N_g]}{E[N_g]^2}\ .
\end{aligned}
$$

Then, by Slutsky's theorem, Lemma C.2.3 and Lemma B.3 of Bugni et al. (2018a),

$$\begin{aligned}&\frac{1}{G_1}\sum_{1\leq g\leq G}\bar{Y}_g(1,\pi_2)N_g\left(\hat{Y}_g^1(\pi_2)-\tilde{Y}_g(1,\pi_2)\right)I\{H_g=\pi_2\}\\&=\left(\frac{1}{E[N_g]}-\frac{1}{\frac{1}{G}\sum_{1\leq g\leq G}N_g}\right)\frac{1}{G_1}\sum_{1\leq g\leq G}\bar{Y}_g(1,\pi_2)^2N_g^2I\{H_g=\pi_2\}\xrightarrow{P}0\ .\end{aligned}$$

Again, by Lemma B.3 of Bugni et al. (2018a), and

$$E[\bar{Y}_g(1,\pi_2)N_g\tilde{Y}_g(1,\pi_2)]=\frac{E[\bar{Y}_g(1,\pi_2)^2N_g^2]}{E[N_g]}-\frac{E[\bar{Y}_g(1,\pi_2)N_g]E[\bar{Y}_g(1,\pi_2)N_g^2]}{E[N_g]^2}<\infty,$$

We conclude that (C.3) holds, and then

$$\frac{1}{G_1}\sum_{1\leq g\leq G}\left(\frac{1}{E[N_g]}-\frac{1}{\frac{1}{G}\sum_{1\leq g\leq G}N_g}\right)\bar{Y}_g(1,\pi_2)N_g\left(\tilde{Y}_g(1,\pi_2)+\hat{Y}_g^1(\pi_2)\right)I\{H_g=\pi_2\}\xrightarrow{P}0\ .$$

Therefore, (C.2) holds. ■

C.1.4 Proof of Theorem 3.4.1

To preserve space, I only present the proof for primary effect as the proof for spillover effect follows the same argument. First, I analyze the equally-weighted estimator. Note that

$$\sqrt{G}(\hat{\theta}_1^P-\theta_1^P)=(\mathbb{L}_G^{\mathrm{Y}1},\mathbb{L}_G^{\mathrm{Y}0})D_h\ ,$$

where $D_h=\left(\frac{1}{\sqrt{\pi_1}},-\frac{1}{\sqrt{1-\pi_1}}\right)'$ and $\mathbb{L}_G^{\mathrm{Y}1},\mathbb{L}_G^{\mathrm{Y}0}$ are defined in Lemma C.2.4. Thus, by Lemma C.2.4,

$$\sqrt{G}(\hat{\theta}_1^P-\theta_1^P)\to\mathcal{N}(0,D_h'\mathbf{V}^{\mathrm{e}}D_h),$$

where

$$\mathbf{V}^{\mathrm{e}}=\begin{pmatrix}E[\mathrm{Var}[\bar{Y}_g(1,\pi_2)|S_g]] & 0\\ 0 & E[\mathrm{Var}[\bar{Y}_g(0,0)|S_g]]\end{pmatrix}+\mathrm{Var}\left[\begin{pmatrix}\sqrt{\pi_1}E[\bar{Y}_g(1,\pi_2)|S_g]\\ \sqrt{1-\pi_1}E[\bar{Y}_g(0,0)|S_g]\end{pmatrix}\right]\ ,$$

By simple calculation, we conclude that $D_h' \mathbf{V}^{\mathrm{e}} D_h = V_3(1)$. In order to calculate the variance of size-weighted estimator, I follow the same argument in the end of C.1.1. Note that

$$\sqrt{G}(\hat{\beta}-\beta) = \sqrt{G}\left(\frac{\mathbb{L}_G^{\mathrm{YN1}}}{\sqrt{G_1}}, \frac{\mathbb{L}_G^{\mathrm{N1}}}{\sqrt{G_1}}, \frac{\mathbb{L}_G^{\mathrm{YN0}}}{\sqrt{G_0}}, \frac{\mathbb{L}_G^{\mathrm{N0}}}{\sqrt{G_0}}\right) = \left(\frac{1}{\sqrt{\pi_1}}, \frac{1}{\sqrt{\pi_1}}, \frac{1}{\sqrt{1-\pi_1}}, \frac{1}{\sqrt{1-\pi_1}}\right)\begin{pmatrix} \mathbb{L}_G^{\mathrm{YN1}} \\ \mathbb{L}_G^{\mathrm{N1}} \\ \mathbb{L}_G^{\mathrm{YN0}} \\ \mathbb{L}_G^{\mathrm{N0}} \end{pmatrix} .$$

By a similar calculation and argument in C.1.2 and Lemma C.2.4, the final results is obtained. ■

C.1.5 Proof of Theorem 3.4.2

First, note that we can write the variance expression as follows:

$$\begin{aligned} V_3(z) &= \frac{1}{\pi_1}\operatorname{Var}\left[\bar{Y}_g(z,\pi_2)\right] + \frac{1}{1-\pi_1}\operatorname{Var}\left[\bar{Y}_g(0,0)\right] - \pi_1(1-\pi_1)E\left[\left(\frac{1}{\pi_1}m_{z,\pi_2}(S_g) + \frac{1}{1-\pi_1}m_{0,0}(S_g)\right)^2\right] \\ &= \frac{1}{\pi_1}E\left[\operatorname{Var}\left[\bar{Y}_g(z,\pi_2) \mid S_g\right]\right] + \frac{1}{1-\pi_1}E\left[\operatorname{Var}\left[\bar{Y}_g(0,0) \mid S_g\right]\right] + \operatorname{Var}\left[E\left[\bar{Y}_g(z,\pi_2) \mid S_g\right]\right] \\ &\quad + \operatorname{Var}\left[E\left[\bar{Y}_g(0,0) \mid S_g\right]\right] - 2\cdot\operatorname{Cov}\left[E\left[\bar{Y}_g(z,\pi_2) \mid S_g\right], E\left[\bar{Y}_g(0,0) \mid S_g\right]\right] . \end{aligned}$$

By Slutsky's theorem and Lemma C.3.2-C.3.4, we conclude that $\hat{V}_3(z) \xrightarrow{P} V_3(z)$. Similarly, by Slutsky's theorem and Lemma C.3.5-C.3.7, we conclude that $\hat{V}_4(z) \xrightarrow{P} V_4(z)$.

C.1.6 Proof of Theorem 3.5.1

To begin with, observe that it is equivalent to show $g_z^{\mathrm{e}}(C_g, N_g) = E\left[\frac{\bar{Y}_g(z,\pi_2)}{\pi_1} + \frac{\bar{Y}_g(0,0)}{1-\pi_1} \mid C_g, N_g\right]$ maximizes

$$\begin{aligned} &E\left[\left(\frac{m_{z,\pi_2}(S_g)}{\pi_1} + \frac{m_{0,0}(S_g)}{1-\pi_1}\right)^2\right] \\ &= E\left[\left(\frac{E[\bar{Y}_g(z,\pi_2) \mid S_g] - E[\bar{Y}_g(z,h)]}{\pi_1} + \frac{E[\bar{Y}_g(0,0) \mid S_g] - E[\bar{Y}_g(z,h)]}{1-\pi_1}\right)^2\right] , \end{aligned}$$

and $g_z^{\mathrm{s}}(C_g, N_g) = E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1} + \frac{\tilde{Y}_g(0,0)}{1-\pi_1} \mid C_g, N_g\right]$ maximizes

$$E\left[\left(\frac{1}{\pi_1}E[\tilde{Y}_g(z,\pi_2) \mid S_g] + \frac{1}{1-\pi_1}E[\tilde{Y}_g(0,0) \mid S_g]\right)^2\right] .$$

By Theorem C.2. of Bai et al. (2021b), the result for equally-weighted estimators follow directly. In terms of the size-weighted estimators, first observe that

$$
\begin{aligned}
&E\left[\left(g_z^{\mathrm{s}}(C_g,N_g)-E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1}+\frac{\tilde{Y}_g(0,0)}{1-\pi_1}\middle|S_g\right]\right)E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1}+\frac{\tilde{Y}_g(0,0)}{1-\pi_1}\middle|S_g\right]\right]\\
&=E\left[E\left[\left(g_z^{\mathrm{s}}(C_g,N_g)-E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1}+\frac{\tilde{Y}_g(0,0)}{1-\pi_1}\middle|S_g\right]\right)\middle|S_g\right]E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1}+\frac{\tilde{Y}_g(0,0)}{1-\pi_1}\middle|S_g\right]\right]\\
&=0\ ,
\end{aligned}
$$

by law of iterated expectation. Therefore,

$$
\begin{aligned}
&E\left[g_z^{\mathrm{s}}(C_g,N_g)^2\right]\\
&=E\left[\left(g_z^{\mathrm{s}}(C_g,N_g)-E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1}+\frac{\tilde{Y}_g(0,0)}{1-\pi_1}\middle|S_g\right]+E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1}+\frac{\tilde{Y}_g(0,0)}{1-\pi_1}\middle|S_g\right]\right)^2\right]\\
&=E\left[\left(g_z^{\mathrm{s}}(C_g,N_g)-E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1}+\frac{\tilde{Y}_g(0,0)}{1-\pi_1}\middle|S_g\right]\right)^2\right]+E\left[E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1}+\frac{\tilde{Y}_g(0,0)}{1-\pi_1}\middle|S_g\right]^2\right]\\
&\geq E\left[E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1}+\frac{\tilde{Y}_g(0,0)}{1-\pi_1}\middle|S_g\right]^2\right]\ .
\end{aligned}
$$

Thus, it is optimal to match on

$$
g_z^{\mathrm{s}}(C_g,N_g)=E\left[\frac{\tilde{Y}_g(z,\pi_2)}{\pi_1}+\frac{\tilde{Y}_g(0,0)}{1-\pi_1}\mid C_g,N_g\right]\ .
$$

■

C.1.7 Proof of Theorem 3.5.2

In this section, I show the optimality result holds for $V_1(z)$ first. To begin with, observe that the second stage design enters the variance formula only through $Z_{i,g}$, or in other words $\bar{Y}_g(z,\pi_2)$. Moreover, the conditional expectations, $m_{1,\pi_2}(C_g), m_{0,\pi_2}(C_g), m_{1,\pi_2}(S_g), m_{0,\pi_2}(S_g)$, do not depend on the stratification strategy. Take

$m_{1,\pi_2}(C_g)$ as an example:

$$\begin{aligned} m_{1,\pi_2}(C_g) &= E\left[\frac{1}{M_g^1}\sum_{i\in\mathcal{M}_g} Y_{i,g}(1,\pi_2)Z_{i,g}(\pi_2) \mid C_g\right] - E\left[\frac{1}{M_g^1}\sum_{i\in\mathcal{M}_g} Y_{i,g}(1,\pi_2)Z_{i,g}(\pi_2)\right] \\ &= E\left[\frac{1}{M_g^1}\sum_{i\in\mathcal{M}_g} E\left[Y_{i,g}(1,\pi_2)Z_{i,g}(\pi_2) \mid C_g, \mathcal{M}_g, B_g\right] \mid C_g\right] - E\left[\frac{1}{N_g}\sum_{1\le i\le N_g} Y_{i,g}(1,\pi_2)\right] \\ &= E\left[\frac{1}{M_g}\sum_{i\in\mathcal{M}_g} Y_{i,g}(1,\pi_2) \mid C_g\right] - E\left[\frac{1}{N_g}\sum_{1\le i\le N_g} Y_{i,g}(1,\pi_2)\right] \ , \end{aligned}$$

where the last inequality holds by Assumption 3.3.2. Therefore, only the first term is likely to depend on stratification strategy. In addition,

$$\operatorname{Var}\left[\bar{Y}_g(1,\pi_2)\right] = E\left[\bar{Y}_g(1,\pi_2)^2\right] - E\left[\bar{Y}_g(1,\pi_2)\right]^2 \ ,$$

for which I only need to focus on the first term. Let $X_g = (X_{i,g} : 1 \le i \le N_g)$.

$$\begin{aligned} E\left[\bar{Y}_g(1,\pi_2)^2\right] &= E\left[\left(\frac{1}{M_g^1}\sum_{i\in\mathcal{M}_g} Y_{i,g}(1,\pi_2)Z_{i,g}(\pi_2)\right)^2\right] \\ &= E\left[\frac{1}{\left(M_g^1\right)^2} E\left[\left(\sum_{i\in\mathcal{M}_g} Y_{i,g}(1,\pi_2)Z_{i,g}(\pi_2)\right)^2 \mid X_g, \mathcal{M}_g\right]\right] \ . \end{aligned}$$

In fact, it is equivalent to consider

$$
\begin{aligned}
&E\left[\frac{1}{\left(M_g^1\right)^2}E\left[\left(\sum_{i\in\mathcal{M}_g}Y_{i,g}(1,\pi_2)Z_{i,g}(\pi_2)\right)^2-\left(\sum_{i\in\mathcal{M}_g}Y_{i,g}(1,\pi_2)\pi_2\right)^2\mid X_g,\mathcal{M}_g\right]\right]\\
&=E\left[\frac{1}{\left(M_g^1\right)^2}E\left[\sum_{i,j\in\mathcal{M}_g:B_{i,g}=B_{j,g}}Y_{i,g}(1,\pi_2)Y_{j,g}(1,\pi_2)\left(Z_{i,g}(\pi_2)Z_{j,g}(\pi_2)-\pi_2^2\right)\mid X_g,\mathcal{M}_g\right]\right]\\
&=E\left[\frac{1}{\left(M_g^1\right)^2}\sum_{b\in\mathcal{B}}\sum_{B_{i,g}=B_{j,g}=b}E\left[Y_{i,g}(1,\pi_2)Y_{j,g}(1,\pi_2)\mid X_g\right]E\left[Z_{i,g}(\pi_2)Z_{j,g}(\pi_2)-\pi_2^2\mid X_g\right]\right]\\
&=E\left[\frac{1}{\left(M_g^1\right)^2}\sum_{b\in\mathcal{B}}\sum_{i:B_{i,g}=b}E\left[Y_{i,g}^2(1,\pi_2)\mid X_g\right](\pi_2-\pi_2^2)\right]\\
&\qquad+E\Bigg[\frac{1}{\left(M_g^1\right)^2}\sum_{b\in\mathcal{B}}\sum_{i\neq j:B_{i,g}=B_{j,g}=b}\Big(E\left[Y_{i,g}(1,\pi_2)\mid X_g\right]E\left[Y_{j,g}(1,\pi_2)\mid X_g\right]\\
&\qquad\qquad+\operatorname{Cov}(Y_{i,g}(1,\pi_2),Y_{j,g}(1,\pi_2))\Big)\times E\left[Z_{i,g}(\pi_2)Z_{j,g}(\pi_2)-\pi_2^2\mid X_g\right]\Bigg]\ ,
\end{aligned}
$$

where the last inequality holds by Assumption 3.5.2. Note the last term with $\operatorname{Cov}(Y_{i,g}(1,\pi_2),Y_{j,g}(1,\pi_2))$ does not affect the optimization problem and can be dropped since it is invariant across units. By Lemma II.2 of Bai (2022b), we only need to consider matched-group design with group size k when $\pi_2=l/k$ with $l<k$ being positive integers.[1] Note that the first term does not depend on stratification, for which we can replace $E\left[Y_{i,g}^2(1,\pi_2)\mid X_g\right]$ with $E\left[Y_{i,g}(1,\pi_2)\mid X_g\right]^2$ without affecting the optimzation problem. Then, by Lemma C.2.2 and Assumption 3.5.1, we can write the objective above as

$$
\begin{aligned}
&E\left[\frac{1}{\left(M_g^1\right)^2}\sum_{b\in\mathcal{B}}\sum_{i:B_{i,g}=b}E\left[Y_{i,g}(1,\pi_2)\mid X_g\right]^2\left(\pi_2-\pi_2^2\right)\right]\\
&\quad+E\left[\frac{1}{\left(M_g^1\right)^2}\sum_{b\in\mathcal{B}}\sum_{i\neq j:B_{i,g}=B_{j,g}=b}E\left[Y_{i,g}(1,\pi_2)\mid X_g\right]E\left[Y_{j,g}(1,\pi_2)\mid X_g\right]\left(\frac{\pi_2^2-\pi_2}{k-1}\right)\right]\\
&=E\left[\frac{1}{\left(M_g^1\right)^2}\sum_{b\in\mathcal{B}}\sum_{i:B_{i,g}=b}\left(E\left[Y_{i,g}(1,\pi_2)\mid X_g\right]-\bar{\mu}^b(X_g)\right)^2\frac{k\left(\pi_2-\pi_2^2\right)}{k-1}\right]\ ,
\end{aligned}
$$

where

$$
\bar{\mu}^b(X_g)=\frac{1}{k}\sum_{i:B_{i,g}=b}E\left[Y_{i,g}(1,\pi_2)\mid X_g\right]\ .
$$

[1] Without loss of generality, I implicitly assume that N_g/k is an integer.

Therefore, the optimal matching strategy matches on $E\left[Y_{i,g}(1,\pi_2) \mid X_g\right]$.

Now, let's turn to $V_2(z)$ for $z \in \{0,1\}$. Follow the same argument to conclude that $E[\tilde{Y}_g(z,\pi_2) \mid S_g]$ is invariant to stratification strategy. Then, only the first term is likely to be affected by stratification.

$$\mathrm{Var}[\tilde{Y}_g(z,\pi_2)] = E\left[\frac{N_g^2}{E[N_g]^2}\left(\bar{Y}_g(z,\pi_2)^2 - 2\bar{Y}_g(z,\pi_2)\frac{E[\bar{Y}_g(z,\pi_2)N_g]}{E[N_g]} + \frac{E[\bar{Y}_g(z,\pi_2)N_g]^2}{E[N_g]^2}\right)\right],$$

for which we only need to focus on

$$E\left[N_g^2\bar{Y}_g(z,\pi_2)^2\right] = E\left[N_g^2 E\left[\bar{Y}_g(z,\pi_2)^2 \mid N_g\right]\right],$$

which is also minimized by a matched-group design that matches on $E\left[Y_{i,g}(z,\pi_2) \mid X_g\right]$. ■

C.2 Auxiliary Lemmas

Lemma C.2.1. *If cluster size is fixed for all $1 \le g \le G$, i.e. $N_g = N$, then, $V_1(z) = V_2(z)$ for $z \in \{0,1\}$.*

Proof. Note that when $N_g = N$,

$$\tilde{Y}_g(z,h) = \bar{Y}_g(z,h) - E[\bar{Y}_g(z,h)]\ .$$

Then,

$$\begin{aligned}
V_2(z) = &\frac{1}{\pi_1}\mathrm{Var}[Y_g(z,\pi_2)] + \frac{1}{1-\pi_1}\mathrm{Var}[Y_g(0,0)] \\
&- E\left[\left(\sqrt{\frac{1-\pi_1}{\pi_1}}m_{z,\pi_2}(S_g) + \sqrt{\frac{\pi_1}{1-\pi_1}}m_{0,0}(S_g)\right)^2\right] \\
&+ E\left[\tau(S_g)\left(\frac{1}{\pi_1}m_{z,\pi_2}(S_g) + \frac{1}{1-\pi_1}m_{0,0}(S_g)\right)^2\right]\ .
\end{aligned}$$

By law of iterated expectation, we have $E\left[m_{z,h}\left(C_g\right) \mid S_g\right] = m_{z,h}\left(S_g\right)$. Thus,

$$\begin{aligned}
V_1(z) &= \frac{1}{\pi_1}\operatorname{Var}\left[\tilde{Y}_g(z,\pi_2)\right] + \frac{1}{1-\pi_1}\operatorname{Var}\left[\tilde{Y}_g(0,0)\right] + E\left[\left(m_{z,\pi_2}\left(S_g\right) - m_{0,0}\left(S_g\right)\right)^2\right] \\
&\quad + E\left[\tau\left(S_g\right)\left(\frac{1}{\pi_1} m_{z,\pi_2}\left(S_g\right) + \frac{1}{1-\pi_1} m_{0,0}\left(S_g\right)\right)^2\right] \\
&= \frac{E\left[\bar{Y}_g^2(z,\pi_2)\right] - E[\bar{Y}_g(z,\pi_2)]^2}{\pi_1} + \frac{E\left[\bar{Y}_g^2(0,0)\right] - E[\bar{Y}_g(0,0)]^2}{1-\pi_1} - 2E\left[m_{z,\pi_2}\left(S_g\right) m_{0,0}\left(S_g\right)\right] \\
&\quad - \frac{1-\pi_1}{\pi_1}\left(E\left[E[Y_g(z,\pi_2) \mid S_g]^2\right] - E[Y_g(z,\pi_2)]^2\right) - \frac{\pi_1}{1-\pi_1}\left(E\left[E[Y_g(0,0) \mid S_g]^2\right] - E[Y_g(0,0)]^2\right) \\
&= V_2(z) \ .
\end{aligned}$$

■

Lemma C.2.2. *Given a sequence of binary random variables $A^{(n)} = (A_i : 1 \leq i \leq n)$ with the joint distribution*

$$P\left(A^{(n)} = a^{(n)}\right) = \frac{1}{\begin{pmatrix} n \\ n\pi \end{pmatrix}} \text{ for all } a^{(n)} = (a_i : 1 \leq i \leq n) \text{ such that } \sum_{1 \leq i \leq n} a_i = n\pi \ ,$$

where $n\pi \in \mathbb{N}$ is an integer, otherwise $P\left(A^{(n)} = a^{(n)}\right) = 0$. We have $E[A_iA_j] = \pi^2 - \frac{\pi(1-\pi)}{n-1}$ for all $i \neq j \in [1,n]$.

Proof. Note that

$$\operatorname{Var}\left[\sum_{1 \leq i \leq n} A_i\right] = 0 = \sum_{1 \leq i \leq n} \operatorname{Var}\left[A_i\right] + \sum_{i \neq j} \operatorname{Cov}(A_i, A_j) = n\pi(1-\pi) + n(n-1)\operatorname{Cov}(A_i, A_j) \ ,$$

for any $i \neq j \in [1,n]$, which implies

$$E[A_iA_j] = \operatorname{Cov}(A_i, A_j) + E[A_i]E[A_j] = \pi^2 - \frac{\pi(1-\pi)}{n-1} \ .$$

■

Lemma C.2.3. *Suppose Assumption 3.2.2 holds, then*

$$E[\bar{Y}_g^r(z,\pi_2)|C_g, N_g] \leq C \quad a.s. \ ,$$

for $r \in \{1,2\}, z \in \{0,1\}$ *for some constant* $C > 0$,

$$E\left[\bar{Y}_g^r(z,\pi_2)N_g^\ell\right] < \infty \ ,$$

for $r \in \{1,2\}, \ell \in \{0,1,2\}, z \in \{0,1\}$, *and*

$$E\left[E[\bar{Y}_g(z,\pi_2)N_g|S_g]^2\right] < \infty \ .,$$

for $z \in \{0,1\}$. *In addition, suppose Assumption 3.4.1 (b) holds, then*

$$E[\bar{Y}_g(z,\pi_2)^r N_g^\ell \mid S_g] \leq C \quad a.s. \ ,$$

for $z \in \{0,1\}$.

Proof. We show the first statement for $r = 2$ and $z = 1$, since the case $r = 1$ follows similarly. By the Cauchy-Schwarz inequality,

$$\bar{Y}_g(1,\pi_2)^2 = \left(\frac{1}{M_g}\sum_{i\in\mathcal{M}_g} Y_{i,g}(1,\pi_2)Z_{i,g}(\pi_2)\right)^2 \leq \frac{1}{M_g}\sum_{i\in\mathcal{M}_g} Y_{i,g}(1,\pi_2)^2 \ ,$$

and hence

$$E[\bar{Y}_g(1,\pi_2)^2|C_g,N_g] \leq E\left[\frac{1}{M_g}\sum_{i\in\mathcal{M}_g} Y_{i,g}(1,\pi_2)^2 \mid C_g,N_g\right] \leq \sum_{1\leq i\leq N_g} E\left[\frac{1\{i\in\mathcal{M}_g\}}{M_g} \mid C_g,N_g\right] C \leq C \ ,$$

where the first inequality follows from the above derivation, Assumption 3.2.2(e) and the law of iterated expectations, and final inequality follows from Assumption 3.2.2(d). I show the next statement for $r = \ell = 2$, since the other cases follow similarly. By the law of iterated expectations,

$$\begin{aligned} E\left[\bar{Y}_g^2(1,\pi_2)N_g^2\right] &= E\left[N_g^2 E[\bar{Y}_g^2(1,\pi_2)|C_g,N_g]\right] \\ &\lesssim E\left[N_g^2\right] < \infty \ , \end{aligned}$$

where the final line follows by Assumption 3.2.2 (c). Next,

$$\begin{aligned} E\left[E[\bar{Y}_g(1,\pi_2)N_g|S_g]^2\right] &= E\left[E[N_g E[\bar{Y}_g(1,\pi_2)|C_g,N_g]|S_g]^2\right] \\ &\lesssim E\left[E[N_g|C_g]^2\right] < \infty \ , \end{aligned}$$

where the final line follows from Jensen's inequality and Assumption 3.2.2(c). Finally,

$$E[\bar{Y}_g(z,\pi_2)^r N_g^\ell \mid S_g] = E[N_g^\ell E[\bar{Y}_g(z,\pi_2)^r \mid C_g.N_g] \mid S_g] \lesssim E[N_g^\ell \mid S_g] \leq C \ ,$$

where the last inequality follows by Assumption 3.4.1 (b). ■

Lemma C.2.4. *Suppose Q_G satisfies Assumptions 3.2.2 and 3.4.1 and the treatment assignment mechanism satisfies Assumptions 3.3.2 and 3.4.2-3.4.3. Define*

$$\begin{aligned} \mathbb{L}_G^{\mathrm{Y1}} &= \frac{1}{\sqrt{nl}} \sum_{1\leq g\leq nk} (\bar{Y}_g(1,\pi_2) - E[\bar{Y}_g(1,\pi_2)]) I\{H_g = \pi_2\} \\ \mathbb{L}_G^{\mathrm{YN1}} &= \frac{1}{\sqrt{nl}} \sum_{1\leq g\leq nk} (\bar{Y}_g(1,\pi_2)N_g - E[\bar{Y}_g(1,\pi_2)N_g]) I\{H_g = \pi_2\} \\ \mathbb{L}_G^{\mathrm{N1}} &= \frac{1}{\sqrt{nl}} \sum_{1\leq g\leq 2G} (N_g - E[N_g]) I\{H_g = \pi_2\} \\ \mathbb{L}_G^{\mathrm{Y0}} &= \frac{1}{\sqrt{n(k-l)}} \sum_{1\leq g\leq nk} (\bar{Y}_g(0,0) - E[\bar{Y}_g(0,0)N_g]) I\{H_g = 0\} \\ \mathbb{L}_G^{\mathrm{YN0}} &= \frac{1}{\sqrt{n(k-l)}} \sum_{1\leq g\leq 2G} (\bar{Y}_g(0,0)N_g - E[\bar{Y}_g(0,0)N_g]) I\{H_g = 0\} \\ \mathbb{L}_G^{\mathrm{N0}} &= \frac{1}{\sqrt{n(k-l)}} \sum_{1\leq g\leq 2G} (N_g - E[N_g]) I\{H_g = 0\} \ . \end{aligned}$$

Then, as $n \to \infty$,

$$\left(\mathbb{L}_G^{\mathrm{Y1}}, \mathbb{L}_G^{\mathrm{YN1}}, \mathbb{L}_G^{\mathrm{N1}}, \mathbb{L}_G^{\mathrm{Y0}}, \mathbb{L}_G^{\mathrm{YN0}}, \mathbb{L}_G^{\mathrm{N0}}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}) \ ,$$

where

$$\mathbf{V} = \mathbf{V}_1 + \mathbf{V}_2$$

for

$$\mathbf{V}_1 = \begin{pmatrix} \mathbf{V}_1^1 & 0 \\ 0 & \mathbf{V}_1^0 \end{pmatrix}$$

$$\mathbf{V}_1^1 = \begin{pmatrix} E[\text{Var}[\bar{Y}_g(1,\pi_2)|S_g]] & E[\text{Cov}[\bar{Y}_g(1,\pi_2), \bar{Y}_g(1,\pi_2)N_g|S_g]] & E[\text{Cov}[\bar{Y}_g(1,\pi_2), N_g|S_g]] \\ E[\text{Cov}[\bar{Y}_g(1,\pi_2), \bar{Y}_g(1,\pi_2)N_g|S_g]] & E[\text{Var}[\bar{Y}_g(1,\pi_2)N_g|S_g]] & E[\text{Cov}[\bar{Y}_g(1,\pi_2)N_g, N_g|S_g]] \\ E[\text{Cov}[\bar{Y}_g(1,\pi_2), N_g|S_g]] & E[\text{Cov}[\bar{Y}_g(1,\pi_2)N_g, N_g|S_g]] & E[\text{Var}[N_g|S_g]] \end{pmatrix}$$

$$\mathbf{V}_1^0 = \begin{pmatrix} E[\text{Var}[\bar{Y}_g(0,0)|S_g]] & E[\text{Cov}[\bar{Y}_g(0,0), \bar{Y}_g(0,0)N_g|S_g]] & E[\text{Cov}[\bar{Y}_g(0,0), N_g|S_g]] \\ E[\text{Cov}[\bar{Y}_g(0,0), \bar{Y}_g(0,0)N_g|S_g]] & E[\text{Var}[\bar{Y}_g(0,0)N_g|S_g]] & E[\text{Cov}[\bar{Y}_g(0,0)N_g, N_g|S_g]] \\ E[\text{Cov}[\bar{Y}_g(0,0), N_g|S_g]] & E[\text{Cov}[\bar{Y}_g(0,0)N_g, N_g|S_g]] & E[\text{Var}[N_g|S_g]] \end{pmatrix}$$

$$\mathbf{V}_2 = \text{Var}\left[\begin{pmatrix} \sqrt{\pi_1}E[\bar{Y}_g(1,\pi_2)|S_g] \\ \sqrt{\pi_1}E[\bar{Y}_g(1,\pi_2)N_g|S_g] \\ \sqrt{\pi_1}E[N_g|S_g] \\ \sqrt{1-\pi_1}E[\bar{Y}_g(0,0)|S_g] \\ \sqrt{1-\pi_1}E[\bar{Y}_g(0,0)N_g|S_g] \\ \sqrt{1-\pi_1}E[N_g|S_g] \end{pmatrix}\right] .$$

Proof. Note

$$\begin{aligned}(\mathbb{L}_G^{\text{Y1}}, \mathbb{L}_G^{\text{YN1}}, \mathbb{L}_G^{\text{N1}}, \mathbb{L}_G^{\text{Y0}}, \mathbb{L}_G^{\text{YN0}}, \mathbb{L}_G^{\text{N0}}) &= (\mathbb{L}_{1,G}^{\text{Y1}}, \mathbb{L}_{1,G}^{\text{YN1}}, \mathbb{L}_{1,G}^{\text{N1}}, \mathbb{L}_{1,G}^{\text{Y0}}, \mathbb{L}_{1,G}^{\text{YN0}}, \mathbb{L}_{1,G}^{\text{N0}}) \\ &\quad + (\mathbb{L}_{2,G}^{\text{Y1}}, \mathbb{L}_{2,G}^{\text{YN1}}, \mathbb{L}_{2,G}^{\text{N1}}, \mathbb{L}_{2,G}^{\text{Y0}}, \mathbb{L}_{2,G}^{\text{YN0}}, \mathbb{L}_{2,G}^{\text{N0}}) ,\end{aligned}$$

where

$$\begin{aligned}\mathbb{L}_{1,G}^{\text{YN1}} &= \frac{1}{\sqrt{nl}} \sum_{1 \le g \le 2G} (\bar{Y}_g(1,\pi_2)N_g I\{H_g = \pi_2\} - E[\bar{Y}_g(1,\pi_2)N_g I\{H_g = \pi_2\}|S^{(G)}, H^{(G)}]) \\ \mathbb{L}_{2,G}^{\text{YN1}} &= \frac{1}{\sqrt{nl}} \sum_{1 \le g \le 2G} (E[\bar{Y}_g(1,\pi_2)N_g I\{H_g = \pi_2\}|S^{(G)}, H^{(G)}] - E[\bar{Y}_g(1,\pi_2)N_g]I\{H_g = \pi_2\})\end{aligned}$$

and similarly for the rest. Next, note $(\mathbb{L}_{1,G}^{\text{Y1}}, \mathbb{L}_{1,G}^{\text{YN1}}, \mathbb{L}_{1,G}^{\text{N1}}, \mathbb{L}_{1,G}^{\text{Y0}}, \mathbb{L}_{1,G}^{\text{YN0}}, \mathbb{L}_{1,G}^{\text{N0}}), n \ge 1$ is a triangular array of normalized sums of random vectors. We will apply the Lindeberg central limit theorem for random vectors, i.e., Proposition 2.27 of van der Vaart (1998), to this triangular array. Conditional on $S^{(G)}, H^{(G)}$, $(\mathbb{L}_{1,G}^{\text{Y1}}, \mathbb{L}_{1,G}^{\text{YN1}}, \mathbb{L}_{1,G}^{\text{N1}}) \perp (\mathbb{L}_{1,G}^{\text{Y0}}, \mathbb{L}_{1,G}^{\text{YN0}}, \mathbb{L}_{1,G}^{\text{N0}})$. Moreover, it follows from $Q_G = Q^G$ (by Lemma 5.1 of Bugni et al. (2022b) and Assumption 3.2.2 (a)-(b)) and Assumption 3.3.2, 3.4.2 that

$$\begin{aligned}&\text{Var}\left[\left(\mathbb{L}_{1,G}^{\text{Y1}}, \mathbb{L}_{1,G}^{\text{YN1}}, \mathbb{L}_{1,G}^{\text{N1}}\right)' |S^{(G)}, H^{(G)}\right] \\ &= \begin{pmatrix} \frac{1}{nl}\sum_{g=1}^G \text{Var}[\bar{Y}_g(1,\pi_2)|S_g]\tilde{H}_g & \frac{1}{nl}\sum_{g=1}^G \text{Cov}[\bar{Y}_g(1,\pi_2), \bar{Y}_g(1,\pi_2)N_g|S_g]\tilde{H}_g & \frac{1}{nl}\sum_{g=1}^G \text{Cov}[\bar{Y}_g(1,\pi_2), N_g|S_g]\tilde{H}_g \\ \frac{1}{nl}\sum_{g=1}^G \text{Cov}[\bar{Y}_g(1,\pi_2), \bar{Y}_g(1,\pi_2)N_g|S_g]\tilde{H}_g & \frac{1}{nl}\sum_{g=1}^G \text{Var}[\bar{Y}_g(1,\pi_2)N_g|S_g]\tilde{H}_g & \frac{1}{nl}\sum_{g=1}^G \text{Cov}[\bar{Y}_g(1,\pi_2)N_g, N_g|S_g]\tilde{H}_g \\ \frac{1}{nl}\sum_{g=1}^G \text{Cov}[\bar{Y}_g(1,\pi_2), N_g|S_g]\tilde{H}_g & \frac{1}{nl}\sum_{g=1}^G \text{Cov}[\bar{Y}_g(1,\pi_2)N_g, N_g|S_g]\tilde{H}_g & \frac{1}{nl}\sum_{g=1}^G \text{Var}[N_g|S_g]\tilde{H}_g \end{pmatrix} ,\end{aligned}$$

where $\tilde{H}_g = I\{H_g = \pi_2\}$. For the upper left component, we have

$$\frac{1}{G_1}\sum_{1\le g\le G}\mathrm{Var}[\bar{Y}_g(1,\pi_2)|S_g]\tilde{H}_g = \frac{1}{G_1}\sum_{1\le g\le G}E[\bar{Y}_g^2(1,\pi_2)|S_g]\tilde{H}_g - \frac{1}{G_1}\sum_{1\le g\le G}E[\bar{Y}_g(1,\pi_2)|S_g]^2\tilde{H}_g\ , \tag{C.4}$$

where $G_1 = nl$. Note

$$\begin{aligned}\frac{1}{G_1}\sum_{1\le g\le G}E[\bar{Y}_g^2(1,\pi_2)|S_g]\tilde{H}_g &= \frac{1}{G}\sum_{1\le g\le G}E[\bar{Y}_g^2(1,\pi_2)|S_g]\\ &\quad+(1-\pi_2)\left(\frac{1}{G_1}\sum_{1\le g\le G:\tilde{H}_g=1}E[\bar{Y}_g^2(1,\pi_2)|S_g]-\frac{1}{G_0}\sum_{1\le g\le G:\tilde{H}_g=0}E[\bar{Y}_g^2(1,\pi_2)|S_g]\right)\ .\end{aligned}$$

It follows from the weak law of large numbers, and Lemma C.2.3, that

$$\frac{1}{G}\sum_{1\le g\le G}E[\bar{Y}_g^2(1,\pi_2)|S_g]\xrightarrow{P}E[\bar{Y}_g^2(1,\pi_2)]\ .$$

On the other hand, it follows from Assumption 3.4.1(b) and 3.4.3 that

$$\begin{aligned}&\left|\frac{1}{G_1}\sum_{1\le g\le G:\tilde{H}_g=1}E[\bar{Y}_g^2(1,\pi_2)|S_g]-\frac{1}{G_0}\sum_{1\le g\le G:\tilde{H}_g=0}E[\bar{Y}_g^2(1,\pi_2)|S_g]\right|\\ &=\frac{1}{G}\left|\frac{1}{\pi_2}\sum_{1\le g\le G:\tilde{H}_g=1}E[\bar{Y}_g^2(1,\pi_2)|S_g]-\frac{1}{1-\pi_2}\sum_{1\le g\le G:\tilde{H}_g=0}E[\bar{Y}_g^2(1,\pi_2)|S_g]\right|\\ &\le\frac{1}{G}\sum_{1\le j\le n}k\cdot\max_{i,k\in\lambda_j}|E[\bar{Y}_i^2(1,\pi_2)|S_i]-E[\bar{Y}_k^2(1,\pi_2)|S_k]|\\ &\lesssim\frac{1}{n}\sum_{1\le j\le n}\max_{i,k\in\lambda_j}|S_i-S_k|\ .\end{aligned}$$

Therefore,

$$\frac{1}{G_1}\sum_{1\le g\le G}E[\bar{Y}_g^2(1,\pi_2)|S_g]\tilde{H}_g\xrightarrow{P}E[\bar{Y}_g^2(1,\pi_2)]\ .$$

Meanwhile,

$$\begin{aligned}\frac{1}{G_1}\sum_{1\le g\le G}E[\bar{Y}_g(1,\pi_2)|S_g]^2\tilde{H}_g &= \frac{1}{G}\sum_{1\le g\le G}E[\bar{Y}_g(1,\pi_2)|S_g]^2\\ &\quad+(1-\pi_2)\left(\frac{1}{G_1}\sum_{1\le g\le G:\tilde{H}_g=1}E[\bar{Y}_g(1,\pi_2)|S_g]^2-\frac{1}{G_0}\sum_{1\le g\le G:\tilde{H}_g=0}E[\bar{Y}_g(1,\pi_2)|S_g]^2\right)\ .\end{aligned}$$

Jensen's inequality implies $E[E[\bar{Y}_g(1,\pi_2)|S_g]^2] \leq E[\bar{Y}_g^2(1,\pi_2)] < E[\bar{Y}_g^2(1,\pi_2)] < \infty$ by Assumption 3.2.2(d), so it follows from the weak law of large numbers as above that

$$\frac{1}{G}\sum_{1\leq g\leq G} E[\bar{Y}_g(1,\pi_2)|S_g]^2 \xrightarrow{P} E[E[\bar{Y}_g(1,\pi_2)|S_g]^2] \ .$$

Next, by Assumption 3.4.1 and 3.4.3, the Cauchy-Schwarz inequality, and the fact that $(a+b)^2 \leq 2a^2+2b^2$,

$$\begin{aligned}
&\left|\frac{1}{G_1}\sum_{1\leq g\leq G:\tilde{H}_g=1} E[\bar{Y}_g(1,\pi_2)|S_g]^2 - \frac{1}{G_0}\sum_{1\leq g\leq G:\tilde{H}_g=0} E[\bar{Y}_g(1,\pi_2)|S_g]^2\right| \\
&= \frac{1}{G}\left|\frac{1}{\pi_2}\sum_{1\leq g\leq G:\tilde{H}_g=1} E[\bar{Y}_g(1,\pi_2)|S_g]^2 - \frac{1}{1-\pi_2}\sum_{1\leq g\leq G:\tilde{H}_g=0} E[\bar{Y}_g(1,\pi_2)|S_g]^2\right| \\
&\leq \frac{1}{G}\sum_{1\leq j\leq n}\left(\max_{i,j\in\lambda_j}|E[\bar{Y}_i(1,\pi_2)|S_i] - E[\bar{Y}_k(1,\pi_2)|S_k]|\right)\left(\sum_{k\in\lambda_j} E[\bar{Y}_k(1,\pi_2)|S_k]\right) \\
&\lesssim \left(\frac{1}{G}\sum_{1\leq j\leq n}\max_{i,j\in\lambda_j}|E[\bar{Y}_i(1,\pi_2)|S_i] - E[\bar{Y}_k(1,\pi_2)|S_k]|^2\right)^{1/2}\left(\frac{1}{G}\sum_{1\leq j\leq n}\left(\sum_{k\in\lambda_j} E[\bar{Y}_k(1,\pi_2)|S_k]\right)^2\right)^{1/2} \\
&\lesssim \left(\frac{1}{G}\sum_{1\leq j\leq n}\max_{i,j\in\lambda_j}|E[\bar{Y}_i(1,\pi_2)|S_i] - E[\bar{Y}_k(1,\pi_2)|S_k]|^2\right)^{1/2}\left(\frac{1}{G}\sum_{1\leq j\leq n}\sum_{k\in\lambda_j} E[\bar{Y}_k(1,\pi_2)|S_k]^2\right)^{1/2} \ .
\end{aligned}$$

Therefore, it follows from (C.4) that

$$\frac{1}{G_1}\sum_{1\leq g\leq G}\mathrm{Var}[\bar{Y}_g(1,\pi_2)|S_g]\tilde{H}_g \xrightarrow{P} E[\mathrm{Var}[\bar{Y}_g(1,\pi_2)|S_g]] \ .$$

Similar arguments together with Assumption 3.4.1(a)-(b) and Lemma C.2.3 imply that

$$\mathrm{Var}\left[\left.\begin{pmatrix}\mathbb{L}^{\mathrm{Y1}}_{1,G}\\ \mathbb{L}^{\mathrm{YN1}}_{1,G}\\ \mathbb{L}^{\mathrm{N1}}_{1,G}\end{pmatrix}\right| S^{(G)}, H^{(G)}\right] \xrightarrow{P} \mathbf{V}_1^1 \ .$$

Similarly,

$$\mathrm{Var}\left[\left.\begin{pmatrix}\mathbb{L}^{\mathrm{Y0}}_{1,G}\\ \mathbb{L}^{\mathrm{YN0}}_{1,G}\\ \mathbb{L}^{\mathrm{N0}}_{1,G}\end{pmatrix}\right| S^{(G)}, H^{(G)}\right] \xrightarrow{P} \mathbf{V}_1^0 \ .$$

If $E[\mathrm{Var}[\bar{Y}_g(1,\pi_2)N_g|S_g]] = E[\mathrm{Var}[N_g|S_g]] = E[\mathrm{Var}[\bar{Y}_g(0,0)N_g|S_g]] = 0$, then it follows from Markov's inequality conditional on $S^{(G)}$ and $H^{(G)}$, and the fact that probabilities are bounded and hence uniformly integrable, that $(\mathbb{L}_{1,G}^{\mathrm{Y1}}, \mathbb{L}_{1,G}^{\mathrm{YN1}}, \mathbb{L}_{1,G}^{\mathrm{N1}}, \mathbb{L}_{1,G}^{\mathrm{Y0}}, \mathbb{L}_{1,G}^{\mathrm{YN0}}, \mathbb{L}_{1,G}^{\mathrm{N0}}) \xrightarrow{P} 0$. Otherwise, it follows from similar arguments to those in the proof of Lemma S.1.5 of Bai et al. (2021b) that

$$\rho(\mathcal{L}((\mathbb{L}_{1,G}^{\mathrm{Y1}}, \mathbb{L}_{1,G}^{\mathrm{YN1}}, \mathbb{L}_{1,G}^{\mathrm{N1}}, \mathbb{L}_{1,G}^{\mathrm{Y0}}, \mathbb{L}_{1,G}^{\mathrm{YN0}}, \mathbb{L}_{1,G}^{\mathrm{N0}})' | S^{(G)}, H^{(G)}), N(0, \mathbf{V}_1)) \xrightarrow{P} 0 \ , \tag{C.5}$$

where $\mathcal{L}$ denotes the distribution and ρ is any metric that metrizes weak convergence.

Next, I study $(\mathbb{L}_{2,G}^{\mathrm{Y1}}, \mathbb{L}_{2,G}^{\mathrm{YN1}}, \mathbb{L}_{2,G}^{\mathrm{N1}}, \mathbb{L}_{2,G}^{\mathrm{Y0}}, \mathbb{L}_{2,G}^{\mathrm{YN0}}, \mathbb{L}_{2,G}^{\mathrm{N0}})$. It follows from $Q_G = Q^G$ (by Lemma 5.1 of Bugni et al. (2022b) and Assumption 3.2.2 (a)-(b)) and Assumption 3.4.2 that

$$\begin{pmatrix} \mathbb{L}_{2,G}^{\mathrm{Y1}} \\ \mathbb{L}_{2,G}^{\mathrm{YN1}} \\ \mathbb{L}_{2,G}^{\mathrm{N1}} \\ \mathbb{L}_{2,G}^{\mathrm{Y0}} \\ \mathbb{L}_{2,G}^{\mathrm{YN0}} \\ \mathbb{L}_{2,G}^{\mathrm{N0}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{G_1}} \sum_{1 \leq g \leq G} \tilde{H}_g(E[\bar{Y}_g(1,\pi_2)|S_g] - E[\bar{Y}_g(1,\pi_2)]) \\ \frac{1}{\sqrt{G_1}} \sum_{1 \leq g \leq G} \tilde{H}_g(E[\bar{Y}_g(1,\pi_2)N_g|S_g] - E[\bar{Y}_g(1,\pi_2)N_g]) \\ \frac{1}{\sqrt{G_1}} \sum_{1 \leq g \leq G} \tilde{H}_g(E[N_g|S_g] - E[N_g]) \\ \frac{1}{\sqrt{G_0}} \sum_{1 \leq g \leq G}(1 - \tilde{H}_g)(E[\bar{Y}_g(0,0)|S_g] - E[\bar{Y}_g(0,0)]) \\ \frac{1}{\sqrt{G_0}} \sum_{1 \leq g \leq G}(1 - \tilde{H}_g)(E[\bar{Y}_g(0,0)N_g|S_g] - E[\bar{Y}_g(0,0)N_g]) \\ \frac{1}{\sqrt{G_0}} \sum_{1 \leq g \leq G}(1 - \tilde{H}_g)(E[N_g|S_g] - E[N_g]) \end{pmatrix} .$$

For $\mathbb{L}_{2,G}^{\mathrm{Y1}}$, note it follows from Assumption 3.4.2 and Lemma C.2.2 that

$$\begin{aligned}
&\mathrm{Var}[\mathbb{L}_{2,G}^{\mathrm{Y1}} | S^{(G)}] = \frac{1}{G_1} \sum_{1 \leq j \leq n} \mathrm{Var}\left[\sum_{i=1}^{k} \tilde{H}_i (E[\bar{Y}_g(1,\pi_2)|S_g] - E[\bar{Y}_g(1,\pi_2)]) \right] \\
&= \frac{1}{G_1} \sum_{1 \leq j \leq n} \pi_1(1-\pi_1) \Bigg((k-1) \sum_{i \in \lambda_j} (E[\bar{Y}_g(1,\pi_2)|S_i] - E[\bar{Y}_g(1,\pi_2)])^2 \\
&\qquad - \sum_{a \neq b} (E[\bar{Y}_g(1,\pi_2)|S_a] - E[\bar{Y}_g(1,\pi_2)])(E[\bar{Y}_g(1,\pi_2)|S_b] - E[\bar{Y}_g(1,\pi_2)]) \Bigg) \\
&\lesssim \frac{1}{n} \sum_{1 \leq j \leq n} \sum_{i \in \lambda_j} \sum_{j \neq i} (E[\bar{Y}_g(1,\pi_2)|S_i] - E[\bar{Y}_g(1,\pi_2)])(E[\bar{Y}_g(1,\pi_2)|S_i] - E[\bar{Y}_g(1,\pi_2)|S_j]) \\
&\lesssim \frac{1}{n} \sum_{1 \leq j \leq n} \sum_{i \in \lambda_j} (E[\bar{Y}_g(1,\pi_2)|S_i] - E[\bar{Y}_g(1,\pi_2)]) \left(\max_{i,k \in \lambda_j} \left| E[\bar{Y}_g(1,\pi_2)|S_i] - E[\bar{Y}_g(1,\pi_2)|S_k] \right| \right) \\
&\lesssim \left(\frac{1}{n} \sum_{1 \leq j \leq n} \max_{i,k \in \lambda_j} \left| E[\bar{Y}_g(1,\pi_2)|S_i] - E[\bar{Y}_g(1,\pi_2)|S_k] \right|^2 \right)^{1/2} \\
&\lesssim \frac{1}{n} \sum_{1 \leq j \leq G} \max_{i,k \in \lambda_j} |S_i - S_k|^2 \xrightarrow{P} 0 \ .
\end{aligned}$$

Therefore, it follows from Markov's inequality conditional on $S^{(G)}$ and $H^{(G)}$, and the fact that probabilities are bounded and hence uniformly integrable, that

$$\mathbb{L}_{2,G}^{\mathrm{Y1}} = E[\mathbb{L}_{2,G}^{\mathrm{Y1}} | S^{(G)}] + o_P(1) \ .$$

Similarly,

$$\begin{pmatrix} \mathbb{L}_{2,G}^{\mathrm{Y1}} \\ \mathbb{L}_{2,G}^{\mathrm{YN1}} \\ \mathbb{L}_{2,G}^{\mathrm{N1}} \\ \mathbb{L}_{2,G}^{\mathrm{Y0}} \\ \mathbb{L}_{2,G}^{\mathrm{YN0}} \\ \mathbb{L}_{2,G}^{\mathrm{N0}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{G}}\sqrt{\pi_1}\sum_{1\le g\le G}(E[\bar{Y}_g(1,\pi_2)|S_g] - E[\bar{Y}_g(1,\pi_2)]) \\ \frac{1}{\sqrt{G}}\sqrt{\pi_1}\sum_{1\le g\le G}(E[\bar{Y}_g(1,\pi_2)N_g|S_g] - E[\bar{Y}_g(1,\pi_2)N_g]) \\ \frac{1}{\sqrt{G}}\sqrt{\pi_1}\sum_{1\le g\le G}(E[N_g|S_g] - E[N_g]) \\ \frac{1}{\sqrt{G}}\sqrt{1-\pi_1}\sum_{1\le g\le G}(E[\bar{Y}_g(0,0)|S_g] - E[\bar{Y}_g(0,0)]) \\ \frac{1}{\sqrt{G}}\sqrt{1-\pi_1}\sum_{1\le g\le G}(E[\bar{Y}_g(0,0)N_g|S_g] - E[\bar{Y}_g(0,0)N_g]) \\ \frac{1}{\sqrt{G}}\sqrt{1-\pi_1}\sum_{1\le g\le G}(E[N_g|S_g] - E[N_g]) \end{pmatrix} + o_P(1) \ .$$

It then follows from Assumption 3.2.2(c)-(d) and 3.4.1(a) and the central limit theorem that

$$(\mathbb{L}_{2,G}^{\mathrm{Y1}}, \mathbb{L}_{2,G}^{\mathrm{YN1}}, \mathbb{L}_{2,G}^{\mathrm{N1}}, \mathbb{L}_{2,G}^{\mathrm{Y0}}, \mathbb{L}_{2,G}^{\mathrm{YN0}}, \mathbb{L}_{2,G}^{\mathrm{N0}})' \xrightarrow{d} N(0, \mathbf{V}_2) \ .$$

Because (C.5) holds and $(\mathbb{L}_{2,G}^{\mathrm{Y1}}, \mathbb{L}_{2,G}^{\mathrm{YN1}}, \mathbb{L}_{2,G}^{\mathrm{N1}}, \mathbb{L}_{2,G}^{\mathrm{Y0}}, \mathbb{L}_{2,G}^{\mathrm{YN0}}, \mathbb{L}_{2,G}^{\mathrm{N0}})$ is deterministic conditional on $S^{(G)}, H^{(G)}$, the conclusion of the theorem follows from Lemma S.1.3 in Bai et al. (2021b). ■

C.3 Lemmas for Proof of Theorem 3.4.2

Lemma C.3.1. *If Assumption 3.3.2-3.2.2, 3.4.1(a) and 3.4.2 hold, then*

1. *$E[\bar{Y}_g^r(z,h) \mid S_g = s]$ and $E[\tilde{Y}_g^r(z,h) \mid S_g = s]$ are Lipschitz in s for $(z,h) \in \{(1,\pi_2),(0,\pi_2),(0,0)\}$ and $r \in \{1,2\}$.*
2. *$E\left[\bar{Y}_g^2(z,h)\right] < \infty$ and $E\left[\tilde{Y}_g^2(z,h)\right] < \infty$ for $(z,h) \in \{(1,\pi_2),(0,\pi_2),(0,0)\}$.*
3. *$((\bar{Y}_g(1,\pi_2), \bar{Y}_g(0,\pi_2), \bar{Y}_g(0,0)) : 1 \le g \le G) \perp H^{(G)} \mid S^{(G)}$ and $((\tilde{Y}_g(1,\pi_2), \tilde{Y}_g(0,\pi_2), \tilde{Y}_g(0,0)) : 1 \le g \le G) \perp H^{(G)} \mid S^{(G)}$.*

Proof. First, (a) is an immediate consequence of Assumption 3.4.1(a). Also, (b) is an immediate consequence of Lemma C.2.3 with Assumption 3.2.2. Finally, (c) follows directly by inspection and Assumption 3.3.2 and 3.4.2. ■

Lemma C.3.2. *Suppose Q_G satisfies Assumptions 3.2.2 and 3.4.1 and the treatment assignment mechanism satisfies Assumptions 3.3.2, 3.4.2-3.4.3 and 3.4.4. Then, for $r = 1, 2$,*

$$\frac{1}{nk(h)} \sum_{1 \le g \le G} \left(\bar{Y}_g^z\right)^r I\{H_g = h\} \xrightarrow{P} E[\bar{Y}_g^r(z,h)] \ .$$

Proof. I only prove the conclusion for $r = 1$ and the proof for $r = 2$ follows similarly. Note that

$$\frac{1}{nk(h)} \sum_{1 \le g \le G} \bar{Y}_g^z I\{H_g = h\} = \frac{1}{nk(h)} \sum_{1 \le g \le G} (\bar{Y}_g(z,h) I\{H_g = h\} - E[\bar{Y}_g(z,h) I\{H_g = h\} | S^{(G)}, H^{(G)}])$$
$$+ \frac{1}{nk(h)} \sum_{1 \le g \le G} E[\bar{Y}_g(z,h) I\{H_g = h\} | S^{(G)}, H^{(G)}] \ .$$

By Lemma C.3.1 (c), Assumption 3.4.3 and similar arguments to those used in the proof of Lemma C.2.4,

$$\frac{1}{nk(h)} \sum_{1 \le g \le G} E[\bar{Y}_g(z,h) I\{H_g = h\} | S^{(G)}, H^{(G)}] = \frac{1}{nk(h)} \sum_{1 \le g \le G} I\{H_g = h\} E[\bar{Y}_g(z,h) | S_g]$$
$$\xrightarrow{P} E[E[\bar{Y}_g(z,h) | S_g]] = E[\bar{Y}_g(z,h)] \ .$$

By following the argument in Lemma S.1.5 of Bai et al. (2021b), we conclude that

$$\frac{1}{nk(h)} \sum_{1 \le g \le G} (\bar{Y}_g(z,h) I\{H_g = h\} - E[\bar{Y}_g(z,h) I\{H_g = h\} | S^{(G)}, H^{(G)}]) \xrightarrow{P} 0 \ .$$

Therefore, the results hold. ■

Lemma C.3.3. *Suppose Q_G satisfies Assumptions 3.2.2 and 3.4.1 and the treatment assignment mechanism satisfies Assumptions 3.3.2, 3.4.2-3.4.3 and 3.4.4. Then, as $n \to \infty$,*

$$\hat{\rho}_n^z(\pi_2, 0) \xrightarrow{P} E[E[\bar{Y}_g(z,\pi_2) \mid S_g] E[\bar{Y}_g(z,0) \mid S^{(G)}]] \ .$$

Proof. To begin with, by Assumption 3.4.2,

$$
\begin{aligned}
&E[\hat{\rho}_n^z(\pi_2, 0) \mid S^{(G)}] \\
&= \frac{1}{n} \sum_{1 \le j \le n} \frac{1}{l(k-l)} E\left[\Big(\sum_{i \in \lambda_j} \bar{Y}_i^z I\{H_i = \pi_2\}\Big)\Big(\sum_{i \in \lambda_j} \bar{Y}_i^z I\{H_i = 0\}\Big) \mid S^{(G)}\right] \\
&= \frac{1}{n} \sum_{1 \le j \le n} \frac{1}{l(k-l)} \sum_{i \neq m \in \lambda_j} E\left[\bar{Y}_i(z, \pi_2) \mid S_i\right] E\left[\bar{Y}_m(z, 0) \mid S_m\right] E\left[I\{H_i = \pi_2\} I\{H_m = 0\} \mid S^{(G)}\right] \\
&= \frac{1}{n} \sum_{1 \le j \le n} \frac{1}{l(k-l)} \sum_{i < m \in \lambda_j} (E\left[\bar{Y}_i(z, \pi_2) \mid S_i\right] E\left[\bar{Y}_i(z, 0) \mid S_i\right] + E\left[\bar{Y}_m(z, \pi_2) \mid S_m\right] E\left[\bar{Y}_m(z, 0) \mid S_m\right] \\
&\quad - (E\left[\bar{Y}_i(z, \pi_2) \mid S_i\right] - E\left[\bar{Y}_m(z, \pi_2) \mid S_m\right])(E\left[\bar{Y}_i(z, 0) \mid S_i\right] - E\left[\bar{Y}_m(z, 0) \mid S_m\right])) \frac{l(k-l)}{k(k-1)} \\
&= \frac{1}{n} \sum_{1 \le j \le n} \frac{1}{k} \sum_{i \in \lambda_j} E\left[\bar{Y}_i(z, \pi_2) \mid S_i\right] E\left[\bar{Y}_i(z, 0) \mid S_i\right] \\
&\quad - \frac{1}{n} \sum_{1 \le j \le n} \frac{1}{k(k-1)} \sum_{i < m \in \lambda_j} (E\left[\bar{Y}_i(z, \pi_2) \mid S_i\right] - E\left[\bar{Y}_m(z, \pi_2) \mid S_m\right])(E\left[\bar{Y}_i(z, 0) \mid S_i\right] - E\left[\bar{Y}_m(z, 0) \mid S_m\right]) .
\end{aligned}
$$

Then, by Lipschitz condition from Lemma C.3.1(a), Lemma C.2.3 and Assumption 3.4.3, we conclude that $E[\hat{\rho}_n^z(\pi_2, 0) \mid S^{(G)}] \xrightarrow{P} E[E[\bar{Y}_g(z, \pi_2) \mid S_g] E[\bar{Y}_g(z, 0) \mid S_g]]$. To conclude the proof, we need to show

$$
\hat{\rho}_n^z(\pi_2, 0) - E[\hat{\rho}_n^z(\pi_2, 0) \mid S_g] \xrightarrow{P} 0 .
$$

Define

$$
\hat{\rho}_{n,j}^z(\pi_2, 0) = \frac{1}{l(k-l)} \Big(\sum_{i \in \lambda_j} \bar{Y}_i^z I\{H_i = \pi_2\}\Big)\Big(\sum_{i \in \lambda_j} \bar{Y}_i^z I\{H_i = 0\}\Big) .
$$

Note that

$$
\begin{aligned}
&\left|E[\hat{\rho}_{n,j}^z(\pi_2, 0) \mid S^{(G)}]\right| I\left\{\left|E[\hat{\rho}_{n,j}^z(\pi_2, 0) \mid S^{(G)}]\right| > \lambda\right\} \\
&= \left|\frac{1}{k(k-1)} \sum_{i \neq m \in \lambda_j} E\left[\bar{Y}_i(z, \pi_2) \mid S_i\right] E\left[\bar{Y}_m(z, 0) \mid S_m\right]\right| \\
&\qquad \times I\left\{\left|\frac{1}{k(k-1)} \sum_{i \neq m \in \lambda_j} E\left[\bar{Y}_i(z, \pi_2) \mid S_i\right] E\left[\bar{Y}_m(z, 0) \mid S_m\right]\right| > \lambda\right\} \\
&\le \sum_{i \neq m \in \lambda_j} \left|E\left[\bar{Y}_i(z, \pi_2) \mid S_i\right] E\left[\bar{Y}_m(z, 0) \mid S_m\right]\right| I\left\{\left|E\left[\bar{Y}_i(z, \pi_2) \mid S_i\right] E\left[\bar{Y}_m(z, 0) \mid S_m\right]\right| > \lambda\right\} .
\end{aligned}
$$

Then, the conclusion follows by repeating the same arguments from Lemma C.2 of Bai et al. (2022b). ■

Lemma C.3.4. *Suppose Q_G satisfies Assumptions 3.2.2 and 3.4.1 and the treatment assignment mechanism* *satisfies Assumptions 3.3.2, 3.4.2-3.4.3 and 3.4.4. Then, as* $n \to \infty$,

$$\hat{\rho}_n^z(h,h) \xrightarrow{P} E[E[\bar{Y}_g(z,h) \mid S_g]^2] \ .$$

Proof. To begin with, by Assumption 3.4.2,

$$\begin{aligned}
&E[\hat{\rho}_n^z(h,h) \mid S^{(G)}] \\
&= \frac{2}{n} \sum_{1 \le j \le \lfloor n/2 \rfloor} \frac{1}{k^2(h)} E\left[\Big(\sum_{i \in \lambda_{2j-1}} \bar{Y}_i^z I\{H_i = h\} \Big) \Big(\sum_{i \in \lambda_{2j}} \bar{Y}_i^z I\{H_i = h\} \Big) \mid S^{(G)} \right] \\
&= \frac{2}{n} \sum_{1 \le j \le \lfloor n/2 \rfloor} \frac{1}{k^2(h)} \frac{k^2(h)}{k^2} \sum_{i \in \lambda_{2j-1}, k \in \lambda_{2j}} E[\bar{Y}_i^z(z,h) \mid S_i] E[\bar{Y}_k^z(z,h) \mid S_k] \\
&= \frac{1}{nk^2} \sum_{1 \le j \le \lfloor n/2 \rfloor} \sum_{i \in \lambda_{2j-1}, k \in \lambda_{2j}} \big(E[\bar{Y}_i^z(z,h) \mid S_i]^2 + E[\bar{Y}_k^z(z,h) \mid S_k]^2 - (E[\bar{Y}_i^z(z,h) \mid S_i] - E[\bar{Y}_k^z(z,h) \mid S_k])^2 \big) \\
&= \frac{1}{G} \sum_{1 \le g \le G} E[\bar{Y}_g^z(z,h) \mid S_g]^2 - \frac{1}{nk^2} \sum_{1 \le j \le \lfloor n/2 \rfloor} \sum_{i \in \lambda_{2j-1}, k \in \lambda_{2j}} (E[\bar{Y}_i^z(z,h) \mid S_i] - E[\bar{Y}_k^z(z,h) \mid S_k])^2 \\
&\xrightarrow{P} E[E[\bar{Y}_g(z,h) \mid S_g]^2] \ ,
\end{aligned}$$

where the convergence in probability follows from Lemma C.3.1(a), Assumption 3.4.3, Lemma C.2.3 and weak law of large numbers. To conclude the proof, we need to show

$$\hat{\rho}_n^z(h,h) - E[\hat{\rho}_n^z(h,h) \mid S^{(G)}] \xrightarrow{P} 0 \ .$$

Define

$$\hat{\rho}_{n,j}^z(h,h) = \frac{1}{k^2(h)} \Big(\sum_{i \in \lambda_{2j-1}} \bar{Y}_i^z I\{H_i = h\} \Big) \Big(\sum_{i \in \lambda_{2j}} \bar{Y}_i^z I\{H_i = h\} \Big) \ .$$

Note that

$$\begin{aligned}
&\Big| E[\hat{\rho}_{n,j}^z(\pi_2, 0) \mid S^{(G)}] \Big| I\left\{ \Big| E[\hat{\rho}_{n,j}^z(\pi_2, 0) \mid S^{(G)}] \Big| > \lambda \right\} \\
&= \left| \frac{1}{k^2} \sum_{i \in \lambda_{2j-1}, k \in \lambda_{2j}} E[\bar{Y}_i^z(z,h) \mid S_i] E[\bar{Y}_k^z(z,h) \mid S_k] \right| \\
&\qquad \times I\left\{ \left| \frac{1}{k^2} \sum_{i \in \lambda_{2j-1}, k \in \lambda_{2j}} E[\bar{Y}_i^z(z,h) \mid S_i] E[\bar{Y}_k^z(z,h) \mid S_k] \right| > \lambda \right\} \\
&\le \sum_{i \in \lambda_{2j-1}, k \in \lambda_{2j}} \big| E[\bar{Y}_i^z(z,h) \mid S_i] E[\bar{Y}_k^z(z,h) \mid S_k] \big| I\left\{ \big| E[\bar{Y}_i^z(z,h) \mid S_i] E[\bar{Y}_k^z(z,h) \mid S_k] \big| > \lambda \right\} \ .
\end{aligned}$$

Then, the conclusion follows by repeating the same arguments from Lemma C.3 of Bai et al. (2022b). ■

Lemma C.3.5. *Suppose Q_G satisfies Assumptions 3.2.2 and 3.4.1 and the treatment assignment mechanism satisfies Assumptions 3.3.2, 3.4.2-3.4.3 and 3.4.4. Then, for $r = 1, 2$,*

$$\frac{1}{nk(h)} \sum_{1 \leq g \leq G} \left(\tilde{Y}_g^z\right)^r I\{H_g = h\} \xrightarrow{P} E[\tilde{Y}_g^r(z,h)] \ .$$

Proof. I only prove the conclusion for $r = 1$ and the proof for $r = 2$ follows similarly. Note that

$$\begin{aligned}\frac{1}{nk(h)} \sum_{1 \leq g \leq G} \tilde{Y}_g^z I\{H_g = h\} = \frac{1}{nk(h)} &\sum_{1 \leq g \leq G} \tilde{Y}_g(z,h) I\{H_g = h\} \\ &+ \frac{1}{nk(h)} \sum_{1 \leq g \leq G} \left(\hat{Y}_g^z(h) - \tilde{Y}_g(z,h)\right) I\{H_g = h\} \ ,\end{aligned}$$

where $\hat{Y}_g^z(h)$ is defined in (C.1). Note that

$$\begin{aligned}&\frac{1}{nk(h)} \sum_{1 \leq g \leq G} \left(\hat{Y}_g^z(h) - \tilde{Y}_g(z,h)\right) I\{H_g = h\} \\ &= \left(\frac{1}{\frac{1}{G}\sum_{1 \leq g \leq G} N_g} - \frac{1}{E[N_g]}\right) \left(\frac{1}{nk(h)} \sum_{1 \leq g \leq G} \bar{Y}_g(z,h) N_g I\{H_g = h\}\right) \\ &\quad - \left(\frac{\frac{1}{G}\sum_{1 \leq g \leq G} \bar{Y}_g(z,h) I\{H_g = h\} N_g}{\left(\frac{1}{G}\sum_{1 \leq g \leq G} N_g\right)^2} - \frac{E[\bar{Y}_g(z,h) N_g]}{E[N_g]^2}\right) \left(\frac{1}{nk(h)} \sum_{1 \leq g \leq G} N_g I\{H_g = h\}\right)\end{aligned}$$

By weak law of large number, Lemma C.2.4 and Slutsky's theorem, we have

$$\frac{1}{nk(h)} \sum_{1 \leq g \leq G} \left(\hat{Y}_g^z(h) - \tilde{Y}_g(z,h)\right) I\{H_g = h\} \xrightarrow{P} 0 \ .$$

By Lemma C.3.1 and repeating the arguments in Lemma C.3.2 with $\tilde{Y}_g(z,h)$ in the place of $\bar{Y}_g(z,h)$, we have

$$\frac{1}{nk(h)} \sum_{1 \leq g \leq G} \tilde{Y}_g(z,h) I\{H_g = h\} \xrightarrow{P} E[\tilde{Y}_g^r(z,h)] \ .$$

Thus, the result follows. ■

Lemma C.3.6. *Suppose Q_G satisfies Assumptions 3.2.2 and 3.4.1 and the treatment assignment mechanism* *satisfies Assumptions 3.3.2, 3.4.2-3.4.3 and 3.4.4. Then, as* $n \to \infty$*,*

$$\frac{1}{n}\sum_{1\le j\le n}\frac{1}{l(k-l)}\Big(\sum_{i\in\lambda_j}\tilde{Y}_i^z I\{H_i=\pi_2\}\Big)\Big(\sum_{i\in\lambda_j}\tilde{Y}_i^z I\{H_i=0\}\Big) \xrightarrow{P} E[E[\bar{Y}_g(z,\pi_2)\mid S_g]E[\bar{Y}_g(z,0)\mid S^{(G)}]] \ .$$

Proof. Note that

$$\begin{aligned}
&\frac{1}{n}\sum_{1\le j\le n}\frac{1}{l(k-l)}\Big(\sum_{i\in\lambda_j}\tilde{Y}_i^z I\{H_i=\pi_2\}\Big)\Big(\sum_{i\in\lambda_j}\tilde{Y}_i^z I\{H_i=0\}\Big)\\
&=\frac{1}{n}\sum_{1\le j\le n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}\hat{Y}_i^z(\pi_2)\hat{Y}_m^z(0)I\{H_i=\pi_2,H_m=0\}\\
&=\frac{1}{n}\sum_{1\le j\le n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}\tilde{Y}_i(z,\pi_2)\tilde{Y}_m(z,0)I\{H_i=\pi_2,H_m=0\}\\
&\quad+\frac{1}{n}\sum_{1\le j\le n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}\Big(\hat{Y}_i^z(\pi_2)\hat{Y}_m^z(0)-\tilde{Y}_i(z,\pi_2)\tilde{Y}_m(z,0)\Big)I\{H_i=\pi_2,H_m=0\} \ .
\end{aligned}$$

The second term can be written as

$$\begin{aligned}
&\frac{1}{n}\sum_{1\le j\le n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}\Big(\hat{Y}_i^z(\pi_2)-\tilde{Y}_i(z,\pi_2)\Big)\tilde{Y}_m(z,0)I\{H_i=\pi_2,H_m=0\}\\
&\qquad+\frac{1}{n}\sum_{1\le j\le n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}\Big(\hat{Y}_m^z(0)-\tilde{Y}_m(z,0)\Big)\tilde{Y}_i(z,\pi_2)I\{H_i=\pi_2,H_m=0\}\\
&\quad+\frac{1}{n}\sum_{1\le j\le n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}\Big(\hat{Y}_i^z(\pi_2)-\tilde{Y}_i(z,\pi_2)\Big)\Big(\hat{Y}_m^z(0)-\tilde{Y}_m(z,0))\Big)I\{H_i=\pi_2,H_m=0\} \ . \qquad \text{(C.6)}
\end{aligned}$$

We show that the first term of (C.6) converges to zero in probability and the other two terms should follow the same arguments:

$$\begin{aligned}
&\frac{1}{n}\sum_{1\le j\le n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}\Big(\hat{Y}_i^z(\pi_2)-\tilde{Y}_i(z,\pi_2)\Big)\tilde{Y}_m(z,0)I\{H_i=\pi_2,H_m=0\}\\
&=\left(\frac{1}{\frac{1}{G}\sum_{1\le g\le G}N_g}-\frac{1}{E[N_g]}\right)\left(\frac{1}{n}\sum_{1\le j\le n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}\bar{Y}_i(z,\pi_2)N_i\tilde{Y}_m(z,0)I\{H_i=\pi_2,H_m=0\}\right)\\
&\quad-\left(\frac{\frac{1}{G}\sum_{1\le g\le G}\bar{Y}_g(z,h)I\{H_g=h\}N_g}{\left(\frac{1}{G}\sum_{1\le g\le G}N_g\right)^2}-\frac{E[\bar{Y}_g(z,h)N_g]}{E[N_g]^2}\right)\\
&\qquad\times\left(\frac{1}{n}\sum_{1\le j\le n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}N_i\tilde{Y}_m(z,0)I\{H_i=\pi_2,H_m=0\}\right)
\end{aligned}$$

By following the same argument in Lemma S.1.6 from Bai et al. (2021b), we have

$$\frac{1}{n}\sum_{1\leq j\leq n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}\bar{Y}_g(z,\pi_2)N_g\tilde{Y}_m(z,0)I\{H_i=\pi_2,H_m=0\}\xrightarrow{P}E[E[N_g\bar{Y}_g(z,\pi_2)\mid S_g]E[\tilde{Y}_m(z,0)\mid S_g]]$$

$$\frac{1}{G}\sum_{1\leq j\leq G}N_{\pi(2j)}\tilde{Y}_{\pi(2j-1)}(0)I\{H_i=\pi_2,H_m=0\}\xrightarrow{P}E[E[N_g\mid S_g]E[\tilde{Y}_m(z,0)\mid S_g]]\ .$$

By weak law of large number, Lemma C.2.4 and Slutsky's theorem, we have

$$\frac{1}{n}\sum_{1\leq j\leq n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}\left(\hat{Y}_i^z(\pi_2)-\tilde{Y}_i(z,\pi_2)\right)\tilde{Y}_m(z,0)I\{H_i=\pi_2,H_m=0\}\xrightarrow{P}0\ .$$

Similarly, the convergence in probability to zero should hold for all three terms in (C.6). Thus, we have

$$\frac{1}{n}\sum_{1\leq j\leq n}\frac{1}{l(k-l)}\sum_{i\neq m\in\lambda_j}\left(\hat{Y}_i^z(\pi_2)\hat{Y}_m^z(0)-\tilde{Y}_i(z,\pi_2)\tilde{Y}_m(z,0)\right)I\{H_i=\pi_2,H_m=0\}\to 0\ .$$

By Lemma C.3.1 and repeating the arguments in Lemma C.3.3 with $\tilde{Y}_g(z,h)$ in the place of $\bar{Y}_g(z,h)$, we conclude the result. ■

Lemma C.3.7. *Suppose Q_G satisfies Assumptions 3.2.2 and 3.4.1 and the treatment assignment mechanism satisfies Assumptions 3.3.2, 3.4.2-3.4.3 and 3.4.4. Then, as $n\to\infty$,*

$$\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\Big(\sum_{i\in\lambda_{2j-1}}\tilde{Y}_i^z I\{H_i=h\}\Big)\Big(\sum_{i\in\lambda_{2j}}\tilde{Y}_i^z I\{H_i=h\}\Big)\xrightarrow{P}E[E[\tilde{Y}_g(z,h)\mid S_g]^2]\ .$$

Proof. Note that

$$\begin{aligned}
&\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\Big(\sum_{i\in\lambda_{2j-1}}\tilde{Y}_i^z I\{H_i=h\}\Big)\Big(\sum_{i\in\lambda_{2j}}\tilde{Y}_i^z I\{H_i=h\}\Big)\\
&=\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},m\in\lambda_{2j}}\hat{Y}_i^z(h)\hat{Y}_m^z(h)I\{H_i=H_m=h\}\\
&=\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},m\in\lambda_{2j}}\tilde{Y}_i(z,h)\tilde{Y}_m(z,h)I\{H_i=H_m=h\}\\
&\quad+\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},m\in\lambda_{2j}}\left(\hat{Y}_i^z(h)\hat{Y}_m^z(h)-\tilde{Y}_i(z,h)\tilde{Y}_m(z,h)\right)I\{H_i=H_m=h\}
\end{aligned}$$

The second term can be written as

$$
\begin{aligned}
&\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}\left(\hat{Y}_i^z(h)-\tilde{Y}_i(z,h)\right)\tilde{Y}_m(z,h)I\{H_i=H_m=h\}\\
&\qquad\qquad+\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}\left(\hat{Y}_m^z(h)-\tilde{Y}_m(z,h)\right)\tilde{Y}_i(z,h)I\{H_i=H_m=h\}\\
&\qquad+\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}\left(\hat{Y}_i^z(h)-\tilde{Y}_i(z,h)\right)\left(\hat{Y}_m^z(h)-\tilde{Y}_m(z,h))\right)I\{H_i=H_m=h\}\ . \quad \text{(C.7)}
\end{aligned}
$$

We show that the first term of (C.7) converges to zero in probability and the other two terms should follow the same arguments:

$$
\begin{aligned}
&\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}\left(\hat{Y}_i^z(h)-\tilde{Y}_i(z,h)\right)\tilde{Y}_m(z,h)I\{H_i=H_m=h\}\\
&=\left(\frac{1}{\frac{1}{G}\sum_{1\leq g\leq G}N_g}-\frac{1}{E[N_g]}\right)\left(\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}\bar{Y}_i(z,h)N_i\tilde{Y}_m(z,h)I\{H_i=H_m=h\}\right)\\
&-\left(\frac{\frac{1}{G}\sum_{1\leq g\leq G}\bar{Y}_g(z,h)I\{H_g=h\}N_g}{\left(\frac{1}{G}\sum_{1\leq g\leq G}N_g\right)^2}-\frac{E[\bar{Y}_g(z,h)N_g]}{E[N_g]^2}\right)\\
&\qquad\times\left(\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}N_i\tilde{Y}_m(z,h)I\{H_i=H_m=h\}\right)
\end{aligned}
$$

By following the same argument in Lemma S.1.6 from Bai et al. (2021b), we have

$$
\begin{aligned}
&\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}\bar{Y}_i(z,h)N_i\tilde{Y}_m(z,h)I\{H_i=H_m=h\}\xrightarrow{P}E[E[Y_g(z,h)N_g\mid S_g]E[Y_g(z,h)\mid S_g]]\\
&\qquad\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}N_i\tilde{Y}_m(z,h)I\{H_i=H_m=h\}\xrightarrow{P}E[E[N_g\mid S_g]E[Y_g(z,h)\mid S_g]]\ .
\end{aligned}
$$

By weak law of large number, Lemma C.2.4 and Slutsky's theorem, we have

$$
\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}\left(\hat{Y}_i^z(h)-\tilde{Y}_i(z,h)\right)\tilde{Y}_m(z,h)I\{H_i=H_m=h\}\xrightarrow{P}0\ .
$$

Similarly, the convergence in probability to zero should hold for all three terms in (C.6). Thus, we have

$$
\frac{2}{n}\sum_{1\leq j\leq\lfloor n/2\rfloor}\frac{1}{k^2(h)}\sum_{i\in\lambda_{2j-1},m\in\lambda_{2j}}\left(\hat{Y}_i^z(h)\hat{Y}_m^z(h)-\tilde{Y}_i(z,h)\tilde{Y}_m(z,h)\right)I\{H_i=H_m=h\}\to 0\ .
$$

By Lemma C.3.1 and repeating the arguments in Lemma C.3.4 with $\tilde{Y}_g(z,h)$ in the place of $\bar{Y}_g(z,h)$, we conclude the result. ■

C.4 Details for Weighted OLS

In this section, let's consider estimator of the coefficient of $Z_{i,g}$ and $L_{i,g}$ in a weighted least squares regression of $Y_{i,g}$ on a constant and $Z_{i,g}$ and $L_{i,g}$ with weights equal to $\sqrt{N_g/M_g}$. The results for weights equal to $\sqrt{1/M_g}$ are similar and omitted here. First, I provide some notatiosn as follows:

$$\begin{aligned}
T_{i,g} &:= \left(\sqrt{\frac{N_g}{|\mathcal{M}_g|}} \quad \sqrt{\frac{N_g}{|\mathcal{M}_g|}}Z_{i,g} \quad \sqrt{\frac{N_g}{|\mathcal{M}_g|}}L_{i,g}\right)' \\
T_g &:= (T_{i,g} : i \in \mathcal{M}_g)' \\
\hat{\epsilon}_g &:= \left(Y_{i,g} - \hat{\alpha} - \hat{\beta}_1 Z_{i,g} - \hat{\beta}_2 L_{i,g} : i \in \mathcal{M}_g\right)' \ ,
\end{aligned}$$

where $\hat{\alpha}, \hat{\beta}_1$ and $\hat{\beta}_2$ are the corresponding estimated coefficients. By doing some algebra, it follows that

$$\sum_{1\le g\le G}\sum_{i\in\mathcal{M}_g} T_{i,g}T_{i,g}' = \begin{pmatrix} \sum_{1\le g\le G} N_g & \sum_{1\le g\le G} N_g\pi_2 I\{H_g=\pi_2\} & \sum_{1\le g\le G} N_g(1-\pi_2)I\{H_g=\pi_2\} \\ \sum_{1\le g\le G} N_g\pi_2 I\{H_g=\pi_2\} & \sum_{1\le g\le G} N_g\pi_2 I\{H_g=\pi_2\} & 0 \\ \sum_{1\le g\le G} N_g(1-\pi_2)I\{H_g=\pi_2\} & 0 & \sum_{1\le g\le G} N_g(1-\pi_2)I\{H_g=\pi_2\} \end{pmatrix}$$

and

$$\begin{aligned}
&\sum_{1\le g\le G}\sum_{i\in\mathcal{M}_g} T_{i,g}\sqrt{\frac{N_g}{|\mathcal{M}_g|}}Y_{i,g} \\
&= \left(\sum_{1\le g\le G}\frac{N_g}{M_g}\sum_{i\in\mathcal{M}_g}Y_{i,g} \quad \sum_{1\le g\le G}\frac{N_g}{M_g}\sum_{i\in\mathcal{M}_g}Y_{i,g}Z_{i,g} \quad \sum_{1\le g\le G}\frac{N_g}{M_g}\sum_{i\in\mathcal{M}_g}Y_{i,g}L_{i,g}\right)' \\
&= \left(\sum_{1\le g\le G}\frac{N_g}{M_g}\sum_{i\in\mathcal{M}_g}Y_{i,g} \quad \sum_{1\le g\le G} I\{H_g=\pi_2\}N_g\bar{Y}_g^1\pi_2 \quad \sum_{1\le g\le G} I\{H_g=\pi_2\}N_g\bar{Y}_g^0(1-\pi_2)\right)'
\end{aligned}$$

Note that

$$\left(\sum_{1\le g\le G}\sum_{i\in\mathcal{M}_g} T_{i,g}T_{i,g}'\right)^{-1} = \begin{pmatrix} \frac{1}{N_0} & -\frac{1}{N_0} & -\frac{1}{N_0} \\ -\frac{1}{N_0} & \frac{1}{N_0}+\frac{1}{N_1\pi_2} & \frac{1}{N_0} \\ -\frac{1}{N_0} & \frac{1}{N_0} & \frac{1}{N_0}+\frac{1}{N_1(1-\pi_2)} \end{pmatrix}$$

Then, it follows that

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\theta}_2^P \\ \hat{\theta}_2^S \end{pmatrix} = \left(\sum_{1 \le g \le G} \sum_{i \in \mathcal{M}_g} T_{i,g} T_{i,g}' \right)^{-1} \left(\sum_{1 \le g \le G} \sum_{i \in \mathcal{M}_g} T_{i,g} \sqrt{\frac{N_g}{|\mathcal{M}_g|}} Y_{i,g} \right)$$

$$= \left(\frac{1}{N_0} \sum_{1 \le g \le N_g} I\{H_g = 0\} N_g \bar{Y}_g^1 \quad \hat{\theta}_2^P \quad \hat{\theta}_2^S \right)' \ .$$

Therefore, we conclude that this weighted OLS regression results in the same estimators as $\hat{\theta}_2^P, \hat{\theta}_2^S$. Next, I consider t-tests based on cluster-robust variance estimator. Note taht the cluster-robust variance estimator can be written as

$$\hat{\mathbf{V}}_{\text{CR}} = G \left(\sum_{1 \le g \le G} T_g' T_g \right)^{-1} \left(\sum_{1 \le g \le G} T_g' \hat{\epsilon}_{i,g} \hat{\epsilon}_{i,g}' T_{i,g} \right) \left(\sum_{1 \le g \le G} T_g' T_g \right)^{-1} \ ,$$

where $\sum_{1 \le g \le G} T_g' T_g$ should be identical to $\sum_{1 \le g \le G} \sum_{i \in \mathcal{M}_g} T_{i,g} T_{i,g}'$. By doing some algebra, if follows that

$$\sum_{1 \le g \le G} T_g' \hat{\epsilon}_{i,g} \hat{\epsilon}_{i,g}' T_{i,g} = \sum_{1 \le g \le G} \left(\frac{N_g}{M_g} \right)^2 \begin{pmatrix} \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} \\ \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} Z_{i,g} \\ \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} L_{i,g} \end{pmatrix} \begin{pmatrix} \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} \\ \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} Z_{i,g} \\ \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} L_{i,g} \end{pmatrix}' \ .$$

And thus cluster-robust variance estimator can be written as $\sum_{1 \le g \le G} \tilde{\epsilon}_g \tilde{\epsilon}_g'$, where

$$\tilde{\epsilon}_g = \begin{pmatrix} \frac{1}{N_0} \frac{1}{M_g} \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} N_g I\{H_g = 0\} \\ \frac{1}{N_1} \frac{1}{M_g^1} \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} N_g Z_{i,g} - \frac{1}{N_0} \frac{1}{M_g} \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} N_g I\{H_g = 0\} \\ \frac{1}{N_1} \frac{1}{M_g^0} \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} N_g L_{i,g} - \frac{1}{N_0} \frac{1}{M_g} \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} N_g I\{H_g = 0\} \end{pmatrix} \ .$$

Take the second diagonal element (primary effect) as an example. Its cluster-robust variance estimator is given by

$$\hat{V}_{\text{CR}}(1) = G \sum_{1 \le g \le G} \left(\frac{1}{N_1} \frac{1}{M_g^1} \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} N_g Z_{i,g} - \frac{1}{N_0} \frac{1}{M_g} \sum_{i \in \mathcal{M}_g} \hat{\epsilon}_{i,g} N_g I\{H_g = 0\} \right)^2$$

$$= \frac{1}{(N_1/G)^2} \frac{1}{G} \sum_{1 \le g \le G} N_g^2 \left(\bar{Y}_g(1, \pi_2) - \hat{\alpha} - \hat{\theta}_2^P \right)^2 I\{H_g = \pi_2\} + \frac{1}{(N_0/G)^2} \frac{1}{G} \sum_{1 \le g \le G} N_g^2 \left(\bar{Y}_g(0,0) - \hat{\alpha} \right)^2 I\{H_g = 0\} \ .$$

In both small and strata cases, by repeating arguments made in the Section C.1.3 and C.1.5, we have the following asymptotic results:

$$\frac{1}{G}\sum_{1\leq g\leq G} N_g^2\left(\bar{Y}_g(1,\pi_2)-\hat{\alpha}-\hat{\theta}_2^P\right)^2 I\{H_g=\pi_2\}\xrightarrow{p}\pi_1 E\left[\left(N_g\bar{Y}_g(1,\pi_2)-N_g\frac{E[N_g\bar{Y}_g(1,\pi_2)]}{E[N_g]}\right)^2\right]$$

$$\frac{1}{G}\sum_{1\leq g\leq G} N_g^2\left(\bar{Y}_g(0,0)-\hat{\alpha}\right)^2\xrightarrow{p}(1-\pi_1)E\left[\left(N_g\bar{Y}_g(0,0)-N_g\frac{E[N_g\bar{Y}_g(0,0)]}{E[N_g]}\right)^2\right],$$

which implies

$$\hat{V}_{\mathrm{CR}}(1)\xrightarrow{p}\frac{1}{\pi_1}\operatorname{Var}[\tilde{Y}_g(z,\pi_2)]+\frac{1}{1-\pi_1}\operatorname{Var}[\tilde{Y}_g(0,0)]\ .$$

REFERENCES

Abadie, A. and Imbens, G. W. (2008). Estimation of the Conditional Variance in Paired Experiments. *Annales d'Économie et de Statistique*, (91/92):175–187.

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., and Tobias, J. (2012). Targeting the poor: evidence from a field experiment in indonesia. *American Economic Review*, 102(4):1206–40.

Aramburu, J., Garone, L. F., Maffioli, A., Salazar, L., and Lopez, C. A. (2019). Direct and spillover effects of agricultural technology adoption programs: Experimental evidence from the dominican republic. IDB Working Paper Series IDB-WP-971, Washington, DC.

Athey, S. and Imbens, G. W. (2017a). The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier.

Athey, S. and Imbens, G. W. (2017b). The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier.

Bai, Y. (2022a). Optimality of Matched-Pair Designs in Randomized Controlled Trials. *American Economic Review*, 112(12):3911–3940.

Bai, Y. (2022b). Optimality of matched-pair designs in randomized controlled trials.

Bai, Y., Jiang, L., Romano, J. P., Shaikh, A. M., and Zhang, Y. (2023a). Covariate Adjustment in Experiments with Matched Pairs. arXiv:2302.04380 [econ].

Bai, Y., Jiang, L., Romano, J. P., Shaikh, A. M., and Zhang, Y. (2023b). Covariate adjustment in experiments with matched pairs.

Bai, Y., Liu, J., Shaikh, A. M., and Tabord-Meehan, M. (2022a). Inference in cluster randomized trials with matched pairs.

Bai, Y., Liu, J., and Tabord-Meehan, M. (2022b). Inference for matched tuples and fully blocked factorial designs.

Bai, Y., Liu, J., and Tabord-Meehan, M. (2023c). Inference for Matched Tuples and Fully Blocked Factorial Designs. arXiv:2206.04157 [econ, math, stat].

Bai, Y., Romano, J. P., and Shaikh, A. M. (2021a). Inference in Experiments with Matched Pairs*. *Journal of the American Statistical Association*, 0(ja):1–37. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2021.1883437.

Bai, Y., Romano, J. P., and Shaikh, A. M. (2021b). Inference in experiments with matched pairs. *Journal of the American Statistical Association*, 0(0):1–12.

Bai, Y., Romano, J. P., and Shaikh, A. M. (2022c). Inference in Experiments With Matched Pairs. *Journal of the American Statistical Association*, 117(540):1726–1737. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2021.1883437.

Banerjee, A., Chattopadhyay, R., Duflo, E., Keniston, D., and Singh, N. (2021). Improving police performance in rajasthan, india: Experimental evidence on incentives, managerial autonomy, and training. *American Economic Journal: Economic Policy*, 13(1):36–66.

Banerjee, A., Duflo, E., Glennerster, R., and Kinnan, C. (2015). The miracle of microfinance? evidence from a randomized evaluation. *American economic journal: Applied economics*, 7(1):22–53.

Basse, G. and Feller, A. (2018). Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113(521):41–55.

Basse, G. W., Feller, A., and Toulis, P. (2019). Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494.

Besedeš, T., Deck, C., Sarangi, S., and Shor, M. (2012). Age effects and heuristics in decision making. *Review of Economics and Statistics*, 94(2):580–595.

Beuermann, D. W., Cristia, J., Cueto, S., Malamud, O., and Cruz-Aguayo, Y. (2015). One laptop per child at home: Short-term impacts from a randomized experiment in peru. *American Economic Journal: Applied Economics*, 7(2):53–80.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., and Sandefur, J. (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics*, 168:1–20.

Branson, Z., Dasgupta, T., and Rubin, D. B. (2016). Improving covariate balance in 2K factorial designs via rerandomization with an application to a New York City Department of Education High School Study. *The Annals of Applied Statistics*, 10(4):1958 – 1976.

Brown, C. and Andrabi, T. (2020). Inducing positive sorting through performance pay: Experimental evidence from Pakistani schools.

Bruhn, M. and McKenzie, D. (2009a). In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, 1(4):200–232.

Bruhn, M. and McKenzie, D. (2009b). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4):200–232.

Bugni, F., Canay, I., Shaikh, A., and Tabord-Meehan, M. (2022a). Inference for Cluster Randomized Experiments with Non-ignorable Cluster Sizes. arXiv:2204.08356 [econ, stat].

Bugni, F., Canay, I., Shaikh, A., and Tabord-Meehan, M. (2022b). Inference for cluster randomized experiments with non-ignorable cluster sizes.

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2018a). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, 113(524):1784–1796. PMID: 30906087.

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2018b). Inference Under Covariate-Adaptive Randomization. *Journal of the American Statistical Association*, 113(524):1784–1796. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1375934.

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2019a). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, 10(4):1747–1785. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE1150.

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2019b). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, 10(4):1747–1785.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *Annals of Statistics*, 41(2):484–507. Publisher: Institute of Mathematical Statistics.

Cruces, G., Tortarolo, D., and Vazquez-Bare, G. (2022). Design of two-stage experiments with an application to spillovers in tax compliance.

Crépon, B., Devoto, F., Duflo, E., and Parienté, W. (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(1):123–150.

Cytrynbaum, M. (2021). Designing representative and balanced experiments by local randomization. *arXiv preprint arXiv:2111.08157*.

Cytrynbaum, M. (2023a). Covariate adjustment in stratified experiments.

Cytrynbaum, M. (2023b). Optimal stratification of survey experiments.

Dasgupta, T., Pillai, N. S., and Rubin, D. B. (2015). Causal inference from 2^k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):727–753.

de Chaisemartin, C. and Ramirez-Cuellar, J. (2019). At what level should one cluster standard errors in paired and small-strata experiments? *arXiv preprint arXiv:1906.00288*.

de Chaisemartin, C. and Ramirez-Cuellar, J. (2022a). At what level should one cluster standard errors in paired and small-strata experiments?

de Chaisemartin, C. and Ramirez-Cuellar, J. (2022b). At what level should one cluster standard errors in paired and small-strata experiments?

de Mel, S., McKenzie, D., and Woodruff, C. (2013). The demand for, and consequences of, formalization among informal firms in sri lanka. *American Economic Journal: Applied Economics*, 5(2):122–50.

DellaVigna, S., List, J. A., Malmendier, U., and Rao, G. (2016). Voting to tell others. *The Review of Economic Studies*, 84(1):143–181.

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

Duflo, E. and Saez, E. (2003). The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment*. *The Quarterly Journal of Economics*, 118(3):815–842.

Fafchamps, M., McKenzie, D., Quinn, S., and Woodruff, C. (2014). Microenterprise growth and the flypaper effect: Evidence from a randomized experiment in ghana. *Journal of Development Economics*, 106:211–226.

Fogarty, C. B. (2018). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105(4):994–1000.

Foos, F. and de Rooij, E. A. (2017). All in the family: Partisan disagreement and electoral mobilization in intimate networks—a spillover experiment. *American Journal of Political Science*, 61(2):289–304.

Forastiere, L., Airoldi, E. M., and Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534):901–918.

Glennerster, R. and Takavarasha, K. (2013). *Running Randomized Evaluations: A Practical Guide*. Princeton University Press.

Hansen, B. E. and Lee, S. (2019). Asymptotic theory for clustered samples. *Journal of econometrics*, 210(2):268–290.

Haushofer, J., Ringdal, C., Shapiro, J. P., and Wang, X. Y. (2019). Income changes and intimate partner violence: Evidence from unconditional cash transfers in kenya. Working Paper 25627, National Bureau of Economic Research.

Haushofer, J. and Shapiro, J. (2016). The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya*. *The Quarterly Journal of Economics*, 131(4):1973–2042.

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

Hidrobo, M., Peterman, A., and Heise, L. (2016). The effect of cash, vouchers, and food transfers on intimate partner violence: Evidence from a randomized experiment in northern ecuador. *American Economic Journal: Applied Economics*, 8(3):284–303.

Hu, Y., Li, S., and Wager, S. (2021). Average direct and indirect causal effects under interference.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842. PMID: 19081744.

Ichino, N. and Schündeln, M. (2012). Deterring or displacing electoral irregularities? spillover effects of observers in a randomized field experiment in ghana. *The Journal of Politics*, 74(1):292–307.

Imai, K., Jiang, Z., and Malani, A. (2021). Causal inference with interference and noncompliance in two-stage randomized experiments. *Journal of the American Statistical Association*, 116(534):632–644.

Imai, K., King, G., and Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science*, 24(1):29–53.

Janssen, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Statistics & Probability Letters*, 36(1):9–21.

Jiang, L., Linton, O. B., Tang, H., and Zhang, Y. (2022a). Improving estimation efficiency via regression-adjustment in covariate-adaptive randomizations with imperfect compliance.

Jiang, L., Liu, X., Phillips, P. C., and Zhang, Y. (2020). Bootstrap inference for quantile treatment effects in randomized experiments with matched pairs. *The Review of Economics and Statistics*, pages 1–47.

Jiang, L., Phillips, P. C. B., Tao, Y., and Zhang, Y. (2021). Regression-adjusted estimation of quantile treatment effects under covariate-adaptive randomizations.

Jiang, Z., Imai, K., and Malani, A. (2022b). Statistical inference and power analysis for direct and spillover effects in two-stage randomized experiments. *Biometrics*, 00(1-12):1–12.

Karlan, D., Osei, R., Osei-Akoto, I., and Udry, C. (2014). Agricultural decisions after relaxing credit and risk constraints. *The Quarterly Journal of Economics*, 129(2):597–652.

Kaur, S., Kremer, M., and Mullainathan, S. (2015). Self-control at work. *Journal of Political Economy*, 123(6):1227–1277.

Kinnan, C., Malani, A., Voena, A., Conti, G., and Imai, K. (2020). Adverse selection does not explain why utilization rises with premiums: evidence from a health insurance experiment in india.

Li, X., Ding, P., and Rubin, D. B. (2020). Rerandomization in 2^K factorial experiments. *The Annals of Statistics*, 48(1):43–63.

Liu, H., Ren, J., and Yang, Y. (2022). Randomization-based joint central limit theorem and efficient covariate adjustment in randomized block 2^k factorial experiments. *Journal of the American Statistical Association*, pages 1–15.

Liu, L. and Hudgens, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association*, 109(505):288–301. PMID: 24659836.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325.

Malani, A., Holtzman, P., Imai, K., Kinnan, C., Miller, M., Swaminathan, S., Voena, A., Woda, B., and Conti, G. (2021). Effect of health insurance in india: A randomized controlled trial. Working Paper 29576, National Bureau of Economic Research.

McKenzie, D. and Puerto, S. (2021). Growing markets through business training for female entrepreneurs: A market-level randomized experiment in kenya. *American Economic Journal: Applied Economics*, 13(2):297–332.

Middleton, J. A. and Aronow, P. M. (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy*, 6(1-2):39–75.

Muralidharan, K., Romero, M., and Wüthrich, K. (2019). Factorial designs, model selection, and (incorrect) inference in randomized experiments. Technical report, National Bureau of Economic Research.

Muralidharan, K. and Sundararaman, V. (2015). The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India *. *The Quarterly Journal of Economics*, 130(3):1011–1066.

Negi, A. and Wooldridge, J. M. (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, 40(5):504–534.

Pashley, N. E. and Bind, M.-A. C. (2019). Causal inference for multiple treatments using fractional factorial designs. *arXiv preprint arXiv:1905.07596.*

Rigdon, J. and Hudgens, M. G. (2015). Exact confidence intervals in the presence of interference. *Stat Probab Lett*, 105:130–135.

Rogers, T. and Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour*, 2(5):335–342.

Schochet, P. Z., Pashley, N. E., Miratrix, L. W., and Kautz, T. (2021). Design-based ratio estimators and central limit theorems for clustered, blocked rcts. *Journal of the American Statistical Association*, pages 1–12.

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

Toulis, P. and Kao, E. (2013). Estimation of causal peer influence effects. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1489–1497, Atlanta, Georgia, USA. PMLR.

van der Laan, M. J., Balzer, L. B., and Petersen, M. L. (2012). Adaptive matching in randomized trials and observational studies. *Journal of statistical research*, 46(2):113.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Vazquez-Bare, G. (2022). Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics*.

Wang, B., Park, C., Small, D. S., and Li, F. (2022). Model-robust and efficient inference for cluster-randomized experiments. *arXiv preprint arXiv:2210.07324*.

Wu, C. J. and Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons.