Bugni, Federico A.; Canay, Ivan A.; Shaikh, Azeem M.

Working Paper

Inference under covariate-adaptive randomization

cemmap working paper, No. CWP25/17

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Bugni, Federico A.; Canay, Ivan A.; Shaikh, Azeem M. (2017) : Inference under covariate-adaptive randomization, cemmap working paper, No. CWP25/17, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2017.2517

This Version is available at:
https://hdl.handle.net/10419/189730

cemmap
centre for microdata methods and practice

Inference under covariate-adaptive randomization

Federico A. Bugni
Ivan A. Canay
Azeem M. Shaikh

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP25/17

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Inference under Covariate-Adaptive Randomization*

Federico A. Bugni
Department of Economics
Duke University
federico.bugni@duke.edu

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

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

May 19, 2017

*We would like to thank the Co-Editor, the Associate Editor, and three anonymous referees for useful comments and suggestions. We additionally thank Lori Beaman, Robert Garlick, Raymond Guiteras, Aprajit Mahajan, Joseph Romano, Andres Santos, and seminar participants at various institutions for helpful comments on this paper. We finally thank Yuehao Bai and Winnie van Dijk for excellent research assistance. The research of the first author was supported by National Institutes of Health Grant 40-4153-00-0-85-399. The research of the second author was supported by National Science Foundation Grant SES-1530534. The research of the third author was supported by National Science Foundation Grants DMS-1308260, SES-1227091, and SES-1530661.

Abstract

This paper studies inference for the average treatment effect in randomized controlled trials with covariate-adaptive randomization. Here, by covariate-adaptive randomization, we mean randomization schemes that first stratify according to baseline covariates and then assign treatment status so as to achieve "balance" within each stratum. Our main requirement is that the randomization scheme assigns treatment status within each stratum so that the fraction of units being assigned to treatment within each stratum has a well behaved distribution centered around a proportion π as the sample size tends to infinity. Such schemes include, for example, Efron's biased-coin design and stratified block randomization. When testing the null hypothesis that the average treatment effect equals a pre-specified value in such settings, we first show the usual two-sample t-test is conservative in the sense that it has limiting rejection probability under the null hypothesis no greater than and typically strictly less than the nominal level. We show, however, that a simple adjustment to the usual standard error of the two-sample t-test leads to a test that is exact in the sense that its limiting rejection probability under the null hypothesis equals the nominal level. Next, we consider the usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes on treatment assignment and indicators for each of the strata. We show that this test is exact for the important special case of randomization schemes with $\pi = \frac{1}{2}$, but is otherwise conservative. We again provide a simple adjustment to the standard error that yields an exact test more generally. Finally, we study the behavior of a modified version of a permutation test, which we refer to as the covariate-adaptive permutation test, that only permutes treatment status for units within the same stratum. When applied to the usual two-sample t-statistic, we show that this test is exact for randomization schemes with $\pi = \frac{1}{2}$ and that additionally achieve what we refer to as "strong balance." For randomization schemes with $\pi \neq \frac{1}{2}$, this test may have limiting rejection probability under the null hypothesis strictly greater than the nominal level. When applied to a suitably adjusted version of the two-sample t-statistic, however, we show that this test is exact for all randomization schemes that achieve "strong balance," including those with $\pi \neq \frac{1}{2}$. A simulation study confirms the practical relevance of our theoretical results. We conclude with recommendations for empirical practice and an empirical illustration.

KEYWORDS: Covariate-adaptive randomization, stratified block randomization, Efron's biased-coin design, treatment assignment, randomized controlled trial, permutation test, two-sample t-test, strata fixed effects

JEL classification codes: C12, C14

1 Introduction

This paper studies inference for the average treatment effect in randomized controlled trials with covariate-adaptive randomization. Here, by covariate-adaptive randomization, we mean randomization schemes that first stratify according to baseline covariates and then assign treatment status so as to achieve "balance" within each stratum. Many such methods are used routinely when assigning treatment status in randomized controlled trials in all parts of the sciences. See, for example, Rosenberger and Lachin (2016) for a textbook treatment focused on clinical trials and Duflo et al. (2007) and Bruhn and McKenzie (2008) for reviews focused on development economics. In this paper, we take as given the use of such a treatment assignment mechanism and study its consequences for testing the null hypothesis that the average treatment effect equals a pre-specified value in such settings. Our main requirement is that the randomization scheme assigns treatment status within each stratum so that the fraction of units being assigned to treatment within each stratum has a well behaved distribution centered around a proportion π as the sample size tends to infinity. Importantly, as explained in Section 3 below, our results apply to most commonly used treatment assignment mechanisms, including simple random sampling, Efron's biased coin design, Wei's adaptive biased coin design, and stratified block randomization. The latter treatment assignment scheme is especially noteworthy because of its widespread use recently in development economics. See, for example, Dizon-Ross (2014, footnote 13), Duflo et al. (2014, footnote 6), Callen et al. (2015, page 24), and Berry et al. (2015, page 6). A caveat to our analysis, however, is that we require the proportion π to be constant across the strata. For an analysis of settings where π is allowed to vary across strata, see Imbens and Rubin (2015, Chapter 9) and Bugni et al. (2016).

Our first result establishes that the usual two-sample t-test is conservative in the sense that it has limiting rejection probability under the null hypothesis no greater than and typically strictly less than the nominal level. We additionally provide a characterization of when the limiting rejection probability under the null hypothesis is in fact strictly less than the nominal level. As explained further in Remark 4.3 below, our result substantially generalizes a related result obtained by Shao et al. (2010), who established this phenomenon under much stronger assumptions and for only one specific randomization scheme. In a simulation study, we find that the rejection probability of these tests may in fact be dramatically less than the nominal level, and, as a result, they may have very poor power when compared to other tests. Intuitively, the conservative feature of these tests is a consequence of the dependence in treatment status across units and between treatment status and baseline covariates resulting from covariate-adaptive randomization. We show, however, that a simple adjustment to the usual standard error of the two-sample t-test leads to a test that is exact in the sense that its limiting rejection probability under the null hypothesis equals the nominal level.

Next, we consider the usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes on treatment assignment and indicators for each of the strata. We refer to this test as the t-test with strata fixed effects. Based on simulation evidence and earlier assertions by Kernan et al. (1999), the use of this test has been recommended by Bruhn and McKenzie (2008), but, to the best of our knowledge, there has not yet been any formal analysis of its properties. Our results show that this test is typically conservative as well. As in the case of the two-sample t-test, we additionally provide a characterization of when the limiting rejection probability under the null hypothesis is in fact strictly less than the nominal level. This characterization reveals that the test is exact for the the important special case of randomization

schemes with $\pi = \frac{1}{2}$. We again provide a simple adjustment to the standard errors that yields an exact test more generally.

Finally, we study the behavior of a modified version of a permutation test, which we refer to as the covariate-adaptive permutation test, that only permutes treatment status for units within the same stratum. When applied to the usual two-sample t-statistic, we show that this test is exact for randomization schemes with $\pi = \frac{1}{2}$ and that additionally achieve what we refer to as "strong balance," defined formally in Section 2 below. For randomization schemes with $\pi \neq \frac{1}{2}$, this test may have limiting rejection probability under the null hypothesis strictly greater than the nominal level. When applied to a suitably adjusted version of the two-sample t-statistic, however, we show that this test is exact for all randomization schemes that achieve "strong balance," including those with $\pi \neq \frac{1}{2}$. As explained further in Remark 4.11 below, this test or closely related tests have been previously proposed and justified in finite samples for testing much more narrowly defined versions of the null hypothesis, including what is sometimes referred to as the "sharp null hypothesis." See, for example, Rosenbaum (2007), Heckman et al. (2011), Lee and Shaikh (2014), Rosenberger and Lachin (2016, Section 6.4), and, more recently, Young (2016). Exploiting recent results on the large-sample behavior of permutation tests by Chung and Romano (2013), our results, in contrast, asymptotically justify the use of the covariate-adaptive permutation test for testing the null hypothesis that the average treatment effect equals a pre-specified value for randomization schemes satisfying our assumptions below, while retaining in some cases the finite-sample validity for the narrower version of the null hypothesis.

The remainder of the paper is organized as follows. In Section 2, we describe our setup and notation. In particular, there we describe the assumptions we impose on the treatment assignment mechanism and what we mean by randomization schemes that achieve "strong balance." In Section 3, we discuss several examples of treatment assignment mechanisms satisfying these assumptions. Our main results concerning the two-sample t-test, the t-test with strata fixed effects, and the covariate-adaptive permutation test are contained in Section 4. In Section 5, we examine the finite-sample behavior of these tests as well as some other tests via a small simulation study. We discuss recommendations for empirical practice based on our theoretical results in Section 6. Finally, in Section 7, we provide an empirical illustration of our methods. Proofs of all results are provided in the Appendix.

2 Setup and Notation

Let Y_i denote the (observed) outcome of interest for the ith unit, A_i denote an indicator for whether the ith unit is treated or not, and Z_i denote observed, baseline covariates for the ith unit. Further denote by $Y_i(1)$ the potential outcome of the ith unit if treated and by $Y_i(0)$ the potential outcome of the ith unit if not treated. As usual, the (observed) outcome and potential outcomes are related to treatment assignment by the relationship

$$Y_i = Y_i(1)A_i + Y_i(0)(1 - A_i) \ . \tag{1}$$

Denote by P_n the distribution of the observed data

$$X^{(n)} = \{(Y_i, A_i, Z_i) : 1 \leq i \leq n\}$$

and denote by Q_n the distribution of

$$W^{(n)} = \{(Y_i(1), Y_i(0), Z_i) : 1 \leq i \leq n\} \ .$$

Note that P_n is jointly determined by (1), Q_n, and the treatment assignment mechanism. We therefore state our assumptions below in terms of assumptions on Q_n and assumptions on the treatment assignment mechanism. Indeed, we will not make reference to P_n in the sequel and all operations are understood to be under Q_n and the treatment assignment mechanism.

Strata are constructed from the observed, baseline covariates Z_i using a function $S : \text{supp}(Z_i) \to \mathcal{S}$, where $\mathcal{S}$ is a finite set. For $1 \leq i \leq n$, let $S_i = S(Z_i)$ and denote by $S^{(n)}$ the vector of strata $(S_1, \ldots, S_n)$.

We begin by describing our assumptions on Q_n. We assume that $W^{(n)}$ consists of n i.i.d. observations, i.e., $Q_n = Q^n$, where Q is the marginal distribution of $(Y_i(1), Y_i(0), Z_i)$. We further restrict Q to satisfy the following mild requirement:

Assumption 2.1. Q satisfies

$$E[Y_i^2(1)] < \infty \text{ and } E[Y_i^2(0)] < \infty$$

and

$$\min\Big\{\text{Var}[Y_i(0) - E[Y_i(0)|S_i]], \text{Var}[Y_i(1) - E[Y_i(1)|S_i]]\Big\} > 0 \ .$$

We note that the second requirement in Assumption 2.1 is made only to rule out degenerate situations and is stronger than required for our results.

Next, we describe our assumptions on the mechanism determining treatment assignment. In order to describe these assumptions more formally, we require some further notation. To this end, denote by $A^{(n)}$ the vector of treatment assignments $(A_1, \ldots, A_n)$ and, for $s \in \mathcal{S}$, let

$$D_n(s) = \sum_{1 \leq i \leq n} (A_i - \pi) I\{S_i = s\} \ , \tag{2}$$

where $\pi \in (0, 1)$ is the "target" proportion of units to assign to treatment in each stratum. Note that $D_n(s)$ as defined in (2) measures the amount of imbalance in stratum s relative to this "target" proportion. In order to rule out trivial strata, we henceforth assume that $p(s) = P\{S_i = s\} > 0$ for all $s \in \mathcal{S}$. Our other requirements on the treatment assignment mechanism are summarized in the following assumption:

Assumption 2.2. The treatment assignment mechanism is such that

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

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

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

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

Assumption 2.2.(a) simply requires that the treatment assignment mechanism is a function only of the vector of strata and an exogenous randomization device. Assumption 2.2.(b) formalizes our requirement that the randomization scheme assigns treatment status within each stratum so that the fraction of units being assigned to treatment within each stratum has a well behaved distribution centered around the "target" proportion π as the sample size tends to infinity. In the following section, we provide several important examples of treatment assignment mechanisms satisfying this assumption, including many that are used routinely in the sciences, including clinical trials and development economics. When Assumption 2.2.(b) holds with $\tau(s) = 0$ for all $s \in \mathcal{S}$, we say that the randomization scheme achieves "strong balance." This terminology is intended to reflect the fact that the measure of imbalance $D_n(s)$ is less dispersed around zero when Assumption 2.2.(b) holds with $\tau(s) = 0$ when compared with the case where it holds with $\tau(s) > 0$.

Our object of interest is the average effect of the treatment on the outcome of interest, defined to be

$$\theta(Q) = E[Y_i(1) - Y_i(0)] \ . \tag{3}$$

For a pre-specified choice of θ_0, the testing problem of interest is

$$H_0 : \theta(Q) = \theta_0 \text{ versus } H_1 : \theta(Q) \neq \theta_0 \tag{4}$$

at level $\alpha \in (0, 1)$.

Remark 2.1. The term "balance" is often used in a different way to describe whether the distributions of baseline covariates Z_i in the treatment and control groups are similar. For example, this might be measured according to the difference in the means of Z_i in the treatment and control groups. Our usage follows the usage in Efron (1971) or Hu and Hu (2012), where "balance" refers to the extent to which the of fraction of treated units within a strata differs from the target proportion π. ■

3 Examples

In this section, we briefly describe several different randomization schemes that satisfy our Assumption 2.2. A more detailed review of these methods and their properties can be found in Rosenberger and Lachin (2016). In our descriptions, we make use of the notation $A^{(k-1)} = (A_1, \ldots, A_{k-1})$ and $S^{(k)} = (S_1, \ldots, S_k)$ for $1 \leq k \leq n$, where $A^{(0)}$ is understood to be a constant.

Example 3.1. *(Simple Random Sampling)* Simple random sampling, also known as Bernoulli trials, refers to the case where $A^{(n)}$ consists of n i.i.d. random variables with

$$P\{A_k = 1 | S^{(k)}, A^{(k-1)}\} = P\{A_k = 1\} = \pi \tag{5}$$

for $1 \leq k \leq n$. In this case, Assumption 2.2.(a) follows immediately from (5), and Assumption 2.2.(b) follows from the central limit theorem with $\tau(s) = \pi(1 - \pi)$ for all $s \in \mathcal{S}$. Note that $E[D_n(s)] = 0$ for all $s \in \mathcal{S}$, so SRS ensures "balance" on average, yet in finite samples $D_n(s)$ may be far from zero. ■

Example 3.2. *(Biased-Coin Design)* A biased-coin design is a generalization of simple random sampling with $\pi = \frac{1}{2}$ originally proposed by Efron (1971) with the aim of improving "balance" in finite samples. In

this randomization scheme, treatment assignment is determined recursively for $1 \leq k \leq n$ as follows:

$$P\{A_k = 1 | S^{(k)}, A^{(k-1)}\} = \begin{cases} \frac{1}{2} & \text{if } D_{k-1}(S_k) = 0 \\ \lambda & \text{if } D_{k-1}(S_k) < 0 \\ 1-\lambda & \text{if } D_{k-1}(S_k) > 0 \end{cases}, \tag{6}$$

where $D_{k-1}(S_k) = \sum_{1 \leq i \leq k-1} \left(A_i - \frac{1}{2}\right) I\{S_i = S_k\}$, and $\frac{1}{2} < \lambda \leq 1$. Here, $D_0(S_1)$ is understood to be zero. The randomization scheme adjusts the probability with which the kth unit is assigned to treatment in an effort to improve "balance" in the corresponding stratum in finite samples. It follows from Lemma B.11 in the Appendix that this treatment assignment mechanism satisfies Assumption 2.2 with $\pi = \frac{1}{2}$. In particular, it is an example of a randomization scheme that achieves "strong balance" in that Assumption 2.2.(b) holds with $\tau(s) = 0$ for all $s \in \mathcal{S}$. In this sense, we see that biased-coin design provides improved "balance" relative to simple random sampling. ■

Example 3.3. *(Adaptive Biased-Coin Design)* An adaptive biased-coin design, also known as Wei's urn design, is an alternative generalization of simple random sampling with $\pi = \frac{1}{2}$ originally proposed by Wei (1978). This randomization scheme is similar to a biased-coin design, except that the probability λ in (6) depends on $D_{k-1}(S_k)$ defined in Example 3.2, the magnitude of imbalance among the first $k-1$ units in the corresponding stratum. More precisely, in this randomization scheme, treatment assignment is determined recursively for $1 \leq k \leq n$ as follows:

$$P\{A_k = 1 | S^{(k)}, A^{(k-1)}\} = \varphi\left(\frac{D_{k-1}(S_k)}{k-1}\right), \tag{7}$$

where $\varphi(x) : [-1, 1] \to [0, 1]$ is a pre-specified non-increasing function satisfying $\varphi(-x) = 1 - \varphi(x)$. Here, $\frac{D_0(S_1)}{0}$ is understood to be zero. It follows from Lemma B.12 in the Appendix that this treatment assignment mechanism satisfies Assumption 2.2 with $\pi = \frac{1}{2}$. In particular, Assumption 2.2.(b) holds with $\tau(s) = \frac{1}{4}(1 - 4\varphi'(0))^{-1}$, which lies in the interval $(0, \frac{1}{4})$ for the choice of $\varphi(x)$ recommended by Wei (1978) and used in Section 5. In this sense, adaptive biased-coin designs provide improved "balance" relative to simple random sampling with $\pi = \frac{1}{2}$ (i.e., $\tau(s) < \frac{1}{4}$), but to a lesser extent than biased-coin designs (i.e., $\tau(s) > 0$). ■

Example 3.4. *(Stratified Block Randomization)* An early discussion of stratified block randomization is provided by Zelen (1974). This randomization scheme is sometimes also referred to as block randomization or permuted blocks within strata. In order to describe this treatment assignment mechanism, for $s \in \mathcal{S}$, denote by $n(s)$ the number of units in stratum s and let $m(s) \leq n(s)$ be given. In this randomization scheme, $m(s)$ units in stratum s are assigned to treatment and the remainder are assigned to control, where all

$$\binom{n(s)}{m(s)}$$

possible assignments are equally likely and treatment assignment across strata are independent. By setting

$$m(s) = \lfloor \pi n(s) \rfloor, \tag{8}$$

this scheme ensures $|D_n(s)| \leq 1$ for all $s \in \mathcal{S}$ and therefore exhibits the best "balance" in finite samples among the methods discussed here. It follows from Lemma B.13 in the Appendix that this treatment assignment mechanism satisfies Assumption 2.2. In particular, as in Example 3.2, it is also an example of a randomization scheme that achieves "strong balance" in that Assumption 2.2.(b) holds with $\tau(s) = 0$ for all $s \in \mathcal{S}$. ■

Remark 3.1. Another treatment assignment mechanism for randomized controlled trials that has received considerable attention is re-randomization. See, for example, Bruhn and McKenzie (2008) and Lock Morgan and Rubin (2012). In this case, as explained by Lock Morgan and Rubin (2012), the properties of $D_n(s)$ depend on the rule used to decide whether to re-randomize and how to re-randomize. As a result, the analysis of such randomization schemes is necessarily case-by-case, and we do not consider them further in this paper. See instead **?** for an asymptotic analysis in this type of setting. ■

Remark 3.2. Another treatment assignment mechanism that has been used in clinical trials is minimization methods. These methods were originally proposed by Pocock and Simon (1975) and more recently extended and further studied by Hu and Hu (2012). In Hu and Hu (2012), treatment assignment is determined recursively for $1 \leq k \leq n$ as follows:

$$P\{A_k = 1 | S^{(k)}, A^{(k-1)}\} = \begin{cases} \frac{1}{2} & \text{if } \text{Imb}_k = 0 \\ \lambda & \text{if } \text{Imb}_k < 0 \\ 1 - \lambda & \text{if } \text{Imb}_k > 0 \end{cases}, \tag{9}$$

where $\frac{1}{2} \leq \lambda \leq 1$ and $\text{Imb}_k = \text{Imb}_k(S^{(k)}, A^{(k-1)})$ is a weighted average of different measures of imbalance. See Hu and Hu (2012) for expressions of these quantities. The analysis of this randomization scheme is relatively more involved than those in Examples 3.1–3.3 as it introduces dependence across different strata. We therefore do not consider it further in this paper. ■

Remark 3.3. Our framework does not accommodate response-adaptive randomization schemes. In such randomization schemes, units are assigned to treatment sequentially and treatment assignment for the ith unit, A_i, depends on $Y_1, \ldots, Y_{i-1}$. This feature leads to a violation of part (a) of our Assumption 2.2. It is worth emphasizing that response-adaptive randomization schemes are only feasible when at least some of the outcomes are observed at some point of the treatment assignment process, which is unusual in experiments in economics and other social sciences. ■

4 Main Results

4.1 Two-Sample t-Test

In this section, we consider using the two-sample t-test to test (4) at level $\alpha \in (0,1)$. In order to define this test, for $a \in \{0,1\}$, let

$$\begin{aligned} \bar{Y}_{n,a} &= \frac{1}{n_a} \sum_{1 \le i \le n} Y_i I\{A_i = a\} \\ \hat{\sigma}^2_{n,a} &= \frac{1}{n_a} \sum_{1 \le i \le n} (Y_i - \bar{Y}_{n,a})^2 I\{A_i = a\} \ , \end{aligned}$$

where $n_a = |\{1 \le i \le n : A_i = a\}|$. The two-sample t-test is given by

$$\phi_n^{t\text{-test}}(X^{(n)}) = I\{|T_n^{t\text{-stat}}(X^{(n)})| > z_{1-\frac{\alpha}{2}}\} \ , \tag{10}$$

where

$$T_n^{t\text{-stat}}(X^{(n)}) = \frac{\bar{Y}_{n,1} - \bar{Y}_{n,0} - \theta_0}{\sqrt{\frac{\hat{\sigma}^2_{n,1}}{n_1} + \frac{\hat{\sigma}^2_{n,0}}{n_0}}} \tag{11}$$

and $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of a standard normal random variable. This test may equivalently be described as the usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes on treatment assignment with heteroskedasticity-robust standard errors. It is used routinely throughout economics and the social sciences, including settings with covariate-adaptive randomization (see Duflo et al. (2007, Section 4), Bruhn and McKenzie (2008, Section E), and references therein). Note that further results on linear regression are developed in Section 4.2 below.

The following theorem describes the asymptotic behavior of the two-sample t-statistic defined in (11) and, as a consequence, the two-sample t-test defined in (10) under covariate-adaptive randomization. In particular, the theorem shows that the limiting rejection probability of the two-sample t-test under the null hypothesis is generally strictly less than the nominal level.

Theorem 4.1. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Then,*

$$\frac{\bar{Y}_{n,1} - \bar{Y}_{n,0} - \theta(Q)}{\sqrt{\frac{\hat{\sigma}^2_{n,1}}{n_1} + \frac{\hat{\sigma}^2_{n,0}}{n_0}}} \xrightarrow{d} N(0, \varsigma^2_{t\text{-}test}) \ ,$$

where $\varsigma^2_{t\text{-}test} \le 1$. Furthermore, $\varsigma^2_{t\text{-}test} < 1$ unless

$$(\pi(1-\pi) - \tau(s)) \left(\frac{1}{\pi} E[m_1(Z_i)|S_i = s] + \frac{1}{1-\pi} E[m_0(Z_i)|S_i = s] \right)^2 = 0 \text{ for all } s \in \mathcal{S} \ , \tag{12}$$

where

$$m_a(Z_i) = E[Y_i(a)|Z_i] - E[Y_i(a)] \tag{13}$$

for $a \in \{0,1\}$. Thus, for the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n^{t\text{-}test}(X^{(n)})$ defined in (10) satisfies

$$\lim_{n\to\infty} E[\phi_n^{t\text{-}test}(X^{(n)})] = P\{\varsigma_{t\text{-}test}|Z| > z_{1-\frac{\alpha}{2}}\} \leq \alpha \ , \tag{14}$$

where $Z \sim N(0,1)$, whenever Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$. Furthermore, the inequality in (14) is strict unless (12) holds.

Remark 4.1. Note that the two-sample t-test defined in (10) uses the $1-\frac{\alpha}{2}$ quantile of a standard normal random variable instead of the corresponding quantile of a t-distribution. Theorem 4.1 remains true with such a choice of critical value. See Imbens and Kolesar (2012) for a recent review of some such degrees of freedom adjustments. ■

Remark 4.2. While we generally expect that (12) will fail to hold, there are some important cases in which it does hold. First, as explained in Example 3.1, for simple random sampling Assumption 2.2 holds with $\tau(s) = \pi(1-\pi)$ for all $s \in \mathcal{S}$. Hence, (12) holds, and Theorem 4.1 implies, as one would expect, that the two-sample t-test is not conservative under simple random sampling. Second, if stratification is irrelevant for potential outcomes in the sense that $E[Y_i(a)|S_i] = E[Y_i(a)]$ for all $a \in \{0,1\}$, then $E[m_a(Z_i)|S_i] = 0$ for $a \in \{0,1\}$. Hence, (12) again holds, and Theorem 4.1 implies that the two-sample t-test is not conservative when stratification is irrelevant for potential outcomes. Note that a special case of irrelevant stratification is simply no stratification, i.e., S_i is constant. ■

Remark 4.3. Under substantially stronger assumptions than those in Theorem 4.1, Shao et al. (2010) also establish conservativeness of the two-sample t-test for a specific covariate-adaptive randomization scheme. For a textbook summary of the results in Shao et al. (2010), see Section 9.5 of Rosenberger and Lachin (2016). Shao et al. (2010) require, in particular, that $m_a(Z_i) = \gamma' Z_i$, that $\text{Var}[Y_i(a)|Z_i]$ does not depend on Z_i, and that the treatment assignment rule is a biased-coin design, as described in Example 3.2. Theorem 4.1 relaxes all of these requirements. ■

Remark 4.4. While Theorem 4.1 characterizes when the limiting rejection probability of the two-sample t-test under the null hypothesis is strictly less than the nominal level, it does not reveal how significant this difference might be. The magnitude of this difference will, of course, depend on the value of $\varsigma^2_{t\text{-test}}$, which will in turn depend on Q and the treatment assignment mechanism. In our simulation study in Section 5, we find that the rejection probability may in fact be dramatically less than the nominal level and that this difference translates into substantial power losses when compared with exact tests studied below. ■

We now provide an adjustment to the two-sample t-test that leads to a test that is exact in the sense that its limiting rejection probability under the null hypothesis equals the nominal level. In order to describe the test, we first observe that in the proof of Theorem 4.1 in the Appendix, it is shown that

$$\sqrt{n}(\bar{Y}_{n,1} - \bar{Y}_{n,0} - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\tilde{Y}}(\pi) + \varsigma^2_H + \varsigma^2_A(\pi)) \ , \tag{15}$$

where

$$\begin{aligned}
\varsigma^2_{\tilde{Y}}(\pi) &= \frac{1}{\pi}\mathrm{Var}[\tilde{Y}_i(1)] + \frac{1}{1-\pi}\mathrm{Var}[\tilde{Y}_i(0)] && (16)\\
\varsigma^2_H &= E[(E[m_1(Z_i)|S_i] - E[m_0(Z_i)|S_i])^2] && (17)\\
\varsigma^2_A(\pi) &= E\left[\tau(S_i)\left(\frac{1}{\pi}E[m_1(Z_i)|S_i] + \frac{1}{1-\pi}E[m_0(Z_i)|S_i]\right)^2\right] && (18)
\end{aligned}$$

with $\tilde{Y}_i(a) = Y_i(a) - E[Y_i(a)|S_i]$ for $a \in \{0,1\}$. Natural estimators of the quantities in (16)–(18) may be constructed by replacing population quantities with their sample counterparts. In order to define these estimators formally, it is useful to introduce some further notation. For $a \in \{0,1\}$, let

$$\hat{\mu}_{n,a}(s) = \frac{1}{n_a(s)}\sum_{1\le i\le n} Y_i I\{A_i = a, S_i = s\} \quad (19)$$

where $n_a(s) = |\{1 \le i \le n : A_i = a, S_i = s\}|$. In terms of this notation, we may define the following estimators:

$$\begin{aligned}
\hat{\varsigma}^2_{\tilde{Y}}(\pi) &= \frac{1}{\pi}\left(\frac{1}{n_1}\sum_{1\le i\le n} Y_i^2 A_i - \sum_{s\in\mathcal{S}}\frac{n(s)}{n}\hat{\mu}_{n,1}(s)^2\right) + \\
&\qquad \frac{1}{1-\pi}\left(\frac{1}{n_0}\sum_{1\le i\le n} Y_i^2(1-A_i) - \sum_{s\in\mathcal{S}}\frac{n(s)}{n}\hat{\mu}_{n,0}(s)^2\right) && (20)\\
\hat{\varsigma}^2_H &= \sum_{s\in\mathcal{S}}\frac{n(s)}{n}\left((\hat{\mu}_{n,1}(s) - \bar{Y}_{n,1}) - (\hat{\mu}_{n,0}(s) - \bar{Y}_{n,0})\right)^2 && (21)\\
\hat{\varsigma}^2_A(\pi) &= \sum_{s\in\mathcal{S}}\tau(s)\frac{n(s)}{n}\left(\frac{1}{\pi}(\hat{\mu}_{n,1}(s) - \bar{Y}_{n,1}) + \frac{1}{1-\pi}(\hat{\mu}_{n,0}(s) - \bar{Y}_{n,0})\right)^2 . && (22)
\end{aligned}$$

In (20)–(22), recall that $n(s)$, as in Example 3.4, denotes the number of units in stratum s. The "adjusted" two-sample t-test is given by

$$\phi_n^{t\text{-test,adj}}(X^{(n)}) = I\{|T_n^{t\text{-stat,adj}}(X^{(n)})| > z_{1-\frac{\alpha}{2}}\}\ , \quad (23)$$

where

$$T_n^{t\text{-stat,adj}}(X^{(n)}) = \frac{\bar{Y}_{n,1} - \bar{Y}_{n,0} - \theta_0}{\sqrt{\frac{\hat{\varsigma}^2_{\tilde{Y}}(\pi) + \hat{\varsigma}^2_H + \hat{\varsigma}^2_A(\pi)}{n}}}\ . \quad (24)$$

The following theorem establishes the desired result about the asymptotic behavior of the "adjusted" two-sample t-test.

Theorem 4.2. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. For the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n^{t\text{-test,adj}}(X^{(n)})$ defined in (23) satisfies*

$$\lim_{n\to\infty} E[\phi_n^{t\text{-test,adj}}(X^{(n)})] = \alpha \quad (25)$$

whenever Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$.

4.2 t-Test with Strata Fixed Effects

In this section, we consider using the usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes on treatment assignment and indicators for each of the strata. As mentioned previously, we refer to this test as the t-test with strata fixed effects. For concreteness, we use the usual heteroskedasticity-robust standard errors. Note that the two-sample t-test studied in Section 4.1 can be viewed as the usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes on treatment assignment only with heteroksedasticity-robust standard errors. It follows from Theorem 4.1 and Remark 4.2 that such a test is conservative in the sense that the limiting rejection probability under the null hypothesis may be strictly less than the nominal level. In this section, we first show that the addition of strata fixed effects results in a test is exact in the important special case of randomization schemes with $\pi = \frac{1}{2}$, but remains conservative otherwise.

In order to define the test, consider estimation of the equation

$$Y_i = \beta A_i + \sum_{s \in \mathcal{S}} \delta_s I\{S_i = s\} + u_i \tag{26}$$

by ordinary least squares. Denote by $\hat{\beta}_n$ the resulting estimator of β in (26). Let

$$T_n^{\text{sfe}}(X^{(n)}) = \frac{\hat{\beta}_n - \theta_0}{\sqrt{\frac{\hat{V}_{n,\beta}}{n}}} \ , \tag{27}$$

where $\hat{V}_{n,\beta}$ equals the usual heteroskedasticity-robust standard error for $\hat{\beta}_n$. See (A-60) in the Appendix for an exact expression. Using this notation, the test of interest is given by

$$\phi_n^{\text{sfe}}(X^{(n)}) = I\{|T_n^{\text{sfe}}(X^{(n)})| > z_{1-\frac{\alpha}{2}}\} \ . \tag{28}$$

The following theorem describes the asymptotic behavior of the proposed test. In particular, it shows that its limiting rejection probability under the null hypothesis equals the nominal level for randomization schemes with $\pi = \frac{1}{2}$ and is generally strictly less than the nominal level otherwise.

Theorem 4.3. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Then,*

$$\frac{\hat{\beta}_n - \theta_0}{\sqrt{\frac{\hat{V}_{n,\beta}}{n}}} \xrightarrow{d} N(0, \varsigma_{sfe}^2) \ , \tag{29}$$

where $\varsigma_{sfe}^2 \leq 1$. Furthermore, $\varsigma_{sfe}^2 < 1$ unless

$$(1-2\pi)\left(\pi(1-\pi) - \tau(s)\right)\left(E[m_1(Z_i)|S_i = s] - E[m_0(Z_i)|S_i = s]\right)^2 = 0 \text{ for all } s \in \mathcal{S} \ . \tag{30}$$

Thus, for the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n^{sfe}(X^{(n)})$ defined in (28) satisfies

$$\lim_{n\to\infty} E[\phi_n^{sfe}(X^{(n)})] = P\{\varsigma_{sfe}|Z| > z_{1-\frac{\alpha}{2}}\} \leq \alpha \tag{31}$$

where $Z \sim N(0,1)$, for Q additionally satisfying the null hypothesis, i.e., $\theta(Q) = \theta_0$. Furthermore, the inequality in (31) is strict unless (30) holds.

Remark 4.5. As in the case of the two-sample t-test, we generally expect that (30) will fail to hold, but there are again some important cases in which it does hold. As one would expect, it again holds in the case of simple random sampling and when stratification is irrelevant for potential outcomes in the sense that $E[Y_i(a)|S_i] = E[Y_i(a)]$ for all $a \in \{0,1\}$, but it additionally holds when $\pi = \frac{1}{2}$. ■

Remark 4.6. In Lemma B.10 in the Appendix, we show that replacing the heteroskedasticity-robust standard error with the homoskedasticity-only standard error also leads to an exact test when $\pi = \frac{1}{2}$. This result may seem surprising at first, but it may be viewed as a generalization of the following familiar fact in the usual two-sample t-test: even if the variances in the two samples are different, one may use either the pooled or unpooled estimate of the variance whenever the ratio of the two sample sizes tends to one. When $\pi \neq \frac{1}{2}$, however, using the homoskedasticity-only standard error leads to a test whose limiting rejection probability under the null hypothesis may strictly exceed the nominal level. ■

Remark 4.7. As in the literature on linear panel data models with fixed effects, $\hat{\beta}_n$ may be equivalently computed using ordinary least squares and the deviations of Y_i and A_i from their respective means within strata. However, it is important to note that the resulting standard errors are not equivalent to the usual heteroksedasticity-robust standard errors associated with ordinary least squares estimation of (26). ■

As in the case of the two-sample t-test studied previously, it is possible to provide an adjustment to the t-test with strata fixed effects that leads to a test that is exact. In order to describe the test, we first observe that in the proof of Theorem 4.3 in the Appendix, it is shown that

$$\sqrt{n}(\hat{\beta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\tilde{Y}}(\pi) + \varsigma^2_H + \varsigma^2_\pi) \ , \tag{32}$$

where $\varsigma^2_{\tilde{Y}}(\pi)$ and ς^2_H are defined as in (16) and (17) and

$$\varsigma^2_\pi = \frac{(1-2\pi)^2}{\pi^2(1-\pi)^2} E[\tau(S_i)(E[m_1(Z_i)|S_i] - E[m_0(Z_i)|S_i])^2] \ . \tag{33}$$

As before, replacing population quantities with their sample counterparts leads to a natural estimator of (33), specifically,

$$\hat{\varsigma}^2_\pi = \frac{(1-2\pi)^2}{\pi^2(1-\pi)^2} \sum_{s \in \mathcal{S}} \tau(s) \frac{n(s)}{n} \left((\hat{\mu}_{n,1}(s) - \bar{Y}_{n,1}) - (\hat{\mu}_{n,0}(s) - \bar{Y}_{n,0}) \right)^2 \ , \tag{34}$$

where $\hat{\mu}_{n,a}(s)$ is defined as in (19). The "adjusted" t-test with strata fixed effects is given by

$$\phi^{\text{sfe,adj}}_n(X^{(n)}) = I\{|T^{\text{sfe,adj}}_n(X^{(n)})| > z_{1-\frac{\alpha}{2}}\} \ , \tag{35}$$

where

$$T^{\text{sfe,adj}}_n(X^{(n)}) = \frac{\hat{\beta}_n - \theta_0}{\sqrt{\frac{\hat{\varsigma}^2_{\tilde{Y}}(\pi) + \hat{\varsigma}^2_H + \hat{\varsigma}^2_\pi}{n}}} \tag{36}$$

and $\hat{\varsigma}^2_{\tilde{Y}}(\pi)$ and $\hat{\varsigma}^2_H$ are defined as in (20) and (21). The following theorem establishes the desired result about

the asymptotic behavior of the "adjusted" t-test with strata fixed effects.

Theorem 4.4. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. For the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n^{sfe,adj}(X^{(n)})$ defined in (35) satisfies*

$$\lim_{n\to\infty} E[\phi_n^{sfe,adj}(X^{(n)})] = \alpha \tag{37}$$

whenever Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$.

Remark 4.8. In the leading case of randomization schemes that achieve "strong balance," i.e., satisfying $\tau(s) = 0$ for all $s \in \mathcal{S}$, it is worth emphasizing that the limiting variances in (15) and (32) are identical. In this sense, there is no reason to prefer one of $\phi_n^{t\text{-test,adj}}(X^{(n)})$ or $\phi_n^{\text{sfe,adj}}(X^{(n)})$ over the other for such randomization schemes. As explained in Section 3, examples of randomization schemes that achieve "strong balance" include biased-coin designs and stratified block randomization. More generally, the limiting variances in (15) and (32) are not ordered unambiguously. ■

Remark 4.9. Imbens and Rubin (2015, Ch. 9.6) examine the limit in probability of $\hat{\beta}_n$ under a specific randomization scheme, namely, stratified block randomization; see Example 3.4. In contrast to our results, they do not impose the requirement that $m(s)$ is chosen as in (8). In particular, they allow the proportion π to vary across strata. As a result, Assumption 2.2.(b) does not necessarily hold, and they conclude that $\hat{\beta}_n$ is generally not consistent for the average treatment effect, $\theta(Q)$. By exploiting Assumption 2.2.(b), we show that that $\hat{\beta}_n$ is in fact consistent for $\theta(Q)$. Imbens and Rubin (2015, Theorem 9.1) also analyze the limiting behavior of $\sqrt{n}(\hat{\beta}_n - \theta(Q))$. While they do not formally study $\phi_n^{\text{sfe}}(X^{(n)})$, their expression for the limiting variance suggests that this test would be exact when $m(s)$ is chosen as in (8). Our results show that this is generally not the case. In our simulation study in Section 5, we find that the rejection probability may in fact be dramatically less than the nominal level and that this difference translates into substantial power loss when compared with exact tests. For further discussion and results for the case of randomization schemes where the proportion π to vary across strata, see Bugni et al. (2016). ■

4.3 Covariate-Adaptive Permutation Test

In this section, we study the properties of a modified version of the permutation test, which we term the covariate-adaptive permutation test. In order to define the test, we require some further notation. Define

$$\mathbf{G}_n(S^{(n)}) = \{g \in \mathbf{G}_n : S_{g(i)} = S_i \text{ for all } 1 \leq i \leq n\} , \tag{38}$$

i.e., $\mathbf{G}_n(S^{(n)})$ is the subgroup of permutations of n elements that only permutes indices within strata. Define the action of $g \in \mathbf{G}_n(S^{(n)})$ on $X^{(n)}$ as follows:

$$gX^{(n)} = \{(Y_i, A_{g(i)}, Z_i) : 1 \leq i \leq n\} ,$$

i.e., $g \in \mathbf{G}_n$ acts on $X^{(n)}$ by permuting treatment assignment. For a given choice of test statistic $T_n(X^{(n)})$, the covariate-adaptive permutation test is given by

$$\phi_n^{\text{cap}}(X^{(n)}) = I\{T_n(X^{(n)}) > \hat{c}_n^{\text{cap}}(1-\alpha)\} , \tag{39}$$

where

$$\hat{c}_n^{\text{cap}}(1-\alpha) = \inf\left\{ x \in \mathbf{R} : \frac{1}{|\mathbf{G}_n(S^{(n)})|} \sum_{g \in \mathbf{G}_n(S^{(n)})} I\{T_n(gX^{(n)}) \leq x\} \geq 1-\alpha \right\} . \tag{40}$$

The following theorem describes the asymptotic behavior of the covariate-adaptive permutation test defined in (39) with $T_n(X^{(n)})$ given by the absolute value of $T_n^{t\text{-stat}}(X^{(n)})$ in (11). In particular, it shows that the limiting rejection probability of the proposed test under the null hypothesis equals the nominal level for randomization schemes with $\pi = \frac{1}{2}$ and $\tau(s) = 0$ for all $s \in \mathcal{S}$. As explained in Section 3, examples of such randomization schemes include biased-coin designs and stratified block randomization with $\pi = \frac{1}{2}$.

Theorem 4.5. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2 with $\pi = \frac{1}{2}$ and $\tau(s) = 0$ for all $s \in \mathcal{S}$. For the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n^{cap}(X^{(n)})$ defined in (39) with $T_n(X^{(n)})$ given by the absolute value of $T_n^{t\text{-}stat}(X^{(n)})$ in (11) satisfies*

$$\lim_{n\to\infty} E[\phi_n^{cap}(X^{(n)})] = \alpha \tag{41}$$

whenever Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$.

A by-product of the proof of Theorem 4.5 in the Appendix is that (41) holds even if $\pi \neq \frac{1}{2}$ provided that

$$\text{Var}[Y_i(1)] = \text{Var}[Y_i(0)] \text{ and } \text{Var}[\tilde{Y}_i(1)] = \text{Var}[\tilde{Y}_i(0)] ,$$

where $\tilde{Y}_i(a) = Y_i(a) - E[Y_i(a)|S_i]$ for $a \in \{0,1\}$. Outside of this exceptional circumstance, the limiting rejection probability of the test considered in Theorem 4.5 may strictly exceed the nominal level when $\pi \neq \frac{1}{2}$, as is evident in our simulation study in Section 5. The following theorem shows that this shortcoming of the covariate-adaptive permutation test can be removed by applying it with a more suitable choice of $T_n(X^{(n)})$, namely the absolute value of $T_n^{t\text{-stat,adj}}(X^{(n)})$ in (36). In this way, the theorem highlights the importance of Studentizing appropriately when applying the covariate-adaptive permutation test. See also Remark 4.13 below.

Theorem 4.6. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2 with $\tau(s) = 0$ for all $s \in \mathcal{S}$. For the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n^{cap}(X^{(n)})$ defined in (39) with $T_n(X^{(n)})$ given by the absolute value of $T_n^{t\text{-}stat,adj}(X^{(n)})$ in (24) satisfies*

$$\lim_{n\to\infty} E[\phi_n^{cap}(X^{(n)})] = \alpha$$

whenever Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$.

Remark 4.10. Note that $\{D_n(s)/\sqrt{n} : s \in \mathcal{S}\}$ is invariant with respect to transformations $g \in \mathbf{G}_n(S^{(n)})$. For this reason, it is not surprising that the validity of the covariate-adaptive permutation test in Theorems 4.5–4.6 requires that there is no (limiting) variation in this quantity in the sense that $\tau(s) = 0$ for all $s \in \mathcal{S}$. ■

Remark 4.11. By arguing as in Heckman et al. (2011) or Lee and Shaikh (2014), it is possible to show that (39) with any choice of $T_n(X^{(n)})$ is level α in finite samples for testing the much more narrowly defined null

hypothesis that specifies

$$Y_i(0)|S_i \stackrel{d}{=} Y_i(1)|S_i \tag{42}$$

whenever the treatment assignment mechanism is such that

$$gA^{(n)}|S^{(n)} \stackrel{d}{=} A^{(n)}|S^{(n)} \text{ for all } g \in \mathbf{G}_n(S^{(n)}) \ . \tag{43}$$

The property in (43) clearly holds, for example, for simple random sampling and stratified block randomization. Note further that (42) is implied by what is sometimes referred to as the "sharp null hypothesis," which specifies that

$$Y_i(0) = Y_i(1) \tag{44}$$

with probability one. Of course, Fisher-type tests of the null hypothesis (44) may be available even when (43) does not hold, though in that case they may not take the form of the covariate-adaptive permutation test studied here. For additional discussion, see Rosenbaum (2007) as well as Young (2016). By contrast, Theorems 4.5–4.6 asymptotically justify the use of (39) with certain choices of $T_n(X^{(n)})$ for testing the null hypothesis defined in (4) for randomization schemes satisfying our assumptions. By construction, this test additionally retains the finite-sample validity described above whenever (43) holds. The proof of Theorems 4.5–4.6 exploit recent developments in the literature on the asymptotic behavior of permutation tests. In particular, we employ a novel coupling construction following the approach put forward by Chung and Romano (2013) to verify conditions similar to those in Hoeffding (1952). ■

Remark 4.12. It may often be the case that $\mathbf{G}_n(S^{(n)})$ is too large to permit computation of $\hat{c}_n^{\text{cap}}(1-\alpha)$ defined in (40). In such situations, a stochastic approximation to the test may again be used by replacing $\mathbf{G}_n(S^{(n)})$ with $\hat{\mathbf{G}}_n = \{g_1, \ldots, g_B\}$, where g_1 equals the identity permutation and $g_2, \ldots, g_B$ are i.i.d. Unif$(\mathbf{G}_n(S^{(n)}))$. Theorems 4.5–4.6 remain true with such an approximation provided that $B \to \infty$ as $n \to \infty$. ■

Remark 4.13. One may, of course, also consider the behavior of the covariate-adaptive permutation test defined in (39) with other choices of $T_n(X^{(n)})$. For example, one may consider $T_n^{\text{sfe}}(X^{(n)})$ or $T_n^{\text{sfe, adj}}(X^{(n)})$ defined in (27) or (36), respectively. While we do not provide any details here, it is possible to show using arguments similar to those used in establishing Theorems 4.5–4.6 that Theorem 4.5 continues to hold with $T_n(X^{(n)})$ given by $|T_n^{\text{sfe}}(X^{(n)})|$ and that Theorem 4.6 continues to hold with $T_n(X^{(n)})$ given by $|T_n^{\text{sfe,adj}}(X^{(n)})|$. ■

Remark 4.14. Replacing $\mathbf{G}_n(S^{(n)})$ in the definition of the covariate-adaptive permutation test with $\mathbf{G}_n$, the set of all permutations of n elements, leads to what we refer to as a "naïve" permutation test. Rosenberger and Lachin (2016, page 105) argue that the use such tests in the presence of covariate-adaptive randomization is "inappropriate" because it does not respect stratification. In particular, it does not have the finite-sample validity of the covariate-adaptive permutation test described in Remark 4.11. For that reason, we do not consider it further in this paper. ■

5 Simulation Study

In this section, we examine the finite-sample performance of several different tests of (4), including those introduced in Section 4, with a simulation study. For $a \in \{0,1\}$ and $1 \leq i \leq n$, potential outcomes are generated in the simulation study according to the equation:

$$Y_i(a) = \mu_a + m_a(Z_i) + \sigma_a(Z_i)\epsilon_{a,i} \ . \tag{45}$$

where μ_a, $m_a(Z_i)$, $\sigma_a(Z_i)$, and $\epsilon_{a,i}$ are specified as follows. In each of the following specifications, $n = 200$, $\{(Z_i, \epsilon_{0,i}, \epsilon_{1,i}) : 1 \leq i \leq n\}$ are i.i.d., and the functions $m_a(Z_i)$ have been re-centered to have zero mean.

Model 1: $Z_i \sim$ Beta$(2,2)$ (re-centered and re-scaled to have mean zero and variance one); $\sigma_0(Z_i) = \sigma_0 = 1$ and $\sigma_1(Z_i) = \sigma_1$; $\epsilon_{0,i} \sim N(0,1)$ and $\epsilon_{1,i} \sim N(0,1)$; $m_0(Z_i) = m_1(Z_i) = \gamma Z_i$. In this case

$$Y_i = \mu_0 + (\mu_1 - \mu_0)A_i + \gamma Z_i + \eta_i \ ,$$

where

$$\eta_i = \sigma_1 A_i \epsilon_{1,i} + \sigma_0(1 - A_i)\epsilon_{0,i}$$

and $E[\eta_i | A_i, Z_i] = 0$.

Model 2: As in Model 1, but $m_0(Z_i) = -\gamma \log(Z_i + 3) I\{Z_i \leq \frac{1}{2}\}$.

Model 3: $Z_i \sim$ Unif$(-2,2)$; $\epsilon_{0,i} \sim \frac{1}{3}t_3$ and $\epsilon_{1,i} \sim \frac{1}{3}t_3$; $\sigma_0(Z_i) = Z_i^2$; $\sigma_1(Z_i) = Z_i^2\sigma_1$; and

$$m_0(Z_i) = m_1(Z_i) = \begin{cases} \gamma Z_i^2 & \text{if } Z_i \in [-1,1] \\ \gamma(2 - Z_i^2) & \text{otherwise} \end{cases} .$$

Model 4: As in Model 3, but

$$m_0(Z_i) = \begin{cases} \gamma Z_i^2 & \text{if } Z_i \in [-1,1] \\ \gamma Z_i & \text{otherwise} \end{cases} \quad \text{and} \quad m_1(Z_i) = \begin{cases} \gamma Z_i & \text{if } Z_i \in [-1,1] \\ \gamma Z_i^2 & \text{otherwise} \end{cases} .$$

When $\pi = \frac{1}{2}$, treatment status is determined according to one of the following four different covariate-adaptive randomization schemes:

SRS: Treatment assignment is generated as in Example 3.1.

BCD: Treatment assignment is generated as in Example 3.2 with $\lambda = \frac{3}{4}$.

WEI: Treatment assignment is generated as in Example 3.3 with $\varphi(x) = \frac{1}{2}(1 - x)$.

SBR: Treatment assignment is generated as in Example 3.4.

When $\pi \neq \frac{1}{2}$, we only consider simple random sampling and stratified block randomization. In each case, strata are determined by dividing the support of Z_i into $|\mathcal{S}|$ intervals of equal length and letting $S(Z_i)$ be

the function that returns the interval in which Z_i lies. In all cases, observed outcomes Y_i are generated according to (1). Finally, for each of the above specifications, we consider different values of $(|\mathcal{S}|, \pi, \gamma, \sigma)$ and consider both $(\mu_0, \mu_1) = (0, 0)$ (i.e., under the null hypothesis) and $(\mu_0, \mu_1) = (0, \frac{1}{2})$ (i.e., under the alternative hypothesis).

The results of our simulations are presented in Tables 1–5 below. Rejection probabilities are computed using 10^4 replications. Columns are labeled in the following way:

> ***t*-test**: The usual two-sample t-test as well as the "adjusted" version studied in Section 4.1.
>
> **Reg**: The usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes Y_i on treatment assignment A_i and covariates Z_i using heteroskedasticity-robust standard errors.
>
> **SFE**: The t-test with strata fixed effects (using heteroskedasticity-robust standard errors) as well as the "adjusted" version studied in Section 4.2.
>
> **CAP**: The covariate-adaptive permutation test applied to the usual two-sample t-statistic as well as the "adjusted" statistic studied in Section 4.3.
>
> **SFEP**: The covariate-adaptive permutation test applied to the t-test with strata fixed effects (using heteroskedasticity-robust standard errors) as well as the "adjusted" statistic described in Remark 4.13.

Note that whenever a column corresponds to more than one test, the "adjusted" version is listed second.

Table 1 displays the results of our baseline specification, where $(|\mathcal{S}|, \pi, \gamma, \sigma) = (4, 0.5, 2, 1)$. Table 2 displays the results for $(|\mathcal{S}|, \pi, \gamma, \sigma) = (4, 0.5, 4, \sqrt{2})$, to explore sensitivity to changes in (γ, σ_1). Tables 3 and 4 replace $\pi = 0.5$ with $\pi = 0.7$, so $(|\mathcal{S}|, \pi, \gamma, \sigma) = (4, 0.7, 2, 1)$ and $(|\mathcal{S}|, \pi, \gamma, \sigma) = (4, 0.7, 4, \sqrt{2})$. Finally, Table 5 is our baseline specification with a higher number of strata, so $(|\mathcal{S}|, \pi, \gamma, \sigma) = (10, 0.5, 2, 1)$. We organize our discussion of the results by test:

> ***t*-test**: As expected in light of Theorem 4.1 and Remark 4.2, we see the usual two-sample t-test has rejection probability under the null hypothesis very close to the nominal level under simple random sampling, but has rejection probability under the null hypothesis strictly less than the nominal level under more complicated randomization schemes. Indeed, in some instances, the rejection probability under the null hypothesis is close to zero. Moreover, for all specifications, the two-sample t-test has nearly the lowest rejection probability under the alternative hypothesis. Remarkably, this difference in power is pronounced even under simple random sampling. These results do not depend on the value of π or the number of strata, so Tables 1–5 show qualitatively similar results.
>
> Consistent with Theorem 4.2, the "adjusted" two-sample t-test has rejection probability under the null hypothesis close to the nominal level in nearly all specifications. An exception is when the treatment assignment mechanism is Efron's biased coin design, particularly in Model 1 (e.g., 8.72 in Table 2 and 12.73 in Table 5). The reason for the over-rejection in these cases appears to be purely a small sample phenomenon, consistent with the observation that it worsens when the number of strata is larger (and therefore the number of observations per stratum is smaller). In fact, in some simulations, the number of observations in a stratum may be as small as ten when there are four strata and as small as four

		Rejection rate under null - $\theta = 0$					Rejection rate under alternative - $\theta = 1/2$				
M	CAR	t-test	Reg	SFE	CAP	SFEP	t-test	Reg	SFE	CAP	SFEP
1	SRS	5.58/5.29	5.18	5.08/5.49	5.19/5.20	5.07/5.44	36.00/35.98	93.95	85.04/85.95	60.83/60.91	84.69/83.15
	WEI	0.77/5.59	4.94	5.08/5.64	5.28/5.28	5.17/4.70	31.39/59.99	94.12	85.21/86.14	68.53/68.96	84.75/84.59
	BCD	0.01/6.91	4.61	4.68/5.37	4.89/4.92	4.60/4.85	25.70/84.75	94.15	85.36/86.79	80.36/80.69	84.89/84.70
	SBR	0.02/5.45	4.81	4.86/5.40	4.77/4.78	4.89/4.95	24.68/86.09	93.96	85.42/86.12	84.50/84.54	85.03/84.82
2	SRS	5.50/5.31	5.30	4.89/5.84	5.26/5.20	5.15/5.09	48.33/48.50	69.22	65.96/68.00	54.90/55.12	66.67/66.38
	WEI	2.59/5.37	4.81	4.98/5.47	4.91/4.91	4.90/5.14	48.27/60.26	69.44	66.78/67.98	60.51/60.96	66.91/66.38
	BCD	1.46/5.92	4.64	4.88/5.03	4.98/5.09	4.97/4.90	47.97/68.27	69.72	67.32/68.54	65.29/65.37	66.83/66.31
	SBR	1.49/5.46	4.60	4.78/5.34	4.73/4.84	4.71/4.95	46.94/67.31	68.71	66.20/68.09	64.99/65.53	66.10/65.84
3	SRS	5.25/5.25	5.08	4.92/5.93	4.98/4.97	5.09/4.80	52.67/52.82	51.96	57.43/59.41	53.14/53.25	57.49/56.83
	WEI	4.33/5.60	4.13	5.45/5.44	5.42/5.36	5.61/5.00	52.99/57.23	52.27	57.68/59.18	56.03/56.27	57.94/57.89
	BCD	3.55/5.37	3.33	4.87/5.11	5.06/5.07	5.04/5.27	53.28/59.80	52.69	58.59/60.02	58.28/58.34	58.40/58.65
	SBR	3.87/5.51	3.73	5.07/5.50	5.24/5.23	5.25/4.96	53.42/59.98	52.89	58.73/59.13	58.43/58.42	58.58/58.68
4	SRS	5.53/5.41	5.51	5.26/5.80	5.36/5.36	5.86/5.61	29.69/29.54	32.70	38.52/41.67	34.04/34.02	39.43/40.28
	WEI	2.75/5.26	3.94	4.90/5.57	5.10/5.10	5.01/4.96	27.37/36.49	32.17	39.28/41.19	36.47/36.50	39.76/40.47
	BCD	2.22/5.82	3.38	5.23/5.50	5.16/5.17	5.29/5.23	25.50/41.68	32.45	40.09/41.38	39.25/39.31	39.95/40.00
	SBR	1.81/5.51	3.18	5.11/5.08	4.96/4.99	5.03/5.10	25.52/42.40	32.53	40.90/41.47	40.49/40.56	40.48/40.64

Table 1: Parameter values: $|\mathcal{S}| = 4$, $\pi = 0.5$, $\gamma = 2$, $\sigma_1 = 1$.

		Rejection rate under null - $\theta = 0$					Rejection rate under alternative - $\theta = 1/2$				
M	CAR	t-test	Reg	SFE	CAP	SFEP	t-test	Reg	SFE	CAP	SFEP
1	SRS	5.29/5.02	5.35	5.14/5.41	5.23/5.29	4.97/5.03	13.89/13.60	81.60	52.39/53.77	36.30/36.34	51.99/52.22
	WEI	0.57/5.65	4.96	5.40/4.96	5.18/5.21	5.28/5.10	5.35/25.57	81.71	52.64/54.30	39.90/40.03	52.28/52.61
	BCD	0.02/8.54	5.06	4.95/5.16	5.22/5.20	4.88/5.09	0.82/54.71	82.37	53.80/55.27	47.69/48.12	53.58/52.33
	SBR	0.00/5.64	5.03	4.87/5.44	4.76/4.79	4.89/5.30	0.64/54.81	82.16	53.13/54.53	51.91/52.08	53.31/53.48
2	SRS	5.24/5.04	5.32	4.92/6.35	5.53/5.36	5.54/5.51	18.92/18.63	32.30	29.16/32.04	24.82/24.81	30.13/29.82
	WEI	2.10/5.58	5.01	5.26/5.57	5.25/5.15	5.42/4.74	14.27/24.78	31.34	28.89/31.72	26.01/25.93	29.26/29.76
	BCD	1.10/5.98	4.68	5.02/5.59	4.90/4.99	5.05/4.98	12.27/30.77	31.20	29.01/31.30	28.07/28.03	29.02/29.83
	SBR	1.06/6.07	4.77	5.30/5.45	5.20/5.41	5.24/5.12	11.36/30.60	31.20	29.64/31.05	28.50/29.08	29.54/29.33
3	SRS	5.26/5.19	4.94	5.27/5.80	5.15/5.13	5.34/5.13	20.19/20.20	19.65	22.96/26.07	20.87/20.93	22.65/23.04
	WEI	3.82/5.20	3.64	4.81/4.97	4.94/4.95	4.82/5.19	19.05/22.97	18.67	23.09/24.84	22.24/22.29	22.86/22.68
	BCD	3.64/5.95	3.49	5.40/5.51	5.36/5.41	5.36/5.02	19.24/25.34	18.87	24.24/25.27	23.92/23.97	24.05/23.11
	SBR	3.54/5.63	3.40	5.04/5.18	5.37/5.37	5.30/5.07	18.72/24.84	18.48	23.66/25.96	23.63/23.62	23.55/23.60
4	SRS	4.98/4.88	4.83	4.47/6.06	4.85/4.76	5.07/5.33	11.77/11.58	11.48	12.99/15.97	12.72/12.63	13.67/14.02
	WEI	2.62/4.89	3.15	4.43/5.51	4.84/4.76	4.54/5.02	7.39/12.45	9.180	12.73/15.36	12.26/12.35	13.13/13.68
	BCD	1.59/5.25	2.82	4.52/5.42	4.63/4.65	4.58/4.66	6.25/14.28	8.640	13.04/15.45	12.93/12.90	12.96/13.69
	SBR	1.57/4.95	2.74	4.50/5.25	4.48/4.52	4.45/4.94	6.15/14.20	8.890	13.07/15.14	13.29/13.35	13.24/13.74

Table 2: Parameter values: $|\mathcal{S}| = 4$, $\pi = 0.5$, $\gamma = 4$, $\sigma_1 = \sqrt{2}$.

when there are ten strata. In unreported simulations, we find that this phenomenon disappears with a larger sample size. It is worth noting that while $\tau(s) = 0$ for Efron's biased coin design, as discussed in Example 3.2, the distribution of $D_n(s)$ may exhibit considerable variation in small samples. For this reason, estimating $\varsigma_A(\pi)$ with zero, as suggested by $\hat{\varsigma}_A(\pi)$ defined in (22), may be misleading in small samples. As expected, the "adjusted" two-sample t-test is considerably more powerful than the usual two-sample t-test. On the other hand, in comparison with other exact tests, it is generally less powerful under simple random sampling and Wei's adaptive biased coin design, but among the most powerful if not the most powerful under Efron's biased coin design and stratified block randomization.

Reg: The usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes Y_i on treatment assignment A_i and covariates Z_i using heteroskedasticity-robust standard errors has rejection probability under the null hypothesis very close to the nominal level for Model 1, i.e., when the

		Rejection rate under null - $\theta = 0$				Rejection rate under alternative - $\theta = 1/2$			
M	CAR	t-test	SFE	CAP	SFEP	t-test	SFE	CAP	SFEP
1	SRS	5.43/4.94	5.30/6.35	4.85/4.82	5.37/5.37	31.30/31.38	77.61/78.85	54.73/54.97	76.32/76.32
	SBR	0.03/5.56	5.20/5.32	4.90/5.07	5.02/5.01	17.38/79.02	79.13/79.63	77.58/77.42	78.61/78.61
2	SRS	5.00/4.92	5.12/5.98	6.94/7.06	4.72/4.72	50.14/50.38	53.65/55.73	58.53/58.91	51.90/51.90
	SBR	2.46/5.57	3.34/5.49	10.25/4.98	3.24/5.05	49.55/62.61	54.76/63.31	73.11/60.62	53.65/61.59
3	SRS	5.45/5.54	4.80/5.61	4.89/4.89	4.85/4.85	45.25/45.59	50.67/52.90	45.78/45.93	49.90/49.90
	SBR	3.83/5.64	5.21/5.64	5.00/5.06	5.07/5.07	46.13/52.69	51.67/52.93	51.09/50.85	51.15/50.85
4	SRS	5.51/5.07	5.09/6.07	7.01/6.79	4.94/4.94	27.70/27.73	27.03/28.51	33.11/33.18	25.87/25.87
	SBR	1.20/5.34	1.65/5.30	5.08/4.63	1.53/4.61	22.66/43.12	24.98/43.09	41.99/40.52	23.70/40.70

Table 3: Parameter values: $|\mathcal{S}| = 4$, $\pi = 0.7$, $\gamma = 2$, $\sigma_1 = 1$.

		Rejection rate under null - $\theta = 0$				Rejection rate under alternative - $\theta = 1/2$			
M	CAR	t-test	SFE	CAP	SFEP	t-test	SFE	CAP	SFEP
1	SRS	5.22/4.67	4.97/6.44	4.17/4.24	4.85/4.85	13.11/12.58	48.18/48.82	32.12/32.28	46.30/46.30
	SBR	0.00/5.70	5.00/4.99	3.79/4.84	4.89/4.88	0.32/49.26	49.25/49.23	42.67/46.72	48.29/48.37
2	SRS	5.04/4.90	5.11/5.99	7.95/8.05	4.84/4.84	19.47/19.46	21.80/23.87	28.72/28.94	21.00/21.00
	SBR	1.85/5.57	2.83/5.31	10.87/4.90	2.60/4.81	13.95/27.64	18.60/27.56	39.48/25.34	17.77/25.93
3	SRS	5.70/5.72	4.94/5.67	4.92/4.88	5.01/5.01	17.70/17.70	20.52/22.21	18.18/18.22	19.91/19.91
	SBR	3.31/5.67	5.09/5.56	4.81/4.85	4.96/4.96	14.81/21.33	20.57/21.44	19.24/20.13	20.31/20.08
4	SRS	5.42/5.06	5.28/6.46	7.38/7.12	5.21/5.21	11.53/11.11	10.27/11.03	13.77/13.61	9.16/9.16
	SBR	1.16/5.75	1.60/5.58	5.43/5.09	1.43/4.83	5.00/15.67	5.76/15.43	14.64/14.33	5.42/14.25

Table 4: Parameter values: $|\mathcal{S}| = 4$, $\pi = 0.7$, $\gamma = 4$, $\sigma_1 = \sqrt{2}$.

linear regression is correctly specified. Interestingly, even though the linear regression is incorrectly specified for all other models, the rejection probability of the test under the null hypothesis never exceeds the nominal level, though it is frequently much less than the nominal level. Not surprisingly, for Model 1, the test also has the highest rejection probability under the alternative hypothesis. For all other models, the rejection probability of the test under the alternative hypothesis is lower than that of some of the exact tests considered.

SFE: As expected in light of Theorem 4.3 and Remark 4.5, the t-test with strata fixed effects has rejection probability under the null hypothesis very close to the nominal level under simple random sampling or when $\pi = 0.5$, but has rejection probability under the null hypothesis strictly less than the nominal level under stratified block randomization with $\pi = 0.7$. Under simple random sampling or when $\pi = 0.5$, it is among the most powerful if not the most powerful test considered here.

When $\pi = 0.5$, the "adjusted" t-test with strata fixed effects behaves similarly to the t-test with strata fixed effects, though it exhibits some mild over-rejection, especially when the number of strata is larger (e.g., Models 2 and 4 under simple random sampling in Table 5). As in the case of the "adjusted" t-test above, this reflects a small sample phenomenon that disappears in unreported simulations with a larger sample size. In contrast to the t-test with strata fixed effects and consistent with Theorem 4.4, the "adjusted" t-test with strata fixed effects has rejection probability under the null hypothesis very close to the nominal level even when $\pi = 0.7$. As a result, it also has much greater power than the t-test with strata fixed effects when $\pi = 0.7$. Indeed, among all tests we consider, it appears to be among the most powerful if not the most powerful test considered here. It also appears to be less susceptible to over-rejection than the "adjusted" two-sample t-test.

		Rejection rate under null - $\theta = 0$					Rejection rate under alternative - $\theta = 1/2$				
M	CAR	t-test	Reg	SFE	CAP	SFEP	t-test	Reg	SFE	CAP	SFEP
1	SRS	5.71/5.45	5.18	5.25/5.82	5.18/5.18	5.24/4.90	35.59/35.56	93.65	90.94/92.73	64.97/65.04	90.74/90.98
	WEI	0.62/5.78	5.24	4.80/5.52	4.91/4.89	4.86/5.22	30.70/63.36	93.98	91.84/93.27	73.18/73.70	91.69/91.84
	BCD	0.07/12.73	4.95	4.89/5.74	5.15/5.17	4.90/4.79	26.48/89.07	93.92	92.80/93.42	81.37/82.45	92.53/92.69
	SBR	0.01/6.81	5.21	4.98/5.74	5.06/5.01	5.05/4.87	22.02/92.51	94.00	92.68/93.78	90.25/90.32	92.59/92.44
2	SRS	5.19/5.03	4.74	5.18/7.18	5.07/5.10	5.81/6.10	48.41/48.64	68.38	68.78/74.23	57.99/58.05	70.34/70.58
	WEI	2.11/5.44	4.21	4.62/5.89	4.66/4.64	4.76/5.56	47.74/63.56	70.34	71.26/74.04	63.68/63.81	71.34/70.67
	BCD	1.31/6.49	3.92	4.68/5.71	4.48/4.36	4.55/5.20	47.86/72.85	69.35	71.39/73.72	67.09/67.15	70.96/71.26
	SBR	0.94/6.26	4.11	4.84/5.67	4.70/4.99	4.89/5.27	46.96/73.03	68.57	71.62/74.55	67.74/69.22	71.36/71.24
3	SRS	5.41/5.17	5.18	4.51/5.88	4.76/4.70	4.81/5.10	52.42/52.53	52.02	86.93/87.81	68.75/68.95	86.91/87.17
	WEI	1.29/5.57	1.17	5.10/5.86	5.27/5.32	5.38/4.83	54.11/74.77	53.33	87.39/88.91	77.06/77.70	87.55/87.30
	BCD	0.33/7.56	0.30	4.53/5.49	4.80/4.78	4.87/4.77	54.40/87.29	53.31	87.36/88.73	83.67/84.07	87.82/87.81
	SBR	0.12/5.98	0.11	4.68/5.18	5.09/5.08	5.02/5.07	54.32/88.33	53.57	87.70/89.47	87.15/87.23	87.83/88.19
4	SRS	5.47/5.28	5.47	4.92/7.15	5.76/5.75	6.14/5.53	30.60/30.48	33.46	38.34/46.98	36.43/36.43	41.45/41.31
	WEI	2.47/5.42	3.37	4.43/5.46	4.90/4.82	4.90/4.73	27.48/39.47	32.95	40.92/46.03	39.71/39.51	42.58/41.75
	BCD	1.60/6.16	2.78	4.89/5.21	4.87/4.86	4.97/4.96	25.39/46.70	32.31	43.02/46.17	41.83/41.79	43.28/42.67
	SBR	1.44/5.58	2.46	4.66/5.01	4.66/4.72	4.75/4.33	24.14/46.40	31.74	43.13/45.70	42.71/43.08	43.27/42.28

Table 5: Parameter values: $|\mathcal{S}| = 10$, $\pi = 0.5$, $\gamma = 2$, $\sigma_1 = 1$.

CAP: As expected in light of Theorems 4.5–4.6, both the covariate-adaptive permutation test applied to the usual two-sample t-statistic and applied to the "adjusted" statistic studied in Section 4.1 have rejection probability under the null hypothesis very close to the nominal level when $\pi = 0.5$. Remarkably, the rejection probabilities are close to the nominal level even for treatment assignment mechanisms with $\tau(s) > 0$, such as simple random sampling and Wei's adaptive biased coin design. The rejection probability under the alternative hypothesis, on the other hand, is typically lower than that of some of the other exact tests considered, including the "adjusted" two-sample t-test and "adjusted" t-test with strata fixed effects, though the difference in power is small under stratified block randomization.

When $\pi = 0.7$, the covariate-adaptive permutation test applied to the usual two-sample t-statistic may under-reject (e.g., Model 1 with stratified block randomization in Table 4) or over-reject (e.g., Model 2 with stratified block randomization in Table 5) under the null hypothesis, which illustrates the importance of the requirement $\pi = 0.5$ in Theorem 4.5. On the other hand, consistent with Theorem 4.6, when $\pi = 0.7$ the covariate-adaptive permutation test applied to the "adjusted" statistic studied in Section 4.1 has rejection probability under the null hypothesis very close to the nominal level under stratified block randomization. The rejection probability under the alternative hypothesis, on the other hand, is again lower than that of some of the other exact tests considered, including the "adjusted" two-sample t-test and "adjusted" t-test with strata fixed effects.

SFEP: The results for covariate-adaptive permutation tests applied to a t-test with strata fixed effects or applied to the "adjusted" statistic described in Section 4.2 are qualitatively the same as those above. This phenomenon is consistent with Remark 4.13. The rejection probability under the alternative hypothesis is occasionally greater than that of the covariate-adaptive permutation test applied to the "adjusted" two-sample t-statistic described in Section 4.1 (e.g., Model 1 with stratified block randomization in Table 4).

6 Recommendations for Empirical Practice

According to our theoretical results, the "adjusted" two-sample t-test and the "adjusted" t-test with strata fixed effects are both exact under the most general conditions considered here. In particular, they are both exact for any value of π and regardless of whether $\tau(s) = 0$ for all $s \in \mathcal{S}$ or not. In our simulations, the "adjusted" t-test with strata fixed effects appears to be more powerful than the "adjusted" two-sample t-test whenever it is not the case that $\tau(s) = 0$ for all $s \in \mathcal{S}$. It also appears to be less prone to over-rejection than the "adjusted" two-sample t-test. For these reasons, we recommend it over the "adjusted" two-sample t-test.

When $\tau(s) = 0$ for all $s \in \mathcal{S}$, the covariate-adaptive permutation tests described in Theorems 4.5–4.6 and Remark 4.13 are available. These tests have some finite-sample validity, as described in Remark 4.11. When applied with the "adjusted" t-statistic with strata fixed effects described in Section 4.2, the power of the test in our simulations is close to, but typically lower than, that of the "adjusted" t-test with strata fixed effects. To the extent that the finite-sample validity described in Remark 4.11 is deemed important, this tests may be preferred to the "adjusted" t-test with strata fixed effects.

Finally, we note some implications of our results for the design of experiments. To the extent that researchers wish to use any of the exact tests in this paper, our results imply that randomization schemes that achieve "strong balance," i.e., that satisfy $\tau(s) = 0$ for all $s \in \mathcal{S}$, have two advantages. First, these tests have higher power (at least in large samples) under such randomization schemes. In the case of the "adjusted" two-sample t-test, this feature follows from the expression for the limiting variance in (15), but similar conclusions hold for any of the exact tests considered in this paper. Second, as mentioned above, such randomization schemes permit the use of covariate-adaptive permutation tests, which have some finite-sample validity.

7 Empirical Illustration

We conclude with an empirical illustration of our theoretical results using data from Chong et al. (2016), who study the effect of iron deficiency on educational attainment and cognitive ability using a randomized controlled trial.

7.1 Empirical Setting

We provide here only a brief description of the empirical setting and refer the reader to Chong et al. (2016) for a more detailed description. The units in the experiment consist of 215 students in a rural secondary school in the Cajamara district of Peru between October and December in 2009. During this period of time, each student was assigned at random to one of two treatments and a control. Assignment was stratified by the number of years of secondary school completed, which took values in $\mathcal{S} = \{1, \ldots, 5\}$. The number of students in a stratum ranges from 30 to 58. Within each stratum, (approximately) one third of the students were assigned to each of treatment one, treatment two, and a control. Students assigned to treatment one were shown upon logging into their school computer an educational video in which a physician explained the

Outcome	p-values			
	t-test	SFE	CAP	SFEP
1: # of Pills	0.063/0.062	0.102/0.070	0.599/0.599	0.599/0.599
2: Grade Point Avg.	6.494/5.304	4.991/4.206	5.788/5.788	5.190/4.790
3: Cognitive Ability	6.466/5.273	5.305/4.355	6.387/6.387	5.389/5.389

Table 6: Application using data from Chong et al. (2016)

importance of iron for overall health and encouraged iron supplementation; students assigned to treatment two were instead exposed to a similar educational video in which the physician was replaced by a well known soccer player; students assigned to the control were exposed to a video unrelated to iron featuring a dentist encouraging good oral hygiene. Importantly, throughout the experiment, the researchers stocked the local clinic with iron supplements, which were given at no cost to any student who requested them. Chong et al. (2016) examine the effect of both treatments on a variety of different outcomes. The following three outcomes are among the most important for their conclusions: number of iron supplement pills taken between October and December in 2009; grade point average between June and December in 2009; and cognitive ability as measured by the average score across different Nintendo Wii games (labeled "identify", "memorize", "analyze", "compute", and "visualize"). Below, we revisit the effect of treatment one (relative to the control) on each of these three outcomes using the tests described in Section 4. As a result, Assumption 2.2 is satisfied with $\pi = \frac{1}{2}$ and $\tau(s) = 0$ for all $s \in \mathcal{S}$.

7.2 Results

The results of our exercise are presented in Table 6 below. The rows of Table 6 correspond to the three different outcomes mentioned above, whereas the columns of Table 6 correspond to the same eight tests presented in Tables 3–4. Each entry in the table is a pair of p-values for testing the null hypothesis that treatment one has no average effect on the outcome of interest using the corresponding test; as in Section 5, the second p-value in each case corresponds to the "adjusted" version of the corresponding test.

In the case of Outcome 1, we see that the p-value from the two-sample t-test and the "adjusted" two-sample t-test are nearly the same. Since $\tau(s) \neq \pi(1-\pi)$, this suggests that it may be the case that stratification is irrelevant for this outcome. In the case of Outcome 2 and Outcome 3, however, we see that the p-values from the "adjusted" two-sample t-test are nearly 20% lower. In all cases, the "adjusted" t-test with strata fixed effects leads to lower p-values than the t-test with strata fixed effects. Since in this case $\pi = \frac{1}{2}$, however, this may simply be a small sample phenomenon that would disappear with a larger sample size. Finally, the p-values obtained by the four different covariate-adaptive permutation tests considered here are all larger than the corresponding p-values considered above.

Appendix A Proof of the main results

Throughout the Appendix we employ the following notation, not necessarily introduced in the text.

$\sigma^2_X(s)$	For a random variable X, $\sigma^2_X(s) = \text{Var}[X\|S=s]$
σ^2_X	For a random variable X, $\sigma^2_X = \text{Var}[X]$
μ_a	For $a \in \{0,1\}$, $E[Y_i(a)]$
$\tilde{Y}_i(a)$	For $a \in \{0,1\}$, $Y_i(a) - E[Y_i(a)\|S_i]$
$m_a(Z_i)$	For $a \in \{0,1\}$, $E[Y_i(a)\|Z_i] - \mu_a$
$\varsigma^2_Y(\pi)$	$\frac{1}{\pi}\sigma^2_{Y(1)} + \frac{1}{1-\pi}\sigma^2_{Y(0)}$
$\varsigma^2_{\tilde{Y}}(\pi)$	$\frac{1}{\pi}\sigma^2_{\tilde{Y}(1)} + \frac{1}{1-\pi}\sigma^2_{\tilde{Y}(0)}$
$\varsigma^2_A(\pi)$	$\sum_{s\in\mathcal{S}} p(s)\tau(s)(\frac{1}{\pi}E[m_1(Z_i)\|S_i=s] + \frac{1}{1-\pi}E[m_0(Z_i)\|S_i=s])^2$
ς^2_H	$\sum_{s\in\mathcal{S}} p(s)(E[m_1(Z_i)\|S_i=s] - E[m_0(Z_i)\|S_i=s])^2$
ς^2_π	$\frac{(1-2\pi)^2}{\pi^2(1-\pi)^2}\sum_{s\in\mathcal{S}} p(s)\tau(s)\left(E[m_1(Z_i)\|S_i=s] - E[m_0(Z_i)\|S_i=s]\right)^2$
$n(s)$	Number of individuals in stratum $s \in \mathcal{S}$
$n_a(s)$	For $a \in \{0,1\}$, number of individuals with $A_i = a$ in stratum $s \in \mathcal{S}$

Table 7: Useful notation

A.1 Proof of Theorem 4.1

We start the proof by showing that

$$\sqrt{n}\left(\bar{Y}_{n,1} - \bar{Y}_{n,0} - \theta(Q)\right) \xrightarrow{d} N\left(0, \varsigma^2_{\tilde{Y}}(\pi) + \varsigma^2_H + \varsigma^2_A(\pi)\right) . \tag{A-46}$$

Consider the following derivation:

$$\begin{aligned}
\sqrt{n}(\bar{Y}_{n,1} - \bar{Y}_{n,0} - \theta(Q)) &= \sqrt{n}\left(\frac{1}{n_1}\sum_{i=1}^n (Y_i(1)-\mu_1)A_i - \frac{1}{n_0}\sum_{i=1}^n (Y_i(0)-\mu_0)(1-A_i)\right) \\
&= R^*_{n,1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\left(1-\pi-\frac{D_n}{n}\right)(Y_i(1)-\mu_1)A_i - \left(\pi+\frac{D_n}{n}\right)(Y_i(0)-\mu_0)(1-A_i)\right) \\
&= R^*_{n,1}(R^*_{n,2} + R^*_{n,3}) ,
\end{aligned}$$

where we used $D_n = \sum_{s\in\mathcal{S}} D_n(s)$, $\frac{n_1}{n} = \frac{D_n}{n} + \pi$, and the following definitions:

$$\begin{aligned}
R^*_{n,1} &\equiv \left(\frac{D_n}{n} + \pi\right)^{-1}\left(1 - \pi - \frac{D_n}{n}\right)^{-1} , \\
R^*_{n,2} &\equiv \frac{1}{2\sqrt{n}}\sum_{i=1}^n \left((Y_i(1)-\mu_1)(1-\pi)A_i - (Y_i(0)-\mu_0)\pi(1-A_i)\right) , \\
R^*_{n,3} &\equiv -\frac{D_n}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^n \left((Y_i(1)-\mu_1)A_i + (Y_i(0)-\mu_0)(1-A_i)\right) .
\end{aligned}$$

By Assumption 2.2.(b), $\frac{D_n}{n} \xrightarrow{P} 0$, which in turn implies that $R^*_{n,1} \xrightarrow{P} (\pi(1-\pi))^{-1}$. Lemma B.1 implies $R^*_{n,2} \xrightarrow{d} \pi(1-\pi)N(0, \varsigma^2_{\tilde{Y}}(\pi) + \varsigma^2_H + \varsigma^2_A(\pi))$. Lemma B.3 and Assumption 2.2.(b) imply $R^*_{n,3} \xrightarrow{P} 0$. The desired conclusion thus follows from the continuous mapping theorem.

We next prove that

$$\sqrt{n}\sqrt{\frac{\hat{\sigma}^2_{n,1}}{n_1} + \frac{\hat{\sigma}^2_{n,0}}{n_0}} \xrightarrow{P} \varsigma_Y(\pi) \ . \tag{A-47}$$

This follows from showing that $\frac{n\hat{\sigma}^2_{n,1}}{n_1} \xrightarrow{P} \frac{1}{\pi}\sigma^2_{Y(1)}$ and $\frac{n\hat{\sigma}^2_{n,0}}{n_0} \xrightarrow{P} \frac{1}{1-\pi}\sigma^2_{Y(0)}$. We only show the first result; the proof of the second one is analogous. Start by writing $\bar{Y}_{n,1}$ as follows:

$$\bar{Y}_{n,1} \equiv \frac{1}{n_1}\sum_{i=1}^n A_i Y_i = \mu_1 + \frac{n}{n_1}\frac{1}{n}\sum_{i=1}^n A_i(Y_i(1) - \mu_1) \ . \tag{A-48}$$

Then consider the following derivation:

$$\begin{aligned}
\frac{n\hat{\sigma}^2_{n,1}}{n_1} &= \frac{n}{n_1}\frac{1}{n_1}\sum_{i=1}^n (Y_i - \bar{Y}_{n,1})^2 A_i \\
&= \frac{n}{n_1}\frac{1}{n_1}\sum_{i=1}^n (\mu_1 - \bar{Y}_{n,1} + Y_i(1) - \mu_1)^2 A_i \\
&= \frac{n}{n_1}\left(\frac{n}{n_1}\frac{1}{n}\sum_{i=1}^n (Y_i(1) - \mu_1)^2 A_i - (\mu_1 - \bar{Y}_{n,1})^2\right) \\
&= \left(\frac{n}{n_1}\right)^2 R^\star_{n,4} - \left(\frac{n}{n_1}\right)^3 R^{\star 2}_{n,5} \ ,
\end{aligned}$$

where we used (A-48) and the following definitions:

$$\begin{aligned}
R^\star_{n,4} &\equiv \frac{1}{n}\sum_{i=1}^n (Y_i(1) - \mu_1)^2 A_i \ , \\
R^\star_{n,5} &\equiv \frac{1}{n}\sum_{i=1}^n (Y_i(1) - \mu_1) A_i \ .
\end{aligned}$$

Since $\frac{n}{n_1} = (\frac{D_n}{n} + \pi)^{-1}$ and $\frac{D_n}{n} \xrightarrow{P} 0$ by Assumption 2.2.(b), it follows that $\frac{n}{n_1} \xrightarrow{P} \frac{1}{\pi}$. The result follows from showing that $R^\star_{n,4} \xrightarrow{P} \pi\sigma^2_{Y(1)}$ and $R^\star_{n,5} \xrightarrow{P} 0$. Since $E[(Y_i(1) - \mu_1)^2] = \sigma^2_{Y(1)}$ and $E[(Y_i(1) - \mu_1)] = 0$, this follows immediately from Lemma B.3. Finally, note that Assumption 2.1 implies that $\varsigma^2_{\tilde{Y}}(\pi) > 0$.

To prove that $\varsigma^2_{\tilde{Y}}(\pi) + \varsigma^2_H + \varsigma^2_A(\pi) \leq \varsigma^2_Y(\pi)$ holds with strict inequality unless (12) holds, notice that for $a \in \{0, 1\}$,

$$\sigma^2_{\tilde{Y}(a)} = \sigma^2_{Y(a)} - \sum_{s\in\mathcal{S}} E[(Y_i(1) - \mu_1)|S_i = s]^2 p(s) = \sigma^2_{Y(a)} - \sum_{s\in\mathcal{S}} E[m_a(Z_i)|S_i = s]^2 p(s) \ . \tag{A-49}$$

Using (A-49), we see that

$$\begin{aligned}
\varsigma^2_Y(\pi) - \varsigma^2_{\tilde{Y}}(\pi) - \varsigma^2_H - \varsigma^2_A(\pi) &= \frac{1}{\pi}(\sigma^2_{Y(1)} - \sigma^2_{\tilde{Y}(1)}) + \frac{1}{1-\pi}(\sigma^2_{Y(0)} - \sigma^2_{\tilde{Y}(0)}) \\
&\quad - \sum_{s\in\mathcal{S}} p(s)(E[m_1(Z_i)|S_i = s] - E[m_0(Z_i)|S_i = s])^2 \\
&\quad - \sum_{s\in\mathcal{S}} p(s)\tau(s)\left(\frac{1}{\pi}E[m_1(Z_i)|S_i = s] + \frac{1}{1-\pi}E[m_0(Z_i)|S_i = s]\right)^2 \\
&= \sum_{s\in\mathcal{S}} p(s)\left(\pi(1-\pi) - \tau(s)\right)\left(\frac{1}{\pi}E[m_1(Z_i)|S_i = s] + \frac{1}{1-\pi}E[m_0(Z_i)|S_i = s]\right)^2 \ ,
\end{aligned}$$

where, by Assumption 2.2.(b), $0 \leq \tau(s) \leq \pi(1-\pi)$. The right-hand side of this last display is non-negative and it is zero if and only if (12) holds, as required.

A.2 Proof of Theorem 4.2

As noted in the text, the proof of Theorem 4.1 establishes that (15) holds. Note further that Assumption 2.1 implies that $\varsigma^2_{\tilde{Y}}(\pi) > 0$. To complete the proof, we argue that $\hat{\varsigma}^2_{\tilde{Y}}(\pi)$, $\hat{\varsigma}^2_H$, and $\hat{\varsigma}^2_A(\pi)$ defined in (20)–(22) satisfy $\hat{\varsigma}^2_{\tilde{Y}}(\pi) \xrightarrow{P} \varsigma^2_{\tilde{Y}}(\pi)$, $\hat{\varsigma}^2_H \xrightarrow{P} \varsigma^2_H$, and $\hat{\varsigma}^2_A(\pi) \xrightarrow{P} \varsigma^2_A(\pi)$.

We begin by noting some preliminary facts that will be useful in the analysis of these estimators. First, the weak law of large numbers implies that $\frac{n(s)}{n} = \frac{1}{n}\sum_{1\leq i\leq n} I\{S_i = s\} \xrightarrow{P} p(s)$. Second, for $a \in \{0,1\}$, $\bar{Y}_{n,a} \xrightarrow{P} \mu_a$. For the case of $\bar{Y}_{n,1}$, this follows by writing $\bar{Y}_{n,1}$ as in (A-48), arguing as in the proof of Theorem 4.1 to establish that $\frac{n}{n_1} \xrightarrow{P} \frac{1}{\pi}$, and applying Lemma B.3. A similar argument establishes the result for $\bar{Y}_{n,0}$. Finally, for $a \in \{0,1\}$, $\hat{\mu}_{n,a}(s)$ defined in (19) converges in probability to $E[Y_i(a)|S_i = s]$. We prove this for the case of $\hat{\mu}_{n,1}(s)$; an analogous argument establishes it for $\hat{\mu}_{n,0}(s)$. Begin by writing

$$\hat{\mu}_{n,1}(s) = \frac{1}{n_1(s)} \sum_{1\leq i\leq n} Y_i I\{A_i = 1, S_i = s\} = \frac{n}{n_1(s)} \frac{1}{n} \sum_{1\leq i\leq n} Y_i(1) I\{S_i = s\} A_i \ .$$

From Assumption 2.2.(b), $\frac{n_1(s)}{n} = \frac{D_n(s)}{n} + \pi\frac{n(s)}{n} \xrightarrow{P} \pi p(s)$. Lemma B.3 implies that

$$\frac{1}{n} \sum_{1\leq i\leq n} Y_i(1) I\{S_i = s\} A_i \xrightarrow{P} \pi E[Y_i(1) I\{S_i = s\}] \ .$$

The desired conclusion about $\hat{\mu}_{n,0}(s)$ now follows immediately.

Now consider $\hat{\varsigma}^2_{\tilde{Y}}(\pi)$. For $a \in \{0,1\}$, note that

$$\begin{aligned}
\text{Var}[\tilde{Y}_i(a)] &= E[(Y_i(a) - E[Y_i(a)|S_i])^2] \\
&= E[Y_i(a)^2] - 2E[Y_i(a)E[Y_i(a)|S_i]] + E[E[Y_i(a)|S_i]^2] \\
&= E[Y_i(a)^2] - E[E[Y_i(a)|S_i]^2] \\
&= E[Y_i(a)^2] - \sum_{s\in\mathcal{S}} p(s) E[Y_i(a)|S_i = s]^2 \ .
\end{aligned}$$

Next, note that

$$\frac{1}{n_1} \sum_{1\leq i\leq n} Y_i^2 A_i = \frac{n}{n_1} \frac{1}{n} \sum_{1\leq i\leq n} Y_i(1)^2 A_i \ . \tag{A-50}$$

By arguing as in the proof of Theorem 4.1 to establish that $\frac{n}{n_1} \xrightarrow{P} \frac{1}{\pi}$ and applying Lemma B.3, we see that (A-50) converges in proability to $E[Y_i(1)^2]$. An analogous argument shows that

$$\frac{1}{n_0} \sum_{1\leq i\leq n} Y_i^2 (1 - A_i) \xrightarrow{P} E[Y_i(0)^2] \ .$$

Combining these convergences with the preliminary facts above, we see that $\hat{\varsigma}^2_{\tilde{Y}}(\pi) \xrightarrow{P} \varsigma^2_{\tilde{Y}}(\pi)$.

Next, consider $\hat{\varsigma}^2_H$. For $a \in \{0,1\}$, note that

$$E[m_a(Z_i)|S_i] = E[Y_i(a)|S_i] - \mu_a \ . \tag{A-51}$$

Hence,

$$\varsigma_H^2 = \sum_{s\in\mathcal{S}} p(s)((E[Y_i(1)|S_i = s] - \mu_1) - (E[Y_i(0)|S_i = s] - \mu_0))^2 \ .$$

Using the preliminary facts above, we see that $\hat{\varsigma}_H^2 \xrightarrow{P} \varsigma_H^2$.

Finally, consider $\hat{\varsigma}_A^2(\pi)$. Using (A-51), we see that

$$\varsigma_A^2(\pi) = \sum_{s\in\mathcal{S}} p(s)\tau(s)\left(\frac{1}{\pi}(E[Y_i(1)|S_i = s] - \mu_1) - \frac{1}{1-\pi}(E[Y_i(0)|S_i = s] - \mu_0)\right)^2 \ .$$

Using the preliminary facts above, we again see that $\hat{\varsigma}_A^2(\pi) \xrightarrow{P} \varsigma_A^2(\pi)$.

A.3 Proof of Theorem 4.3

We start the proof by showing that

$$\sqrt{n}(\hat{\beta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma_Y^2(\pi) + \varsigma_H^2 + \varsigma_\pi^2) \ . \tag{A-52}$$

To this end, write $\hat{\beta}_n$ as

$$\hat{\beta}_n = \frac{\sum_{i=1}^n \tilde{A}_i Y_i}{\sum_{i=1}^n \tilde{A}_i^2} \ ,$$

where $\tilde{A}_i$ is the projection of A_i on the strata indicators, i.e., $\tilde{A}_i = A_i - n_1(S_i)/n(S_i)$, where

$$\frac{n_1(S_i)}{n(S_i)} = \sum_{s\in\mathcal{S}} I\{S_i = s\}\frac{n_1(s)}{n(s)} \ .$$

Next, note that

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_n - \theta(Q)) &= \frac{\sqrt{n}}{\frac{1}{n}\sum_{i=1}^n \tilde{A}_i^2}\left(\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i Y_i\right) - \theta(Q)\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i^2\right)\right) \\
&= \frac{1}{\frac{1}{n}\sum_{i=1}^n \tilde{A}_i^2}\left(\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i Y_i - \pi(1-\pi)\theta(Q)\right) - \theta(Q)\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i^2 - \pi(1-\pi)\right)\right) \ .
\end{aligned}$$

Below we argue that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i^2 - \pi(1-\pi)\right) = (1-2\pi)\sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}} + o_P(1) \tag{A-53}$$

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i Y_i - \pi(1-\pi)\theta(Q)\right) = R_{n,1} + R_{n,3} + R_{n,4} + \theta(Q)(1-2\pi)\sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}} + o_P(1) \ , \tag{A-54}$$

where $R_{n,1}$ and $R_{n,3}$ are defined as in (B-73) and (B-75) in Lemma B.1, and

$$R_{n,4} \equiv \sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}}(1-2\pi)(E[m_1(Z)|S = s] - E[m_0(Z)|S = s]) \ . \tag{A-55}$$

Step 1: To see that (A-53) holds, let $A_i^* = A_i - \pi$ and note that

$$\begin{aligned}
\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i^2 - \pi(1-\pi)\right) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i A_i - \pi(1-\pi)\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(1 - \frac{n_1(S_i)}{n(S_i)}\right)(A_i - \pi) - \frac{\pi}{\sqrt{n}}\sum_{i=1}^n \left(\frac{n_1(S_i)}{n(S_i)} - \pi\right) \\
&= \sum_{s\in\mathcal{S}} \left(1 - \frac{n_1(s)}{n(s)}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^n A_i^* I\{S_i = s\} - \frac{\pi}{\sqrt{n}}\sum_{i=1}^n \left(\frac{D_n(S_i)}{n(S_i)}\right) \\
&= \sum_{s\in\mathcal{S}} \frac{D_n(s)}{\sqrt{n}}(1-\pi) - \frac{\pi}{\sqrt{n}}\sum_{s\in\mathcal{S}}\left(\frac{D_n(s)}{n(s)}\right)\sum_{i=1}^n I\{S_i = s\} + o_P(1) \\
&= (1-2\pi)\sum_{s\in\mathcal{S}} \frac{D_n(s)}{\sqrt{n}} + o_P(1)\ , \qquad \text{(A-56)}
\end{aligned}$$

where the first equality follows from $\sum_{i=1}^n \tilde{A}_i \frac{n_1(S_i)}{n(S_i)} = 0$, the third equality follows from $\frac{n_1(s)}{n(s)} = \frac{D_n(s)}{n(s)} + \pi$, and the fourth equality follows from $\frac{n_1(s)}{n(s)} \xrightarrow{P} \pi$ for all $s \in \mathcal{S}$, which in turn follows from the preliminary facts noted at the beginning of the proof of Thoerem 4.2.

Step 2: To see that (A-54) holds, note that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i Y_i - \pi(1-\pi)\theta(Q)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n A_i Y_i - \pi(1-\pi)\sqrt{n}\theta(Q) - \frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=1}^n \frac{n_1(s)}{n(s)} I\{S_i = s\} Y_i\ . \qquad \text{(A-57)}$$

Consider the first two terms. Use that $A_i = A_i^* + \pi$ and the definition of Y_i to get

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^n A_i Y_i - \pi(1-\pi)\sqrt{n}\theta(Q) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n A_i^* Y_i + \frac{\pi}{\sqrt{n}}\sum_{i=1}^n Y_i - \pi(1-\pi)\sqrt{n}\theta(Q) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n [(Y_i(1) - \mu_1)(1-\pi)A_i - (Y_i(0) - \mu_0)\pi(1 - A_i)] + \frac{\pi}{\sqrt{n}}\sum_{i=1}^n Y_i \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^n [A_i(1-\pi)\mu_1 - (1-A_i)\pi\mu_0] - \sqrt{n}\pi(1-\pi)(\mu_1 - \mu_0) \\
&= R_{n,1} + R_{n,2} + R_{n,3} + \frac{\pi}{\sqrt{n}}\sum_{i=1}^n Y_i + \frac{1}{\sqrt{n}}\sum_{i=1}^n A_i^*[(1-\pi)\mu_1 + \pi\mu_0]\ , \qquad \text{(A-58)}
\end{aligned}$$

where the last equality follows from the proof of Lemma B.1. Now consider the third term in (A-57). Again using

the fact that $\frac{n_1(s)}{n(s)} = \frac{D_n(s)}{n(s)} + \pi$, we see that

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n}\frac{n_1(s)}{n(s)}I\{S_i=s\}Y_i &= \frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n}\frac{D_n(s)}{n(s)}I\{S_i=s\}Y_i + \frac{\pi}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n}I\{S_i=s\}Y_i \\
&= \sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}}\frac{n}{n(s)}\frac{1}{n}\sum_{i=1}^{n}I\{S_i=s\}Y_i + \frac{\pi}{\sqrt{n}}\sum_{i=1}^{n}Y_i \\
&= \sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}}\frac{n}{n(s)}p(s)\left(\pi(\mu_1+E[m_1(Z)|S=s]) + (1-\pi)(\mu_0+E[m_0(Z)|S=s])\right) \\
&\quad + \frac{\pi}{\sqrt{n}}\sum_{i=1}^{n}Y_i + o_P(1) \\
&= R_{n,2} + \frac{\pi}{\sqrt{n}}\sum_{i=1}^{n}Y_i - \sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}}(1-2\pi)(E[m_1(Z)|S=s]-E[m_0(Z)|S=s]) \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i^*(\pi\mu_1+(1-\pi)\mu_0) + o_P(1)\ . \qquad \text{(A-59)}
\end{aligned}$$

In the preceeding display, the third equality follows from

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}I\{S_i=s\}Y_i &= \frac{1}{n}\sum_{i=1}^{n}I\{S_i=s\}(Y_i(1)A_i + Y_i(0)(1-A_i)) \\
&= \pi E[Y_i(1)I\{S_i=s\}] + (1-\pi)E[Y_i(0)I\{S_i=s\}] \\
&= p(s)\left(\pi(\mu_1+E[m_1(Z)|S=s]) + (1-\pi)(\mu_0+E[m_0(Z)|S=s])\right) + o_P(1)\ ,
\end{aligned}$$

which in turn follows from Lemma B.3, and the last equality follows from the fact that the weak law of large numbers implies that $\frac{n}{n(s)} = \frac{1}{p(s)} + o_P(1)$, Assumption 2.2.(b) implies that $\frac{D_n(s)}{\sqrt{n}} = O_P(1)$ for all $s\in\mathcal{S}$, and the definition of $R_{n,2}$ in (B-74). Combining (A-57)–(A-59) and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i^*((1-\pi)\mu_1+\pi\mu_0) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i^*(\pi\mu_1+(1-\pi)\mu_0) = \theta(Q)(1-2\pi)\sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}}\ ,$$

we conclude that

$$\sqrt{n}(\hat{\beta}_n - \theta(Q)) = \frac{1}{\pi(1-\pi)}(R_{n,1}+R_{n,3}+R_{n,4}) + o_P(1)\ .$$

The result (A-52) now follows from the continuous mapping theorem and Lemma B.2.

Next, we prove that

$$\hat{V}_{n,\beta} = \left[\left(\frac{\mathbb{C}_n'\mathbb{C}_n}{n}\right)^{-1}\left(\frac{\mathbb{C}_n'\,\mathrm{diag}\{\hat{u}_i^2 : 1\le i\le n\}\mathbb{C}_n}{n}\right)\left(\frac{\mathbb{C}_n'\mathbb{C}_n}{n}\right)^{-1}\right]_{[1,1]} \xrightarrow{P} \varsigma_{\tilde{Y}}^2(\pi) + \left(\frac{1}{\pi(1-\pi)}-3\right)\varsigma_H^2\ , \qquad \text{(A-60)}$$

where $\mathbb{C}_n$ is a $n\times|\mathcal{S}|+1$ matrix with the treatment assignment vector $\mathbb{A}_n$ in the first column and the strata indicators vector in the rest of the columns, and $\hat{u}_i$ is the least squares residual of the regression in (26). Note that $\frac{1}{n}\mathbb{C}_n'\mathbb{C}_n \xrightarrow{P} \Sigma_{\mathbb{C}}$, where

$$\Sigma_{\mathbb{C}} \equiv \begin{bmatrix} \pi & \pi p(1) & \pi p(2) & \cdots & \pi p(|\mathcal{S}|) \\ \pi p(1) & p(1) & 0 & \cdots & 0 \\ \pi p(2) & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \pi p(|\mathcal{S}|) & 0 & \cdots & \cdots & p(|\mathcal{S}|) \end{bmatrix}$$

and

$$\Sigma_{\mathbb{C}}^{-1} = \begin{bmatrix} \frac{1}{\pi(1-\pi)} & -\frac{1}{1-\pi} & -\frac{1}{1-\pi} & \cdots & -\frac{1}{1-\pi} \\ -\frac{1}{1-\pi} & \frac{\pi}{1-\pi} + \frac{1}{p(1)} & \frac{\pi}{1-\pi} & \cdots & \frac{\pi}{1-\pi} \\ -\frac{1}{1-\pi} & \frac{\pi}{1-\pi} & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ -\frac{1}{1-\pi} & \frac{\pi}{1-\pi} & \cdots & \cdots & \frac{\pi}{1-\pi} + \frac{1}{p(|\mathcal{S}|)} \end{bmatrix} . \tag{A-61}$$

The convergence in probability follows from $\frac{n_1}{n} = \frac{D_n}{n} + \pi \xrightarrow{P} \pi$, $\frac{n_1(s)}{n} = \frac{D_n(s)}{n} + \pi\frac{n(s)}{n} \xrightarrow{P} \pi p(s)$, and $\frac{n(s)}{n} \xrightarrow{P} p(s)$ for all $s \in \mathcal{S}$. The second result follows from analytically computing the inverse matrix, which we omit here. It follows that the $[1,1]$ component of $\Sigma_{\mathbb{C}}^{-1}$ equals $(\pi(1-\pi))^{-1}$. Note further that

$$\frac{\mathbb{C}_n' \operatorname{diag}\{\hat{u}_i^2 : 1 \le i \le n\}\mathbb{C}_n}{n} = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^n \hat{u}_i^2 A_i & \sum_{i=1}^n \hat{u}_i^2 A_i I\{S_i = 1\} & \cdots & \sum_{i=1}^n \hat{u}_i^2 A_i I\{S_i = |\mathcal{S}|\} \\ \sum_{i=1}^n \hat{u}_i^2 A_i I\{S_i = 1\} & \sum_{i=1}^n \hat{u}_i^2 I\{S_i = 1\} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n \hat{u}_i^2 A_i I\{S_i = |\mathcal{S}|\} & 0 & \cdots & \sum_{i=1}^n \hat{u}_i^2 I\{S_i = |\mathcal{S}|\} \end{bmatrix} .$$

It follows from Lemma B.9 that

$$\frac{\mathbb{C}_n' \operatorname{diag}\{\hat{u}_i^2 : 1 \le i \le n\}\mathbb{C}_n}{n} \xrightarrow{P} \Omega$$

where each component of the matrix Ω corresponds to the respective limits in Lemma B.9. It follows that

$$\left(\frac{\mathbb{C}_n'\mathbb{C}_n}{n}\right)^{-1}\left(\frac{\mathbb{C}_n' \operatorname{diag}\{\hat{u}_i^2 : 1 \le i \le n\}\mathbb{C}_n}{n}\right)\left(\frac{\mathbb{C}_n'\mathbb{C}_n}{n}\right)^{-1} \xrightarrow{P} \Sigma_{\mathbb{C}}^{-1}\Omega\Sigma_{\mathbb{C}}^{-1} = \begin{bmatrix} \varsigma_{\tilde{Y}}^2(\pi) + \left(\frac{1}{\pi(1-\pi)} - 3\right)\varsigma_H^2 & \mathbb{V}_{12} \\ \mathbb{V}_{12}' & \mathbb{V}_{22} \end{bmatrix} ,$$

where we omit the expressions of $\mathbb{V}_{12}$ and $\mathbb{V}_{22}$ as we do not need them for our arguments. From here, (A-60) follows immediately. Finally, note that Assumption 2.1 implies that $\varsigma_{\tilde{Y}}^2(\pi) > 0$.

To prove that $\varsigma_{\tilde{Y}}^2(\pi) + \varsigma_H^2 + \varsigma_\pi^2 \le \varsigma_{\tilde{Y}}^2(\pi) + \left(\frac{1}{\pi(1-\pi)} - 3\right)\varsigma_H^2$ holds with strict inequality unless (30) holds, note that

$$\begin{aligned} &\left(\varsigma_{\tilde{Y}}^2(\pi) + \left(\frac{1}{\pi(1-\pi)} - 3\right)\varsigma_H^2\right) - \left(\varsigma_{\tilde{Y}}^2(\pi) + \varsigma_H^2 + \varsigma_\pi^2\right) \\ &\quad = \left(\frac{1}{\pi(1-\pi)} - 4\right)\varsigma_H^2 - \varsigma_\pi^2 \\ &\quad = \sum_{s\in\mathcal{S}} p(s)\left(\left(\frac{1}{\pi(1-\pi)} - 4\right) - \tau(s)\frac{(1-2\pi)^2}{\pi^2(1-\pi)^2}\right)\left(E[m_1(Z_i)|S_i = s] - E[m_0(Z_i)|S_i = s]\right)^2 \\ &\quad = p(s)\frac{(1-2\pi)^2}{\pi^2(1-\pi)^2}(\pi(1-\pi) - \tau(s))\left(E[m_1(Z_i)|S_i = s] - E[m_0(Z_i)|S_i = s]\right)^2 . \end{aligned}$$

The right-hand side of this last display is non-negative and it is zero if and only if (30) holds, as required.

A.4 Proof of Theorem 4.4

As noted in the text, the proof of Theorem 4.3 establishes that (32) holds. Note further that Assumption 2.1 implies that $\varsigma_{\tilde{Y}}^2(\pi) > 0$. The proof of Theorem 4.2 establishes that $\hat{\varsigma}_{\tilde{Y}}^2(\pi) \xrightarrow{P} \varsigma_{\tilde{Y}}^2(\pi)$ and $\hat{\varsigma}_H^2 \xrightarrow{P} \varsigma_H^2$. To complete the proof, we argue that $\hat{\varsigma}_\pi^2 \xrightarrow{P} \varsigma_\pi^2$. To see this, note using (A-51) that

$$\varsigma_\pi^2 = \frac{(1-2\pi)^2}{\pi^2(1-\pi)^2}\sum_{s\in\mathcal{S}} p(s)\tau(s)\left(\left(E[Y_i(1)|S_i = s] - \mu_1\right) - \left(E[Y_i(0)|S_i = s] - \mu_0\right)\right)^2 .$$

Using the preliminary facts noted at the beginning of the proof of Theorem 4.2, the desired result follows.

A.5 Proof of Theorem 4.5

Below we assume without loss of generality that $\theta_0 = 0$; the general case follows from the same arguments with Y_i replaced by $Y_i - \theta_0 A_i$.

Let $G_n|S^{(n)}$ and $G'_n|S^{(n)} \sim \text{Unif}(\mathbf{G}_n(S^{(n)}))$ with G_n, G'_n, and $X^{(n)}$ independent conditional on $S^{(n)}$. Define

$$\varsigma^2_{\text{cap}} = \frac{\varsigma^2_Y(1-\pi) + \varsigma^2_H}{\varsigma^2_Y(1-\pi)} \ . \tag{A-62}$$

We first argue that for Q such that $\theta(Q) = 0$,

$$(T_n(G_n X^{(n)}), T_n(G'_n X^{(n)})) \xrightarrow{d} (T, T') \ , \tag{A-63}$$

where T and T' are independent with common c.d.f. $\Phi(t/\varsigma_{\text{cap}})$.

Step 1: Following Chung and Romano (2013), we start the proof of (A-63) by coupling the data $X^{(n)}$ with auxiliary "data" $\tilde{X}^{(n)} = \{(V_i, A_i, Z_i) : 1 \leq i \leq n\}$ constructed according to the following algorithm. Set $\mathcal{I} = \{1, \ldots, n\}$ and $K_n = 0$. For each $s \in \mathcal{S}$, repeat the following two stages $n(s)$ times:

1. *Stage 1*: Draw $C_j \in \{0, 1\}$ such that $P\{C_j = 1\} = \pi$.
2. *Stage 2*: If there exists $i \in \mathcal{I}$ such that $A_i = C_j$ and $S_i = s$, set $V_j = Y_i$ and set $\mathcal{I} = \mathcal{I} \setminus \{i\}$; otherwise, draw a new, independent observation from the distribution of $Y_i(C_j)|S_i = s$, set it equal to V_j and set $K_n = K_n + 1$.

Note that K_n constructed in this way is an upper bound on the number of elements in $\{V_i : 1 \leq i \leq n\}$ that are not identically equal to elements in $\{Y_i : 1 \leq i \leq n\}$. Indeed, there exists $g_0 \in \mathbf{G}_n(S^{(n)})$ such that

$$K_n \geq \sum_{i=1}^{n} I\{V_{g_0(i)} \neq Y_i\} \ .$$

Step 2: We now prove that for Q such that $\theta(Q) = 0$ and

$$T^U_n(X^{(n)}) \equiv \sqrt{n}\left(\frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1 - A_i)\right) \ , \tag{A-64}$$

it follows that

$$(T^U_n(G_n X^{(n)}), T^U_n(G'_n X^{(n)})) \xrightarrow{d} (T^U, T^{U\prime}) \ , \tag{A-65}$$

where T^U and $T^{U\prime}$ are independent with common distribution given by $N(0, \varsigma^2_Y(1-\pi) + \varsigma^2_H)$ when $\tau(s) = 0$ for all $s \in \mathcal{S}$. Arguing as in the proof of Lemma 5.1 in Chung and Romano (2013), (A-65) follows by verifying the following two conditions:

$$(T^U_n(G_n \tilde{X}^{(n)}), T^U_n(G'_n \tilde{X}^{(n)})) \quad \xrightarrow{d} \quad (T^U, T^{U\prime}) \tag{A-66}$$

$$T^U_n(G_n g_0 \tilde{X}^{(n)}) - T^U_n(G_n X^{(n)}) \quad = \quad o_P(1) \ . \tag{A-67}$$

Lemma B.4 establishes (A-66) and Lemma B.5 establishes (A-67).

Step 3: Note that we can write $T_n(G_n X^{(n)})$ as

$$T_n(G_n X^{(n)}) = \frac{T_n^U(G_n X^{(n)})}{T_n^L(G_n X^{(n)})} ,$$

where $T_n^L(X^{(n)}) = \sqrt{T_n^{L,1}(X^{(n)}) + T_n^{L,0}(X^{(n)})}$ with

$$\begin{aligned} T_n^{L,1}(X^{(n)}) &= \frac{n}{n_1}\frac{1}{n_1}\sum_{i=1}^n (Y_i - \bar{Y}_{n,1})^2 A_i \\ T_n^{L,0}(X^{(n)}) &= \frac{n}{n_0}\frac{1}{n_0}\sum_{i=1}^n (Y_i - \bar{Y}_{n,0})^2 (1-A_i) . \end{aligned}$$

It follows that

$$T_n^{L,1}(G_n X^{(n)}) = \frac{1}{\left(\frac{n_1}{n}\right)^2}\left(\frac{1}{n}\sum_{i=1}^n (Y_i(1)A_i + Y_i(0)(1-A_i))^2 A_{G_n(i)} - \frac{\left(\frac{1}{n}\sum_{i=1}^n (Y_i(1)A_i + Y_i(0)(1-A_i))A_{G_n(i)}\right)^2}{\frac{n_1}{n}}\right)$$

Assumption 2.2.(b) implies that $\frac{n_1}{n} = \frac{D_n}{n} + \pi \xrightarrow{P} \pi$. Lemma B.3 implies that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n (Y_i(1)A_i + Y_i(0)(1-A_i))^2 A_{G_n(i)} &= \frac{1}{n}\sum_{i=1}^n (Y_i(1)^2 A_i + Y_i(0)^2(1-A_i))A_{G_n(i)} \\ &\xrightarrow{P} \pi^2 E[Y_i(1)^2] + \pi(1-\pi)E[Y_i(0)^2] \qquad \text{(A-68)} \\ \frac{1}{n}\sum_{i=1}^n (Y_i(1)A_i + Y_i(0)(1-A_i))A_{G_n(i)} &\xrightarrow{P} \pi^2 E[Y_i(1)] + \pi(1-\pi)E[Y_i(0)] . \end{aligned}$$

Using the fact that $E[Y_i(1)] = E[Y_i(0)]$, it follows that

$$T_n^{L,1}(G_n X^{(n)}) \xrightarrow{P} \frac{1}{\pi}(\pi\sigma^2_{Y(1)} + (1-\pi)\sigma^2_{Y(0)}) .$$

A similar argument shows that $T_n^{L,0}(G_n X^{(n)}) \xrightarrow{P} \frac{1}{1-\pi}(\pi\sigma^2_{Y(1)} + (1-\pi)\sigma^2_{Y(0)})$. It thus follows that

$$T_n^L(G_n X^{(n)}) \xrightarrow{P} \varsigma_Y(1-\pi) . \qquad \text{(A-69)}$$

Finally, note that Assumption 2.1 implies that $\varsigma^2_Y(1-\pi) > 0$. Combining (A-65) and (A-69), we see that (A-63) holds.

It now follows from Lemma B.7 that $\hat{c}_n^{\text{cap}}(1-\alpha) \xrightarrow{P} \varsigma_{\text{cap}}\Phi^{-1}(1-\alpha)$. When $\tau(s) = 0$ for all $s \in \mathcal{S}$, it follows that $\varsigma_A^2 = 0$ and so $\varsigma^2_{\text{cap}} = \varsigma^2_{\text{t-test}}$ whenever $\pi = \frac{1}{2}$ or $\sigma^2_{Y(1)} = \sigma^2_{Y(0)}$ and $\sigma^2_{Y(1)} = \sigma^2_{Y(0)}$. Combining this last result with Theorem 4.1, we see that $\lim_{n\to\infty} E[\phi_n^{\text{cap}}(X^{(n)})] = \alpha$ whenever $\theta(Q) = \theta_0$, completing the proof of the theorem.

A.6 Proof of Theorem 4.6

The proof follows closely the proof of Theorem 4.5, we therefore omit some details. As in the proof of Theorem 4.5, we assume without loss of generality that $\theta_0 = 0$; the general case then follows from the same arguments with Y_i replaced by $Y_i - \theta_0 A_i$. Let G_n and G_n' be defined as in the proof of Theorem 4.5. We first argue that for Q such that $\theta(Q) = 0$,

$$(T_n(G_n X^{(n)}), T_n(G_n' X^{(n)})) \xrightarrow{d} (T, T') , \qquad \text{(A-70)}$$

where T and T' are independent with common c.d.f. $\Phi(t)$. Note that we can write $T_n(G_n X^{(n)})$ as

$$T_n(G_n X^{(n)}) = \frac{T_n^U(G_n X^{(n)})}{\tilde{T}_n^L(G_n X^{(n)})} ,$$

where $T_n^U(X^{(n)})$ is defined as in (A-64) and

$$\tilde{T}_n^L(X^{(n)}) = \sqrt{\hat{\varsigma}_{\dot{Y}}^2(\pi) + \hat{\varsigma}_H^2} .$$

Recall that $\hat{\varsigma}_{\dot{Y}}^2(\pi)$ and $\hat{\varsigma}_H^2$ are defined in (20) and (21). Note also that $\hat{\varsigma}_A(\pi)$ is not included in the definition of $\tilde{T}_n^L(X^{(n)})$ because it is assumed that $\tau(s) = 0$ for all $s \in \mathcal{S}$. The proof of Theorem 4.5 shows that (A-65) holds. We now argue that

$$\tilde{T}_n^L(G_n X^{(n)}) \xrightarrow{P} \sqrt{\varsigma_{\dot{Y}}^2(1-\pi) + \varsigma_H^2} . \tag{A-71}$$

To see this, first note the following consequences of Lemma B.3:

$$\begin{aligned}
\frac{1}{n_1} \sum_{1 \le i \le n} Y_i^2 A_{G_n(i)} &\xrightarrow{P} \pi E[Y_i(1)^2] + (1-\pi) E[Y_i(0)^2] \qquad \text{(A-72)} \\
\frac{1}{n_0} \sum_{1 \le i \le n} Y_i^2 (1 - A_{G_n(i)}) &\xrightarrow{P} \pi E[Y_i(1)^2] + (1-\pi) E[Y_i(0)^2] \\
\frac{1}{n_1} \sum_{1 \le i \le n} Y_i A_{G_n(i)} &\xrightarrow{P} \pi E[Y_i(1)] + (1-\pi) E[Y_i(0)] \\
\frac{1}{n_0} \sum_{1 \le i \le n} Y_i (1 - A_{G_n(i)}) &\xrightarrow{P} \pi E[Y_i(1)] + (1-\pi) E[Y_i(0)] \\
\frac{1}{n_1(s)} \sum_{1 \le i \le n} Y_i I\{S_i = s\} A_{G_n(i)} &\xrightarrow{P} \pi E[Y_i(1) | S_i = s] + (1-\pi) E[Y_i(0) | S_i = s] \\
\frac{1}{n_0(s)} \sum_{1 \le i \le n} Y_i I\{S_i = s\} (1 - A_{G_n(i)}) &\xrightarrow{P} \pi E[Y_i(1) | S_i = s] + (1-\pi) E[Y_i(0) | S_i = s] .
\end{aligned}$$

The convergence (A-72) follows from Assumption 2.2.(b), which implies that $\frac{n_1}{n} = \frac{D_n}{n} + \pi \xrightarrow{P} \pi$, and (A-68), established in the proof of Thoerem 4.5. The remaining convergences can be established in a similar fashion. Using these convergences, some calculation shows that (A-71) holds. Note further that Assumption 2.1 implies that $\varsigma_{\dot{Y}}^2(1-\pi) > 0$. Combining (A-65) and (A-71), we see that (A-70) holds. It thus follows from Lemma B.7 that $\hat{c}_n^{\text{cap}}(1-\alpha) \xrightarrow{P} \Phi^{-1}(1-\alpha)$. Combining this last result with Theorem 4.2, we see that $\lim_{n\to\infty} E[\phi_n^{\text{cap}}(X^{(n)})] = \alpha$ whenever $\theta(Q) = \theta_0$, completing the proof of the theorem.

Appendix B Auxiliary Results

Lemma B.1. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Then,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} ((Y_i(1) - \mu_1)(1-\pi) A_i - (Y_i(0) - \mu_0)\pi(1 - A_i)) \xrightarrow{d} \pi(1-\pi) N(0, \varsigma_{\dot{Y}}^2(\pi) + \varsigma_H^2 + \varsigma_A^2(\pi)) ,$$

where $\varsigma_{\dot{Y}}^2(\pi)$, ς_H^2, and $\varsigma_A^2(\pi)$ are defined in Table 7.

Proof. Let $\tilde{Y}_i(a) \equiv Y_i(a) - E[Y_i(a)|S_i]$, $m_a(Z_i) \equiv E[Y_i(a)|Z_i] - \mu_a$, and consider the following derivation:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left((Y_i(1)-\mu_1)(1-\pi)A_i - (Y_i(0)-\mu_0)\pi(1-A_i)\right) \\
&\quad= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\tilde{Y}_i(1)(1-\pi)A_i - \tilde{Y}_i(0)\pi(1-A_i)\right) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(E[m_1(Z_i)|S_i](1-\pi)A_i - E[m_0(Z_i)|S_i]\pi(1-A_i)\right) \\
&\quad= R_{n,1} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i^*\left(\sum_{s\in\mathcal{S}}(1-\pi)E[m_1(Z_i)|S_i=s]I\{S_i=s\} + \sum_{s\in\mathcal{S}}\pi E[m_0(Z_i)|S_i=s]I\{S_i=s\}\right) \\
&\qquad+ \frac{\pi(1-\pi)}{\sqrt{n}}\sum_{i=1}^{n}\left(\sum_{s\in\mathcal{S}}E[m_1(Z_i)|S_i=s]I\{S_i=s\} - \sum_{s\in\mathcal{S}}E[m_0(Z_i)|S_i=s]I\{S_i=s\}\right) \\
&\quad= R_{n,1} + R_{n,2} + R_{n,3}\ ,
\end{aligned}$$

where we used $A_i^* = A_i - \pi$ and the following definitions:

$$R_{n,1} \equiv \frac{\pi(1-\pi)}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{1}{\pi}\tilde{Y}_i(1)A_i - \frac{1}{1-\pi}\tilde{Y}_i(0)(1-A_i)\right) \tag{B-73}$$

$$R_{n,2} \equiv \pi(1-\pi)\sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}}\left(\frac{1}{\pi}E[m_1(Z_i)|S_i=s] + \frac{1}{1-\pi}E[m_0(Z_i)|S_i=s]\right) \tag{B-74}$$

$$R_{n,3} \equiv \pi(1-\pi)\sum_{s\in\mathcal{S}}\sqrt{n}\left(\frac{n(s)}{n} - p(s)\right)\left(E[m_1(Z_i)|S_i=s] - E[m_0(Z_i)|S_i=s]\right)\ . \tag{B-75}$$

The result now follows immediately from Lemma B.2 and the continuous mapping theorem. ■

Lemma B.2. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Let $R_{n,1}$, $R_{n,2}$, $R_{n,3}$, and $R_{n,4}$ be defined as in (B-73), (B-74), (B-75), and (A-55), respectively. Then,*

$$(R_{n,1}, R_{n,2}, R_{n,3}) \xrightarrow{d} (\zeta_{R_1}, \zeta_{R_2}, \zeta_{R_3})\ , \tag{B-76}$$

where ζ_{R_1}, ζ_{R_2}, and ζ_{R_3} are independent and satisfy $\zeta_{R_1} \sim \pi(1-\pi)N(0, \varsigma_Y^2(\pi))$, $\zeta_{R_2} \sim \pi(1-\pi)N(0, \varsigma_A^2(\pi))$, and $\zeta_{R_3} \sim \pi(1-\pi)N(0, \varsigma_H^2)$. Furthermore,

$$(R_{n,1}, R_{n,4}, R_{n,3}) \xrightarrow{d} (\zeta_{R_1}, \zeta_{R_4}, \zeta_{R_3})\ , \tag{B-77}$$

where ζ_{R_1}, ζ_{R_4}, and ζ_{R_3} are independent and satisfy $\zeta_{R_1} \sim \pi(1-\pi)N(0, \varsigma_Y^2(\pi))$, $\zeta_{R_4} \sim \pi(1-\pi)N(0, \varsigma_\pi^2$, and $\zeta_{R_3} \sim \pi(1-\pi)N(0, \varsigma_H^2)$.

Proof. We prove only (B-76); an analogous argument establishes (B-77). To this end, we first argue that

$$(R_{n,1}, R_{n,2}, R_{n,3}) \stackrel{d}{=} (R_{n,1}^*, R_{n,2}, R_{n,3}) + o_P(1) \tag{B-78}$$

for a random variable $R_{n,1}^*$ that satisfies $R_{n,1}^* \perp\!\!\!\perp (R_{n,2}, R_{n,3})$ and $R_{n,1}^* \xrightarrow{d} \zeta_{R_1}$. To this end, note that under the assumption that $W^{(n)}$ is i.i.d. and Assumption 2.2(a), the distribution of $R_{n,1}$ is the same as the distribution of the same quantity where units are ordered by strata and then ordered by $A_i = 1$ first and $A_i = 0$ second within strata. In order to exploit this observation, it is useful to introduce some further notation. Define $N(s) \equiv \sum_{i=1}^{n} I\{S_i < s\}$ and $F(s) \equiv P\{S_i < s\}$ for all $s \in \mathcal{S}$. Furthermore, independently for each $s \in \mathcal{S}$ and independently of $(A^{(n)}, S^{(n)})$, let $\{(\tilde{Y}_i^s(1), \tilde{Y}_i^s(0)) : 1 \le i \le n\}$ be i.i.d. with marginal distribution equal to the distribution of $(\tilde{Y}_i(1), \tilde{Y}_i(0))|S_i = s$.

With this notation, define

$$\tilde{R}_{n,1} \equiv \pi(1-\pi)\sum_{s\in\mathcal{S}}\left(\frac{1}{\sqrt{n}}\sum_{i=n\frac{N(s)}{n}+1}^{n\left(\frac{N(s)}{n}+\frac{n_1(s)}{n}\right)}\frac{1}{\pi}\tilde{Y}_i^s(1)-\frac{1}{\sqrt{n}}\sum_{i=n\left(\frac{N(s)}{n}+\frac{n_1(s)}{n}\right)+1}^{n\left(\frac{N(s)}{n}+\frac{n(s)}{n}\right)}\frac{1}{1-\pi}\tilde{Y}_i^s(0)\right) . \tag{B-79}$$

By construction, $\{R_{n,1}|S^{(n)},A^{(n)}\}\stackrel{d}{=}\{\tilde{R}_{n,1}|S^{(n)},A^{(n)}\}$, and so $R_{n,1}\stackrel{d}{=}\tilde{R}_{n,1}$. Since $R_{n,2}$ and $R_{n,3}$ are functions of $S^{(n)}$ and $A^{(n)}$, we have further that $(R_{n,1},R_{n,2},R_{n,3})\stackrel{d}{=}(\tilde{R}_{n,1},R_{n,2},R_{n,3})$. Next, define

$$R^*_{n,1} \equiv \pi(1-\pi)\sum_{s\in\mathcal{S}}\left(\frac{1}{\sqrt{n}}\sum_{i=\lfloor nF(s)\rfloor+1}^{\lfloor n(F(s)+\pi p(s))\rfloor}\frac{1}{\pi}\tilde{Y}_i^s(1)-\frac{1}{\sqrt{n}}\sum_{i=\lfloor n(F(s)+\pi p(s))\rfloor+1}^{\lfloor n(F(s)+p(s))\rfloor}\frac{1}{1-\pi}\tilde{Y}_i^s(0)\right) . \tag{B-80}$$

Since $R^*_{n,1}$ is a function of $\{(\tilde{Y}_i^s(1),\tilde{Y}_i^s(0)):1\le i\le n,s\in\mathcal{S}\}\perp\!\!\!\perp(S^{(n)},A^{(n)})$, and $(R_{n,2},R_{n,3})$ is a function of $(S^{(n)},A^{(n)})$, we see that $R^*_{n,1}\perp\!\!\!\perp(R_{n,2},R_{n,3})$.

To complete the proof of (B-78), we establish that $R^*_{n,1}\stackrel{d}{\to}\zeta_{R_1}$ and $\Delta_n\equiv\tilde{R}_{n,1}-R^*_{n,1}\stackrel{P}{\to}0$. To this end, consider an arbitrary $s\in\mathcal{S}$ and define the following partial sum process:

$$g_n(u) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nu\rfloor}\tilde{Y}_i^s(1) .$$

Under our assumptions, this converges weakly to a suitably scaled Brownian motion (see, e.g., Shorack and Wellner (2009, Theorem 1, page 53) or Durrett (2010, Theorem 8.6.5, page 328)). Indeed, by elementary properties of Brownian motion, we have that

$$\frac{1}{\sqrt{n}}\sum_{i=\lfloor nF(s)\rfloor+1}^{\lfloor n(F(s)+\pi p(s))\rfloor}\frac{1}{\pi}\tilde{Y}_i^s(1)\stackrel{d}{\to}N\left(0,\frac{p(s)\sigma^2_{\tilde{Y}(1)}(s)}{\pi}\right) ,$$

where we have used that $\sigma^2_{\tilde{Y}^s(1)}=\sigma^2_{\tilde{Y}(1)}(s)$. Furthermore, since

$$\left(\frac{N(s)}{n},\frac{n_1(s)}{n}\right)\stackrel{P}{\to}(F(s),\pi p(s)) ,$$

it follows that

$$g_n\left(\frac{N(s)+n_1(s)}{n}\right)-g_n\left(\frac{N(s)}{n}\right)-\left(g_n\left(F(s)+\pi p(s)\right)-g_n(F(s))\right)\stackrel{P}{\to}0 ,$$

where the convergence follows from elementary properties of Brownian motion and the continuous mapping theorem. Repeating an analogous argument for $\tilde{Y}_i^s(0)$ and using the independence of $\{(\tilde{Y}_i^s(1),\tilde{Y}_i^s(0)):1\le i\le n,s\in\mathcal{S}\}$ across both i and s, we conclude that $R^*_{n,1}\stackrel{d}{\to}\zeta_{R_1}$ and $\Delta_n\equiv\tilde{R}_{n,1}-R^*_{n,1}\stackrel{P}{\to}0$.

From Assumption 2.2.(b) and the continuous mapping theorem,

$$\{R_{n,2}|S^{(n)}\}\stackrel{d}{\to}\zeta_{R_2}\text{ a.s.} \tag{B-81}$$

Also, the central limit theorem and continuous mapping theorem imply that

$$R_{n,3}\stackrel{d}{\to}\zeta_{R_3} . \tag{B-82}$$

To complete the proof, we show that $(R_{n,1},R_{n,2},R_{n,3})\stackrel{d}{\to}(\zeta_{R_1},\zeta_{R_2},\zeta_{R_3})$ with $\{\zeta_{R_1},\zeta_{R_2},\zeta_{R_3}\}$ independent. From

(B-78), it suffices to show that $(R^*_{n,1}, R_{n,2}, R_{n,3}) \xrightarrow{d} (\zeta_{R_1}, \zeta_{R_2}, \zeta_{R_3})$, i.e.,

$$P\{R^*_{n,1} \leq h_1\}P\{R_{n,2} \leq h_2, R_{n,2} \leq h_3\} \to P\{\zeta_{R_1} \leq h_1\}P\{\zeta_{R_2} \leq h_2\}P\{\zeta_{R_3} \leq h_3\} \ , \tag{B-83}$$

for any $h = (h_1, h_2, h_3) \in \mathbf{R}^3$ s.t. $P\{\zeta_{R_1} \leq h_1\}P\{\zeta_{R_2} \leq h_2\}P\{\zeta_{R_3} \leq h_3\}$ is continuous.

As a first case, we assume that $P\{\zeta_{R_1} \leq \cdot\}$, $P\{\zeta_{R_2} \leq \cdot\}$, and $P\{\zeta_{R_3} \leq \cdot\}$ are continuous at h_1, h_2, h_3, respectively. Then, $R^*_{n,1} \xrightarrow{d} \zeta_{R_1}$ implies $P\{R^*_{n,1} \leq h_1\} \to P\{\zeta_{R_1} \leq h_1\}$ and (B-83) follows from the following argument:

$$\begin{aligned}
P\{R_{n,2} \leq h_2, R_{n,2} \leq h_3\} &= E[P\{R_{n,2} \leq h_2, R_{n,3} \leq h_3 | S^{(n)}\}] \\
&= E[P\{R_{n,2} \leq h_2 | S^{(n)}\} I\{R_{n,3} \leq h_3\}] \\
&= E[(P\{R_{n,2} \leq h_2 | S^{(n)}\} - P\{\zeta_{R_2} \leq h_2\}) I\{R_{n,3} \leq h_3\}] \\
&\quad + E[P\{\zeta_{R_2} \leq h_2\} I\{R_{n,3} \leq h_3\}] \\
&= E[(P\{R_{n,2} \leq h_2 | S^{(n)}\} - P\{\zeta_{R_2} \leq h_2\}) I\{R_{n,3} \leq h_3\}] + P\{\zeta_{R_2} \leq h_2\}P\{R_{n,3} \leq h_3\} \\
&\to P\{\zeta_{R_2} \leq h_2\}P\{\zeta_{R_3} \leq h_3\} \ ,
\end{aligned}$$

where the convergence follows from the dominated convergence theorem, (B-81), and (B-82).

Finally, we now consider the case in which $P\{\zeta_{R_j} \leq \cdot\}$ is discontinuous at h_j for some $1 \leq j \leq 3$. Since ζ_{R_j} is normally distributed, this implies that ζ_{R_j} must be degenerate and equal to zero and thus $h_j = 0$. In turn, since $P\{\zeta_{R_1} \leq \cdot\}P\{\zeta_{R_2} \leq \cdot\}P\{\zeta_{R_3} \leq \cdot\}$ is continuous at (h_1, h_2, h_3), then this implies that $\Pi_{k \neq j} P\{\zeta_{R_k} \leq h_k\} = 0$. Since ζ_{R_k} for $k \neq j$ are also normally distributed, this implies for some $k \neq j$ that ζ_{R_k} is also degenerate and equal to zero and $h_k < 0$. Then, (B-83) follows from the following argument:

$$\begin{aligned}
P\{R_{n,1} \leq h_1, R_{n,2} \leq h_2, R_{n,2} \leq h_3\} &\leq P\{R_{n,k} \leq h_k\} \\
&\to 0 \\
&= P\{\zeta_{R_1} \leq h_1\}P\{\zeta_{R_2} \leq h_2\}P\{\zeta_{R_3} \leq h_3\} \ ,
\end{aligned}$$

where the convergence follows from $h_k < 0$ and the fact that $R_{n,k} \xrightarrow{d} \zeta_{R_k}$ with ζ_{R_k} degenerate and equal to zero, and the final equality uses that $P\{\zeta_{R_k} \leq h_k\} = 0$. ■

Lemma B.3. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Let $W_i = f(Y_i(1), Y_i(0), S_i)$ for some function $f(\cdot)$ satisfy $E[|W_i|] < \infty$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} W_i A_i \xrightarrow{P} \pi E[W_i] \ . \tag{B-84}$$

Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n} W_i A_{G_n(i)} \xrightarrow{P} \pi E[W_i] \tag{B-85}$$

$$\frac{1}{n}\sum_{i=1}^{n} W_i A_i A_{G_n(i)} \xrightarrow{P} \pi^2 E[W_i] \tag{B-86}$$

for $G_n | S^{(n)} \sim$ Unif$(\mathbf{G}_n(S^{(n)}))$ with G_n and $X^{(n)}$ independent conditional on $S^{(n)}$.

Proof. We first prove (B-84). By arguing as in the proof of Lemma B.1, note that

$$\frac{1}{n}\sum_{i=1}^{n} W_i A_i \overset{d}{=} \sum_{s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n_1(s)} W_i^s \ ,$$

where, independently for each $s \in S$ and independently of $(A^{(n)}, S^{(n)})$, $\{W_i^s : 1 \leq i \leq n\}$ are i.i.d. with marginal distribution equal to the distribution of $W_i | S_i = s$. In order to establish the desired result, it suffices to show that

$$\frac{1}{n} \sum_{i=1}^{n_1(s)} W_i^s \xrightarrow{P} \pi p(s) E[W_i^s] \ . \tag{B-87}$$

From Assumption 2.2.(b), $\frac{n_1(s)}{n} = \frac{D_n(s)}{n} + \pi \frac{n(s)}{n} \xrightarrow{P} \pi p(s)$, so (B-87) follows from

$$\frac{1}{n_1(s)} \sum_{i=1}^{n_1(s)} W_i^s \xrightarrow{P} E[W_i^s] \ . \tag{B-88}$$

To establish (B-88), use the almost sure representation theorem to construct $\frac{\tilde{n}_1(s)}{n}$ such that $\frac{\tilde{n}_1(s)}{n} \stackrel{d}{=} \frac{n_1(s)}{n}$ and $\frac{\tilde{n}_1(s)}{n} \to \pi p(s)$ a.s. Using the independence of $(A^{(n)}, S^{(n)})$ and $\{W_i^s : 1 \leq i \leq n\}$, we see that for any $\epsilon > 0$,

$$\begin{aligned}
P\left\{ \left| \frac{1}{n_1(s)} \sum_{i=1}^{n_1(s)} W_i^s - E[W_i^s] \right| > \epsilon \right\} &= P\left\{ \left| \frac{1}{n\frac{n_1(s)}{n}} \sum_{i=1}^{n\frac{n_1(s)}{n}} W_i^s - E[W_i^s] \right| > \epsilon \right\} \\
&= P\left\{ \left| \frac{1}{n\frac{\tilde{n}_1(s)}{n}} \sum_{i=1}^{n\frac{\tilde{n}_1(s)}{n}} W_i^s - E[W_i^s] \right| > \epsilon \right\} \\
&= E\left[P\left\{ \left| \frac{1}{n\frac{\tilde{n}_1(s)}{n}} \sum_{i=1}^{n\frac{\tilde{n}_1(s)}{n}} W_i^s - E[W_i^s] \right| > \epsilon \middle| \frac{\tilde{n}_1(s)}{n} \right\} \right] \\
&\to 0 \ ,
\end{aligned}$$

where the convergence follows from the dominated convergence theorem and

$$P\left\{ \left| \frac{1}{n\frac{\tilde{n}_1(s)}{n}} \sum_{i=1}^{n\frac{\tilde{n}_1(s)}{n}} W_i^s - E[W_i^s] \right| > \epsilon \middle| \frac{\tilde{n}_1(s)}{n} \right\} \to 0 \text{ a.s. .} \tag{B-89}$$

To see that the convergence (B-89) holds, note that the weak law of large numbers implies that

$$\frac{1}{n_k} \sum_{i=1}^{n_k} W_i^s \xrightarrow{P} E[W_i^s] \tag{B-90}$$

for any subsequence $n_k \to \infty$ as $k \to \infty$. Since $n\frac{\tilde{n}_1(s)}{n} \to \infty$ a.s., (B-89) follows from the independence of $\frac{\tilde{n}_1(s)}{n}$ and $\{W_i^s : 1 \leq i \leq n\}$ and (B-90).

The convergence (B-85) follows from an analogous argument after conditioning on G_n. The convergence (B-86) follows in the same way using additionally the convergence (B-97) established in the proof of Lemma B.4. ■

Lemma B.4. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2 with $\tau(s) = 0$ for all $s \in \mathcal{S}$. Let G_n, G_n', and $\tilde{X}^{(n)}$ be defined as in the proof of Theorem 4.5. For $T_n^U(X^{(n)})$ defined in (A-64), we have that (A-66) holds whenever Q additionally satisfies $\theta(Q) = 0$.*

Proof. Let $g = G_n$ and $g' = G_n'$. Note that

$$T_n^U(g\tilde{X}^{(n)}) = \frac{1}{(D_n/n + \pi)(1 - \pi - D_n/n)} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i A^*_{g(i)} - \frac{D_n}{\sqrt{n}} \frac{1}{n} \sum_{i=1}^{n} V_i \right) \ ,$$

where $A_i^* = A_i - \pi$. Since Assumption 2.2.(b) holds with $\tau(s) = 0$ for all $s \in \mathcal{S}$, $\frac{D_n}{\sqrt{n}} \xrightarrow{P} 0$. From the weak law of large numbers, we have further that

$$\frac{1}{n}\sum_{i=1}^{n} V_i \xrightarrow{P} E[V_i] \ .$$

It thus follows that

$$T_n^U(g\tilde{X}^{(n)}) = \frac{1}{\pi(1-\pi)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i A^*_{g(i)} + o_P(1) \ .$$

Repeating the same argument for $T_n^U(g'\tilde{X}^{(n)})$, we see that

$$(T_n^U(g\tilde{X}^{(n)}), T_n^U(g'\tilde{X}^{(n)})) = \frac{1}{\pi(1-\pi)}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i A^*_{g(i)}, \frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i A^*_{g'(i)}\right) + o_P(1) \ .$$

Using the Cramér-Wold device, it suffices to show for real numbers a and b that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(aA^*_{g(i)} + bA^*_{g'(i)}) \xrightarrow{d} \pi(1-\pi)N(0, (a^2+b^2)(\varsigma^2_{\tilde{Y}}(1-\pi) + \varsigma^2_H)) \ . \tag{B-91}$$

Note that the left-hand side of (B-91) equals

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(V_i - E[V_i|S_i])(aA^*_{g(i)} + bA^*_{g'(i)}) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} E[V_i|S_i](aA^*_{g(i)} + bA^*_{g'(i)}) \ . \tag{B-92}$$

Because

$$\sum_{i=1}^{n} A^*_{g(i)} I\{S_i = s\} = \sum_{i=1}^{n} A^*_{g'(i)} I\{S_i = s\} = D_n(s) \ ,$$

the second term in (B-92) equals

$$\frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n} E[V_i|S_i = s]I\{S_i = s\}(aA^*_{g(i)} + bA^*_{g'(i)}) = (a+b)\sum_{s\in\mathcal{S}} E[V_i|S_i = s]\frac{D_n(s)}{\sqrt{n}} = o_P(1) \ ,$$

where in the last equality we again use the fact that $\frac{D_n(s)}{\sqrt{n}} \xrightarrow{P} 0$. To analyze the first term in (B-92), define

$$n_s(d, d') = |\{1 \le i \le n : A_{g(i)} \le d, A_{g'(i)} \le d', S_i = s\}| \ .$$

By arguing as in the proof of Lemma B.1, we see that this term is equal in distribution to the following:

$$\begin{aligned}\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\Bigg(&\sum_{i=1}^{n_s(0,0)} \tilde{V}_i^s(-\pi(a+b)) + \sum_{i=n_s(0,0)+1}^{n_s(0,1)} \tilde{V}_i^s((1-\pi)b - \pi a) \\ &+ \sum_{i=n_s(0,1)+1}^{n_s(1,0)} \tilde{V}_i^s((1-\pi)a - \pi b) + \sum_{i=n_s(1,0)+1}^{n_s(1,1)} \tilde{V}_i^s(1-\pi)(a+b)\Bigg) \ ,\end{aligned} \tag{B-93}$$

where, independently for each $s \in \mathcal{S}$ and independently of $(A^{(n)}, S^{(n)}, g, g')$, $\{\tilde{V}_i^s : 1 \le i \le n\}$ are i.i.d. with marginal

distribution equal to the distribution of $V_i - E[V_i|S_i]|S_i = s$. Next we argue that

$$\frac{n_s(0,0)}{n} \xrightarrow{P} (1-\pi)^2 p(s) \tag{B-94}$$

$$\frac{n_s(0,1) - n_s(0,0)}{n} \xrightarrow{P} \pi(1-\pi)p(s) \tag{B-95}$$

$$\frac{n_s(1,0) - n_s(0,1)}{n} \xrightarrow{P} \pi(1-\pi)p(s) \tag{B-96}$$

$$\frac{n_s(1,1) - n_s(1,0)}{n} \xrightarrow{P} \pi^2 p(s) \ . \tag{B-97}$$

We prove only (B-94); an analogous arugment establishes (B-95)–(B-97). Conditional on $A^{(n)}$ and $S^{(n)}$, $n_s(0,0)$ is a hypergeometric random variable corresponding to $n_0(s)$ draws from an urn with $n(s)$ balls and $n_0(s)$ successes. Hence,

$$\begin{aligned} E\left[\frac{n_s(0,0)}{n}\Big|A^{(n)}, S^{(n)}\right] &= \frac{n_0(s)^2}{n(s)n} \xrightarrow{P} (1-\pi)^2 p(s) \\ \mathrm{Var}\left[\frac{n_s(0,0)}{n}\Big|A^{(n)}, S^{(n)}\right] &= \frac{n_0(s)^2 n_1(s)^2}{n^2 n(s)^2 (n(s)-1)^2} \xrightarrow{P} 0 \ , \end{aligned} \tag{B-98}$$

where the convergences in probability follow, as before, using Assumption 2.2.(b). It therefore follows by Chebychev's inequality (applied conditionally) that

$$P\left\{\left|\frac{n_s(0,0)}{n} - \frac{n_0(s)^2}{n(s)n}\right| > \epsilon \Big| A^{(n)}, S^{(n)}\right\} \xrightarrow{P} 0 \ ,$$

which implies further that

$$\frac{n_s(0,0)}{n} - \frac{n_0(s)^2}{n(s)n} \xrightarrow{P} 0 \ .$$

The convergence (B-94) thus follows from (B-98).

Therefore, again by arguing as in the proof of Lemma B.1, we see that (B-93) converges in distribution to a normal with mean zero and variance given by

$$\begin{aligned} &\sum_{s\in\mathcal{S}} \pi^2(1-\pi)^2 \left\{2\pi(1-\pi)(a+b)^2 + ((1-\pi)b - \pi a)^2 + ((1-\pi)a - \pi b)^2\right\} \frac{p(s)\,\mathrm{Var}[V_i|S_i=s]}{\pi(1-\pi)} \\ &\quad = \pi^2(1-\pi)^2(a^2+b^2) \sum_{s\in\mathcal{S}} \frac{p(s)\mathrm{Var}[V_i|S_i=s]}{\pi(1-\pi)} \ . \end{aligned} \tag{B-99}$$

To complete the proof, note that

$$\begin{aligned} \mathrm{Var}[V_i|S_i=s] &= \pi\mathrm{Var}[Y_i(1)|S_i=s] + (1-\pi)\mathrm{Var}[Y_i(0)|S_i=s] \\ &\quad + \left(\pi E[Y_i(1)|S_i=s]^2 + (1-\pi)E[Y_i(0)|S_i=s]^2\right) \\ &\quad - \left(\pi E[Y_i(1)|S_i=s] + (1-\pi)E[Y_i(0)|S_i=s]\right)^2 \\ &= \pi\sigma^2_{Y(1)}(s) + (1-\pi)\sigma^2_{Y(0)}(s) + \frac{1}{4}\left(E[Y_i(1)|S_i=s] - E[Y_i(0)|S_i=s]\right)^2 \\ &= \pi\sigma^2_{Y(1)}(s) + (1-\pi)\sigma^2_{Y(0)}(s) + \pi(1-\pi)\left(E[m_1(Z_i)|S_i=s] - E[m_0(Z_i)|S_i=s]\right)^2 \ , \end{aligned}$$

where in the final equality we have used the fact that $\mu_1 = \mu_0$ because $\theta(Q) = 0$. It thus follows from the expressions for $\varsigma^2_Y(\pi)$ and ς^2_H in Table 7 that (B-99) equals $\pi^2(1-\pi)^2(a^2+b^2)(\varsigma^2_Y(1-\pi) + \varsigma^2_H)$, and (B-91) follows. ■

Lemma B.5. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Let G_n, g_0 and $\tilde{X}^{(n)}$ be defined as in the proof of Theorem 4.5. Let $T_n^U(X^{(n)})$ be defined as in (A-64). Then*

(A-67) *holds whenever* Q *additionally satisfies* $\theta(Q) = 0$.

Proof. Let $g = G_n$. Note that (A-67) equals

$$\sqrt{n}\left(\frac{1}{n_1}\sum_{i=1}^{n}(V_{gg_0(i)} - Y_{g(i)})A_i - \frac{1}{n_0}\sum_{i=1}^{n}(V_{gg_0(i)} - Y_{g(i)})(1 - A_i)\right) . \tag{B-100}$$

Since

$$Y_{g(i)} = Y_{g(i)}(1)A_{g(i)} + Y_{g(i)}(0)(1 - A_{g(i)}) ,$$

we have that (B-100) equals

$$\begin{aligned}\sqrt{n}\Bigg(&\frac{1}{n_1}\sum_{i=1}^{n}(V_{gg_0(i)} - Y_{g(i)}(1))A_{g(i)}A_i + \frac{1}{n_1}\sum_{i=1}^{n}(V_{gg_0(i)} - Y_{g(i)}(0))(1 - A_{g(i)})A_i \\ &-\frac{1}{n_0}\sum_{i=1}^{n}(V_{gg_0(i)} - Y_{g(i)}(1))A_{g(i)}(1 - A_i) - \frac{1}{n_0}\sum_{i=1}^{n}(V_{gg_0(i)} - Y_{g(i)}(0))(1 - A_{g(i)})(1 - A_i)\Bigg) .\end{aligned} \tag{B-101}$$

By construction, all but at most K_n of the terms in the four summations in (B-101) must be identically equal to zero. Moreover, conditionally on g, g_0, $A^{(n)}$ and K_n, (B-101) has mean equal to zero. This follows from the fact that $E[V_i] = \pi\mu_1 + (1 - \pi)\mu_0$ and $\mu_1 - \mu_0 = 0$ because $\theta(Q) = 0$. Using the fact that $\text{Var}[A + B] \leq 2(\text{Var}[A] + \text{Var}[B])$, we see that, conditionally on g, g_0, $A^{(n)}$ and K_n, (B-101) has variance bounded above by

$$\frac{K_n}{n}\left(\left(\frac{n}{n_1}\right)^2 + \left(\frac{n}{n_0}\right)^2\right)M ,$$

where

$$M = 4\max\{\text{Var}[V_i - Y_i(1)], \text{Var}[V_i - Y_i(0)]\} . \tag{B-102}$$

Lemma B.6 implies that $\frac{K_n}{n} \xrightarrow{P} 0$. It therefore follows by Chebychev's inequality (applied conditionally) that

$$P\{|T_n^U(gg_0\tilde{X}_n) - T_n^U(gX^{(n)})| > \epsilon | g, g_0, A^{(n)}, K_n\} \xrightarrow{P} 0 ,$$

from which the desired unconditional convergence (A-67) also follows. ■

Lemma B.6. *Suppose* Q *satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Let* K_n *be defined as in the proof of Theorem 4.5. Then,*

$$\frac{K_n}{n} \xrightarrow{P} 0 . \tag{B-103}$$

Proof. The argument provided here follows closely arguments in Chung and Romano (2013). For each s, let $N_a(s) = |\{C_j = a : j = 1, \ldots, n(s)\}|$ for C_j as in the proof of Theorem 4.5. In this notation,

$$K_n = \sum_{s\in\mathcal{S}}\sum_{a\in\{0,1\}}\max\{n_a(s) - N_a(s), 0\} .$$

In order to show (B-103), it suffices to show that for all $s \in \mathcal{S}$ and $a \in \{0, 1\}$,

$$\frac{n_a(s) - N_a(s)}{n} \xrightarrow{P} 0 . \tag{B-104}$$

We prove (B-104) for the case $a = 1$; an analogous argument establishes it for the case $a = 0$. To this end, write

$$\begin{aligned}\frac{n_1(s) - N_1(s)}{n} &= \left(\frac{n_1(s)}{n} - \pi\frac{n(s)}{n}\right) - \left(\frac{N_1(s)}{n} - \pi\frac{n(s)}{n}\right) \\ &= \frac{D_n(s)}{n} - \frac{n(s)}{n}\left(\frac{N_1(s)}{n(s)} - \pi\right) .\end{aligned}$$

Under our assumptions, $\frac{D_n(s)}{n} \xrightarrow{P} 0$ and $\frac{n(s)}{n} \xrightarrow{P} p(s)$. Note that for each $s \in \mathcal{S}$, $N_1(s)|n(s)$ is distributed according to a binomial distribution with $n(s)$ trials and probability of success equal to π. It therefore follows from Chebychev's inequality (applied conditionally) that

$$P\left\{\left|\frac{N_1(s)}{n(s)} - \pi\right| > \epsilon \middle| S^{(n)}\right\} \leq \frac{1}{n\frac{n(s)}{n}\epsilon^2} \xrightarrow{P} 0 .$$

Hence,

$$\frac{N_1(s)}{n(s)} - \pi \xrightarrow{P} 0 .$$

The desired result now follows. ■

Lemma B.7. *Let $G_n|S^{(n)}$ and $G'_n|S^{(n)} \sim$ Unif$(\mathbf{G}_n(S^{(n)}))$ with G_n, G'_n, and $X^{(n)}$ independent conditional on $S^{(n)}$. Suppose*

$$(T_n(G_n X^{(n)}), T_n(G'_n X^{(n)})) \xrightarrow{d} (T, T') , \tag{B-105}$$

where $T \sim T'$ with c.d.f. $R(t)$ and T and T' are independent. Then, for any continuity point t of R,

$$\hat{R}_n(t) = \frac{1}{|\mathbf{G}_n(S^{(n)})|} \sum_{g \in \mathbf{G}_n(S^{(n)})} I\{T_n(gX^{(n)}) \leq t\} \xrightarrow{P} R(t) .$$

Proof. Let t be a continuity point of R. Note that

$$\begin{aligned}E[\hat{R}_n(t)] &= E[E[\hat{R}_n(t)|S^{(n)}]] \\ &= E\left[\frac{1}{|\mathbf{G}_n(S^{(n)})|} \sum_{g \in \mathbf{G}_n(S^{(n)})} P\{T_n(gX^{(n)}) \leq t|S^{(n)}\}\right] \\ &= E[P\{T_n(G_n X^{(n)}) \leq t|S^{(n)}\}] \\ &= P\{T_n(G_n X^{(n)}) \leq t\} \\ &\to R(t) ,\end{aligned}$$

where the third equality follows from the distribution for $G_n|S^{(n)}$ and the independence of G_n and $X^{(n)}$ conditional on $S^{(n)}$ and the convergence follows from (B-105). It therefore suffices to show that $\text{Var}[\hat{R}_n(t)] \to 0$. Equivalently, it is enough to show that $E[\hat{R}_n^2(t)] \to R^2(t)$. To this end, note that

$$\begin{aligned}E[\hat{R}_n^2(t)] &= E\left[\frac{1}{|\mathbf{G}_n(S^{(n)})|^2} \sum_{g \in \mathbf{G}_n(S^{(n)})} \sum_{g' \in \mathbf{G}_n(S^{(n)})} I\{T_n(gX^{(n)}) \leq t, T_n(g'X^{(n)}) \leq t\}\right] \\ &= E\left[\frac{1}{|\mathbf{G}_n(S^{(n)})|^2} \sum_{g \in \mathbf{G}_n(S^{(n)})} \sum_{g' \in \mathbf{G}_n(S^{(n)})} P\{T_n(gX^{(n)}) \leq t, T_n(g'X^{(n)}) \leq t|S^{(n)}\}\right] \\ &= E[P\{T_n(G_n X^{(n)}) \leq t, T_n(G'_n X^{(n)}) \leq t|S^{(n)}\}] \\ &= P\{T_n(G_n X^{(n)}) \leq t, T_n(G'_n X^{(n)}) \leq t\} \\ &\to R^2(t) ,\end{aligned}$$

where, as before, the third equality follows from the distributions for $G_n|S^{(n)}$ and $G_n'|S^{(n)}$ and the independence of G_n, G_n', and $X^{(n)}$ conditional on $S^{(n)}$, and the convergence follows from (B-105). ■

Lemma B.8. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Let $\gamma = (\beta, \delta_1, \ldots, \delta_{|\mathcal{S}|})'$ be the parameters in the regression (26) and let $\hat{\gamma}_n$ be the least squares estimator of γ. Then,*

$$\hat{\gamma}_n \xrightarrow{P} \gamma \equiv \begin{bmatrix} \theta(Q) \\ \mu_0 + \pi E[m_1(Z_i)|S_i = 1] + (1-\pi)E[m_0(Z_i)|S_i = 1] \\ \vdots \\ \mu_0 + \pi E[m_1(Z_i)|S_i = |\mathcal{S}|] + (1-\pi)E[m_0(Z_i)|S_i = |\mathcal{S}|] \end{bmatrix} .$$

Proof. First note that $\hat{\gamma}_n = (\mathbb{C}_n'\mathbb{C}_n)^{-1}\mathbb{C}_n'\mathbb{Y}_n$, where $\mathbb{C}_n$ is an $n \times (|\mathcal{S}|+1)$ matrix with the treatment assignment vector $\mathbb{A}_n$ in the first row and the strata indicators vector in the rest of the rows, and $\mathbb{Y}_n$ is an $n \times 1$ vector of outcomes. The $(s+1)$th element of $\frac{1}{n}\mathbb{C}_n'\mathbb{Y}_n$ equals $\frac{1}{n}\sum_{i=1}^n A_i Y_i$ if $s = 0$ and $\frac{1}{n}\sum_{i=1}^n I\{S_i = s\}Y_i$ for $s \in \mathcal{S}$. In turn, this last term satisfies

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n I\{S_i = s\}Y_i &= \frac{n_1(s)}{n}\left(\mu_1 + E[m_1(Z_i)|S_i = s]\right) + \left(\frac{n(s)}{n} - \frac{n_1(s)}{n}\right)\left(\mu_0 + E[m_0(Z_i)|S_i = s]\right) \\ &\quad + \frac{1}{n}\sum_{i=1}^n A_i I\{S_i = s\}\tilde{Y}_i(1) + \frac{1}{n}\sum_{i=1}^n (1 - A_i)\, I\{S_i = s\}\tilde{Y}_i(0) \\ &= p(s)\left(\pi(\mu_1 + E[m_1(Z_i)|S_i = s]) + (1-\pi)(\mu_0 + E[m_0(Z_i)|S_i = s])\right) + o_P(1) , \end{aligned}$$

where in the last step we used $\frac{n_1(s)}{n} = \frac{D_n(s)}{n} + \pi\frac{n(s)}{n} \xrightarrow{P} \pi p(s)$, $\frac{n(s)}{n} \xrightarrow{P} p(s)$, and that $\frac{1}{n}\sum_{i=1}^n A_i I\{S_i = s\}\tilde{Y}_i(a) \xrightarrow{P} 0$ for $a \in \{0, 1\}$ by Lemma B.3. Also by Lemma B.3,

$$\frac{1}{n}\sum_{i=1}^n A_i Y_i = \frac{1}{n}\sum_{i=1}^n A_i Y_i(1) \xrightarrow{P} \pi\mu_1 ,$$

so that we conclude that

$$\frac{1}{n}\mathbb{C}_n'\mathbb{Y}_n \xrightarrow{P} \begin{bmatrix} \pi\mu_1 \\ p(1)\left(\pi(\mu_1 + E[m_1(Z_i)|S_i = 1]) + (1-\pi)(\mu_0 + E[m_0(Z_i)|S_i = 1])\right) \\ \vdots \\ p(|\mathcal{S}|)\left(\pi(\mu_1 + E[m_1(Z_i)|S_i = 1]) + (1-\pi)(\mu_0 + E[m_0(Z_i)|S_i = |\mathcal{S}|])\right) \end{bmatrix} .$$

The result then follows from the above display, (A-61), and some additional algebra. ■

Lemma B.9. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Let $C_i \equiv [A_i, I\{S_i = 1\}, \ldots, I\{S_i = |\mathcal{S}|\}]'$ be the ith row of the matrix $\mathbb{C}_n$ formed by stacking the treatment assignment vector $\mathbb{A}_n$ in the first column and the strata indicators vector in the rest of the columns, $\hat{u}_i$ be the least squares residuals of the regression in (26), and $\hat{\gamma}_n$ be the least squares estimator of the regression coefficients $\gamma = (\beta, \delta_1, \ldots, \delta_{|\mathcal{S}|})'$. Then,*

$$\hat{u}_i = \sum_{s \in \mathcal{S}} I\{S_i = s\}A_i^* E\left[m_1(Z_i) - m_0(Z_i)|S_i = s\right] + \tilde{Y}_i(1)A_i + \tilde{Y}_i(0)(1 - A_i) + C_i(\gamma - \hat{\gamma}_n) . \tag{B-106}$$

Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 \xrightarrow{P} \pi(1-\pi)(\varsigma_H^2 + \varsigma_{\tilde{Y}}^2(1-\pi)) \tag{B-107}$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 A_i \xrightarrow{P} \pi(1-\pi)^2\varsigma_H^2 + \pi\sigma_{\tilde{Y}(1)}^2 \tag{B-108}$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 I\{S_i = s\} \xrightarrow{P} p(s)\left(\pi(1-\pi)\left(E\left[m_1(Z_i) - m_0(Z_i)|S_i = s\right]\right)^2 + \pi\sigma_{\tilde{Y}(1)}^2(s) + (1-\pi)\sigma_{\tilde{Y}(0)}^2(s)\right) \tag{B-109}$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 I\{S_i = s\} A_i \xrightarrow{P} p(s)\left(\pi(1-\pi)^2\left(E\left[m_1(Z_i) - m_0(Z_i)|S_i = s\right]\right)^2 + \pi\sigma_{\tilde{Y}(1)}^2(s)\right) . \tag{B-110}$$

Proof. Consider the following derivation:

$$\begin{aligned} Y_i &= Y_i(1)A_i + Y_i(0)(1-A_i) \\ &= \theta(Q)A_i + \mu_0 + \tilde{Y}_i(1)A_i + \tilde{Y}_i(0)(1-A_i) \\ &\quad + \sum_{s\in\mathcal{S}} I\{S_i = s\}\left(A_i E[m_1(Z_i)|S_i = s] + (1-A_i)E[m_0(Z_i)|S_i = s]\right) . \end{aligned}$$

Using Lemma B.8, we have that

$$C_i\gamma = \theta(Q)A_i + \mu_0 + \sum_{s\in\mathcal{S}} I\{S_i = s\}\left(\pi E[m_1(Z_i)|S_i = s] + (1-\pi)E[m_0(Z_i)|S_i = s]\right) .$$

Hence,

$$\begin{aligned} u_i &= Y_i - C_i\gamma \\ &= \sum_{s\in\mathcal{S}} I\{S_i = s\}A_i^* E\left[m_1(Z_i) - m_0(Z_i)|S_i = s\right] + \tilde{Y}_i(1)A_i + \tilde{Y}_i(0)\left(1-A_i\right) . \end{aligned} \tag{B-111}$$

Since $\hat{u}_i = u_i + C_i(\gamma - \hat{\gamma}_n)$, the desired expression for (B-106) follows.

To prove (B-107)–(B-110), note that for any univariate random variable X_i such that

$$\frac{1}{n}\sum_{i=1}^{n}\left(C_i'C_i \otimes X_i\right) = O_P(1) \text{ and } \frac{1}{n}\sum_{i=1}^{n} C_i u_i X_i = O_P(1) , \tag{B-112}$$

we have that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 X_i &= \frac{1}{n}\sum_{i=1}^{n} u_i^2 X_i + (\gamma - \hat{\gamma}_n)'\frac{1}{n}\sum_{i=1}^{n}\left(C_i'C_i \otimes X_i\right)(\gamma - \hat{\gamma}_n) + 2(\gamma - \hat{\gamma}_n)\frac{1}{n}\sum_{i=1}^{n} C_i u_i X_i \\ &= \frac{1}{n}\sum_{i=1}^{n} u_i^2 X_i + o_P(1) , \end{aligned}$$

where in the second equality we used $\hat{\gamma}_n \xrightarrow{P} \gamma$ from Lemma B.8. Since the condition in (B-112) certainly holds for $X_i = 1$, $I\{S_i = s\}$, A_i, and $I\{S_i = s\}A_i$, it is enough to show that (B-107)–(B-110) holds with u_i^2 in place of $\hat{u}_i^2$.

Using (B-111), we have that

$$
\begin{aligned}
u_i^2 = &\sum_{s\in\mathcal{S}} (A_i^*)^2 I\{S_i = s\} \left(E\left[m_1(Z_i) - m_0(Z_i)|S_i = s\right]\right)^2 + \tilde{Y}_i(1)^2 A_i + \tilde{Y}_i(0)^2 (1 - A_i) \\
&+ 2\sum_{s\in\mathcal{S}} A_i^* I\{S_i = s\} E\left[m_1(Z_i) - m_0(Z_i)|S_i = s\right] \tilde{Y}_i(1) A_i \\
&+ 2\sum_{s\in\mathcal{S}} A_i^* I\{S_i = s\} E\left[m_1(Z_i) - m_0(Z_i)|S_i = s\right] \tilde{Y}_i(0)(1 - A_i) \ .
\end{aligned}
$$

Lemma B.3 thus implies that

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n u_i^2 &\xrightarrow{P} \pi(1-\pi)\sum_{s\in\mathcal{S}} p(s) \left(E\left[m_1(Z_i) - m_0(Z_i)|S_i = s\right]\right)^2 + \pi\sigma^2_{\tilde{Y}(1)} + (1-\pi)\frac{1}{2}\sigma^2_{\tilde{Y}(0)} \\
&= \pi(1-\pi)(\varsigma_H^2 + \varsigma^2_{\tilde{Y}}(1-\pi)) \\
\frac{1}{n}\sum_{i=1}^n u_i^2 A_i &\xrightarrow{P} \pi(1-\pi)^2\sum_{s\in\mathcal{S}} p(s) \left(E\left[m_1(Z_i) - m_0(Z_i)|S_i = s\right]\right)^2 + \pi\sigma^2_{\tilde{Y}(1)} \\
&= \pi(1-\pi)^2\varsigma_H^2 + \pi\sigma^2_{\tilde{Y}(1)} \\
\frac{1}{n}\sum_{i=1}^n u_i^2 I\{S_i = s\} &\xrightarrow{P} \pi(1-\pi)p(s)\left(\left(E\left[m_1(Z_i) - m_0(Z_i)|S_i = s\right]\right)^2 + \frac{1}{1-\pi}\sigma^2_{\tilde{Y}(1)}(s) + \frac{1}{\pi}\sigma^2_{\tilde{Y}(0)}(s)\right) \\
\frac{1}{n}\sum_{i=1}^n u_i^2 I\{S_i = s\} A_i &\xrightarrow{P} p(s)\left(\pi(1-\pi)^2\left(E\left[m_1(Z_i) - m_0(Z_i)|S_i = s\right]\right)^2 + \pi\sigma^2_{\tilde{Y}(1)}(s)\right) \ ,
\end{aligned}
$$

thus completing the proof. ■

Lemma B.10. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Then,*

$$
\hat{V}^{homo}_{n,\beta} = \left(\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2\right)\left(\frac{\mathbb{C}_n'\mathbb{C}_n}{n}\right)^{-1}_{[1,1]} \xrightarrow{P} \varsigma^2_{\tilde{Y}}(1-\pi) + \varsigma_H^2 \ , \tag{B-113}
$$

where $\mathbb{C}_n$ and $\hat{u}_n$ are defined as in the proof of Theorem 4.3.

Proof. By arguing as in the proof of Theorem 4.3, we see that

$$
\left(\frac{\mathbb{C}_n'\mathbb{C}_n}{n}\right)^{-1}_{[1,1]} \xrightarrow{P} [\pi(1-\pi)]^{-1} \ .
$$

Lemma B.9 shows that $\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2 \xrightarrow{P} \pi(1-\pi)\left(\varsigma^2_{\tilde{Y}}(1-\pi) + \varsigma_H^2\right)$. The desired result thus follows. ■

Lemma B.11. *Let $A^{(n)}$ be a treatment assignment generated by the biased coin design mechanism described in Example 3.2. Then, Assumption 2.2 holds.*

Proof. Part (a) holds by definition. For part (b), note that for $s \neq s'$, $D_n(s) \perp\!\!\!\perp D_n(s')|S^{(n)}$. Moreover, within stratum the assignment is exactly the one considered in Markaryan and Rosenberger (2010, Theorem 2.1). It follows from that result that

$$
\{D_n(s)|S^{(n)}\} = O_P(1) \text{ a.s.}
$$

These two properties imply that part (b) holds with $\tau(s) = 0$ for all $s \in \mathcal{S}$. ■

Lemma B.12. *Let $A^{(n)}$ be a treatment assignment generated by the adaptive biased coin design mechanism described in Example 3.3. Then, Assumption 2.2 holds.*

Proof. Part (a) holds by definition. Part (b) holds by Wei (1978, Theorem 3) adapted to account for stratification. This result implies that

$$\left\{ \frac{D_n(s)}{\sqrt{n}} | S^{(n)} \right\} \xrightarrow{d} N\left(0, \frac{1}{4(1-4\varphi'(0))}\right) \text{ a.s.} \tag{B-114}$$

Since $D_n(s) \perp\!\!\!\perp D_n(s') | S^{(n)}$ for $s \neq s'$, part (b) holds with $\tau(s) = \frac{1}{4(1-4\varphi'(0))}$ for all $s \in \mathcal{S}$. ■

Lemma B.13. *Let $A^{(n)}$ be a treatment assignment generated by the stratified block randomization mechanism described in Example 3.4. Then, Assumption 2.2 holds.*

Proof. Part (a) follows by definition. Next note that, conditional on $S^{(n)}$, $n(s) = \sum_{i=1}^{n} I\{S_i = s\}$ and $n_1(s) = \lfloor \frac{n(s)}{2} \rfloor$ are non-random. Thus, conditional on $S^{(n)}$, $\{D_n(s) : s \in \mathcal{S}\}$ is non-random with

$$D_n(s) = \begin{cases} 0 & \text{if } n(s) \text{ is even or } n(s) = 0 \\ -1 & \text{if } n(s) \text{ is odd} \end{cases}$$

for all $s \in \mathcal{S}$. Part (b) then follows with $\tau(s) = 0$ for all $s \in \mathcal{S}$. ■

References

Berry, J., Karlan, D. S. and Pradhan, M. (2015). The impact of financial education for youth in Ghana. *Working paper*.

Bruhn, M. and McKenzie, D. (2008). In pursuit of balance: Randomization in practice in development field experiments. *World Bank Policy Research Working Paper*, **4752**.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2016). Inference under covariate adaptive randomization with multiple treatments. In progress.

Callen, M., Gulzar, S., Hasanain, A., Khan, Y. and Rezaee, A. (2015). Personalities and public sector performance: Evidence from a health experiment in Pakistan. Tech. rep., Working paper.

Chong, A., Cohen, I., Field, E., Nakasone, E. and Torero, M. (2016). Iron deficiency and schooling attainment in peru. *American Economic Journal: Applied Economics*, **8** 222–255.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, **41** 484–507.

Dizon-Ross, R. (2014). Parents' perceptions and children's education: Experimental evidence from Malawi. Manuscript, M.I.T.

Duflo, E., Dupas, P. and Kremer, M. (2014). Education, HIV, and early fertility: Experimental evidence from kenya. Tech. rep., National Bureau of Economic Research.

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

Durrett, R. (2010). *Probability: theory and examples*. 4th ed. Cambridge university press.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, **58** 403–417.

Heckman, J. J., Pinto, R., Shaikh, A. M. and Yavitz, A. (2011). Inference with imperfect randomization: The case of the Perry Preschool. Manuscript.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, **23** pp. 169–192. URL http://www.jstor.org/stable/2236445.

Hu, Y. and Hu, F. (2012). Asymptotic properties of covariate-adaptive randomization. *Annals of Statistics, forthcoming*.

Imbens, G. W. and Kolesar, M. (2012). Robust standard errors in small samples: some practical advice. Tech. rep., National Bureau of Economic Research.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Kernan, W. N., Viscoli, C. M., Makuch, R. W., Brass, L. M. and Horwitz, R. I. (1999). Stratified randomization for clinical trials. *Journal of clinical epidemiology*, **52** 19–26.

Lee, S. and Shaikh, A. M. (2014). Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of progresa on school enrollment. *Journal of Applied Econometrics*, **29** 612–626.

Lock Morgan, K. and Rubin, D. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, **40** 1263–1282.

Markaryan, T. and Rosenberger, W. (2010). Exact properties of efron's biased coin randomization procedure. *The Annals of Statistics*, **38** 1546–1567.

Pocock, S. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 103–115.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, **102**.

Rosenberger, W. F. and Lachin, J. M. (2016). *Randomization in clinical trials: theory and practice*. 2nd ed. John Wiley & Sons.

Shao, J., Yu, X. and Zhong, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*, **97** 347–360.

Shorack, G. R. and Wellner, J. A. (2009). *Empirical processes with applications to statistics*. Siam.

Wei, L. (1978). The adaptive biased coin design for sequential experiments. *The Annals of Statistics*, **6** 92–100.

Young, A. (2016). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. Tech. rep., Technical Report, Working paper.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of chronic diseases*, **27** 365–375.