

# Inference under Covariate-Adaptive Randomization\*

Federico A. Bugni

Department of Economics

Duke University

[federico.bugni@duke.edu](mailto:federico.bugni@duke.edu)

Ivan A. Canay

Department of Economics

Northwestern University

[iacanay@northwestern.edu](mailto:iacanay@northwestern.edu)

Azeem M. Shaikh

Department of Economics

University of Chicago

[amshaikh@uchicago.edu](mailto:amshaikh@uchicago.edu)

August 4, 2016

## 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. 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 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. In a simulation study, we find that the rejection probability may in fact be dramatically less than the nominal level. We show further that these same conclusions remain true for a naïve permutation test, but that a modified version of the permutation test yields a test that is non-conservative in the sense that its limiting rejection probability under the null hypothesis equals the nominal level for a wide variety of randomization schemes that achieve what we refer to as “strong balance.” The modified version of the permutation test has the additional advantage that it has rejection probability exactly equal to the nominal level for some distributions satisfying the null hypothesis and some randomization schemes. Finally, we show that 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 yields a non-conservative test as well under very weak assumptions on the randomization scheme, including randomization schemes that do not achieve “strong balance.” In a simulation study, we find that the non-conservative tests have substantially greater power than the usual two-sample  $t$ -test.

**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

---

\*We would like to thank Lori Beaman, Raymond Guiteras, Aprajit Mahajan, Andres Santos and seminar participants at various institutions for helpful comments on this paper. We additionally thank Yuehao Bai 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. We additionally thank the computational resources of and staff for the Social Sciences Computing cluster (SSCC) at Northwestern University.

# 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 satisfying weak assumptions and study its consequences for testing the null hypothesis that the average treatment effect equals a pre-specified value in such settings. 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. \(2015, page 6\)](#).

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 4.4](#) 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. We show further that these conclusions remain true for a naïve permutation test. 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.

Motivated by these results, we go on to show that a modified version of the permutation test which only permutes treatment status for units within the same stratum yields a test that is non-conservative in the sense that its limiting rejection probability under the null hypothesis equals the nominal level for a wide variety of randomization schemes that achieve what we refer to as “strong balance,” defined formally in [Section 2](#) below. We refer to this test as the covariate-adaptive permutation test. As explained further in [Remark 4.10](#) 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,” when treatment status is determined using very specific randomization schemes. 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 these tests for testing the null hypothesis that the average treatment effect equals a pre-specified value for

a much wider variety of randomization schemes, while retaining the finite-sample validity for the narrower version of the null hypothesis and some randomization schemes.

We additionally consider the usual  $t$ -test (on the coefficient on treatment assignment) in a linear projection 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. Remarkably, our results show that this simple modification of the usual two-sample  $t$ -test yields a test that is non-conservative as well under very weak assumptions on the randomization scheme, including randomization schemes that do not achieve “strong balance.” In contrast to the covariate-adaptive permutation test, however, this test does not have the limited finite-sample validity of the covariate-adaptive permutation test described above.

The remainder of the paper is organized as follows. In [Section 2](#), we describe our setup and notation. In particular, there we describe the weak 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, importantly including stratified block randomization. Our main results about the four tests mentioned above 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. Proofs of all results are provided in the [Appendix](#).

## 2 Setup and Notation

Let  $Y_i$  denote the (observed) outcome of interest for the  $i$ th unit,  $A_i$  denote an indicator for whether the  $i$ th unit is treated or not, and  $Z_i$  denote observed, baseline covariates for the  $i$ th unit. Further denote by  $Y_i(1)$  the potential outcome of the  $i$ th unit if treated and by  $Y_i(0)$  the potential outcome of the  $i$ th 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 mechanism for determining treatment assignment. We therefore state our assumptions below in terms of assumptions on  $Q_n$  and assumptions on the mechanism for determining treatment status. Indeed, we will not make reference to  $P_n$  in the sequel and all operations are understood to be under  $Q_n$  and the mechanism for determining treatment status.

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

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

Next, we describe our assumptions on the mechanism determining treatment assignment. As mentioned previously, in this paper we focus on covariate-adaptive randomization, i.e., randomization schemes that first stratify according to baseline covariates and then assign treatment status so as to achieve “balance” within each stratum. In order to describe our assumptions on the treatment assignment mechanism more formally, we require some further notation. To this end, let  $S : \text{supp}(Z_i) \rightarrow \mathcal{S}$ , where  $\mathcal{S}$  is a finite set, be the function used to construct strata and, for  $1 \leq i \leq n$ , let  $S_i = S(Z_i)$ . Denote by  $S^{(n)}$  the vector of strata  $(S_1, \dots, S_n)$  and denote by  $A^{(n)}$  the vector of treatment assignments  $(A_1, \dots, A_n)$ . For  $s \in \mathcal{S}$ , let  $p(s) = P\{S_i = s\}$  and

$$D_n(s) = \sum_{1 \leq i \leq n} A_i^* I\{S_i = s\} , \quad (2)$$

where

$$A_i^* = A_i - \frac{1}{2} .$$

Note that  $D_n(s)$  as defined in (2) is simply a measure of the imbalance in stratum  $s$ . In order to rule out trivial strata, we, of course, assume that  $p(s) > 0$  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  $\Sigma_D = \text{diag}\{\varsigma_D^2(s) : s \in \mathcal{S}\}$  and

$$\varsigma_D^2(s) = p(s)\tau(s) \text{ with } 0 \leq \tau(s) \leq \frac{1}{4} \text{ for all } s \in \mathcal{S} .$$

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) is an additional requirement that is satisfied by a wide variety of randomization schemes. 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.”

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

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

For a pre-specified choice of  $\theta_0$ , the testing problem of interest is

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

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

### 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  $A^{(k-1)} = (A_1, \dots, A_{k-1})$  and  $S^{(k)} = (S_1, \dots, S_k)$  for  $1 \leq k \leq n$ , where  $A^{(0)}$  is understood to be a constant.

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

$$P\{A_k = 1 | S^{(n)}, A^{(k-1)}\} = P\{A_k = 1\} = \frac{1}{2} \quad (5)$$

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  $\tau(s) = \frac{1}{4}$  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 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:

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

where  $D_{k-1}(S_k) = \sum_{1 \leq i \leq k-1} A_i^* I\{S_i = S_k\}$ , 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  $k$ th unit is assigned to treatment in an effort to improve “balance” in the corresponding stratum in finite samples. It follows from Lemma B.9 that this treatment assignment mechanism satisfies Assumption 2.2. 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 SRS originally proposed by Wei (1978). This randomization scheme is similar to a biased-coin design, except that the probability  $\lambda$  in (6) depends on  $D_{k-1}(S_k)$ , the magnitude of imbalance in the corresponding stratum. More precisely, in this randomization scheme, treatment assignment

is determined recursively for  $1 \leq k \leq n$  as follows:

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

where  $\varphi(x) : [-1, 1] \rightarrow [0, 1]$  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.10 that this treatment assignment mechanism satisfies Assumption 2.2. 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 (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  $n_1(s) \leq n(s)$  be given. In this randomization scheme,  $n_1(s)$  units in stratum  $s$  are assigned to treatment and the remainder are assigned to control, where all

$$\begin{pmatrix} n(s) \\ n_1(s) \end{pmatrix}$$

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

$$n_1(s) = \left\lfloor \frac{n(s)}{2} \right\rfloor, \quad (8)$$

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.11 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. ■

**Remark 3.2.** Another treatment assignment mechanism that has been used in clinical trials is are 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:

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

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  $i$ th unit,  $A_i$ , depends on  $Y_1, \dots, Y_{i-1}$ . 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

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

where  $n_a = \sum_{1 \leq i \leq n} I\{A_i = a\}$ . The two-sample  $t$ -test is given by

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

where

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

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. Note that further results on linear regression are developed in [Section 4.4](#) 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,

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

where  $\varsigma_{t\text{-test}}^2 \leq 1$ . Furthermore,  $\varsigma_{t\text{-test}}^2 < 1$  unless

$$(1 - 4\tau(s)) (E[m_1(Z_i)|S_i = s] + E[m_0(Z_i)|S_i = s])^2 = 0 \text{ for all } s \in \mathcal{S}, \quad (12)$$

where

$$m_a(Z_i) = E[Y_i(a)|Z_i] - E[Y_i(a)] \quad (13)$$

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

$$\limsup_{n \rightarrow \infty} E[\phi_n^{t\text{-test}}(X^{(n)})] \leq \alpha \quad (14)$$

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  $\tau(s) = \frac{1}{4}$  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.** In the proof of Theorem 4.1 in the Appendix, it is shown that

$$\varsigma_{t\text{-test}}^2 = \frac{\varsigma_Y^2 + \varsigma_H^2 + \varsigma_A^2}{\varsigma_Y^2}, \quad (15)$$

where

$$\varsigma_Y^2 = 2\text{Var}[Y_i(1)] + 2\text{Var}[Y_i(0)] \quad (16)$$

$$\varsigma_Y^2 = 2\text{Var}[\tilde{Y}_i(1)] + 2\text{Var}[\tilde{Y}_i(0)] \quad (17)$$

$$\varsigma_H^2 = E[(E[m_1(Z_i)|S_i] - E[m_0(Z_i)|S_i])^2] \quad (18)$$

$$\varsigma_A^2 = E\left[2\tau(S_i) (E[m_1(Z_i)|S_i] + E[m_0(Z_i)|S_i])^2\right] \quad (19)$$

with  $\tilde{Y}_i(a) = Y_i(a) - E[Y_i(a)|S_i]$ . From (15), we see that three different sources of variation contribute to the variance. The first quantity,  $\varsigma_Y^2$ , reflects variation in the potential outcomes; the second quantity,  $\varsigma_H^2$ , reflects variation due to heterogeneity in the responses to treatment, i.e.,  $m_1 \neq m_0$ ; and the third quantity,  $\varsigma_A^2$ , reflects variation due to “imperfectly balanced” treatment assignment, i.e.,  $\tau(s) > 0$  in Assumption 2.2.(b). ■

**Remark 4.4.** 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 4.5.** 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. 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 non-conservative tests studied in Sections 4.3 and 4.4. ■

## 4.2 Naïve Permutation Test

In this section, we consider using a naïve permutation test to test (4) at level  $\alpha \in (0, 1)$ . In order to define this test, let  $\mathbf{G}_n$  to be the group of permutations of  $n$  elements. Define the action of  $g \in \mathbf{G}_n$  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 naïve permutation test is given by

$$\phi_n^{\text{naïve}}(X^{(n)}) = I\{T_n(X^{(n)}) > \hat{c}_n^{\text{naïve}}(1 - \alpha)\} , \quad (20)$$

where

$$\hat{c}_n^{\text{naïve}}(1 - \alpha) = \inf \left\{ x \in \mathbf{R} : \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} I\{T_n(gX^{(n)}) \leq x\} \geq 1 - \alpha \right\} . \quad (21)$$

The following theorem describes the asymptotic behavior of naïve permutation test defined in (20) with  $T_n(X^{(n)})$  given by (11) under covariate-adaptive randomization. In particular, it shows that the naïve permutation test, like the two-sample  $t$ -test, also has limiting rejection probability under the null hypothesis generally strictly less than the nominal level.

**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  $\alpha \in (0, 1)$ ,  $\phi_n^{\text{naïve}}(X^{(n)})$  defined in (20) with  $T_n(X^{(n)})$*

given by  $|T_n^{t-stat}(X^{(n)})|$  in (11) satisfies

$$\limsup_{n \rightarrow \infty} E[\phi_n^{naive}(X^{(n)})] \leq \alpha \quad (22)$$

whenever  $Q$  additionally satisfies the null hypothesis, i.e.,  $\theta(Q) = \theta_0$ . Furthermore, the inequality in (22) is strict unless (12) holds.

**Remark 4.6.** Rosenberger and Lachin (2016, page 105) argue that the use of permutation tests that do not respect stratification in the presence of covariate-adaptive randomization is “inappropriate.” Theorem 4.2 elaborates on this assertion by showing that the specific permutation test defined in (20) is inappropriate in the sense that it may be conservative. We note further that the result essentially follows from Theorem 4.1, which establishes the asymptotic behavior of the two-sample  $t$ -statistic, and results in Janssen (1997) and Chung and Romano (2013), which establish the asymptotic behavior of  $\hat{c}_n^{naive}(1 - \alpha)$  defined in (21). ■

**Remark 4.7.** It may often be the case that  $\mathbf{G}_n$  is too large to permit computation of  $\hat{c}_n^{naive}(1 - \alpha)$  defined in (21). In such situations, a stochastic approximation to the test may be used by replacing  $\mathbf{G}_n$  with  $\hat{\mathbf{G}}_n = \{g_1, \dots, g_B\}$ , where  $g_1$  equals the identity permutation and  $g_2, \dots, g_B$  are i.i.d.  $\text{Unif}(\mathbf{G}_n)$ . Theorem 4.2 remains true with such an approximation provided that  $B \rightarrow \infty$  as  $n \rightarrow \infty$ . ■

**Remark 4.8.** While Theorem 4.2 characterizes when the limiting rejection probability of the naïve permutation test under the null hypothesis is strictly less than the nominal level, it does not reveal how significant this difference might be. In our simulation study in Section 5, we find that, like the two-sample  $t$ -test studied in the previous section, 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 non-conservative tests studied in Sections 4.3 and 4.4. ■

### 4.3 Covariate-Adaptive Permutation Test

It follows from Theorems 4.1-4.2 and Remark 4.2 that the two-sample  $t$ -test and naïve permutation test are conservative in the sense that their limiting rejection probability under the null hypothesis is generally strictly less than the nominal level. As explained in Remarks 4.5 and 4.8, the finite-sample rejection probability may in fact be dramatically less than the nominal level. In this section, we propose a modified version of the permutation test, which we term the covariate-adaptive permutation test, that is not conservative in this way.

In order to define the test, we require some further notation. Define

$$\mathbf{G}_n(S^{(n)}) = \{g \in \mathbf{G}_n : S_{g(i)} = S_i \text{ for all } 1 \leq i \leq n\}, \quad (23)$$

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 before. For a given choice of test statistic  $T_n(X^{(n)})$ , the covariate-adaptive permutation test is given by

$$\phi_n^{\text{cap}}(X^{(n)}) = I\{T_n(X^{(n)}) > \hat{c}_n^{\text{cap}}(1 - \alpha)\}, \quad (24)$$

where

$$\hat{c}_n^{\text{cap}}(1 - \alpha) = \inf \left\{ x \in \mathbf{R} : \frac{1}{|\mathbf{G}_n(S^{(n)})|} \sum_{g \in \mathbf{G}_n(S^{(n)})} I\{T_n(gX^{(n)}) \leq x\} \geq 1 - \alpha \right\}. \quad (25)$$

The following theorem describes the asymptotic behavior of the covariate-adaptive permutation test defined in (24) with  $T_n(X^{(n)})$  given by  $|T_n^{\text{t-stat}}(X^{(n)})|$  in (11) under covariate-adaptive randomization. In particular, it shows that the limiting rejection probability of the proposed test under the null hypothesis equals the nominal level. As a result of this, we show in our simulations that the test has dramatically greater power than either the two-sample  $t$ -test or the naïve permutation test. In comparison with our preceding results, the theorem further requires that the randomization scheme achieves “strong balance,” i.e.,  $\tau(s) = 0$  for all  $s \in \mathcal{S}$ , but, as explained in Section 3, this property holds for a wide variety of treatment assignment mechanisms, including biased-coin designs and stratified block randomization.

**Theorem 4.3.** *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  $\alpha \in (0, 1)$ ,  $\phi_n^{\text{cap}}(X^{(n)})$  defined in (24) with  $T_n(X^{(n)})$  given by  $|T_n^{\text{t-stat}}(X^{(n)})|$  in (11) satisfies*

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

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

**Remark 4.9.** 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 requires that there is no (limiting) variation in this quantity in the sense that  $\tau(s) = 0$  for all  $s \in \mathcal{S}$ . ■

**Remark 4.10.** By arguing as in Heckman et al. (2011) or Lee and Shaikh (2014), it is possible to show that (24) is level  $\alpha$  in finite samples for testing the much more narrowly defined null hypothesis that specifies

$$Y_i(0)|S_i \stackrel{d}{=} Y_i(1)|S_i \quad (26)$$

whenever the treatment assignment mechanism is such that

$$gA^{(n)}|S^{(n)} \stackrel{d}{=} A^{(n)}|S^{(n)} \text{ for all } g \in \mathbf{G}_n(S^{(n)}) . \quad (27)$$

The property in (27) clearly holds, for example, for simple random sampling and stratified block randomization. Note further that (26) is implied by what is sometimes referred to as the “sharp null hypothesis,” which specifies that

$$Y_i(0) = Y_i(1) \quad (28)$$

with probability one. See, for example, Rosenbaum (2007) as well as Young (2016), who discusses Fisher-type tests of the null hypothesis (28) more generally. By contrast, Theorem 4.3 asymptotically justifies the use of (24) for testing the null hypothesis defined in (4) for a wide variety of treatment assignment mechanisms while retaining this finite-sample validity. The proof of Theorem 4.3 exploits 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 \(2013\)](#) in order to show that the test statistic  $T_n(X^{(n)})$  in [\(11\)](#) and the group of permutations  $\mathbf{G}_n(S^{(n)})$  in [\(23\)](#) satisfy the conditions in [Hoeffding \(1952\)](#).

■

**Remark 4.11.** As with the naïve permutation test, 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 [\(25\)](#). In such situations, a stochastic approximation to the test may be used by replacing  $\mathbf{G}_n(S^{(n)})$  with  $\hat{\mathbf{G}}_n = \{g_1, \dots, g_B\}$ , where  $g_1$  equals the identity permutation and  $g_2, \dots, g_B$  are i.i.d.  $\text{Unif}(\mathbf{G}_n(S^{(n)}))$ . [Theorem 4.3](#) remains true with such an approximation provided that  $B \rightarrow \infty$  as  $n \rightarrow \infty$ . ■

#### 4.4 Linear Regression with Strata Indicators

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. We consider tests with both homoskedasticity-only and 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 heteroskedasticity-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. Remarkably, in this section, we show that the addition of strata fixed effects results in a test is not conservative in this way, regardless of whether homoskedasticity-only or heteroskedasticity-robust standard errors are used.

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

$$Y_i = \beta A_i + \sum_{s \in \mathcal{S}} \delta_s I\{S_i = s\} + \epsilon_i \quad (29)$$

by ordinary least squares. Denote by  $\hat{\beta}_n$  the resulting estimator of  $\beta$  in [\(29\)](#). Let

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

where  $\hat{V}_{n,\beta}$  equals either the usual homoskedasticity-only or heteroskedasticity-robust standard error for  $\hat{\beta}_n$ . See [\(A-53\)](#) and [\(A-55\)](#) in the Appendix for exact expressions. Using this notation, the test of interest is given by

$$\phi_n^{\text{sfe}}(X^{(n)}) = I\{|T_n^{\text{sfe}}(X^{(n)})| > z_{1-\frac{\alpha}{2}}\}. \quad (30)$$

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. In the simulation results below, we show that, like the covariate-adaptive permutation test, the test has dramatically greater power than either the two-sample  $t$ -test or the naïve permutation test. Note that, in contrast to our preceding result on the covariate-adaptive permutation test, the theorem does not require that the randomization scheme

achieve “strong balance,” i.e.  $\tau(s) = 0$  for all  $s \in \mathcal{S}$ . On the other hand, the  $t$ -test with strata fixed effects studied here does not share with the covariate-adaptive permutation test the finite-sample validity explained in Remark 4.10.

**Theorem 4.4.** *Suppose  $Q$  satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Then,*

$$\sqrt{n}(\hat{\beta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma_{sfe}^2) . \quad (31)$$

Furthermore,

$$\hat{V}_{n,\beta} \xrightarrow{P} \varsigma_{sfe}^2 , \quad (32)$$

where  $\hat{V}_{n,\beta}$  equals either the usual homoskedasticity-only or heteroskedasticity-robust standard error for  $\hat{\beta}_n$ . Thus, for the problem of testing (4) at level  $\alpha \in (0, 1)$ ,  $\phi_n^{sfe}(X^{(n)})$  defined in (30) with either choice of  $\hat{V}_{n,\beta}$  satisfies

$$\lim_{n \rightarrow \infty} E[\phi_n^{sfe}(X^{(n)})] = \alpha \quad (33)$$

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

**Remark 4.12.** The above result in Theorem 4.4 that one may use either the usual homoskedasticity-only or heteroskedasticity-robust standard error for  $\hat{\beta}_n$  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. ■

**Remark 4.13.** In the proof of Theorem 4.4 in the Appendix, it is shown that

$$\varsigma_{sfe}^2 = \varsigma_Y^2 + \varsigma_H^2 , \quad (34)$$

where  $\varsigma_Y^2$  and  $\varsigma_H^2$  are defined as in (17) and (18), respectively. Remarkably, from (34), we see that variation due to “imperfectly balanced” treatment assignment, i.e.,  $\varsigma_D^2(s) > 0$  in Assumption 2.2, does not contribute to the variance  $\varsigma_{sfe}^2$ . ■

**Remark 4.14.** 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 standard errors associated with ordinary least squares estimation of (29). We therefore do not recommend computing  $\hat{\beta}_n$  using the deviations of  $Y_i$  and  $A_i$  from their respective means within strata. ■

**Remark 4.15.** Imbens and Rubin (2015, Ch. 9.6) examine the limit in probability of  $\hat{\beta}_n$  under a specific randomization assignment, namely, stratified block randomization; see Example 3.4. In contrast to our results, they do not impose the requirement that  $n_1(s)$  is chosen to achieve “balance” as in (8). 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 not only conclude that  $\hat{\beta}_n$  is consistent for  $\theta(Q)$ , but the test  $\phi_n^{sfe}(X^{(n)})$  has limiting rejection probability under the null hypothesis equal to the nominal level. Importantly, Imbens and Rubin (2015) do not include results on  $\phi_n^{sfe}(X^{(n)})$ . Note

that the required arguments are involved due to  $A^{(n)}$  not being i.i.d., relying in particular on non-standard convergence results, such as Lemmas [B.1](#) and [B.2](#) in the Appendix. ■

**Remark 4.16.** Assumption [2.2](#)(b) implies in particular that  $\frac{D_n(s)}{n} \xrightarrow{P} 0$  for all  $s \in \mathcal{S}$ . In other words, for each stratum, the fraction of units assigned to treatment and control are approximately equal in the sense that the ratio tends to one. While researchers often choose to assign units to treatment and control in this way, it may be desired to assign fewer people to treatment than to control when, for instance, treatment is “expensive.” [Bugni et al. \(2016\)](#) study inference on the average treatment effect in such settings. They show, in particular, that none of the tests considered in this paper provide non-conservative inference. Indeed, non-conservative inference based on  $\hat{\beta}_n$  requires a standard error different from the usual heteroskedasticity-robust standard error. A by-product of their analysis is that the expression for the limiting variance of  $\sqrt{n}(\hat{\beta}_n - \theta(Q))$  provided by [Imbens and Rubin \(2015, Theorem 9.1\)](#) is generally incorrect. ■

**Remark 4.17.** It is important to point out that the asymptotic validity of neither the covariate-adaptive permutation test nor the  $t$ -test with strata fixed effects discussed in this section rely on a particular model of (potential) outcomes. In the simulations below, we see that when such additional information is available, it may be possible to exploit it to devise even more powerful methods (e.g., linear regression of outcomes on treatment assignment and covariates). However, these methods may perform quite poorly when this information is incorrect, which is also apparent in the simulations below. ■

## 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:

$$Y_i(a) = \mu_a + m_a(Z_i) + \sigma_a(Z_i)\epsilon_{a,i} . \tag{35}$$

where  $\mu_a$ ,  $m_a(Z_i)$ ,  $\sigma_a(Z_i)$ , and  $\epsilon_{a,i}$  are specified as follows. In each of the following specifications,  $n = 100$  and  $\{(Z_i, \epsilon_{0,i}, \epsilon_{1,i}) : 1 \leq i \leq n\}$  are i.i.d.

**Model 1:**  $Z_i \sim \text{Beta}(2, 2)$  (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  $\epsilon_{1,i} \sim N(0, 1)$ ;  $m_0(Z_i) = m_1(Z_i) = \gamma Z_i$ . Note that 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  $m_0(Z_i) = m_1(Z_i) = \sin(\gamma Z_i)$ .

**Model 3:** As in Model 2, but  $m_1(Z_i) = \sin(\gamma Z_i) + \sqrt{Z_i + 2.25}$ .

**Model 4:** As in Model 3, but  $\sigma_0(Z_i) = Z_i^2$  and  $\sigma_1(Z_i) = Z_i^2 \sigma_1$ .

**Model 5:** As in Model 4, but  $\epsilon_{0,i} \sim \frac{1}{3}t_3$  and  $\epsilon_{1,i} \sim \frac{1}{3}t_3$ ;  $Z_i \sim \text{Unif}(-2, 2)$ ; and

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

For each of the above specifications of  $m_a(Z_i)$ ,  $\sigma_a(Z_i)$ , and  $\epsilon_{i,a}$ , we consider both  $(\gamma, \sigma_1) = (2, 1)$  and  $(\gamma, \sigma_1) = (4, \sqrt{2})$ . For each resulting specifications, we additionally 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).

Treatment assignment is generated according to one of the four different covariate-adaptive randomization schemes. In each of the schemes, strata are determined by dividing the support of  $Z_i$  (which is a closed interval in all specifications) into ten intervals of equal length and having  $S(Z_i)$  be the function that returns the interval in which  $Z_i$  lies. In particular,  $|\mathcal{S}| = 10$  in all specifications.

**SRS:** Treatment assignment is generated as in Example 3.1.

**BCD:** Treatment assignment is generated as in Example 3.2 with  $\lambda = \frac{2}{3}$ .

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

**SBR:** Treatment assignment is generated as in Example 3.4 with blocks of size  $\lfloor \frac{n(s)}{2} \rfloor$ .

In all cases, observed outcomes  $Y_i$  are generated according to (1).

In the simulation results below, we consider the following five different tests:

***t*-test:** The usual two-sample *t*-test studied in Section 4.1.

**Naïve:** The naïve permutation test studied in Section 4.2.

**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.

**SYZ:** The bootstrap-based test proposed by Shao et al. (2010).

**CAP:** The covariate-adaptive permutation test studied in Section 4.3.

**SFE:** The *t*-test with strata fixed effects studied in Section 4.4. In this case, we consider both homoskedasticity-only and heteroskedasticity-robust standard errors: homoskedastic/robust.

In all cases, rejection probabilities are computed using  $10^4$  replications.

Table 1 displays the results of the simulation study for  $(\gamma, \sigma_1) = (2, 1)$  and Table 2 displays the results of the simulation study for  $(\gamma, \sigma_1) = (4, \sqrt{2})$ . In the ‘SFE’ column in both tables, the first number corresponds

Model	CAR	Rejection rate under null - $\theta = 0$						Rejection rate under alternative - $\theta = 1/2$					
		$t$ -test	Naïve	Reg	SYZ	CAP	SFE	$t$ -test	Naïve	Reg	SYZ	CAP	SFE
1	SRS	5.39	4.81	5.45	4.92	5.04	5.07/5.26	20.03	18.86	68.18	19.08	44.94	61.38/61.73
	WEI	0.69	0.57	5.01	5.39	4.98	5.32/5.13	13.25	11.94	69.71	35.7	50.61	64.02/64.02
	BCD	0.90	0.68	4.94	6.05	5.00	4.90/4.99	13.08	11.63	69.50	35.87	50.02	64.96/65.26
	SBR	0.01	0.01	5.20	4.93	5.05	5.13/5.04	5.30	4.42	70.10	59.49	61.45	66.14/66.47
2	SRS	5.69	5.03	5.20	5.11	5.01	5.49/5.29	53.80	51.97	55.13	52.36	57.71	63.85/62.65
	WEI	3.05	2.65	3.23	5.91	4.51	5.15/4.88	55.12	53.00	56.64	65.17	63.32	66.12/66.63
	BCD	3.08	2.65	3.28	5.89	5.00	4.90/4.96	54.99	52.84	56.16	65.19	63.59	67.13/65.94
	SBR	2.19	1.95	2.37	5.63	4.75	5.22/4.57	55.95	53.58	57.12	70.71	67.73	68.09/67.81
3	SRS	5.43	4.80	5.05	4.98	5.03	5.42/4.98	49.32	47.20	54.15	47.89	55.05	63.43/62.15
	WEI	2.74	2.41	3.51	6.10	4.85	4.69/5.02	49.92	47.70	55.06	61.72	59.93	65.79/64.22
	BCD	2.56	2.22	3.39	5.69	4.81	5.16/4.90	48.80	46.47	54.43	61.33	60.76	65.65/64.70
	SBR	1.66	1.45	2.40	6.24	4.98	4.83/4.85	50.08	47.95	55.85	68.92	65.39	66.01/66.44
4	SRS	5.18	4.51	5.09	4.96	4.77	5.03/5.80	34.47	32.59	37.80	32.89	39.11	41.97/46.90
	WEI	3.54	3.03	4.08	6.67	5.18	5.25/5.83	33.92	32.21	36.93	42.79	40.99	44.45/46.43
	BCD	3.17	2.80	3.91	5.78	4.73	5.07/5.84	33.74	31.91	37.31	42.09	40.68	44.69/46.80
	SBR	2.91	2.52	3.75	7.29	5.06	5.88/6.06	33.66	31.69	36.99	47.52	42.20	45.20/46.26
5	SRS	5.68	5.09	5.27	5.32	5.11	4.97/4.79	31.45	29.71	30.84	30.21	48.63	60.88/61.55
	WEI	1.08	0.87	0.94	6.10	4.75	4.79/4.88	26.72	24.80	25.53	48.68	53.46	63.83/64.29
	BCD	1.14	0.94	1.07	5.73	4.99	4.68/4.61	26.61	24.74	25.52	49.26	55.38	64.06/64.44
	SBR	0.19	0.16	0.17	5.62	5.22	4.86/4.86	23.30	20.49	21.83	64.67	63.50	64.63/ 64.68

Table 1: Parameter values:  $\gamma = 2$ ,  $\sigma_1 = 1$ .

to homoskedasticity-only standard errors and the second number corresponds to the heteroskedasticity-robust standard errors. 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 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 the 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.

**Naïve:** The results for the naïve permutation test are very similar to those for the two-sample  $t$ -test.

**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 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 other tests considered below.

**SYZ:** For most specifications, the bootstrap-based test proposed by Shao et al. (2010) has rejection

Model	CAR	Rejection rate under null - $\theta = 0$						Rejection rate under alternative - $\theta = 1/2$					
		$t$ -test	Naïve	Reg	SYZ	CAP	SFE	$t$ -test	Naïve	Reg	SYZ	CAP	SFE
1	SRS	5.89	5.27	5.13	5.37	5.30	4.86/5.26	9.73	8.81	52.02	9.11	31.81	41.80/42.70
	WEI	0.37	0.31	5.15	5.46	4.93	5.20/5.04	2.20	1.80	51.55	15.03	32.97	44.08/43.78
	BCD	0.56	0.50	4.69	5.22	4.61	4.99/4.63	2.32	1.98	52.57	15.20	33.54	43.99/44.43
	SBR	0.00	0.00	5.21	3.62	5.09	5.27/4.98	0.01	0.01	51.95	30.43	39.08	46.90/45.56
2	SRS	5.65	5.14	5.04	5.24	4.96	4.73/4.96	43.22	41.42	41.47	41.97	43.99	45.18/46.52
	WEI	4.24	3.78	3.83	6.43	5.26	4.43/5.22	42.83	40.78	40.74	48.89	46.38	47.32/47.76
	BCD	4.16	3.63	3.65	6.08	5.20	4.88/5.04	42.84	40.57	40.63	49.16	47.04	48.42/47.14
	SBR	3.42	3.22	3.10	6.57	5.01	5.35/5.26	42.82	41.82	40.81	53.62	49.16	49.96/49.27
3	SRS	5.51	4.85	4.89	5.04	5.03	5.36/5.06	42.17	40.45	40.78	40.81	42.93	46.27/45.06
	WEI	4.23	3.73	3.86	6.20	5.20	5.13/5.02	42.26	40.21	40.86	48.50	46.02	48.51/47.78
	BCD	3.73	3.33	3.39	5.67	4.85	5.14/4.77	41.63	39.78	40.44	47.85	45.86	48.10/47.37
	SBR	3.21	2.96	2.83	6.31	4.81	5.31/4.93	42.12	41.14	40.58	53.48	48.86	49.10/48.98
4	SRS	5.44	4.66	4.98	4.79	5.27	5.09/5.95	28.54	26.91	28.06	27.48	29.92	31.53/34.38
	WEI	4.07	3.51	3.77	5.74	4.74	5.43/5.36	27.68	26.02	27.27	31.76	30.26	33.57/34.50
	BCD	4.19	3.65	3.95	6.01	4.86	4.96/5.70	27.97	26.49	27.44	32.25	30.86	33.30/34.53
	SBR	3.80	3.58	3.58	7.53	5.00	6.77/6.23	27.53	26.51	26.68	36.58	31.13	34.91/33.65
5	SRS	5.59	4.96	5.13	5.23	5.22	4.91/4.86	14.18	13.21	13.78	13.40	27.59	35.46/36.10
	WEI	0.75	0.59	0.55	6.51	5.18	4.79/4.81	6.38	5.60	5.81	22.41	29.84	37.17/37.59
	BCD	0.69	0.57	0.70	5.82	4.78	4.56/4.68	5.79	5.03	5.43	22.07	30.65	37.22/37.43
	SBR	0.03	0.01	0.02	5.15	5.23	4.90/4.85	1.66	1.31	1.55	36.18	36.36	38.00/37.57

Table 2: Parameter values:  $\gamma = 4$ ,  $\sigma_1 = \sqrt{2}$ .

probability under the null hypothesis very close to the nominal level, though in some instances the rejection probability under the null hypothesis mildly exceeds the nominal level (e.g., 7.53% under Model 4 and stratified block randomization with  $\gamma = 4$  and  $\sigma_1 = \sqrt{2}$ ). Its rejection probability under the alternative hypothesis is often considerably lower than that of the other tests considered below. Recall, however, that [Shao et al. \(2010\)](#) only justify the use of this test for biased-coin design.

**CAP:** As expected in light of [Theorem 4.3](#), the covariate-adaptive permutation test has rejection probability under the null hypothesis very close to the nominal level in all specifications. Indeed, among all the tests considered here, it arguably has rejection probability under the null hypothesis closest to the nominal level across all specifications. As explained in [Remark 4.10](#), its rejection probability under the null hypothesis even equals the nominal level in finite-samples for some specifications. Furthermore, its rejection probability under the alternative hypothesis typically exceeds that of all the tests considered previously and often by a considerably margin. On the other hand, its rejection probability under the alternative hypothesis is typically less than that of the following test.

**SFE:** As expected in light of [Theorem 4.4](#), the  $t$ -test with strata fixed effects has rejection probability under the null hypothesis very close to the nominal level in nearly all specifications. Its rejection probability under the alternative hypothesis typically exceeds that of all the tests considered previously and often by a considerable margin. Note that the results using homoskedasticity-only and heteroskedasticity-robust standard errors are nearly identical.

## Appendix A Proof of the main results

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

$\sigma_X^2(s)$	For a random variable $X$ , $\sigma_X^2(s) = \text{Var}[X S = s]$
$\sigma_X^2$	For a random variable $X$ , $\sigma_X^2 = \text{Var}[X]$
$\mu_a$	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 $a \in \{0, 1\}$ , $E[Y_i(a) Z_i] - \mu_a$
$\varsigma_Y^2$	$2\sigma_{Y(1)}^2 + 2\sigma_{Y(0)}^2$
$\varsigma_Y^2$	$2\sigma_{Y(1)}^2 + 2\sigma_{Y(0)}^2$
$\varsigma_A^2$	$\sum_{s \in \mathcal{S}} p(s)\tau(s)(2E[m_1(Z_i) S_i = s] + 2E[m_0(Z_i) S_i = s])^2$
$\varsigma_H^2$	$\sum_{s \in \mathcal{S}} p(s)(E[m_1(Z_i) S_i = s] - E[m_0(Z_i) S_i = s])^2$
$n(s)$	Number of individuals in strata $s \in \mathcal{S}$
$n_1(s)$	Number of individuals in the treatment group in strata $s \in \mathcal{S}$

Table 3: Useful notation

### A.1 Proof of Theorem 4.1

We start the proof by showing that the numerator of the root in the statement of the theorem satisfies

$$\sqrt{n}(\bar{Y}_{n,1} - \bar{Y}_{n,0} - \theta(Q)) \xrightarrow{d} N(0, \varsigma_Y^2 + \varsigma_H^2 + \varsigma_A^2) . \quad (\text{A-36})$$

Consider the following derivation:

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

where we used  $D_n = \sum_{s \in \mathcal{S}} D_n(s)$ ,  $\frac{n_1}{n} = \frac{D_n}{n} + \frac{1}{2}$ , and the following definitions:

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

By Assumption 2.2.(b),  $\frac{D_n}{n} \xrightarrow{P} 0$ , which in turn implies that  $R_{n,1}^* \xrightarrow{P} 4$ . Lemma B.1 implies  $R_{n,2}^* \xrightarrow{d} \frac{1}{4} N(0, \varsigma_Y^2 + \varsigma_H^2 + \varsigma_A^2)$ . Lemma B.2 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

$$\sqrt{n} \sqrt{\frac{\hat{\sigma}_{n,1}^2}{n_1} + \frac{\hat{\sigma}_{n,0}^2}{n_0}} \xrightarrow{P} \varsigma_Y . \quad (\text{A-37})$$

This follows from showing that  $\frac{n\hat{\sigma}_{n,a}^2}{n_a} \xrightarrow{P} 2\sigma_{Y(a)}^2$  for  $a \in \{0, 1\}$ . We only show the result for  $a = 1$ ; the proof of the result for  $a = 0$  is analogous. Start by writing  $\bar{Y}_{n,1}$  as follows:

$$\bar{Y}_{n,1} \equiv \frac{1}{n_1} \sum_{i=1}^n A_i Y_i = \mu_1 + \frac{n}{n_1} \frac{1}{n} \sum_{i=1}^n A_i (Y_i(1) - \mu_1) . \quad (\text{A-38})$$

Then consider the following derivation:

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

where we used (A-38) and the following definitions:

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

Since  $\frac{n}{n_1} = (\frac{D_n}{n} + \frac{1}{2})^{-1}$  and  $\frac{D_n}{n} \xrightarrow{P} 0$  by Assumption 2.2.(b), it follows that  $\frac{n}{n_1} \xrightarrow{P} 2$ . The result follows from showing that  $R_{n,4}^* \xrightarrow{P} \frac{1}{2}\sigma_{Y(1)}^2$  and  $R_{n,5}^* \xrightarrow{P} 0$ . Since  $E[(Y_i(1) - \mu_1)^2] = \sigma_{Y(1)}^2$  and  $E[(Y_i(1) - \mu_1)] = 0$ , this follows immediately from Lemma B.2.

To prove that  $\varsigma_Y^2 + \varsigma_H^2 + \varsigma_A^2 \leq \varsigma_Y^2$  holds with strict inequality unless (12) holds, notice that for  $a \in \{0, 1\}$ ,

$$\sigma_{Y(a)}^2 = \sigma_{Y(a)}^2 - \sum_{s \in \mathcal{S}} E[(Y_i(1) - \mu_1) | S_i = s]^2 p(s) = \sigma_{Y(a)}^2 - \sum_{s \in \mathcal{S}} E[m_a(Z_i) | S_i = s]^2 p(s) . \quad (\text{A-39})$$

Using (A-39), we see that

$$\begin{aligned} \varsigma_Y^2 - \varsigma_Y^2 - \varsigma_H^2 - \varsigma_A^2 &= 2(\sigma_{Y(1)}^2 - \sigma_{Y(1)}^2) + 2(\sigma_{Y(0)}^2 - \sigma_{Y(0)}^2) \\ &\quad - \sum_{s \in \mathcal{S}} p(s) (E[m_1(Z_i) | S_i = s] - E[m_0(Z_i) | S_i = s])^2 \\ &\quad - \sum_{s \in \mathcal{S}} p(s) \tau(s) (2E[m_1(Z_i) | S_i = s] + 2E[m_0(Z_i) | S_i = s])^2 \\ &= \sum_{s \in \mathcal{S}} p(s) (1 - 4\tau(s)) (E[m_1(Z_i) | S_i = s] + E[m_0(Z_i) | S_i = s])^2 , \end{aligned}$$

where, by Assumption 2.2.(b),  $\tau(s) \in [0, \frac{1}{4}]$ . 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 4.3

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

$$\varsigma_{\text{cap}}^2 = \frac{\varsigma_Y^2 + \varsigma_H^2}{\varsigma_Y^2}. \quad (\text{A-40})$$

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

$$(T_n(G_n X^{(n)}), T_n(G'_n X^{(n)})) \xrightarrow{d} (T, T'), \quad (\text{A-41})$$

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-41) 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  $\mathcal{I} = \{1, \dots, n\}$  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  $P\{C_j = 1\} = \frac{1}{2}$ .
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

$$T_n^U(X^{(n)}) \equiv \sqrt{n} \left( \frac{1}{n_1} \sum_{i=1}^n Y_i A_i - \frac{1}{n_0} \sum_{i=1}^n Y_i (1 - A_i) \right), \quad (\text{A-42})$$

it follows that

$$(T_n^U(G_n X^{(n)}), T_n^U(G'_n X^{(n)})) \xrightarrow{d} (T^U, T^{U'}), \quad (\text{A-43})$$

where  $T^U$  and  $T^{U'}$  are independent with common distribution given by  $N(0, \varsigma_Y^2 + \varsigma_H^2)$  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-43) follows by verifying the following two conditions:

$$(T_n^U(G_n \tilde{X}^{(n)}), T_n^U(G'_n \tilde{X}^{(n)})) \xrightarrow{d} (T^U, T^{U'}) \quad (\text{A-44})$$

$$T_n^U(G_n g_0 \tilde{X}^{(n)}) - T_n^U(G_n X^{(n)}) = o_P(1). \quad (\text{A-45})$$

Lemma B.3 establishes (A-44) and Lemma B.4 establishes (A-45).

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

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

Since

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

and Assumption 2.2.(b) implies that  $\frac{n_1}{n} = \frac{D_n}{n} + \frac{1}{2} \xrightarrow{P} \frac{1}{2}$ , it follows from Lemma B.2 that  $T_n^{L,1}(G_n X^{(n)}) \xrightarrow{P} 2\sigma_{Y(1)}^2$ . A similar argument shows that  $T_n^{L,0}(G_n X^{(n)}) \xrightarrow{P} 2\sigma_{Y(0)}^2$ . It thus follows that

$$T_n^L(G_n X^{(n)}) \xrightarrow{P} \varsigma_Y . \quad (\text{A-46})$$

Combining (A-43) and (A-46), we see that (A-41) holds.

It now follows from Lemma B.6 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  $\varsigma_{\text{cap}}^2 = \varsigma_{\text{t-test}}^2$ . Combining this last result with Theorem 4.1, we see that  $\lim_{n \rightarrow \infty} E[\phi_n^{\text{cap}}(X^{(n)})] = \alpha$  whenever  $\theta(Q) = \theta_0$ , completing the proof of the theorem.

### A.3 Proof of Theorem 4.4

We first prove that (31) holds with  $\varsigma_{\text{ife}}^2 = \varsigma_Y^2 + \varsigma_H^2$ . 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

$$\begin{aligned} \sqrt{n}(\hat{\beta}_n - \theta(Q)) &= \frac{\sqrt{n}}{\frac{1}{n} \sum_{i=1}^n \tilde{A}_i^2} \left( \left( \frac{1}{n} \sum_{i=1}^n \tilde{A}_i Y_i \right) - \theta(Q) \left( \frac{1}{n} \sum_{i=1}^n \tilde{A}_i^2 \right) \right) \\ &= \frac{1}{\frac{1}{n} \sum_{i=1}^n \tilde{A}_i^2} \left( \frac{1}{4} \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n 4\tilde{A}_i Y_i - \theta(Q) \right) - \theta(Q) \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \tilde{A}_i^2 - \frac{1}{4} \right) \right) . \end{aligned}$$

Below we argue that

$$\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \tilde{A}_i^2 - \frac{1}{4} \right) = o_P(1) \quad (\text{A-47})$$

$$\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n 4\tilde{A}_i Y_i - \theta(Q) \right) = 4R_{n,1} + 4R_{n,3} + o_P(1) \quad (\text{A-48})$$

where  $R_{n,1}$  and  $R_{n,3}$  are defined, respectively, as in (B-58) and (B-60) below. By inspecting the proof of Lemma B.1, we also have that  $4R_{n,1} + 4R_{n,3} \xrightarrow{d} N(0, \varsigma_Y^2 + \varsigma_H^2)$ , from which (31) thus follows.

Step 1: To see that (A-47) holds, note that

$$\begin{aligned}
\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \tilde{A}_i^2 - \frac{1}{4} \right) &= \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \left( A_i - \sum_{s \in \mathcal{S}} I\{S_i = s\} \frac{n_1(s)}{n(s)} \right)^2 - \frac{1}{4} \right) \\
&= \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \left( A_i - 2A_i \sum_{s \in \mathcal{S}} I\{S_i = s\} \frac{n_1(s)}{n(s)} + \sum_{s \in \mathcal{S}} I\{S_i = s\} \left( \frac{n_1(s)}{n(s)} \right)^2 \right) - \frac{1}{4} \right) \\
&= \sqrt{n} \frac{1}{n} \sum_{i=1}^n \left( A_i - \frac{1}{2} \right) - \sqrt{n} \sum_{s \in \mathcal{S}} \frac{\left( \frac{n_1(s)}{n} \right)^2}{\frac{n(s)}{n}} + \sqrt{n} \frac{1}{4} \\
&= \sqrt{n} \sum_{s \in \mathcal{S}} \frac{D_n(s)}{n} - \sum_{s \in \mathcal{S}} \sqrt{n} \left( \frac{n_1(s)}{n} - \frac{p(s)}{2} \right) + o_P(1) \\
&= \sqrt{n} \left[ \sum_{s \in \mathcal{S}} \frac{D_n(s)}{n} - \sum_{s \in \mathcal{S}} \left( \sqrt{n} \left( \frac{n(s)}{2n} - \frac{p(s)}{2} \right) + \frac{D_n(s)}{n} \right) \right] + o_P(1), \tag{A-49}
\end{aligned}$$

where the fourth equality follows from the following:

$$\begin{aligned}
\sqrt{n} \left( \frac{\left( \frac{n_1(s)}{n} \right)^2}{\frac{n(s)}{n} - \frac{p(s)}{4}} \right) &= \sqrt{n} \left( \frac{n_1(s)}{n} - \frac{p(s)}{2} \right) - \frac{\sqrt{n}}{4} \left( \frac{n(s)}{n} - p(s) \right) + o_P(1) \\
\sqrt{n} \left( \frac{n_1(s)}{n} - \frac{p(s)}{2} \right) &= \frac{D_n(s)}{\sqrt{n}} + \sqrt{n} \frac{1}{2} \left( \frac{n(s)}{n} - p(s) \right) = O_P(1) \\
\sqrt{n} \left( \frac{n(s)}{n} - p(s) \right) &= O_P(1).
\end{aligned}$$

Note further that the term in brackets in (A-49) equals zero because  $\sum_{s \in \mathcal{S}} p(s) = 1$  and  $\sum_{s \in \mathcal{S}} n(s) = n$ .

Step 2: To see that (A-48) holds, note that

$$\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n 4\tilde{A}_i Y_i - \theta(Q) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n 4A_i Y_i - \sqrt{n}\theta(Q) - \frac{2}{\sqrt{n}} \sum_{s \in \mathcal{S}} \sum_{i=1}^n \frac{2n_1(s)}{n(s)} I\{S_i = s\} Y_i. \tag{A-50}$$

Consider the first two terms. Use that  $A_i^* = A_i - \frac{1}{2}$  and the definition of  $Y_i$  to see that

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i=1}^n 4A_i Y_i - \sqrt{n}\theta(Q) &= \frac{4}{\sqrt{n}} \sum_{i=1}^n A_i^* Y_i + \frac{2}{\sqrt{n}} \sum_{i=1}^n Y_i - \sqrt{n}\theta(Q) \\
&= \frac{2}{\sqrt{n}} \sum_{i=1}^n ((Y_i(1) - \mu_1)A_i - (Y_i(0) - \mu_0)(1 - A_i)) + \frac{2}{\sqrt{n}} \sum_{i=1}^n Y_i \\
&\quad + \frac{2}{\sqrt{n}} \sum_{i=1}^n (A_i \mu_1 - (1 - A_i)\mu_0) - \sqrt{n}(\mu_1 - \mu_0) \\
&= 4R_{n,1} + 4R_{n,3} + 4R_{n,2} + \frac{2}{\sqrt{n}} \sum_{i=1}^n Y_i + \frac{2}{\sqrt{n}} \sum_{i=1}^n (A_i^* (\mu_1 + \mu_0)), \tag{A-51}
\end{aligned}$$

where  $R_{n,2}$  is defined as in (B-59) below and the last equality follows from the derivation in the beginning of the

proof of Lemma B.1. Now consider the third term in (A-50). Since  $2\frac{n_1(s)}{n(s)} = 2\frac{D_n(s)}{n(s)} + 1$ , we see that

$$\begin{aligned}
\frac{2}{\sqrt{n}} \sum_{s \in \mathcal{S}} \sum_{i=1}^n \frac{2n_1(s)}{n(s)} I\{S_i = s\} Y_i &= \frac{2}{\sqrt{n}} \sum_{s \in \mathcal{S}} \sum_{i=1}^n \frac{2D_n(s)}{n(s)} I\{S_i = s\} Y_i + \frac{2}{\sqrt{n}} \sum_{s \in \mathcal{S}} \sum_{i=1}^n I\{S_i = s\} Y_i \\
&= \sum_{s \in \mathcal{S}} \frac{2D_n(s)}{\sqrt{n}} \frac{2n}{n(s)} \frac{1}{n} \sum_{i=1}^n I\{S_i = s\} Y_i + \frac{2}{\sqrt{n}} \sum_{i=1}^n Y_i \\
&= \sum_{s \in \mathcal{S}} \frac{2D_n(s)}{\sqrt{n}} \frac{2n}{n(s)} \left( \frac{p(s)}{2} (\mu_1 + \mu_0 + E[m_1(Z_i) + m_0(Z_i)|S_i = s]) + o_P(1) \right) \\
&\quad + \frac{2}{\sqrt{n}} \sum_{i=1}^n Y_i \\
&= 4R_{n,2} + \frac{2}{\sqrt{n}} \sum_{i=1}^n Y_i + \frac{2}{\sqrt{n}} \sum_{i=1}^n (A_i^* (\mu_1 + \mu_0)) + o_P(1), \tag{A-52}
\end{aligned}$$

where the third equality follows from  $\frac{1}{n} \sum_{i=1}^n I\{S_i = s\} Y_i = \frac{p(s)}{2} (\mu_1 + \mu_0 + E[m_1(Z_i) + m_0(Z_i)|S_i = s]) + o_P(1)$ , and the last equality follows from  $\frac{n}{n(s)} = \frac{1}{p(s)} + o_P(1)$ ,  $\frac{D_n(s)}{\sqrt{n}} = O_P(1)$  from Assumption 2.2.(b), and the definition of  $R_{n,2}$  in (B-59). The desired result (A-48) thus follows from (A-50)–(A-52).

Next we prove (32). We first prove the result for homoskedasticity-only standard errors, i.e.,

$$\hat{V}_{\text{homo}} = \left( \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 \right) \left( \frac{\mathbf{C}'_n \mathbf{C}_n}{n} \right)_{[1,1]}^{-1} \xrightarrow{P} \varsigma_Y^2 + \varsigma_H^2, \tag{A-53}$$

where  $\mathbf{C}_n$  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 (29).

Next note that  $\frac{1}{n} \mathbf{C}'_n \mathbf{C}_n \xrightarrow{P} \Sigma_C$  where

$$\Sigma_C \equiv \begin{bmatrix} 1/2 & \frac{1}{2}p(1) & \frac{1}{2}p(2) & \cdots & \frac{1}{2}p(|\mathcal{S}|) \\ \frac{1}{2}p(1) & p(1) & 0 & \cdots & 0 \\ \frac{1}{2}p(2) & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \frac{1}{2}p(|\mathcal{S}|) & 0 & \cdots & \cdots & p(|\mathcal{S}|) \end{bmatrix} \text{ and } \Sigma_C^{-1} = \begin{bmatrix} 4 & -2 & -2 & \cdots & -2 \\ -2 & 1 + \frac{1}{p(1)} & 1 & \cdots & 1 \\ -2 & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ -2 & 1 & \cdots & \cdots & 1 + \frac{1}{p(|\mathcal{S}|)} \end{bmatrix}. \tag{A-54}$$

The convergence in probability follows from  $\frac{n_1}{n} = \frac{D_n}{n} + \frac{1}{2} \xrightarrow{P} \frac{1}{2}$ ,  $\frac{n_1(s)}{n} = \frac{D_n(s)}{n} + \frac{1}{2} \frac{n(s)}{n} \xrightarrow{P} \frac{p(s)}{2}$ , 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  $\Sigma_C^{-1}$  equals 4. By Lemma B.8, we have that  $\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 \xrightarrow{P} \frac{1}{4} (\varsigma_Y^2 + \varsigma_H^2)$ . The result in (A-53) immediately follows.

We now prove the result for heteroskedasticity-robust standard errors, i.e.,

$$\hat{V}_{\text{hc}} = \left[ \left( \frac{\mathbf{C}'_n \mathbf{C}_n}{n} \right)^{-1} \left( \frac{\mathbf{C}'_n \text{diag}\{\hat{u}_i^2 : 1 \leq i \leq n\} \mathbf{C}_n}{n} \right) \left( \frac{\mathbf{C}'_n \mathbf{C}_n}{n} \right)^{-1} \right]_{[1,1]} \xrightarrow{P} \varsigma_Y^2 + \varsigma_H^2. \tag{A-55}$$

First note that

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

It follows from Lemma B.8 that

$$\frac{\mathbf{C}'_n \text{diag}\{\hat{u}_i^2 : 1 \leq i \leq n\} \mathbf{C}_n}{n} \xrightarrow{P} \Omega \quad (\text{A-56})$$

where each component of the matrix  $\Omega$  corresponds to the respective limits in Lemma B.8. It follows that

$$\left(\frac{\mathbf{C}'_n \mathbf{C}_n}{n}\right)^{-1} \left(\frac{\mathbf{C}'_n \text{diag}\{\hat{u}_i^2 : 1 \leq i \leq n\} \mathbf{C}_n}{n}\right) \left(\frac{\mathbf{C}'_n \mathbf{C}_n}{n}\right)^{-1} \xrightarrow{P} \Sigma_{\mathbf{C}}^{-1} \Omega \Sigma_{\mathbf{C}}^{-1} = \begin{bmatrix} \varsigma_{\tilde{Y}}^2 + \varsigma_H^2 & \mathbb{V}_{12} \\ \mathbb{V}'_{12} & \mathbb{V}_{22} \end{bmatrix}, \quad (\text{A-57})$$

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-55) follows immediately.

From the previous results, the final conclusion of the theorem, (33), follows immediately.

## Appendix B Auxiliary Results

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

$$\frac{1}{2\sqrt{n}} \sum_{i=1}^n ((Y_i(1) - \mu_1)A_i - (Y_i(0) - \mu_0)(1 - A_i)) \xrightarrow{d} \frac{1}{4} N(0, \varsigma_{\tilde{Y}}^2 + \varsigma_H^2 + \varsigma_A^2),$$

where  $\varsigma_{\tilde{Y}}^2$ ,  $\varsigma_H^2$ , and  $\varsigma_A^2$  are defined in Table 3.

*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:

$$\begin{aligned} & \frac{1}{2\sqrt{n}} \sum_{i=1}^n ((Y_i(1) - \mu_1)A_i - (Y_i(0) - \mu_0)(1 - A_i)) \\ &= \frac{1}{2\sqrt{n}} \sum_{i=1}^n (\tilde{Y}_i(1)A_i - \tilde{Y}_i(0)(1 - A_i)) + \frac{1}{2\sqrt{n}} \sum_{i=1}^n (E[m_1(Z_i)|S_i]A_i - E[m_0(Z_i)|S_i](1 - A_i)) \\ &= R_{n,1} + \frac{1}{2\sqrt{n}} \sum_{i=1}^n A_i^* \left( \sum_{s \in \mathcal{S}} E[m_1(Z_i)|S_i = s] I\{S_i = s\} + \sum_{s \in \mathcal{S}} E[m_0(Z_i)|S_i = s] I\{S_i = s\} \right) \\ & \quad + \frac{1}{4\sqrt{n}} \sum_{i=1}^n \left( \sum_{s \in \mathcal{S}} E[m_1(Z_i)|S_i = s] I\{S_i = s\} - \sum_{s \in \mathcal{S}} E[m_0(Z_i)|S_i = s] I\{S_i = s\} \right) \\ &= R_{n,1} + R_{n,2} + R_{n,3}, \end{aligned}$$

where we used  $A_i^* = A_i - \frac{1}{2}$  and the following definitions:

$$R_{n,1} \equiv \frac{1}{2\sqrt{n}} \sum_{i=1}^n (\tilde{Y}_i(1)A_i - \tilde{Y}_i(0)(1 - A_i)) , \quad (\text{B-58})$$

$$R_{n,2} \equiv \frac{1}{2} \sum_{s \in \mathcal{S}} \frac{D_n(s)}{\sqrt{n}} (E[m_1(Z_i)|S_i = s] + E[m_0(Z_i)|S_i = s]) , \quad (\text{B-59})$$

$$R_{n,3} \equiv \frac{1}{4} \sum_{s \in \mathcal{S}} \sqrt{n} \left( \frac{n(s)}{n} - p(s) \right) (E[m_1(Z_i)|S_i = s] - E[m_0(Z_i)|S_i = s]) . \quad (\text{B-60})$$

The result follows from the continuous mapping theorem once we show that  $R_n \equiv (R_{n,1}, R_{n,2}, R_{n,3}) \xrightarrow{d} (\zeta_{R_1}, \zeta_{R_2}, \zeta_{R_3})$  where  $\zeta_{R_1}$ ,  $\zeta_{R_2}$ , and  $\zeta_{R_3}$  are independent and satisfy  $\zeta_{R_1} \sim \frac{1}{4}N(0, \zeta_Y^2)$ ,  $\zeta_{R_2} \sim \frac{1}{4}N(0, \zeta_A^2)$ , and  $\zeta_{R_3} \sim \frac{1}{2}N(0, \zeta_H^2)$ .

We first show that

$$(R_{n,1}, R_{n,2}, R_{n,3}) \stackrel{d}{=} (R_{n,1}^*, R_{n,2}, R_{n,3}) + o_P(1) \quad (\text{B-61})$$

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  $\{(\tilde{Y}_i^s(1), \tilde{Y}_i^s(0)) : 1 \leq i \leq n\}$  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

$$\tilde{R}_{n,1} \equiv \frac{1}{2} \sum_{s \in \mathcal{S}} \left( \frac{1}{\sqrt{n}} \sum_{i=n\frac{N(s)}{n}+1}^{n\left(\frac{N(s)}{n} + \frac{n_1(s)}{n}\right)} \tilde{Y}_i^s(1) + \frac{1}{\sqrt{n}} \sum_{i=n\left(\frac{N(s)}{n} + \frac{n_1(s)}{n}\right)+1}^{n\left(\frac{N(s)}{n} + \frac{n(s)}{n}\right)} \tilde{Y}_i^s(0) \right) . \quad (\text{B-62})$$

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

$$R_{n,1}^* \equiv \frac{1}{2} \sum_{s \in \mathcal{S}} \left( \frac{1}{\sqrt{n}} \sum_{i=\lfloor nF(s) \rfloor + 1}^{\lfloor n(F(s) + \frac{p(s)}{2}) \rfloor} \tilde{Y}_i^s(1) + \frac{1}{\sqrt{n}} \sum_{i=\lfloor n(F(s) + \frac{p(s)}{2}) \rfloor + 1}^{\lfloor n(F(s) + p(s)) \rfloor} \tilde{Y}_i^s(0) \right) . \quad (\text{B-63})$$

Since  $R_{n,1}^*$  is a function of  $\{(\tilde{Y}_i^s(1), \tilde{Y}_i^s(0)) : 1 \leq i \leq n, s \in \mathcal{S}\} \perp\!\!\!\perp (S^{(n)}, A^{(n)})$ , 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-61), we establish that  $R_{n,1}^* \xrightarrow{d} \zeta_{R_1}$  and  $\Delta_n \equiv \tilde{R}_{n,1} - R_{n,1}^* \xrightarrow{P} 0$ . 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

$$\frac{1}{\sqrt{n}} \sum_{i=[nF(s)]+1}^{\lfloor n(F(s)+\frac{p(s)}{2}) \rfloor} \tilde{Y}_i^s(1) \xrightarrow{d} N\left(0, \frac{p(s)\sigma_{\tilde{Y}(1)}^2(s)}{2}\right),$$

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

$$\left(\frac{N(s)}{n}, \frac{n_1(s)}{n}\right) \xrightarrow{P} \left(F(s), \frac{p(s)}{2}\right),$$

it follows that

$$g_n\left(\frac{N(s)+n_1(s)}{n}\right) - g_n\left(\frac{N(s)}{n}\right) - \left(g_n\left(F(s)+\frac{p(s)}{2}\right) - g_n(F(s))\right) \xrightarrow{P} 0,$$

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  $\{(\tilde{Y}_i^s(1), \tilde{Y}_i^s(0)) : 1 \leq i \leq n, s \in \mathcal{S}\}$  across both  $i$  and  $s$ , we conclude that  $R_{n,1}^* \xrightarrow{d} \zeta_{R_1}$  and  $\Delta_n \equiv \tilde{R}_{n,1} - R_{n,1}^* \xrightarrow{P} 0$ .

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

$$\{R_{n,2}|S^{(n)}\} \xrightarrow{d} \zeta_{R_2} \text{ a.s.} \quad (\text{B-64})$$

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

$$R_{n,3} \xrightarrow{d} \zeta_{R_3}. \quad (\text{B-65})$$

To complete the proof, we show that  $(R_{n,1}, R_{n,2}, R_{n,3}) \xrightarrow{d} (\zeta_{R_1}, \zeta_{R_2}, \zeta_{R_3})$  with  $\{\zeta_{R_1}, \zeta_{R_2}, \zeta_{R_3}\}$  independent. From (B-61), 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.,

$$P\{R_{n,1}^* \leq h_1\}P\{R_{n,2} \leq h_2, R_{n,2} \leq h_3\} \rightarrow P\{\zeta_{R_1} \leq h_1\}P\{\zeta_{R_2} \leq h_2\}P\{\zeta_{R_3} \leq h_3\}, \quad (\text{B-66})$$

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  $P\{R_{n,1}^* \leq h_1\} \rightarrow P\{\zeta_{R_1} \leq h_1\}$  and (B-66) follows from the following argument:

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

where the convergence follows from the dominated convergence theorem, (B-64), and (B-65).

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  $\zeta_{R_j}$  is normally distributed, this implies that  $\zeta_j$  must be degenerate and equal to zero and thus  $h_j = 0$ . In turn, since  $P\{\zeta_{R_1} \leq \cdot\}P\{\zeta_2 \leq \cdot\}P\{\zeta_3 \leq \cdot\}$  is continuous at  $(h_1, h_2, h_3)$ , then this implies that  $\prod_{k \neq j} P\{\zeta_k \leq h_k\} = 0$ . Since  $\zeta_k$  for  $k \neq j$  are also normally distributed, this implies for some  $k \neq j$  that  $\zeta_k$  is also degenerate and equal to zero and

$h_k < 0$ . Then, (B-66) follows from the following argument:

$$\begin{aligned} P\{R_{n,1} \leq h_1, R_{n,2} \leq h_2, R_{n,2} \leq h_3\} &\leq P\{R_{n,k} \leq h_k\} \\ &\rightarrow 0 \\ &= P\{\zeta_1 \leq h_1\}P\{\zeta_2 \leq h_2\}P\{\zeta_3 \leq h_3\}, \end{aligned}$$

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

**Lemma B.2.** *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*

$$\frac{1}{n} \sum_{i=1}^n W_i A_i \xrightarrow{P} \frac{1}{2} E[W_i]. \quad (\text{B-67})$$

Furthermore,

$$\frac{1}{n} \sum_{i=1}^n W_i A_{G_n(i)} \xrightarrow{P} \frac{1}{2} E[W_i] \quad (\text{B-68})$$

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

*Proof.* We prove only (B-67); the convergence (B-68) follows from analogous arguments.

By arguing as in the proof of Lemma B.1, note that

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

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

$$\frac{1}{n} \sum_{i=1}^{n_1(s)} W_i^s \xrightarrow{P} \frac{1}{2} p(s) E[W_i^s]. \quad (\text{B-69})$$

From Assumption 2.2.(b),  $\frac{n_1(s)}{n} = \frac{D_n(s)}{n} + \frac{1}{2} \frac{n(s)}{n} \xrightarrow{P} \frac{p(s)}{2}$ , so (B-69) follows from

$$\frac{1}{n_1(s)} \sum_{i=1}^{n_1(s)} W_i^s \xrightarrow{P} E[W_i^s]. \quad (\text{B-70})$$

To establish (B-70), 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

$\frac{\tilde{n}_1(s)}{n} \rightarrow \frac{1}{2}p(s)$  a.s. Using the independence of  $(A^{(n)}, S^{(n)})$  and  $\{W_i^s : 1 \leq i \leq n\}$ , we see that for any  $\epsilon > 0$ ,

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

where the convergence follows from the dominated convergence theorem and

$$P \left\{ \left| \frac{1}{n \frac{\tilde{n}_1(s)}{n}} \sum_{i=1}^{n \frac{\tilde{n}_1(s)}{n}} W_i^s - E[W_i^s] \right| > \epsilon \left| \frac{\tilde{n}_1(s)}{n} \right. \right\} \rightarrow 0 \text{ a.s. .} \quad (\text{B-71})$$

To see that the convergence (B-71) holds, note that the weak law of large numbers implies that

$$\frac{1}{n_k} \sum_{i=1}^{n_k} W_i^s \xrightarrow{P} E[W_i^s] \quad (\text{B-72})$$

for any subsequence  $n_k \rightarrow \infty$  as  $k \rightarrow \infty$ . Since  $n \frac{\tilde{n}_1(s)}{n} \rightarrow \infty$  a.s., (B-71) follows from the independence of  $\frac{\tilde{n}_1(s)}{n}$  and  $\{W_i^s : 1 \leq i \leq n\}$  and (B-72). ■

**Lemma B.3.** *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 4.3. For  $T_n^U(X^{(n)})$  defined in (A-42), we have that (A-44) holds whenever  $Q$  additionally satisfies  $\theta(Q) = 0$ .*

*Proof.* Let  $g = G_n$  and  $g' = G'_n$ . Note that

$$T_n^U(g\tilde{X}^{(n)}) = \frac{2}{1 - \left(\frac{D_n}{n}\right)^2} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i A_{g(i)}^* + \frac{D_n}{\sqrt{n}} \frac{1}{n} \sum_{i=1}^n V_i \right),$$

where  $A_i^* = 2A_i - 1$ . 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

$$T_n^U(g\tilde{X}^{(n)}) = \frac{2}{\sqrt{n}} \sum_{i=1}^n V_i A_{g(i)}^* + o_P(1).$$

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

$$(T_n^U(g\tilde{X}^{(n)}), T_n^U(g'\tilde{X}^{(n)})) = \left( \frac{2}{\sqrt{n}} \sum_{i=1}^n V_i A_{g(i)}^*, \frac{2}{\sqrt{n}} \sum_{i=1}^n V_i A_{g'(i)}^* \right) + o_P(1).$$

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

$$\frac{2}{\sqrt{n}} \sum_{i=1}^n V_i (aA_{g^{(i)}}^* + bA_{g'^{(i)}}^*) \xrightarrow{d} N(0, (a^2 + b^2)(\varsigma_Y^2 + \varsigma_H^2)). \quad (\text{B-73})$$

Note that the left-hand side of (B-73) equals

$$\frac{2}{\sqrt{n}} \sum_{i=1}^n (V_i - E[V_i|S_i]) (aA_{g^{(i)}}^* + bA_{g'^{(i)}}^*) + \frac{2}{\sqrt{n}} \sum_{i=1}^n E[V_i|S_i] (aA_{g^{(i)}}^* + bA_{g'^{(i)}}^*). \quad (\text{B-74})$$

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-74) equals

$$\frac{2}{\sqrt{n}} \sum_{s \in \mathcal{S}} \sum_{i=1}^n E[V_i|S_i = s] I\{S_i = s\} (aA_{g^{(i)}}^* + bA_{g'^{(i)}}^*) = 4(a+b) \sum_{s \in \mathcal{S}} E[V_i|S_i = s] \frac{D_n(s)}{\sqrt{n}} = o_P(1),$$

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-74), define

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

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

$$\sum_{s \in \mathcal{S}} \frac{2}{\sqrt{n}} \left( \sum_{i=1}^{n_s(0,0)} \tilde{V}_i^s(-(a+b)) + \sum_{i=n_s(0,0)+1}^{n_s(0,1)} \tilde{V}_i^s(b-a) + \sum_{i=n_s(0,1)+1}^{n_s(1,0)} \tilde{V}_i^s(a-b) + \sum_{i=n_s(1,0)+1}^{n_s(1,1)} \tilde{V}_i^s(a+b) \right), \quad (\text{B-75})$$

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

$$\frac{n_s(0,0)}{n} \xrightarrow{P} \frac{p(s)}{4} \quad (\text{B-76})$$

$$\frac{n_s(0,1) - n_s(0,0)}{n} \xrightarrow{P} \frac{p(s)}{4} \quad (\text{B-77})$$

$$\frac{n_s(1,0) - n_s(0,1)}{n} \xrightarrow{P} \frac{p(s)}{4} \quad (\text{B-78})$$

$$\frac{n_s(1,1) - n_s(1,0)}{n} \xrightarrow{P} \frac{p(s)}{4}. \quad (\text{B-79})$$

First consider (B-76). 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,

$$E \left[ \frac{n_s(0,0)}{n} \middle| A^{(n)}, S^{(n)} \right] = \frac{n_0(s)^2}{n(s)n} \xrightarrow{P} \frac{p(s)}{4} \quad (\text{B-80})$$

$$\text{Var} \left[ \frac{n_s(0,0)}{n} \middle| A^{(n)}, S^{(n)} \right] = \frac{n_0(s)^2 n_1(s)^2}{n^2 n(s)^2 (n(s) - 1)^2} \xrightarrow{P} 0,$$

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

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

which implies further that

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

The convergence (B-76) thus follows from (B-80). The convergences (B-77)–(B-79) follow in the same way.

Therefore, again by arguing as in the proof of Lemma B.1, we see that (B-75) converges in distribution to a normal with mean zero and variance given by

$$\sum_{s \in \mathcal{S}} 2p(s)((a+b)^2 + (a-b)^2) \text{Var}[V_i | S_i = s] = (a^2 + b^2) \sum_{s \in \mathcal{S}} 4p(s) \text{Var}[V_i | S_i = s]. \quad (\text{B-81})$$

To complete the proof, note that

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

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  $\zeta_{\tilde{Y}}^2$  and  $\zeta_H^2$  in Table 3 that (B-81) equals  $(a^2 + b^2)(\zeta_{\tilde{Y}}^2 + \zeta_H^2)$ , as desired. ■

**Lemma B.4.** *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 4.3. Let  $T_n^U(X^{(n)})$  be defined as in (A-42). Then (A-45) holds whenever  $Q$  additionally satisfies  $\theta(Q) = 0$ .*

*Proof.* Let  $g = G_n$ . Note that (A-45) equals

$$\sqrt{n} \left( \frac{1}{n_1} \sum_{i=1}^n (V_{gg_0(i)} - Y_{g(i)}) A_i - \frac{1}{n_0} \sum_{i=1}^n (V_{gg_0(i)} - Y_{g(i)}) (1 - A_i) \right). \quad (\text{B-82})$$

Since

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

we have that (B-82) equals

$$\begin{aligned} \sqrt{n} \left( \frac{1}{n_1} \sum_{i=1}^n (V_{gg_0(i)} - Y_{g(i)}(1)) A_{g(i)} A_i + \frac{1}{n_1} \sum_{i=1}^n (V_{gg_0(i)} - Y_{g(i)}(0)) (1 - A_{g(i)}) A_i \right. \\ \left. - \frac{1}{n_0} \sum_{i=1}^n (V_{gg_0(i)} - Y_{g(i)}(1)) A_{g(i)} (1 - A_i) - \frac{1}{n_0} \sum_{i=1}^n (V_{gg_0(i)} - Y_{g(i)}(0)) (1 - A_{g(i)}) (1 - A_i) \right). \quad (\text{B-83}) \end{aligned}$$

By construction, all but at most  $K_n$  of the terms in the four summations in (B-83) must be identically equal to zero. Moreover, conditionally on  $g$ ,  $g_0$ ,  $A^{(n)}$  and  $K_n$ , (B-83) has mean equal to zero. This follows from the fact that  $E[V_i] = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_0$  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-83) has variance bounded above by

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

where

$$M = 4 \max\{\text{Var}[V_i - Y_i(1)], \text{Var}[V_i - Y_i(0)]\} . \quad (\text{B-84})$$

Lemma B.5 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-45) also follows. ■

**Lemma B.5.** *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 4.3. Then,*

$$\frac{K_n}{n} \xrightarrow{P} 0 . \quad (\text{B-85})$$

*Proof.* The argument provided here follows closely arguments in Chung and Romano (2013). For each  $s$ , let  $N_a(s) = |\{C_j = a : j = 1, \dots, n(s)\}|$  for  $C_j$  as in the proof of Theorem 4.3. 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-85), it suffices to show that for all  $s \in \mathcal{S}$  and  $a \in \{0, 1\}$ ,

$$\frac{n_a(s) - N_a(s)}{n} \xrightarrow{P} 0 . \quad (\text{B-86})$$

To this end, write

$$\begin{aligned} \frac{n_a(s) - N_a(s)}{n} &= \left( \frac{n_a(s)}{n} - \frac{n(s)}{2n} \right) - \left( \frac{N_a(s)}{n} - \frac{n(s)}{2n} \right) \\ &= \frac{1}{2} \frac{D_n(s)}{n} - \frac{n(s)}{n} \left( \frac{N_a(s)}{n(s)} - \frac{1}{2} \right) . \end{aligned}$$

Under our assumptions,  $\frac{D_n(s)}{n} \xrightarrow{P} 0$  and  $\frac{n(s)}{n} \xrightarrow{P} p(s)$ . Note that for each  $a \in \{0, 1\}$  and  $s \in \mathcal{S}$ ,  $N_a(s)|n(s)$  is distributed according to a binomial distribution with  $n(s)$  trials and probability of success equal to  $\frac{1}{2}$ . It therefore follows from Chebychev's inequality (applied conditionally) that

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

It follows that

$$\frac{N_a(s)}{n(s)} - \frac{1}{2} \xrightarrow{P} 0 ,$$

from which (B-86) follows. ■

**Lemma B.6.** *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  $S^{(n)}$ . Suppose*

$$(T_n(G_n X^{(n)}), T_n(G'_n X^{(n)})) \xrightarrow{d} (T, T') , \quad (\text{B-87})$$

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

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

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-87). It therefore suffices to show that  $\text{Var}[\hat{R}_n(t)] \rightarrow 0$ . Equivalently, it is enough to show that  $E[\hat{R}_n^2(t)] \rightarrow R^2(t)$ . To this end, note that

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

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-87). ■

**Lemma B.7.** *Suppose  $Q$  satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Let  $\gamma = (\beta, \delta_1, \dots, \delta_{|\mathcal{S}|})'$  be the parameters in the regression (29) and let  $\hat{\gamma}_n$  be the least squares estimator of  $\gamma$ . Then,*

$$\hat{\gamma}_n \xrightarrow{P} \gamma \equiv \begin{bmatrix} \theta(Q) \\ \mu_0 + \frac{1}{2}E[m_1(Z_i)|S_i = 1] + \frac{1}{2}E[m_0(Z_i)|S_i = 1] \\ \vdots \\ \mu_0 + \frac{1}{2}E[m_1(Z_i)|S_i = |\mathcal{S}|] + \frac{1}{2}E[m_0(Z_i)|S_i = |\mathcal{S}|] \end{bmatrix}.$$

*Proof.* First note that  $\hat{\gamma}_n = (\mathbf{C}'_n \mathbf{C}_n)^{-1} \mathbf{C}'_n \mathbb{Y}_n$ , where  $\mathbf{C}_n$  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  $\frac{1}{n} \mathbf{C}'_n \mathbb{Y}_n$  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

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n I\{S_i = s\} Y_i &= \frac{n_1(s)}{n} (\mu_1 + E[m_1(Z_i)|S_i = s]) + \left(\frac{n(s)}{n} - \frac{n_1(s)}{n}\right) (\mu_0 + E[m_0(Z_i)|S_i = s]) \\
&\quad + \frac{1}{n} \sum_{i=1}^n A_i I\{S_i = s\} \tilde{Y}_i(1) + \frac{1}{n} \sum_{i=1}^n (1 - A_i) I\{S_i = s\} \tilde{Y}_i(0) \\
&= p(s) \left( \frac{1}{2}(\mu_1 + E[m_1(Z_i)|S_i = s]) + \frac{1}{2}(\mu_0 + E[m_0(Z_i)|S_i = s]) \right) + o_P(1) ,
\end{aligned}$$

where in the last step we used  $\frac{n_1(s)}{n} = \frac{D_n(s)}{n} + \frac{1}{2} \frac{n(s)}{n} \xrightarrow{P} \frac{p(s)}{2}$ ,  $\frac{n(s)}{n} \xrightarrow{P} p(s)$ , 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.2. Also by Lemma B.2,

$$\frac{1}{n} \sum_{i=1}^n A_i Y_i = \frac{1}{n} \sum_{i=1}^n A_i Y_i(1) \xrightarrow{P} \frac{1}{2} \mu_1 ,$$

so that we conclude that

$$\frac{1}{n} \mathbb{C}'_n \mathbb{Y}_n \xrightarrow{P} \begin{bmatrix} \frac{1}{2} \mu_1 \\ p(1) \left( \frac{1}{2} (\mu_1 + E[m_1(Z_i)|S_i = 1]) + \frac{1}{2} (\mu_0 + E[m_0(Z_i)|S_i = 1]) \right) \\ \vdots \\ p(|\mathcal{S}|) \left( \frac{1}{2} (\mu_1 + E[m_1(Z_i)|S_i = 1]) + \frac{1}{2} (\mu_0 + E[m_0(Z_i)|S_i = |\mathcal{S}|]) \right) \end{bmatrix} .$$

The result then follows from the above display, (A-54), and some additional algebra. ■

**Lemma B.8.** *Suppose  $Q$  satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Let  $C_i \equiv [A_i, I\{S_i = 1\}, \dots, I\{S_i = |\mathcal{S}|\}]'$  be the  $i$ th 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 (29), and  $\hat{\gamma}_n$  be the least squares estimator of the regression coefficients  $\gamma = (\beta, \delta_1, \dots, \delta_{|\mathcal{S}|})'$ . Then,*

$$\hat{u}_i = \sum_{s \in \mathcal{S}} I\{S_i = s\} A_i^* E[m_1(Z_i) - m_0(Z_i)|S_i = s] + \tilde{Y}_i(1) A_i + \tilde{Y}_i(0)(1 - A_i) + C_i(\gamma - \hat{\gamma}_n) . \quad (\text{B-88})$$

Furthermore,

$$\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 \xrightarrow{P} \frac{1}{4} (\varsigma_H^2 + \varsigma_Y^2) \quad (\text{B-89})$$

$$\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 A_i \xrightarrow{P} \frac{1}{8} \varsigma_H^2 + \frac{1}{2} \sigma_{\tilde{Y}(1)}^2 \quad (\text{B-90})$$

$$\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 I\{S_i = s\} \xrightarrow{P} p(s) \left( \frac{1}{4} (E[m_1(Z_i) - m_0(Z_i)|S_i = s])^2 + \frac{1}{2} \sigma_{\tilde{Y}(1)}^2(s) + \frac{1}{2} \sigma_{\tilde{Y}(0)}^2(s) \right) \quad (\text{B-91})$$

$$\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 I\{S_i = s\} A_i \xrightarrow{P} p(s) \left( \frac{1}{8} (E[m_1(Z_i) - m_0(Z_i)|S_i = s])^2 + \frac{1}{2} \sigma_{\tilde{Y}(1)}^2(s) \right) . \quad (\text{B-92})$$

*Proof.* Consider the following derivation:

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

Using Lemma B.7, we have that

$$C_i \gamma = \theta(Q)A_i + \mu_0 + \sum_{s \in \mathcal{S}} I\{S_i = s\} \left( \frac{1}{2} E[m_1(Z_i)|S_i = s] + \frac{1}{2} E[m_0(Z_i)|S_i = s] \right) .$$

Hence,

$$\begin{aligned} u_i &= Y_i - C_i \gamma \\ &= \sum_{s \in \mathcal{S}} I\{S_i = s\} A_i^* E[m_1(Z_i) - m_0(Z_i) | S_i = s] + \tilde{Y}_i(1) A_i + \tilde{Y}_i(0) (1 - A_i) . \end{aligned} \quad (\text{B-93})$$

Since  $\hat{u}_i = u_i + C_i(\gamma - \hat{\gamma}_n)$ , the desired expression for (B-88) follows.

To prove (B-89) - (B-92), note that for any univariate random variable  $X_i$  such that

$$\frac{1}{n} \sum_{i=1}^n (C_i' C_i \otimes X_i) = O_P(1) \text{ and } \frac{1}{n} \sum_{i=1}^n C_i u_i X_i = O_P(1) , \quad (\text{B-94})$$

we have that

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

where in the second equality we used  $\hat{\gamma}_n \xrightarrow{P} \gamma$  from Lemma B.7. Since the condition in (B-94) 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-89) - (B-92) holds with  $u_i^2$  in place of  $\hat{u}_i^2$ . Using (B-93), we have that

$$\begin{aligned} u_i^2 &= \sum_{s \in \mathcal{S}} \frac{1}{4} I\{S_i = s\} (E[m_1(Z_i) - m_0(Z_i) | S_i = s])^2 + \tilde{Y}_i(1)^2 A_i + \tilde{Y}_i(0)^2 (1 - A_i) \\ &\quad + \sum_{s \in \mathcal{S}} I\{S_i = s\} E[m_1(Z_i) - m_0(Z_i) | S_i = s] \tilde{Y}_i(1) A_i \\ &\quad - \sum_{s \in \mathcal{S}} I\{S_i = s\} E[m_1(Z_i) - m_0(Z_i) | S_i = s] \tilde{Y}_i(0) (1 - A_i) . \end{aligned}$$

Lemma B.2 thus implies that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n u_i^2 &\xrightarrow{P} \frac{1}{4} \sum_{s \in \mathcal{S}} p(s) (E[m_1(Z_i) - m_0(Z_i) | S_i = s])^2 + \frac{1}{2} \sigma_{\tilde{Y}(1)}^2 + \frac{1}{2} \sigma_{\tilde{Y}(0)}^2 \\ &= \frac{1}{4} (\varsigma_H^2 + \varsigma_Y^2) \\ \frac{1}{n} \sum_{i=1}^n u_i^2 A_i &\xrightarrow{P} \frac{1}{8} \sum_{s \in \mathcal{S}} p(s) (E[m_1(Z_i) - m_0(Z_i) | S_i = s])^2 + \frac{1}{2} \sigma_{\tilde{Y}(1)}^2 \\ &= \frac{1}{8} \varsigma_H^2 + \frac{1}{2} \sigma_{\tilde{Y}(1)}^2 + o_P(1) \\ \frac{1}{n} \sum_{i=1}^n u_i^2 I\{S_i = s\} &\xrightarrow{P} \frac{1}{4} p(s) (E[m_1(Z_i) - m_0(Z_i) | S_i = s])^2 + \frac{1}{2} \sigma_{\tilde{Y}(1)}^2(s) + \frac{1}{2} \sigma_{\tilde{Y}(0)}^2(s) \\ \frac{1}{n} \sum_{i=1}^n u_i^2 I\{S_i = s\} A_i &\xrightarrow{P} \frac{1}{8} p(s) (E[m_1(Z_i) - m_0(Z_i) | S_i = s])^2 + \frac{1}{2} \sigma_{\tilde{Y}(1)}^2(s) , \end{aligned}$$

thus completing the proof. ■

**Lemma B.9.** *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 theorem 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.10.** *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

$$\left\{ \frac{D_n(s)}{\sqrt{n}} | S^{(n)} \right\} \xrightarrow{d} N \left( 0, \frac{1}{4(1 - 4\varphi'(0))} \right) \text{ a.s.} \quad (\text{B-95})$$

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.11.** *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 expensive treatment assignment. In progress.
- CALLEN, M., GULZAR, S., HASANAIN, A., KHAN, Y. and REZAEI, A. (2015). Personalities and public sector performance: Evidence from a health experiment in Pakistan. Tech. rep., Working paper.
- 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.
- JANSSEN, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized Behrens-Fisher problem. *Statistics & probability letters*, **36** 9–21.

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