Inference in Experiments with Matched Pairs[*]

Yuehao Bai
Department of Economics
University of Michigan
yuehaob@umich.edu

Joseph P. Romano
Departments of Economics and Statistics
Stanford University
romano@stanford.edu

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

January 15, 2021

Abstract

This paper studies inference for the average treatment effect in randomized controlled trials where treatment status is determined according to a "matched pairs" design. By a "matched pairs" design, we mean that units are sampled i.i.d. from the population of interest, paired according to observed, baseline covariates and finally, within each pair, one unit is selected at random for treatment. This type of design is used routinely throughout the sciences, but fundamental questions about its implications for inference about the average treatment effect remain. The main requirement underlying our analysis is that pairs are formed so that units within pairs are suitably "close" in terms of the baseline covariates, and we develop novel results to ensure that pairs are formed in a way that satisfies this condition. Under this assumption, we show that, for the problem of testing the null hypothesis that the average treatment effect equals a pre-specified value in such settings, the commonly used two-sample t-test and "matched pairs" t-test are conservative in the sense that these tests have limiting rejection probability under the null hypothesis no greater than and typically strictly less than the nominal level. We show, however, that a simple adjustment to the standard errors of these tests leads to a test that is asymptotically exact in the sense that its limiting rejection probability under the null hypothesis equals the nominal level. We also study the behavior of randomization tests that arise naturally in these types of settings. When implemented appropriately, we show that this approach also leads to a test that is asymptotically exact in the sense described previously, but additionally has finite-sample rejection probability no greater than the nominal level for certain distributions satisfying the null hypothesis. A simulation study and empirical application confirm the practical relevance of our theoretical results.

KEYWORDS: Experiment, matched pairs, matched pairs t-test, permutation test, randomized controlled trial, treatment assignment, two-sample t-test

JEL classification codes: C12, C14

[*] We thank John Duchi for references to the "blossom" algorithm and Panos Toulis for helpful comments. We thank Silvia Barbareschi for excellent research assistance. The research of the second author is supported by NSF Grant MMS-1949845; the research of the third author is supported by NSF Grant SES-1530661.

1 Introduction

This paper studies inference for the average treatment effect in randomized controlled trials where treatment status is determined according to a "matched pairs" design. By a "matched pairs" design, we mean that units are sampled i.i.d. from the population of interest, paired according to observed, baseline covariates and finally, within each pair, one unit is selected at random for treatment. This method is used routinely in all parts of the sciences. Indeed, commands to facilitate its implementation are included in popular software packages, such as `sampsi` in Stata. References to a variety of specific examples can be found, for instance, in the following surveys of various field experiments: Riach and Rich (2002), List and Rasul (2011), White (2013), Crépon et al. (2015), Bertrand and Duflo (2017), and Heard et al. (2017). See also Bruhn and McKenzie (2009), who, based on a survey of selected development economists, report that 56% of researchers have used such a design at some point. Despite the widespread use of "matched pairs" designs, fundamental questions about its implications for inference about the average treatment effect remain. The main requirement underlying our analysis is that pairs are formed so that units within pairs are suitably "close" in terms of the baseline covariates. We develop novel results to ensure that pairs are formed in a way that satisfies this condition. See, in particular, Theorems 4.1–4.3 below. Under this assumption, we derive a variety of results pertaining to the problem of testing the null hypothesis that the average treatment effect equals a pre-specified value in such settings.

We first study the behavior of the two-sample t-test and "matched pairs" t-test, which are both used routinely in the analysis of this type of data. Several specific references are provided in Sections 3.1 and 3.2 below. Our first pair of results establish that these commonly used tests are conservative in the sense that these tests have limiting rejection probability under the null hypothesis no greater than and typically strictly less than the nominal level. For each of these tests, we additionally provide a characterization of when the limiting rejection probability under the null hypothesis is in fact strictly less than the nominal level. 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 the "matched pairs" design. We show, however, that a simple adjustment to the usual standard error of these tests leads to a test that is asymptotically exact in the sense that its limiting rejection probability under the null hypothesis equals the nominal level. We refer to this test as the "adjusted" t-test.

Next, we study the behavior of some randomization tests that arise naturally in these types of settings. More specifically, we study randomization tests based on the idea of permuting only treatment status for units within pairs. When implemented with a suitable choice of test statistic, specifically the test statistic employed in the aforementioned "adjusted" t-test, we show that this approach also leads to a test that is asymptotically exact in the sense described previously. We emphasize, however, that this result relies heavily upon the choice of test statistic. Indeed, as explained further in Remark 3.16, when implemented with other choices of test statistics, randomization tests may behave in large samples like the "matched pairs" t-test described above. On the other hand, regardless of the specific way in which they are implemented, these tests have the attractive feature that they have finite-sample rejection probability no greater than the nominal level for certain distributions satisfying the null hypothesis. We highlight these properties in a simulation study.

The analysis of data from experiments with matched pairs has received considerable attention recently. Much of this literature differs from the present paper in at least one of two important ways: first, the parameter of interest is in many instances not the average treatment effect, but instead the sample average treatment effect or the conditional average treatment effect; second, the sampling scheme differs from ours in that the pairs of units are sampled rather than the units, which are then subsequently paired. We emphasize that the subsequent pairing of the sampled units significantly complicates the analysis in that pairing may only be approximate in the case of continuous covariates and also leads to dependence in the pairs themselves, which must be taken into account, as done in our analysis below. Indeed, Athey and Imbens (2017) advocate against the use of matched pair designs due to difficulties with consistent estimation of the appropriate variance stemming from these complications, which we overcome in our analysis. Examples of papers where at least one of these two distinctions are present include Abadie and Imbens (2008), Imai (2008), Ding (2017) and Fogarty (2018a,b). Remarks 3.7–3.9 below provide further discussion of some of these important, related contributions. A notable exception to this characterization of the literature is van der Laan et al. (2012), who consider a more general framework that includes ours as a special case. By specializing his results to our setting, it is possible to recover results related to some of ours, especially those pertaining to the conservativeness of the "matched pairs" t-test and the possibility of adjusting it to obtain an asymptotically exact test. Even for these results, we note that there are important distinctions between our results and theirs. Furthermore, there is no analysis of the usual two-sample t-test or randomization tests of any kind. We elaborate on these points in Remark 3.12 below.

The remainder of the paper is organized as follows. In Section 2, we describe our setup and notation. In particular, there we describe the precise sense in which we require that units in each pair are "close" in terms of their baseline covariates. Our main results concerning the two-sample t-test, the "matched pairs" t-test, "adjusted" t-test and randomization tests are contained in Section 3. In Section 4, we develop some results that ensure that units in each pair are suitably "close" in terms of their baseline covariates. In Section 5, we examine the finite-sample behavior of these tests via a small simulation study. In Section 6, we provide a brief empirical illustration of our proposed tests using data from an experiment replicating one of the arms in DellaVigna and Pope (2018). Finally, we conclude in Section 7 with some recommendations for empirical practice guided by both our theoretical results and our simulation study. As explained further in that section, for the testing problem considered here, we do not recommend the use of the two-sample t-test or "matched pairs" t-test because they are conservative in the sense described above; we instead encourage the use of the "adjusted" t-test or randomization tests that employ the same test statistic because they are asymptotically exact, and, as a result, considerably more powerful. Proofs of all results are provided in the Supplemental Appendix.

2 Setup and Notation

Let $Y_i \in \mathbf{R}$ denote the (observed) outcome of interest for the ith unit, $D_i \in \{0,1\}$ be an indicator for whether the ith unit is treated, and $X_i \in \mathbf{R}^k$ denote observed, baseline covariates for the ith unit. Further denote by $Y_i(1)$ the potential outcome of the ith unit if treated and by $Y_i(0)$ the potential outcome of the ith unit if not treated. As usual, the (observed) outcome and potential outcomes are related to treatment status by the relationship

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

For a random variable indexed by i, A_i, it will be useful to denote by $A^{(n)}$ the random vector $(A_1, \ldots, A_{2n})$. Denote by P_n the distribution of the observed data $Z^{(n)}$, where $Z_i = (Y_i, D_i, X_i)$, and by Q_n the distribution of $W^{(n)}$, where $W_i = (Y_i(1), Y_i(0), X_i)$. Note that P_n is jointly determined by (1), Q_n, and the mechanism for determining treatment assignment. We assume throughout that $W^{(n)}$ consists of $2n$ i.i.d. observations, i.e., $Q_n = Q^{2n}$, where Q is the marginal distribution of W_i. We therefore state our assumptions below in terms of assumptions on Q and the mechanism for determining treatment assignment. Indeed, we will not make reference to P_n in the sequel and all operations are understood to be under Q and the mechanism for determining treatment assignment.

Our object of interest is the average effect of the treatment on the outcome of interest, which may be expressed in terms of this notation as

$$\Delta(Q) = E[Y_i(1) - Y_i(0)] \ . \tag{2}$$

For a pre-specified choice of Δ_0, the testing problem of interest is

$$H_0 : \Delta(Q) = \Delta_0 \text{ versus } H_1 : \Delta(Q) \neq \Delta_0 \tag{3}$$

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

We now describe our assumptions on Q. We restrict Q to satisfy the following mild requirement:

Assumption 2.1. The distribution Q is such that

(a) $0 < E[\text{Var}[Y_i(d)|X_i]]$ for $d \in \{0,1\}$.

(b) $E[Y_i^2(d)] < \infty$ for $d \in \{0,1\}$.

(c) $E[Y_i(d)|X_i = x]$ and $E[Y_i^2(d)|X_i = x]$ are Lipschitz for $d \in \{0,1\}$.

Assumptions 2.1(a)–(b) are mild restrictions imposed, respectively, to rule out degenerate situations and to permit the application of suitable laws of large numbers and central limit theorems. See, in particular, Lemma S.1.3 in the Supplemental Appendix for a novel law of large numbers for independent and non-identically distributed random variables that is useful in establishing our results. Assumption 2.1(c), on the other hand, is a smoothness requirement that ensures that units that are "close" in terms of their baseline covariates are suitably comparable. Such smoothness requirements have been employed in establishing some types of optimality of "matched pairs" designs. See, in particular, Kallus (2018) and Bai (2020).

Next, we describe our assumptions on the mechanism determining treatment assignment. In order to describe these assumptions more formally, we require some further notation to define the relevant pairs of units. The n pairs may be represented by the sets

$$\{\pi(2j-1), \pi(2j)\} \text{ for } j = 1, \ldots, n \ ,$$

where $\pi = \pi_n(X^{(n)})$ is a permutation of $2n$ elements. Because of its possible dependence on $X^{(n)}$, π encompasses a broad variety of different ways of pairing the $2n$ units according to the observed, baseline covariates $X^{(n)}$. Given such a π, we assume that treatment status is assigned as described in the following assumption:

Assumption 2.2. Treatment status is assigned so that $(Y^{(n)}(1), Y^{(n)}(0)) \perp\!\!\!\perp D^{(n)} \big| X^{(n)}$ and, conditional on $X^{(n)}$, $(D_{\pi(2j-1)}, D_{\pi(2j)}), j = 1, \ldots, n$ are i.i.d. and each uniformly distributed over the values in $\{(0,1), (1,0)\}$.

Our analysis will require some discipline on the way in which the pairs are formed. In particular, we will require that the units in each pair are "close" in terms of their baseline covariates in the sense described by the following assumption:

Assumption 2.3. The pairs used in determining treatment status satisfy

$$\frac{1}{n} \sum_{1 \le j \le n} |X_{\pi(2j)} - X_{\pi(2j-1)}|^r \xrightarrow{P} 0$$

for $r = 1$ and $r = 2$.

It will at times be convenient to require further that units in consecutive pairs are also "close" in terms of their baseline covariates. One may view this requirement, which is formalized in the following assumption, as "pairing the pairs" so that they are "close" in terms of their baseline covariates.

Assumption 2.4. The pairs used in determining treatment status satisfy

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

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

In Section 4 below, we provide results to facilitate constructing pairs satisfying Assumptions 2.3–2.4 under weak assumptions on Q. We emphasize, however, that Assumption 2.4, in contrast to Assumptions 2.1–2.3, will not be required for many of our results. Furthermore, given pairs satisfying Assumption 2.3, it will frequently be possible to "re-order" them so that Assumption 2.4 is satisfied. See Theorem 4.3 below for further details.

Remark 2.1. At the expense of some additional notation, it is straightforward to allow π to depend further on a uniform random variable U that is independent of $(Y^{(n)}(1), Y^{(n)}(0), X^{(n)})$, but we do not pursue this generalization here. ■

Remark 2.2. The treatment assignment scheme described in this section is an example of what is termed in some parts of the literature as a covariate-adaptive randomization scheme, in which treatment status is assigned so as to "balance" units assigned to treatment and the units assigned to control in terms of their baseline covariates. For a review of these types of treatment assignment schemes focused on their use in clinical trials, see Rosenberger and Lachin (2015). In some such schemes, units are sampled i.i.d. from the population of interest, stratified into a finite number of strata according to observed, baseline covariates, and finally, within each stratum, treatment status is assigned so as to achieve "balance" within each stratum. For instance, within each stratum, a researcher may assign (uniformly) at random half of the units to treatment and the remainder to control. Bugni et al. (2018, 2019) develop a variety of results pertaining to these ways of assigning treatment status, but their analysis relies heavily upon the requirement that the units are stratified using the baseline covariates into only a finite number of strata. As a result, their framework cannot accommodate "matched pairs" designs, where the number of strata is equal to the number of pairs and therefore proportional to the sample size. ■

3 Main Results

3.1 Two-Sample t-Test

In this section, we consider using the two-sample t-test to test (3) at level $\alpha \in (0,1)$. In order to define this test, for $d \in \{0,1\}$, define

$$\hat{\mu}_n(d) = \frac{1}{n} \sum_{1 \le i \le 2n : D_i = d} Y_i \tag{4}$$

$$\hat{\sigma}_n^2(d) = \frac{1}{n} \sum_{1 \le i \le 2n : D_i = d} (Y_i - \hat{\mu}_n(d))^2 \tag{5}$$

and let

$$\hat{\Delta}_n = \hat{\mu}_n(1) - \hat{\mu}_n(0) \ . \tag{6}$$

The two-sample t-test is given by

$$\phi_n^{t-\text{test}}(Z^{(n)}) = I\{|T_n^{t-\text{test}}(Z^{(n)})| > z_{1-\frac{\alpha}{2}}\} \ , \tag{7}$$

where

$$T_n^{t-\text{test}}(Z^{(n)}) = \frac{\sqrt{n}(\hat{\Delta}_n - \Delta_0)}{\sqrt{\hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0)}} \tag{8}$$

and $z_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution. While its properties are far from clear in our setting, this classical test is used routinely in the analysis of such data. See, for example, Riach and Rich

(2002), Gelman and Hill (2006, page 174), Duflo et al. (2007), Bertrand and Duflo (2017) and the references therein. See also Imai et al. (2009) for the use of an analgous test in a setting with cluster-level treatment assignment.

The following theorem establishes the asymptotic behavior of the two-sample t-statistic defined in (8) and, as a consequence, the two-sample t-test defined in (7). 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 3.1. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.3. Then,*

$$\frac{\sqrt{n}(\hat{\Delta}_n - \Delta(Q))}{\sqrt{\hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0)}} \xrightarrow{d} N(0, \varsigma^2_{t-test}) \ , \tag{9}$$

where

$$\varsigma^2_{t-test} = 1 - \frac{1}{2}\frac{E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) + (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right]}{Var[Y_i(1)] + Var[Y_i(0)]} \ .$$

Thus, for the problem of testing (3) at level $\alpha \in (0,1)$, $\phi_n^{t-test}(Z^{(n)})$ defined in (7) satisfies

$$\lim_{n\to\infty} E[\phi_n^{t-test}(Z^{(n)})] = P\{\varsigma_{t-test}\,|G| > z_{1-\frac{\alpha}{2}}\} \leq \alpha \ , \tag{10}$$

where $G \sim N(0,1)$, whenever Q additionally satisfies the null hypothesis, i.e., $\Delta(Q) = \Delta_0$. Furthermore, the inequality in (10) is strict unless

$$E[Y_i(1) + Y_i(0)] = E[Y_i(1) + Y_i(0)|X_i] \tag{11}$$

with probability one under Q.

Remark 3.1. Theorem 3.1 shows that the limiting rejection probability of the two-sample t-test under the null hypothesis is strictly less than the nominal level unless the baseline covariates are irrelevant for potential outcomes in the sense described by (11). We note that the conservativeness of the two-sample t-test is mentioned in Athey and Imbens (2017), but without any formal results. The magnitude of the difference between the limiting rejection probability and the nominal level, however, will depend further on Q through the value of $\varsigma^2_{t\text{−test}}$. In our simulation study in Section 5, we find that the rejection probability can be severely less than the nominal level and that this difference translates into significant power losses when compared with tests studied below that are asymptotically exact in the sense that they have limiting rejection probability under the null hypothesis equal to the nominal level. ■

Remark 3.2. In our definition of the two-sample t-test above, we have used the unpooled estimator of the variance rather than the pooled estimator of the variance. Using Lemma S.1.5 in the Supplemental Appendix, it is straightforward to show that the pooled estimator of the variance tends in probability to

$$\frac{\text{Var}[Y_i(1)] + \text{Var}[Y_i(0)]}{2} + \frac{(E[Y_i(1)] - E[Y_i(0)])^2}{4} \ .$$

From this and Lemma S.1.4 in the Supplemental Appendix, it is possible to deduce that with this choice of an estimator of the variance the test may even have limiting rejection probability under the null hypothesis that strictly exceeds the nominal level. ■

3.2 "Matched Pairs" t-Test

Instead of the two-sample t-test studied in the preceding section, it is often recommended to use a "matched pairs" t-test when analyzing such data, which treats the differences of the outcomes within a pair as the observations. This test is also sometimes referred to as the "paired difference-of-means" test. For some examples of its use, see Athey and Imbens (2017), Hsu and Lachenbruch (2007), and Armitage et al. (2008). Formally, this test is given by

$$\phi_n^{\text{paired}}(Z^{(n)}) = I\{|T_n^{\text{paired}}(Z^{(n)})| > z_{1-\frac{\alpha}{2}}\} \ , \tag{12}$$

where

$$T_n^{\text{paired}}(Z^{(n)}) = \frac{\sqrt{n}(\hat{\Delta}_n - \Delta_0)}{\sqrt{\frac{1}{n}\sum_{1 \le j \le n}(Y_{\pi(2j)} - Y_{\pi(2j-1)})^2 - \hat{\Delta}_n^2}} \tag{13}$$

and, as before, $z_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution. Again, despite its widespread use, the properties of this test are not transparent in our setting.

The following theorem describes the asymptotic behavior of the "matched pairs" t-statistic defined in (13), and, as a consequence, the "matched pairs" t-test defined in (12). The theorem shows, in particular, that the behavior of the "matched pairs" t-test is qualitatively similar to that of the two-sample t-test studied in the preceding section.

Theorem 3.2. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.3. Then,*

$$\frac{\sqrt{n}(\hat{\Delta}_n - \Delta(Q))}{\sqrt{\frac{1}{n}\sum_{1 \le j \le n}(Y_{\pi(2j)} - Y_{\pi(2j-1)})^2 - \hat{\Delta}_n^2}} \xrightarrow{d} N(0, \varsigma_{\text{paired}}^2) \ , \tag{14}$$

where

$$\varsigma_{paired}^2 = 1 - \frac{1}{2}\frac{E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) - (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right]}{\begin{array}{l}\Big(E[\mathit{Var}[Y_i(1)|X_i]] + E[\mathit{Var}[Y_i(0)|X_i]] \\ \qquad + E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) - (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right]\Big)\end{array}} \ .$$

Thus, for the problem of testing (3) at level $\alpha \in (0,1)$, $\phi_n^{paired}(Z^{(n)})$ defined in (12) satisfies

$$\lim_{n \to \infty} E[\phi_n^{paired}(Z^{(n)})] = P\{\varsigma_{paired}\,|G| > z_{1-\frac{\alpha}{2}}\} \le \alpha \ , \tag{15}$$

where $G \sim N(0,1)$, whenever Q additionally satisfies the null hypothesis, i.e., $\Delta(Q) = \Delta_0$. Furthermore, *the inequality in (15) is strict unless*

$$E[Y_i(1) - Y_i(0)] = E[Y_i(1) - Y_i(0)|X_i] \tag{16}$$

with probability one under Q.

Remark 3.3. While Theorem 3.2 is qualitatively similar to Theorem 3.1, it is worth emphasizing the difference between (11) and (16). Both conditions determine a sense in which the baseline covariates are irrrelevant for potential outcomes, but the latter condition holds, in particular, whenever the treatment effect $Y_i(1) - Y_i(0)$ is constant. ■

Remark 3.4. The test statistic in (13) is particularly convenient for the purposes of constructing a confidence interval for $\Delta(Q)$, but we note that it is possible to studentize differently if one is only interested in testing (3). In particular, the result in (15) continues to hold for the test formed by replacing the $\hat{\Delta}_n$ in the denominator on the right-hand side of (13) with Δ_0. On the other hand, since $\hat{\Delta}_n$ solves

$$\min_{\Delta} \frac{1}{n} \sum_{1 \leq j \leq n} (Y_{\pi(2j)} - Y_{\pi(2j-1)} - \Delta)^2 ,$$

doing so would lead to a test that is less powerful than the one considered here. ■

Remark 3.5. In the context of observational studies under an unconfoundedness assumption, Abadie and Imbens (2012) also analyze the left-hand side of (14); see, in particular, Theorem 1 and equation (5) in their paper when $M = 1$. In their setting, this quantity converges in distribution to a standard normal distribution. The difference between their result and ours above seems striking when one observes that their framework allows each treated unit to be "matched" to a control unit in a way that satisfies our Assumption 2.3. We emphasize, however, that, in contrast to the setting considered here, treatment status in their framework is i.i.d. This important difference significantly complicates the analysis and explains the diverging results for the same quantities. ■

Remark 3.6. The literature has also at times advocated estimation of $\Delta(Q)$ via estimation by ordinary least squares of the coefficient on D_i in

$$Y_i = \beta D_i + \sum_{1 \leq j \leq n} \lambda_j I\{i \in \{\pi(2j), \pi(2j-1)\}\} + \epsilon_i \ . \tag{17}$$

See, for example, Duflo et al. (2007) and Glennerster and Takavarasha (2013, page 363) as well as Crépon et al. (2015), who estimate $\Delta(Q)$ in the same way, but in a setting with cluster-level treatment assignment. In our setting, it is straightforward to see that the ordinary least squares estimator of β in (17) equals $\hat{\Delta}_n$. It is also possible to show that the usual heteroskedasticity-consistent estimator variance equals

$$\frac{1}{n} \sum_{1 \leq j \leq n} (Y_{\pi(2j)} - Y_{\pi(2j-1)})^2 - \hat{\Delta}_n^2 \ .$$

Hence, the resulting test is identical to the "matched pairs" t-test studied in this section. ■

Remark 3.7. In an asymptotic framework that differs from ours in that pairs of units are sampled rather than the units themselves, Fogarty (2018a) considers the use of the "matched pairs" t-test considered here for inference about the sample average treatment effect as well as the average treatment effect. For inference about the average treatment effect, it is clear that in this different asymptotic framework, the "matched pairs" t-test would in fact be asymptotically exact. In the case of the sample average treatment effect, it is generally conservative. Fogarty (2018a) therefore further suggests some improved tests for the case of the sample average treatment effect, but he notes that those improvements do not carry over to the average treatment effect. See Section 5 of Fogarty (2018a) for further discussion. ■

Remark 3.8. In an asymptotic framework that again differs from ours in that pairs of units are sampled rather than the units themselves, Fogarty (2018b) considers various "regression-assisted" estimators for different treatment effect parameters. For the average treatment effect, he considers, instead of $\hat{\Delta}_n$, the estimator of $\Delta(Q)$ given by the ordinary least squares estimator of the intercept parameter in a linear regression of $(Y_{\pi(2j)} - Y_{\pi(2j-1)})(D_{\pi(2j)} - D_{\pi(2j-1)})$ on a constant and $(X_{\pi(2j)} - X_{\pi(2j-1)})(D_{\pi(2j)} - D_{\pi(2j-1)})$. Denote this estimator by $\hat{\alpha}_n$ and by $\hat{\beta}_n$ the corresponding estimator of the slope parameters. In Section S.1.1 of the Supplemental Appendix, we show that quite generally

$$\sqrt{n}(\hat{\alpha}_n - \Delta(Q)) = \sqrt{n}(\hat{\Delta}_n - \Delta(Q)) + o_P(1) \ . \tag{18}$$

In this sense, there is no benefit to using such estimators instead of $\hat{\Delta}_n$ in our asymptotic framework. Note, however, that this does not preclude the possibility of improvements from using "regression-assisted" estimators that make use of covariates that are not included in X_i. ■

3.3 "Adjusted" t-Test

The proofs of Theorems 3.1 and 3.2 in the Supplemental Appendix rely upon Lemma S.1.4, which establishes that

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

where

$$\nu^2 = \mathrm{Var}[Y_i(1)] + \mathrm{Var}[Y_i(0)] - \frac{1}{2}E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) + (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right] \ . \tag{19}$$

Using this observation, it is possible to provide an adjustment to these tests that leads to a test that is asymptotically exact in the sense that its limiting rejection probability under the null hypothesis equals the nominal level by providing a consistent estimator of (19). As discussed further in Remark 3.11 below, there exist multiple consistent estimators of (19), but a convenient one for our purposes is given by

$$\hat{\nu}_n^2 = \hat{\tau}_n^2 - \frac{1}{2}(\hat{\lambda}_n^2 + \hat{\Delta}_n^2) \ , \tag{20}$$

where

$$
\begin{aligned}
\hat{\tau}_n^2 &= \frac{1}{n} \sum_{1 \leq j \leq n} (Y_{\pi(2j)} - Y_{\pi(2j-1)})^2 && (21) \\
\hat{\lambda}_n^2 &= \frac{2}{n} \sum_{1 \leq j \leq \lfloor \frac{n}{2} \rfloor} \Big((Y_{\pi(4j-3)} - Y_{\pi(4j-2)})(Y_{\pi(4j-1)} - Y_{\pi(4j)}) \\
&\qquad \times (D_{\pi(4j-3)} - D_{\pi(4j-2)})(D_{\pi(4j-1)} - D_{\pi(4j)}) \Big) \ . && (22)
\end{aligned}
$$

In order to provide some intuition for the form of this estimator, note that the expression in (19) qualitatively involves two types of quantities: quantities like $E[Y_i(1)]$, which can be consistently estimated in a straightforward fashion by averaging across pairs outcomes corresponding to treated observations, as explained in Lemma S.1.5 in the Supplemental Appendix; and quantities like $E[E[Y_i(1)|X_i]E[Y_i(0)|X_i]]$, which are more problematic, but can be consistently estimated by averaging across "pairs of pairs" the product of outcomes corresponding to a treated and untreated observation in adjacent pairs. It is for this reason that the estimator in (20) involves a quantity like (22), which averages across "pairs of pairs." The following theorem shows that the "adjusted" t-test, given by

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

with

$$T_n^{t-\text{test,adj}}(Z^{(n)}) = \frac{\sqrt{n}(\hat{\Delta}_n - \Delta_0)}{\hat{\nu}_n} \ , \qquad (24)$$

satisfies the desired property.

Theorem 3.3. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.4. Then,*

$$\frac{\sqrt{n}(\hat{\Delta}_n - \Delta(Q))}{\hat{\nu}_n} \xrightarrow{d} N(0,1) \ . \qquad (25)$$

Thus, for the problem of testing (3) at level $\alpha \in (0,1)$, $\phi_n^{t-test,adj}(Z^{(n)})$ defined in (23) satisfies

$$\lim_{n \to \infty} E[\phi_n^{t-test,adj}(Z^{(n)})] = \alpha \ , \qquad (26)$$

whenever Q additionally satisfies the null hypothesis, i.e., $\Delta(Q) = \Delta_0$.

Remark 3.9. Note that

$$\hat{\nu}_n^2 = \frac{1}{2}(\hat{\tau}_n^2 - \hat{\Delta}_n^2) + \frac{1}{2}\hat{\zeta}_n^2 \ , \qquad (27)$$

where

$$\hat{\zeta}_n^2 = \frac{1}{n} \sum_{1 \leq j \leq \lfloor \frac{n}{2} \rfloor} \Big((Y_{\pi(4j-3)} - Y_{\pi(4j-2)})(D_{\pi(4j-3)} - D_{\pi(4j-2)}) - (Y_{\pi(4j-1)} - Y_{\pi(4j)})(D_{\pi(4j-1)} - D_{\pi(4j)}) \Big)^2 \ .$$

Since $\hat{\tau}_n^2 \geq \hat{\Delta}_n^2$, this decomposition reveals that $\hat{\nu}_n^2$ is non-negative. We note that the quantity $\hat{\zeta}_n^2$ appears in Abadie and Imbens (2008), who show in an asymptotic framework that differs from ours that it is consistent for $\text{Var}[\hat{\Delta}_n|X^{(n)}]$ under Assumption 2.1(c) and Assumption 2.3. They therefore suggest its use for inference about the conditional average treatment effect. The decomposition in (27) reveals, however, that using this estimator of the variance in place of $\hat{\nu}_n^2$ in (24) would lead to a test that in our asymptotic framework for inference about the average treatment effect may even have limiting rejection probability under the null hypothesis that strictly exceeds the nominal level. ■

Remark 3.10. While our discussion has focused on two-sided null hypotheses as described in (3), the convergence in distribution results described in (9), (14) and (25) have straightforward implications for other tests, such related tests of one-sided null hypotheses. ■

Remark 3.11. As mentioned previously, other consistent estimators of (19) exist. For instance, one may consider the estimator given by

$$\tilde{\nu}_n^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0) - \frac{1}{2}\left(\tilde{\lambda}_n^2 - (\hat{\mu}_n(1) + \hat{\mu}_n(0))^2\right) , \tag{28}$$

where

$$\tilde{\lambda}_n^2 = \frac{2}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} (Y_{\pi(4j-3)} + Y_{\pi(4j-2)})(Y_{\pi(4j-1)} + Y_{\pi(4j)}) .$$

Using arguments similar to those used in establishing Theorem 3.3, it is possible to show that Theorem 3.3 remains true when $\hat{\nu}_n^2$ defined in (20) is replaced by $\tilde{\nu}_n^2$ defined in (28). Indeed, one could use the minimum of multiple consistent estimators of (19). One could even include in the minimum any estimators known to converge in probability to something weakly larger, such as the variance estimators employed in the usual two-sample t-test and the "matched pairs" t-test. ■

Remark 3.12. As mentioned previously, van der Laan et al. (2012) consider a more general framework that includes as a special case the one considered here. When specialized to our setting, their results also enable one to obtain the limit in distribution of $\sqrt{n}(\hat{\Delta}_n - \Delta(Q))$, though the expression they provide for the variance of the limiting distribution differs from ours and, in particular, appears to depend on the sample size (van der Laan et al., 2012, Theorem 3). Using this result, they analyze the behavior of the "matched pairs" t-test and find that it should quite generally be conservative in the sense that it has limiting rejection probability no greater than the nominal level under the null hypothesis. Their analysis, however, relies upon higher-level conditions than ours and, as a consequence, they are unable to articulate the circumstances under which the "matched pairs" t-test has limiting rejection probability under the null hypothesis strictly less than the nominal level as succinctly and clearly as in our Theorem 3.2. They go on to suggest a test like our "adjusted" t-test with limiting rejection probability under the null hypothesis equal to nominal level even when these circumstances hold, though no explicit description or formal results about it are provided. In particular, the description of their test relies upon a consistent estimator of the variance of the limiting distribution of $\sqrt{n}(\hat{\Delta}_n - \Delta(Q))$, which in turn depends upon a "super learner" of a reference distribution that they make use of in their analysis. Finally, these similarities not withstanding, we emphasize that van der Laan et al. (2012) provide no analysis of the usual two-sample t-test or any of the randomization tests considered below. ■

3.4 Randomization Tests

In this section, we study the properties of randomization tests based on the idea of permuting treatment status for units within pairs. For ease of exposition, it is convenient to describe the test for the problem of testing (3) with $\Delta_0 = 0$; for the problem of testing (3) more generally, the construction below may be applied with Y_i replaced with $Y_i - D_i\Delta_0$. See Remark 3.15 below for further details.

In order to describe the test formally, it is useful to introduce some further notation. To this end, denote by $\mathbf{G}_n$ the group of all permutations of $2n$ elements and by $\mathbf{G}_n(\pi)$ the subgroup that only permutes elements within the pairs defined by π, i.e.,

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

Define the action of $g \in \mathbf{G}_n(\pi)$ on $Z^{(n)}$ as follows:

$$gZ^{(n)} = \{(Y_i, D_{g(i)}, X_i) : 1 \leq i \leq 2n\} \ ,$$

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

$$\phi_n^{\text{rand}}(Z^{(n)}) = I\{T_n(Z^{(n)}) > \hat{R}_n^{-1}(1-\alpha)\} \ , \tag{29}$$

where

$$\hat{R}_n(t) = \frac{1}{|\mathbf{G}_n(\pi)|} \sum_{g \in \mathbf{G}_n(\pi)} I\{T_n(gZ^{(n)}) \leq t\} \ . \tag{30}$$

Here, $\hat{R}_n^{-1}(1-\alpha)$ is understood to be $\inf\{t \in \mathbf{R} : \hat{R}_n(t) \geq 1-\alpha\}$. We also emphasize that difference choices of $T_n(Z^{(n)})$ lead to different randomization tests and some of our results below will rely upon a particular choice of $T_n(Z^{(n)})$.

Remark 3.13. In some situations, $|\mathbf{G}_n(\pi)| = 2^n$ may be too large to permit computation of $\hat{R}_n(t)$ defined in (30). In such cases, a stochastic approximation to the test may be used by replacing $\mathbf{G}_n(\pi)$ with $\hat{\mathbf{G}}_n = \{g_1, \ldots, g_B\}$, where g_1 is the identity permutation and let $g_2, \ldots, g_B$ are i.i.d. Unif($\mathbf{G}_n(\pi)$). Theorem 3.4 below remains true with such an approximation; Theorem 3.5 below also remains true with such an approximation provided that $B \to \infty$ as $n \to \infty$. ■

3.4.1 Finite-Sample Results

Before developing the large-sample properties of the randomization test given by (29), we present some finite-sample properties of the test. We show, in particular, that for any choice of test statistic the randomization test defined in (29) has rejection probability no greater than the nominal level for the following more restrictive null hypothesis:

$$\tilde{H}_0 : Y_i(1)|X_i \overset{d}{=} Y_i(0)|X_i \ . \tag{31}$$

While the proof of the result follows closely classical arguments that underlie the finite-sample validity of randomization tests more generally, it is presented in the Supplemental Appendix for completeness. Similar results can also be found in Heckman et al. (2011) and Lee and Shaikh (2014).

Theorem 3.4. *Suppose the treatment assignment mechanism satisfies Assumption 2.2. For the problem of testing (31) at level $\alpha \in (0,1)$, $\phi_n^{\text{rand}}(Z^{(n)})$ defined in (29) with any $T_n(Z^{(n)})$ satisfies*

$$E[\phi_n^{\text{rand}}(Z^{(n)})] \leq \alpha \tag{32}$$

whenever Q additionally satisfies the null hypothesis, i.e., $Y_i(1)|X_i \overset{d}{=} Y_i(0)|X_i$.

Remark 3.14. By modifying the test defined in (29) so that it rejects with positive probability when $T_n(Z^{(n)}) = \hat{c}_n^{\text{rand}}(1-\alpha)$, it is possible to ensure that the test has rejection probability exactly equal to α whenever Q satisfies the null hypothesis, rather than simply less than or equal to α, as described in (32). See Lehmann and Romano (2005, Chapter 15) for further details. ■

3.4.2 Large-Sample Properties

In this section, we establish the large-sample validity of the randomization test given by (29) with a suitable choice of test statistic for testing (3). In particular, we show that the limiting rejection probability of the proposed test equals the nominal level under the null hypothesis.

Theorem 3.5. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.4. Let $T_n(Z^{(n)}) = |T_n^{t-test,adj}(Z^{(n)})|$, where $T_n^{t-test,adj}(Z^{(n)})$ is defined in (24). For such a choice of $T_n(Z^{(n)})$,*

$$\sup_{t \in \mathbf{R}} \left| \hat{R}_n(t) - (\Phi(t) - \Phi(-t)) \right| \xrightarrow{P} 0 \ , \tag{33}$$

where $\Phi(\cdot)$ is the standard normal c.d.f. Thus, for the problem of testing (3) with $\Delta_0 = 0$ at level $\alpha \in (0,1)$, $\phi_n^{\text{rand}}(Z^{(n)})$ with such a choice of $T_n(Z^{(n)})$ satisfies

$$\lim_{n \to \infty} E[\phi_n^{\text{rand}}(Z^{(n)})] = \alpha \ , \tag{34}$$

whenever Q additionally satisfies the null hypothesis, i.e., $\Delta(Q) = 0$.

Remark 3.15. For completeness, we briefly describe the way in which Theorem 3.5 extends to testing (3) with $\Delta_0 \neq 0$ in further detail. To this end, let $\tilde{Z}_i = (Y_i - D_i\Delta_0, D_i, X_i)$ and define the action of $g \in \mathbf{G}_n(\pi)$ on $\tilde{Z}^{(n)}$ as follows:

$$g\tilde{Z}^{(n)} = \{(Y_i - D_i\Delta_0, D_{g(i)}, X_i) : 1 \leq i \leq 2n\} \ .$$

Consider the test, $\phi_n^{\text{rand}}(\tilde{Z}^{(n)})$, obtained by replacing $Z^{(n)}$ in the test described in Theorem 3.5 with $\tilde{Z}^{(n)}$. For such a test, we have, under the assumptions of Theorem 3.5, that

$$\lim_{n \to \infty} E[\phi_n^{\text{rand}}(\tilde{Z}^{(n)})] = \alpha$$

whenever Q additionally satisfies the null hypothesis, i.e., $\Delta(Q) = \Delta_0$. ■

Remark 3.16. The conclusion in Theorem 3.5 depends heavily on the choice of test statistic in the definition of (29). In order to illustrate this phenomenon, consider the test defined by (29) with $T_n(Z^{(n)}) = |\sqrt{n}\hat{\Delta}_n|$. Using Lemmas S.1.4 and S.1.8 in the Supplemental Appendix, it is possible to show that this test behaves similarly under the null hypothesis to the "matched pairs" t-test described in Section 3.2. In particular, it has limiting rejection probability under the null hypothesis no greater than α and strictly less than α unless (16) holds. A growing literature suggests that it should be possible to achieve limiting rejection probability under the null hypothesis equal to α by studentizing the test statistic using a consistent estimator of (19). See, for example, Janssen (1997), Chung and Romano (2013, 2016), DiCiccio and Romano (2017) and Bugni et al. (2018). The problem considered here, however, illustrates that this need not be sufficient. To see this, consider the test defined by (29) with $T_n(Z^{(n)}) = \frac{|\sqrt{n}\hat{\Delta}_n|}{\tilde{\nu}_n}$, where $\tilde{\nu}_n^2$ is defined in (28). It is possible to show using arguments similar to those used in establishing Theorem 3.3 that this test also behaves similarly under the null hypothesis to the "matched pairs" t-test described in Section 3.2. The problem can be traced to the following peculiar phenomenon: even though $\tilde{\nu}_n^2$ is consistent for (19), as discussed in Remark 3.11, $\tilde{\nu}_n^2$, in contrast to $\hat{\nu}_n^2$, is not consistent for the variance of the distribution to which the randomization distribution of $\sqrt{n}\hat{\Delta}_n$ converges. See, in particular, Lemmas S.1.8–S.1.9 in the Supplemental Appendix. ■

4 Algorithms for Pairing

In this section, we describe different algorithms for pairing units so that Assumptions 2.3–2.4 are satisfied. For the case where $\dim(X_i) = 1$, a particularly simple algorithm leads to pairs that satisfy these assumptions. In particular, we show that in order to satisfy Assumptions 2.3–2.4 it suffices to pair units simply by first ordering the units from smallest to largest according to X_i and then defining pairs according to adjacent units.

Theorem 4.1. *Suppose $\dim(X_i) = 1$ and $E[X_i^2] < \infty$. Let π be any permutation of $2n$ elements such that that*

$$X_{\pi(1)} \leq \cdots \leq X_{\pi(2n)} .$$

Then, π satisfies Assumptions 2.3–2.4.

For the case where $\dim(X_i) > 1$, it is helpful to assume that $\text{supp}(X_i)$ lies in a known, bounded set, which, without loss of generality, we may assume to be $[0,1]^k$. Because $u^2 \leq u$ for all $0 \leq u \leq 1$, it follows that for any permutation $\tilde{\pi}$ of $2n$ elements

$$\frac{1}{n} \sum_{1 \leq j \leq n} |X_{\tilde{\pi}(2j-1)} - X_{\tilde{\pi}(2j)}|^2 \leq \frac{1}{n} \sum_{1 \leq j \leq n} |X_{\tilde{\pi}(2j-1)} - X_{\tilde{\pi}(2j)}| . \tag{35}$$

In order to satisfy Assumption 2.3, it is therefore natural to choose π so as to minimize the right-hand side of (35). Such minimization problems have been previously considered by Greevy et al. (2004) in the context of "matched pairs" designs in order to achieve "balance" in the sense of Remark 2.2. Algorithms for solving this minimization problem in a polynomial number of operations exist. See, for example, the "blossom" algorithm described in Edmonds (1965) as well as the algorithm described in Derigs (1988) and implemented in the R package `nbpMatching`. The following theorem derives a finite-sample bound on the right-hand side of (35) for π minimizing the right-hand side of (35), which implies, in particular, that pairing units in this way satisfies Assumption 2.3.

Theorem 4.2. *Suppose $supp(X_i) \subseteq [0,1]^k$. Let π be any permutation of $2n$ elements minimizing the right-hand side of (35). Then, for each integer $m > 1$, we have that*

$$\frac{1}{n}\sum_{1\le j\le n} |X_{\pi(2j)} - X_{\pi(2j-1)}| \le \frac{\sqrt{k}}{m} + \frac{m^{k-1}2\sqrt{k}}{n} \ . \tag{36}$$

In particular, if $m \asymp n^{\frac{1}{k}}$, then π satisfies Assumption 2.3.

Given a pairing satisfying Assumption 2.3, we now turn our attention to ensuring that the pairing further satisfies Assumption 2.4. To this end, choose $\bar{\pi}$ so as to minimize

$$\frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} |\bar{X}_{\bar{\pi}(2j)} - \bar{X}_{\bar{\pi}(2j-1)}| \ , \tag{37}$$

where

$$\bar{X}_j = \frac{X_{\pi(2j)} + X_{\pi(2j-1)}}{2} \ . \tag{38}$$

We note that the aforementioned algorithms may also be used to solve this minimization problem in a polynomial number of operations. The following theorem establishes that by re-ordering the pairs according to $\bar{\pi}$, we can ensure that the pairing satisfies Assumption 2.4 in addition to Assumption 2.3.

Theorem 4.3. *Suppose $supp(X_i) \subseteq [0,1]^k$. Let π be a permutation of $2n$ elements such that Assumption 2.3 is satisfied and $\bar{\pi}$ be any permutation of n elements minimizing (37). Define a permutation $\tilde{\pi}$ of $2n$ elements so that*

$$\tilde{\pi}(2j) = \pi(2\bar{\pi}(j)) \quad \textit{and} \quad \tilde{\pi}(2j-1) = \pi(2\bar{\pi}(j)-1) \tag{39}$$

for $1 \le j \le n$. Then, $\tilde{\pi}$ satisfies Assumptions 2.3–2.4.

5 Simulations

In this section, we examine the finite-sample behavior of several different tests of (3) with $\Delta_0 = 0$ at nominal level $\alpha = .05$ with a simulation study. For $d \in \{0,1\}$ and $1 \le i \le 2n$, potential outcomes are generated according to the equation:

$$Y_i(d) = \mu_d + m_d(X_i) + \sigma_d(X_i)\epsilon_{d,i} \ ,$$

where μ_d, $m_d(X_i)$, $\sigma_d(X_i)$ and $\epsilon_{d,i}$ are specified in each model as follows. In each of following specifications, $n = 100$, $(X_i, \epsilon_{0,i}, \epsilon_{1,i}), i = 1\ldots 2n$ are i.i.d., $\mu_0 = 0$ and $\mu_1 = \Delta$, where $\Delta = 0$ to study the behavior of the tests under the null hypothesis and $\Delta = \frac{1}{4}$ to study the behavior of the tests under the alternative hypothesis.

Model 1: $X_i \sim \text{Unif}[0,1]$; $m_1(X_i) = m_0(X_i) = \gamma(X_i - \frac{1}{2})$; $\epsilon_{d,i} \sim N(0,1)$ for $d = 0, 1$; $\sigma_0(X_i) = \sigma_0 = 1$ and $\sigma_1(X_i) = \sigma_1$.

Model 2: As in Model 1, but $m_1(X_i) = m_0(X_i) = \sin(\gamma(X - \frac{1}{2}))$.

Model 3: As in Model 2, but with $m_1(X_i) = m_0(X_i) + X_i^2 - \frac{1}{3}$.

Model 4: As in Model 1, but $m_0(X_i) = 0$ and $m_1(X_i) = 10(X_i^2 - \frac{1}{3})$.

Model 5: As in Model 4, but $m_0(X_i) = -10(X_i^2 - \frac{1}{3})$.

Model 6: As in Model 4, but $\sigma_0(X_i) = X_i^2$ and $\sigma_1(X_i) = \sigma_1 X_i^2$.

Model 7: $X_i = (\Phi(V_{i1}), \Phi(V_{i2}))'$, where $\Phi(\cdot)$ is the standard normal c.d.f. and

$$V_i \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) ;$$

$m_1(X_i) = m_0(X_i) = \gamma' X_i - 1$; $\epsilon_{d,i} \sim N(0,1)$ for $d = 0, 1$; $\sigma_0(X_i) = \sigma_0 = 1$ and $\sigma_1(X_i) = \sigma_1$.

Model 8: As in Model 7, but $m_1(X_i) = m_0(X_i) + 10(\Phi^{-1}(X_{i1})\Phi^{-1}(X_{i2}) - \rho)$.

Model 9: As in Model 7, but $m_0(X_i) = 5(\Phi^{-1}(X_{i1})\Phi^{-1}(X_{i2}) - \rho)$ and $m_1(X_i) = -m_0(X_i)$.

Model 10: $X_i = (\Phi(V_{i1}), \Phi(V_{i2}), \Phi(V_{i3}), \Phi(V_{i4}), \Phi(V_{i5}))'$, where $\Phi(\cdot)$ is the standard normal c.d.f. and $V_i \sim N(0, \Sigma)$, for Σ with 1 on the diagonal and ρ on all other entries. $m_1(X_i) = m_0(X_i) = \gamma'(X_i - \frac{1}{2})$; $\epsilon_{d,i} \sim N(0,1)$ for $d = 0, 1$; $\sigma_0(X_i) = \sigma_0 = 1$ and $\sigma_1(X_i) = \sigma_1$.

Model 11: As in Model 10, but $m_1(X_i) = m_0(X_i) + 10\sum_{j=1}^{5} \Phi^{-1}(X_{ij})$.

Model 12: As in Model 10, but $m_0(X_i) = 5\sum_{j=1}^{5} \Phi^{-1}(X_{ij})$ and $m_1(X_i) = -m_0(X_i)$.

Model 13: As in Model 10 with analogous functional forms for $m_1(X_i)$ and $m_0(X_i)$, but $\dim(X) = 100$.

Model 14: As in Model 11 with analogous functional forms for $m_1(X_i)$ and $m_0(X_i)$, but $\dim(X) = 100$.

Model 15: As in Model 12 with analogous functional forms for $m_1(X_i)$ and $m_0(X_i)$, but $\dim(X) = 100$.

For our subsequent discussion, it is useful to note that Models 5, 9, 12, and 15 satisfy (11), Models 1–2, 7, 10, and 13 satisfy (16), and Models 1–2, 7, 10, and 13 with $\sigma_1 = 1$ satisfy (31) under the null hypothesis.

Treatment status is determined according to Assumption 2.2, where the pairs are calculated as follows. If $\dim(X_i) = 1$, then pairs are calculated by sorting the X_i as described in Theorem 4.1. Note that this ensures that both Assumptions 2.3 and 2.4 are satisfied. If $\dim(X_i) > 1$, then the pairs are calculated by finding π that minimizes the right-hand side of (35) using the `R` package `nbpMatching`. Theorem 4.2 ensures that these pairs satisfy Assumption 2.3. In order to further ensure that the pairs satisfy Assumption 2.4, we re-order the pairs by finding $\bar{\pi}$ that minimizes (37) using the same `R` package and applying Theorem 4.3.

The results of our simulations are presented in Tables 1–3 below. Columns are labeled in the following way:

t-test: The two-sample t-test studied in Theorem 3.1.

naïve: The randomization test defined in (29) with $T_n(Z^{(n)}) = |\sqrt{n}\hat{\Delta}_n|$ and discussed in Remark 3.16. We henceforth refer to this test as the naïve randomization test.

MP-t: The "matched pairs" t-test studied in Theorem 3.2.

t-adj: The "adjusted" t-test studied in Theorem 3.3.

R-adj: The randomization test studied in Theorem 3.5. We henceforth refer to this test as the "adjusted" randomization test.

The tables vary according to the values of γ, σ_1 and ρ, which were not specified in the description of the different models above. Rejection probabilities are calculated using 10^4 replications and presented in percentage points. Because 2^n is large, we employ a stochastic approximation as described in Remark 3.13 with $B = 1000$ when computing each of the randomization tests. We organize our discussion of the results by test:

t-test: As expected in light of Theorem 3.1, the two-sample t-test has rejection probability under the null hypothesis no greater than the nominal level. In some cases, the rejection probability under the null hypothesis is far below the nominal level – see, for instance, Models 4, 6–8, 10–11, and 13–14. In other cases, the rejection probability is close to the nominal level – see, in particular, Models 5, 9, 12, and 15, which satisfy (11) and are therefore expected to exhibit this phenomenon. In almost all cases, the two-sample t-test is among the least powerful tests, but, as expected, this feature is especially acute when it has rejection probability under the null hypothesis severely below the nominal level.

naïve: As expected following the discussion in Remark 3.16, the naïve randomization test has rejection probability under the null hypothesis no greater than the nominal level. In some cases, the rejection probability under the null hypothesis is far below the nominal level – see, for instance, Models 4–6, 8–9, 11–12, and 14–15. In other cases, the rejection probability is close to the nominal level – see, in particular, Models 1–2, 7, 10, and 13, which satisfy (16) and are therefore expected to exhibit this phenomenon. Models 1–2, 7, 10, and 13 with $\sigma_1 = 1$ (corresponding to Tables 1 and 3) in fact satisfy (31) under the null hypothesis, so the rejection probability is exactly equal to the nominal level up to simulation error, in agreement with Theorem 3.4. If its rejection probability is close to the nominal level, then it is also among the most powerful tests, but it otherwise fares poorly in terms of power, especially when compared to the "adjusted" randomization test.

MP-t: As expected in light of Theorem 3.2, the "matched pairs" t-test has rejection probability under the null hypothesis no greater than the nominal level. In some cases, the rejection probability under the null hypothesis is far below the nominal level – see, for instance, Models 4–6, 8–9, 11–12, and 14–15. In other cases, the rejection probability is close to the nominal level – see, in particular, Models 1–2, 7, 10, and 13, which satisfy (16) and are therefore expected to exhibit this phenomenon. In almost all cases, the "matched pairs" t-test is among the least powerful tests, but, as expected, this feature is especially acute when it has rejection probability under the null hypothesis severely below the nominal level.

t-**adj**: As expected in light of Theorem 3.3, the "adjusted" t-test has rejection probability under the null hypothesis close to the nominal level in all cases. In all cases, it is the most powerful test.

R-adj: As expected in light of Theorem 3.5, the "adjusted" randomization test has rejection probability under the null hypothesis close to the nominal level in almost all cases. The exception is Model 8, for which the test exhibits some under-rejection under the null hypothesis. For Models 1–2, 7, 10, and 13 with $\sigma_1 = 1$ (corresponding to Tables 1 and 3), which, as mentioned previously, satisfy (31) under the null hypothesis, the rejection probability is again exactly equal to the nominal level up to simulation error, in agreement with Theorem 3.4. In all cases, it is nearly as powerful as our most powerful test, the "adjusted" t-test.

In Section S.1.11 of the Supplemental Appendix, we present further simulations largely similar to the specifications presented except either $n = 40$ instead of $n = 100$ or $\epsilon_{d,i} \sim t_4$ instead of a standard normal distribution. The results remain qualitatively the same. To further emphasize the important power differences between the tests discussed above, we additionally present in Section S.1.11 of the Supplemental Appendix power curves for the different tests for a single specification.

	Under H_0 — $\Delta = 0$					Under H_1 — $\Delta = 1/4$				
Model	***t*-test**	**näive**	**MP-*t***	***t*-adj**	**R-adj**	***t*-test**	**näive**	**MP-*t***	***t*-adj**	**R-adj**
1	4.25	5.02	5.31	5.29	4.97	40.16	41.87	43.20	43.17	41.44
2	4.32	4.93	5.43	5.42	4.93	39.23	41.37	42.52	42.29	40.78
3	3.51	4.73	5.04	5.15	4.73	35.90	40.09	41.56	42.05	40.67
4	1.28	1.13	1.29	4.89	4.27	5.43	5.12	5.51	15.97	14.45
5	5.69	0.79	0.90	5.68	4.98	9.65	1.94	2.18	9.61	8.60
6	0.87	0.65	0.75	5.33	4.83	4.80	4.03	4.70	19.41	17.36
7	3.29	4.94	5.30	5.44	5.28	35.82	41.56	43.07	43.17	42.16
8	1.00	0.93	1.03	4.56	4.26	0.94	0.93	0.96	4.75	4.37
9	5.30	0.65	0.71	4.28	3.87	7.18	1.52	1.65	6.17	5.83
10	1.20	4.90	5.19	5.23	4.93	22.70	41.39	42.75	42.73	41.62
11	0.66	0.67	0.74	4.42	4.24	0.53	0.58	0.71	4.50	4.17
12	5.05	0.65	0.68	4.18	3.95	5.57	0.79	0.80	4.66	4.32
13	0.00	4.57	4.96	5.00	4.67	0.00	7.93	8.46	8.73	8.26
14	0.75	0.85	0.99	4.76	4.50	0.76	0.84	0.99	4.92	4.63
15	4.93	0.61	0.72	4.77	4.47	4.89	0.62	0.80	4.89	4.74

Table 1: Rej. prob. for Models 1–15 with $\gamma = \iota_{\dim(X)}$ for Models 1–15, $\sigma_1 = 1$, $\rho = 0.2$.

	Under H_0 — $\Delta = 0$					Under H_1 — $\Delta = 1/4$				
Model	***t*-test**	**näive**	**MP-*t***	***t*-adj**	**R-adj**	***t*-test**	**näive**	**MP-*t***	***t*-adj**	**R-adj**
1	4.75	5.11	5.37	5.46	5.06	29.46	30.26	31.51	31.49	30.24
2	4.23	4.59	5.03	5.20	4.70	29.33	29.99	31.39	30.89	29.52
3	4.16	4.84	5.27	5.39	5.09	26.60	28.78	30.07	30.30	29.27
4	1.65	1.53	1.65	5.24	4.74	5.80	5.31	5.91	14.95	13.72
5	5.27	0.68	0.81	5.21	4.67	9.59	2.19	2.53	9.54	8.45
6	0.83	0.81	0.91	5.50	4.86	4.89	4.23	4.66	18.25	16.43
7	0.39	5.21	5.66	5.85	5.54	7.38	30.04	31.01	31.20	30.56
8	1.50	1.58	1.66	5.71	5.27	0.69	0.70	0.77	4.80	4.36
9	5.73	1.34	1.42	5.24	4.87	8.28	2.13	2.22	7.33	6.93
10	0.65	4.99	5.46	5.33	5.12	9.74	29.53	30.67	30.63	29.93
11	0.63	0.71	0.76	5.21	4.98	0.60	0.64	0.77	4.94	4.61
12	5.51	0.72	0.82	5.26	4.89	5.42	0.72	0.76	5.18	4.86
13	0.00	4.93	5.39	5.24	5.20	0.00	10.07	10.75	10.74	10.17
14	0.58	0.60	0.67	5.35	5.03	0.61	0.68	0.78	5.08	4.79
15	5.44	0.66	0.66	5.55	5.20	5.04	0.63	0.73	5.18	4.81

Table 2: Rej. prob. for Models 1–15 with $\gamma = 1$ for Models 1–6, $\gamma' = (1, 4)$ for Models 7–9, $\gamma = \iota_{\dim(X)}$ for Models 10–15, $\sigma_1 = 2$, $\rho = 0.7$.

6 Empirical Application

In this section, we apply several different tests of (3) to a real-world example. For this purpose, we run a button-pressing experiment on Amazon Mechanical Turk (MTurk) as in DellaVigna and Pope (2018). In

	Under H_0 — $\Delta = 0$					Under H_1 — $\Delta = 1/4$				
Model	t-**test**	**näive**	**MP-**t	t-**adj**	**R-adj**	t-**test**	**näive**	**MP-**t	t-**adj**	**R-adj**
1	4.51	5.19	5.62	5.66	5.24	39.09	40.88	42.09	41.92	40.56
2	4.09	4.68	5.03	5.08	4.58	39.95	41.59	42.84	42.43	41.20
3	3.67	4.91	5.26	5.55	5.26	35.10	39.48	40.89	41.48	40.15
4	1.07	0.98	1.13	4.83	4.28	5.43	5.00	5.47	16.52	14.95
5	5.21	0.69	0.79	5.21	4.61	9.98	2.17	2.35	9.93	8.89
6	0.67	0.65	0.69	5.17	4.44	5.11	4.50	4.89	19.03	17.23
7	0.28	4.91	5.19	5.50	5.23	11.20	41.61	43.01	43.18	42.06
8	0.70	0.67	0.81	4.41	4.03	0.95	0.96	1.11	5.26	4.75
9	5.37	0.71	0.79	4.30	4.00	6.93	0.95	1.02	5.52	5.10
10	2.28	5.02	5.42	5.39	5.17	29.53	40.23	41.72	41.84	40.74
11	0.90	1.03	1.10	4.12	3.82	1.07	1.10	1.23	3.99	3.75
12	5.35	0.97	1.03	4.02	3.76	5.01	0.86	0.91	3.60	3.41
13	2.71	4.88	5.22	5.29	5.15	6.51	10.08	10.84	10.78	10.47
14	3.92	3.68	4.04	4.79	4.55	4.38	4.15	4.52	5.19	4.84
15	5.20	2.94	3.18	4.25	3.97	5.61	3.07	3.38	4.36	4.13

Table 3: Rej. prob. for Models 1–12 with $\gamma = 1$ for Models 1–6, $\gamma' = (4, 1)$ for Models 7–9, $\gamma = \iota_{\dim(X)}$ for Models 10–15, $\sigma_1 = 1$, $\rho = 0$.

the experiment, participants are asked to press buttons 'a' and 'b' alternately as much as possible in 5 minutes. The outcome Y_i is number of presses. The treatment D_i is an indicator for whether participants receive financial incentives for the button-pressing task: $D_i = 1$ if a unit is treated, i.e., she receives financial incentives for pressing more buttons, and $D_i = 0$ if a unit is untreated, i.e., she receives no financial incentives. The sample size is $2n = 120$. Following DellaVigna and Pope (2018), each treated unit receives an additional cent for every 100 points they score, where one point corresponds to one alternate press. The covariate X_i is a scalar variable which denotes the individual's performance in prior version of the task with no financial incentives. Units are paired according to X_i as in Theorem 4.1. Using the usual duality between hypothesis testing and constructing confidence intervals, we construct a 95% confidence interval for (2) by inverting the corresponding test. The results are presented in Table 4 below along with the point estimator of $\Delta(Q)$ given by $\hat{\Delta}_n$. In the case of the two-sample t-test, "matched pairs" t-test and "adjusted" t-test, it is possible to describe the corresponding confidence intervals in terms of a standard error, so we include those in the table as well. In the case of the two randomization tests, it is not possible to do this; in these instances, the confidence intervals are instead computed by inverting tests of (3) along a grid of equally spaced points from 0 to 700.

	t-**test**	**näive**	**MP-**t	t-**adj**	**R-adj**
$\hat{\Delta}_n$	219	219	219	219	219
std. errors	81	-	55	52	-
conf. int.	[59, 379]	[109, 332]	[109, 329]	[116, 322]	[115, 322]

Table 4: Inferences about $\Delta(Q)$ using different tests in the empirical application.

We note the following features of our empirical results: (a) the confidence interval given by the "adjusted" t-test is shorter than the one given by the two-sample t-test; (b) the confidence interval given by the "adjusted" t-test is shorter than the one given by the "matched pairs" t-test; and (c) the confidence interval given by the “adjusted” randomization test is shorter than the one given by the näive randomization test. From our theoretical results, (a) suggests (11) is not satisfied, while (b) and (c) suggest that (16) is not satisfied either. Our simulation study confirms that when (11) or (16) are not satisfied, there can be dramatic benefits from using the asymptotically exact methods.

7 Recommendations for Empirical Practice

We conclude with some recommendations for empirical practice based on our theoretical results as well as the simulation study above. For inference about the average treatment effect in the type of “matched pairs” design studied in this paper, we do not recommend the two-sample t-test, the “matched pairs” t-test or the naïve randomization test, which are often considerably less powerful than both the “adjusted” t-test and the “adjusted” randomization test. In our simulations the “adjusted” t-test is always the most powerful among the tests we consider, though sometimes by a small margin in comparison to the “adjusted” randomization test. We also note that the modest gain in power of the “adjusted” t-test is accompanied by the generally higher rejection probability under the null hypothesis of the “adjusted” t-test as well. The “adjusted” randomization test, however, retains the attractive feature that the finite-sample rejection probability under the null hypothesis is no greater than the nominal size for certain distributions satisfying the null hypothesis. To the extent that this feature is deemed important, the “adjusted” randomization test may be preferred to the “adjusted” t-test despite having slightly lower power.

References

Abadie, A. and Imbens, G. W. (2008). Estimation of the conditional variance in paired experiments. *Annales d'Economie et de Statistique* 175–187.

Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, **107** 833–843.

Armitage, P., Berry, G. and Matthews, J. N. S. (2008). *Statistical Methods in Medical Research*. John Wiley & Sons.

Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, vol. 1. Elsevier, 73–140.

Bai, Y. (2020). Optimality of matched-pair designs in randomized controlled trials. Tech. rep.

Bertrand, M. and Duflo, E. (2017). Field experiments on discrimination. In *Handbook of Economic Field Experiments*, vol. 1. Elsevier, 309–393.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, **1** 200–232.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, **113** 1784–1796.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. Becker Friedman Institute for Economics Working Paper 2019-19, University of Chicago.

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

Chung, E. and Romano, J. P. (2016). Multivariate and multiple permutation tests. *Journal of Econometrics*, **193** 76–91.

Crépon, B., Devoto, F., Duflo, E. and Parienté, W. (2015). Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco. *American Economic Journal: Applied Economics*, **7** 123–50.

DellaVigna, S. and Pope, D. (2018). What motivates effort? Evidence and expert forecasts. *The Review of Economic Studies*, **85** 1029–1069.

Derigs, U. (1988). Solving non-bipartite matching problems via shortest path techniques. *Annals of Operations Research*, **13** 225–261.

DiCiccio, C. J. and Romano, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association*, **112** 1211–1220.

Ding, P. (2017). A paradox from randomization-based causal inference. *Statistical science*, **32** 331–345.

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

Edmonds, J. (1965). Maximum matching and a polyhedron with 0, 1-vertices. *Journal of Research of the National Bureau of Standards B*, **69** 55–56.

Fogarty, C. B. (2018a). On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80** 1035–1056.

Fogarty, C. B. (2018b). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, **105** 994–1000.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Glennerster, R. and Takavarasha, K. (2013). *Running randomized evaluations: A practical guide*. Princeton University Press.

Greevy, R., Lu, B., Silber, J. H. and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics*, **5** 263–275.

Heard, K., O'Toole, E., Naimpally, R. and Bressler, L. (2017). Real world challenges to randomization and their solutions.

Heckman, J. J., Pinto, R., Shaikh, A. M. and Yavitz, A. (2011). Inference with imperfect randomization: The case of the Perry preschool program. Tech. rep., National Bureau of Economic Research.

Hsu, H. and Lachenbruch, P. A. (2007). Paired t-test. Wiley Online Library, 1–3.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in medicine*, **27** 4857–4873.

Imai, K., King, G., Nall, C. et al. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science*, **24** 29–53.

Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized behrens-fisher problem. *Statistics & Probability Letters*, **36** 9–21.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80** 85–112.

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

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer, New York.

List, J. A. and Rasul, I. (2011). Field experiments in labor economics. vol. 4. Elsevier, 103–228.

Riach, P. A. and Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal*, **112** 480–518.

Rosenberger, W. F. and Lachin, J. M. (2015). *Randomization in clinical trials: Theory and Practice.* John Wiley & Sons.

van der Laan, M. J., Balzer, L. B. and Petersen, M. L. (2012). Adaptive matching in randomized trials and observational studies. *Journal of Statistical Research*, **46** 113.

White, H. (2013). An introduction to the use of randomised control trials to evaluate development interventions. *Journal of Development Effectiveness*, **5** 30–49.