# Randomization Tests under an Approximate Symmetry Assumption*

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

December 19, 2014

**Abstract**

This paper develops a theory of randomization tests under an approximate symmetry assumption. Randomization tests provide a general means of constructing tests that control size in finite samples whenever the distribution of the observed data exhibits symmetry under the null hypothesis. Here, by exhibits symmetry we mean that the distribution remains invariant under a group of transformations. In this paper, we provide conditions under which the same construction can be used to construct tests that asymptotically control the probability of a false rejection whenever the distribution of the observed data exhibits approximate symmetry in the sense that the limiting distribution of a function of the data exhibits symmetry under the null hypothesis. An important application of this idea is in settings where the data may be grouped into a fixed number of "clusters" with a large number of observations within each cluster. In such settings, we show that the distribution of the observed data satisfies our approximate symmetry requirement under weak assumptions. In particular, our results allow for the clusters to be heterogeneous and also have dependence not only within each cluster, but also across clusters. This approach enjoys several advantages over other approaches in these settings. Among other things, it leads to a test that is asymptotically similar, which, as shown in a simulation study, translates into improved power at many alternatives. Finally, we use our results to revisit the analysis of Angrist and Lavy (2009), who examine the impact of a cash award on exam performance for low-achievement students in Israel.

KEYWORDS: Randomization tests, dependence, heterogeneity, differences-in-differences, clustered data, sign changes, symmetric distribution, weak convergence

JEL classification codes: C12, C14.

*We thank Chris Hansen, Aprajit Mahajan, Ulrich Mueller and Chris Taber for helpful comments. Sergey Gitlin provided excellent research assistance.

†Research supported by NSF Grant DMS-1307973.

‡Research supported by NSF Grants DMS-1227091 and DMS-1307973.

# 1 Introduction

Suppose the researcher observes data $X^{(n)} \sim P_n \in \mathbf{P}_n$, where $\mathbf{P}_n$ is a set of distributions on a sample space $\mathcal{X}_n$, and is interested in testing

$$H_0 : P_n \in \mathbf{P}_{n,0} \text{ versus } H_1 : P_n \in \mathbf{P}_n \setminus \mathbf{P}_{n,0} \ ,$$

where $\mathbf{P}_{n,0} \subset \mathbf{P}_n$, at level $\alpha \in (0,1)$. The index $n$ here will typically denote sample size. The classical theory of randomization tests provides a general way of constructing tests that control size in finite samples provided that the distribution of the observed data exhibits symmetry under the null hypothesis. Here, by exhibits symmetry we mean that the distribution remains invariant under a group of transformations. In this paper, we develop conditions under which the same construction can be used to construct tests that asymptotically control the probability of a false rejection provided that the distribution of the observed data exhibits approximate symmetry. More precisely, the main requirement we impose is that, for a known function $S_n$ from $\mathcal{X}_n$ to a sample space $\mathcal{S}$,

$$S_n(X^{(n)}) \xrightarrow{d} S \tag{1}$$

as $n \to \infty$ under $P_n \in \mathbf{P}_{n,0}$, where $S$ exhibits symmetry in the sense described above. In this way, our results extend the classical theory of randomization tests. Note that in some cases $S_n$ need not be completely known; see Remark 4.4 below.

While they apply more generally, an important application of our results is in settings where the data may be grouped into $q$ "clusters" with a large number of observations within each cluster. A noteworthy feature of our asymptotic framework is that $q$ is fixed and does not depend on $n$. In such environments, it is often the case that the distribution of the observed data satisfies our approximate symmetry requirement under weak assumptions. In particular, it typically suffices to consider

$$S_n(X^{(n)}) = (S_{n,1}(X^{(n)}), \ldots, S_{n,q}(X^{(n)}))' \ , \tag{2}$$

where $S_{n,j}(X^{(n)})$ is an appropriately recentered and rescaled estimator of the parameter of interest based on observations from the $j$th cluster. In this case, the convergence (1) often holds for $S$ that exhibits symmetry in the sense that its distribution remains invariant under the group of sign changes. Importantly, this convergence permits the clusters to be heterogeneous and also have dependence not only within each cluster, but also across clusters. We consider three specific examples of such settings in detail – time series regression, differences-in-differences, and clustered regression.

Our paper is most closely related to the procedure suggested by Ibragimov and Müller (2010). As in our paper, they also consider settings where the data may be grouped into a fixed number of "clusters," $q$, with a large number of observations within each cluster. In order to apply their results, they further assume that the parameter of interest is scalar and that $S_n(X^{(n)})$ defined

in (2) satisfies the convergence (1) with $S$ satisfying additional restrictions beyond our symmetry assumption. Using a result on robustness of the $t$-test established in Bakirov and Székely (2006), they propose an approach that leads to a test that asymptotically controls size for certain values of $q$ and $\alpha$, but may be quite conservative in the sense that its asymptotic rejection probability under the null hypothesis may be much less than $\alpha$. This same result on the $t$-test underlies the approach put forward by Bester et al. (2011), which therefore inherits the same qualifications. The methodology proposed in this paper enjoys several advantages over these approaches, including not requiring the parameter of interest to be scalar, being valid for any values of $q$ and $\alpha$ (thereby permitting in particular the computation of $p$-values), and, perhaps most importantly, being asymptotically similar in the sense of having asymptotic rejection probability under the null hypothesis equal to $\alpha$. As shown in a simulation study, this feature translates into improved power at many alternatives. See Section 2.1.1 and Section 5 for further details.

The remainder of the paper is organized as follows. Section 2 briefly reviews the classical theory of randomization tests. Here, we pay special attention to an example involving the group of sign changes, which, as mentioned previously, underlies many of our later applications and aids comparisons with the approach suggested by Ibragimov and Müller (2010). Our main results are developed in Section 3. Section 4 contains the application of our results to settings where the data may be grouped into a fixed number of "clusters" with a large number of observations within each cluster, emphasizing in particular time series regression, differences-in-differences, and clustered regression. Simulation results based on the time series regression and differences-in-differences examples are presented in Section 5. Finally, in Section 6, we use the clustered regression example to revisit the analysis of Angrist and Lavy (2009), who examine the impact of a cash award on exam performance for low-achievement students in Israel.

## 2 Review of Randomization Tests

In this section, we briefly review the classical theory of randomization tests. Further discussion can be found, for example, in Chapter 15 of Lehmann and Romano (2005). Since the results in this section are non-asymptotic in nature, we omit the index $n$.

Suppose the researcher observes data $X \sim P \in \mathbf{P}$, where $\mathbf{P}$ is a set of distributions on a sample space $\mathcal{X}$, and is interested in testing

$$H_0 : P \in \mathbf{P}_0 \text{ versus } H_1 : P \in \mathbf{P} \setminus \mathbf{P}_0 \ , \tag{3}$$

where $\mathbf{P}_0 \subset \mathbf{P}$, at level $\alpha \in (0, 1)$. Randomization tests require that the distribution of the data, $P$, exhibits symmetry whenever $P \in \mathbf{P}_0$. In order to state this requirement more formally, let $\mathbf{G}$ be a finite group of transformations from $\mathcal{X}$ to $\mathcal{X}$ and denote by $gx$ the action of $g \in \mathbf{G}$ on $x \in \mathcal{X}$.

Using this notation, the classical condition required for a randomization test is

$$X \overset{d}{=} gX \text{ under } P \text{ for any } P \in \mathbf{P}_0 \text{ and } g \in \mathbf{G} \ . \tag{4}$$

We now describe the construction of the randomization test. Let $T(X)$ be a real-valued test statistic such that large values provide evidence against the null hypothesis. Let $M = |\mathbf{G}|$ and denote by

$$T^{(1)}(X) \leq T^{(2)}(X) \leq \cdots \leq T^{(M)}(X)$$

the ordered values of $\{T(gX) : g \in \mathbf{G}\}$. Let $k = \lceil M(1-\alpha) \rceil$ and define

$$\begin{aligned} M^+(X) &= |\{1 \leq j \leq M : T^{(j)}(X) > T^{(k)}(X)\}| \\ M^0(X) &= |\{1 \leq j \leq M : T^{(j)}(X) = T^{(k)}(X)\}| \ . \end{aligned} \tag{5}$$

Using this notation, the randomization test is given by

$$\phi(X) = \begin{cases} 1 & \text{if } T(X) > T^{(k)}(X) \\ a(X) & \text{if } T(X) = T^{(k)}(X) \\ 0 & \text{if } T(X) < T^{(k)}(X) \end{cases} \ , \tag{6}$$

where

$$a(X) = \frac{M\alpha - M^+(X)}{M^0(X)} \ .$$

The following theorem shows that this construction leads to a test that controls size in finite samples whenever (4) holds. In fact, the test in (6) is similar, i.e., has rejection probability exactly equal to $\alpha$ for any $P \in \mathbf{P}_0$ and $\alpha \in (0,1)$.

**Theorem 2.1.** *Suppose $X \sim P \in \mathbf{P}$ and consider the problem of testing (3). Let $\mathbf{G}$ be a group such that (4) holds. Then, for any $\alpha \in (0,1)$, $\phi(X)$ defined in (6) satisfies*

$$E_P[\phi(X)] = \alpha \text{ whenever } P \in \mathbf{P}_0 \ . \tag{7}$$

**Remark 2.1.** Let $\mathbf{G}^x$ denote the $\mathbf{G}$-orbit of $x \in \mathcal{X}$, i.e., $\mathbf{G}^x = \{gx : g \in \mathbf{G}\}$. The result in Theorem 2.1 exploits that, when $\mathbf{G}$ is such that (4) holds, the conditional distribution $X$ given $X \in \mathbf{G}^x$ is uniform on $\mathbf{G}^x$. Since the conditional distribution of $X$ is known for all $P \in \mathbf{P}_0$ (even though $P$ itself is unknown), we can construct a test that is level $\alpha$ conditionally, which leads to a test that is level $\alpha$ unconditionally as well. ■

**Remark 2.2.** In some cases, $M$ is too large to permit computation of $\phi(X)$ defined in (6). When this is the case, the researcher may use a stochastic approximation to $\phi(X)$ without affecting the finite-sample validity of the test. More formally, let

$$\hat{\mathbf{G}} = \{g_1, \ldots, g_B\} \ , \tag{8}$$

where $g_1 =$ the identity transformation and $g_2, \ldots, g_B$ are i.i.d. Uniform($\mathbf{G}$). Theorem 2.1 remains true if, in the construction of $\phi(X)$, $\mathbf{G}$ is replaced by $\hat{\mathbf{G}}$. ■

**Remark 2.3.** One can construct a $p$-value for the test $\phi(X)$ defined in (6) as

$$\hat{p} = \hat{p}(X) = \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} I\{T(gX) \geq T(X)\} \ . \tag{9}$$

When (4) holds, it follows that $P\{\hat{p} \leq u\} \leq u$ for all $0 \leq u \leq 1$ and $P \in \mathbf{P}_0$. This result remains true when $M$ is large and the researcher uses a stochastic approximation, in which case $\hat{\mathbf{G}}$ as defined in (8) replaces $\mathbf{G}$ in (9). ■

**Remark 2.4.** The test in (6) is possibly randomized. In case one prefers not to randomize, note that the non-randomized test that rejects if $T(X) > T^{(k)}(X)$ is level $\alpha$. In our simulations, this test has rejection probability under the null hypothesis only slightly less than $\alpha$ when $M$ is not too small; see Sections 2.1.1, 5.1 and 5.2 below for additional discussion. ■

## 2.1 Symmetric Location Example

In this subsection, we provide an illustration of Theorem 2.1. The example not only makes concrete some of the abstract ideas presented above, but also underlies many of the applications described in Section 4 below.

Suppose $X = (X_1, \ldots, X_q) \sim P \in \mathbf{P}$, where

$$\mathbf{P} = \{\otimes_{j=1}^{q} P_{j,\mu} : P_{j,\mu} \text{ symmetric distribution on } \mathbf{R}^d \text{ about } \mu\} \ .$$

In other words, $X_1, \ldots, X_q$ are independent and each $X_j$ is distributed symmetrically on $\mathbf{R}^d$ about $\mu$, i.e., $X_j - \mu \stackrel{d}{=} \mu - X_j$. The researcher desires to test (3) with

$$\mathbf{P}_0 = \{\otimes_{j=1}^{q} P_{j,\mu} : P_{j,\mu} \text{ a symmetric distribution on } \mathbf{R}^d \text{ about } 0\} \ .$$

In this case, (4) clearly holds with the group of sign changes $\mathbf{G} = \{-1, 1\}^q$, where the action of $g = (g_1, \ldots, g_q) \in \mathbf{G}$ on $x = (x_1, \ldots, x_q) \in \otimes_{j=1}^{q} \mathbf{R}^d$ is defined by $gx = (g_1 x_1, \ldots, g_q x_q)$. As a result, Theorem 2.1 may be applied with any choice of $T(X)$ to construct a test that satisfies (7).

### 2.1.1 Comparison with the $t$-test

Consider the special case of the symmetric location example in which $d = 1$ and $P_{j,\mu} = N(\mu, \sigma_j^2)$, i.e.,

$$\mathbf{P} = \{\otimes_{j=1}^{q} P_{j,\mu} : P_{j,\mu} = N(\mu, \sigma_j^2) \text{ with } \mu \in \mathbf{R} \text{ and } \sigma_j^2 \geq 0\} \tag{10}$$

$$\mathbf{P}_0 = \{\otimes_{j=1}^{q} P_{j,\mu} : P_{j,\mu} = N(\mu, \sigma_j^2) \text{ with } \mu = 0 \text{ and } \sigma_j^2 \geq 0\} \ . \tag{11}$$

For this setting, Bakirov and Székely (2006) show that the usual two-sided $t$-test remains valid despite heterogeneity in the $\sigma_j^2$ for certain values of $\alpha$ and $q$. More formally, they show that for $\alpha \leq 8.3\%$ and $q \geq 2$ or $\alpha \leq 10\%$ and $2 \leq q \leq 14$,

$$P\{T_{|t\text{-stat}|}(X) > c_{q-1,1-\frac{\alpha}{2}}\} \leq \alpha \text{ for any } P \in \mathbf{P}_0 \ ,$$

where $T_{|t\text{-stat}|}(X)$ is the absolute value of the usual $t$-statistic computed using the data $X$ and $c_{q-1,1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the $t$-distribution with $q-1$ degrees of freedom. Bakirov and Székely (2006) go on to show that this result remains true even if each $P_{j,\mu}$ is allowed to be a mixture of normal distributions as well. This result was further explored by Ibragimov and Müller (2010) and Ibragimov and Müller (2013). Ibragimov and Müller (2013) derived a related result for the two-sample problem, while Ibragimov and Müller (2010) showed that the $t$-test is "optimal" in the sense that it is the uniformly most powerful scale invariant level $\alpha$ test against the restricted class of alternatives with $\sigma_j^2 = \sigma^2$ for all $1 \leq j \leq q$. In the Appendix, we establish a similar "optimality" result for the randomization test with $T(X) = T_{|t\text{-stat}|}(X)$ and $\mathbf{G} = \{-1,1\}^q$: we show that it is the uniformly most powerful unbiased level $\alpha$ test against the same class of alternatives.

We compare the randomization test with $T(X) = T_{|t\text{-stat}|}(X)$ and $\mathbf{G} = \{-1,1\}^q$ with the $t$-test. We follow Ibragimov and Müller (2010) and consider the setup in (10)-(11) with $q \in \{8,16\}$ and $\sigma_j^2 = 1$ for $1 \leq j \leq \frac{q}{2}$ and $\sigma_j^2 = a^2$ for $\frac{q}{2} < j \leq q$. Figure 1 shows rejection probabilities under the null hypothesis computed using $100,000$ Monte Carlo repetitions for $\alpha = 5\%$, $a$ ranging over a grid of 50 equally spaced points in $(0.1, 5)$, $q = 8$ (left panel) and $q = 16$ (right panel). As we would expect from Theorem 2.1, the rejection probability of the randomization test equals $\alpha$ for all values of the heterogeneity parameter $a$. The rejection probability of the $t$-test, on the other hand, can be substantially below $\alpha$ when the data are heterogeneous, i.e., $a \neq 1$. Comparing the right and left panels, we see that the performance of the $t$-test improves as $q$ gets larger, but it is worth emphasizing that the results of Bakirov and Székely (2006) do not ensure the validity of the test for $q > 14$ and $\alpha \geq 8.4\%$.

Figure 2 shows rejection probabilities computed using $100,000$ Monte Carlo repetitions for $\alpha = 5\%$, $\mu \in (0, 1.5)$, $q = 8$, $a = 0.1$ (left panel) and $a = 1$ (right panel). The similarity of the randomization test translates into better power for alternatives close to the null hypothesis. When $a = 0.1$, the rejection probability of the randomization test exceeds that of the $t$-test for $\mu$ less than approximately 0.7; for larger values of $\mu$, the situation is reversed, though the difference in power between the two tests is smaller. When $a = 1$, the $t$-test slightly outperforms the randomization test, reflecting the previously mentioned optimality property derived in Ibragimov and Müller (2010). It is important to note that this does not contradict the optimality result for the randomization test established in the Appendix, as the $t$-test is not unbiased. In particular, there are alternatives $P \in \mathbf{P}_1$ under which the $t$-test has rejection probability $< \alpha$. Moreover, the loss in power of the randomization test relative to the $t$-test even in this case is arguably negligible. These comparisons

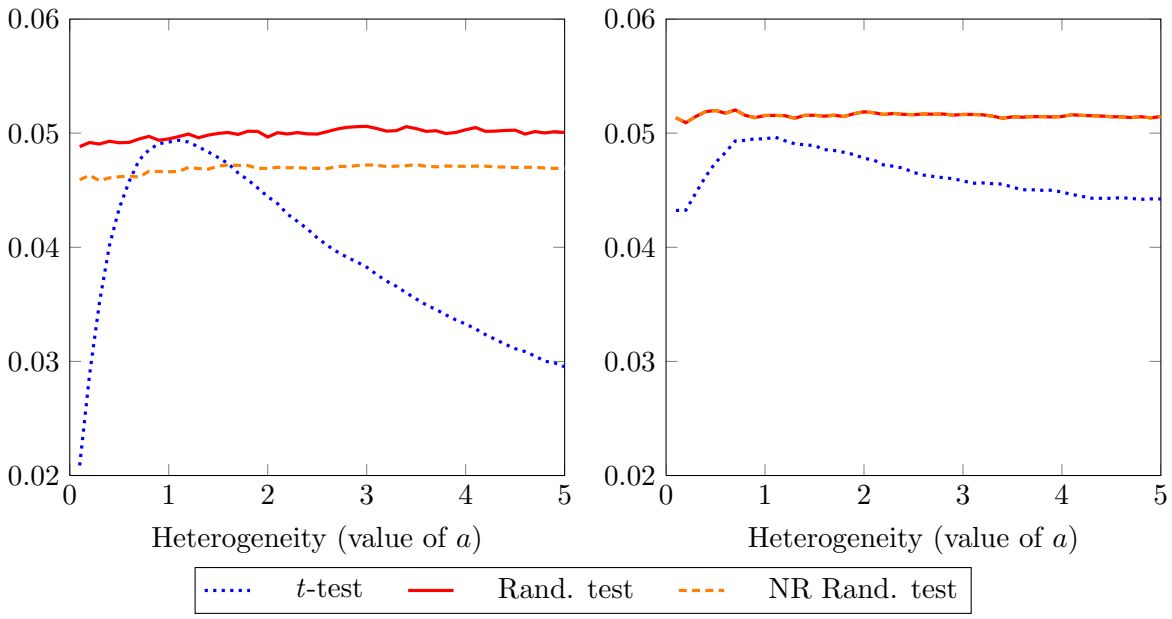


Figure 1: Rejection probabilities under the null hypothesis for different values of $a$ in the symmetric location example. Randomization test (randomized and non-randomized versions) versus $t$-test. $q = 8$ (left panel) and $q = 16$ (right panel).

continue to hold even if the randomization test is replaced with its non-randomized version described in Remark 2.4.

In the context of the symmetric location example, the randomization test provides additional advantages over the $t$-test approach. First, the randomization test works for all levels of $\alpha \in (0, 1)$, which allows for the construction of $p$-values; see Remark 2.3. Second, the randomization test works for vector-valued random variables, i.e., $d > 1$, while the result in Bakirov and Székely (2006) is restricted to scalar random variables. Third, the construction in Theorem 2.1 works for any choice of test statistic $T(X)$. Finally, the condition in (4) is not limited to mixtures of normal distributions and holds for any symmetric distribution, including even distributions with an infinite variance. On the other hand, when $q$ is small the rejection probability of the $t$-test sometimes exceeds that of the non-randomized version of the randomization test described in Remark 2.4; see Figure 1.

## 3 Main Result

In this section, we present our theory of randomization tests under an approximate symmetry assumption. Since our results in this section are asymptotic in nature, we re-introduce the index $n$, which, as mentioned earlier, will typically be used to denote the sample size.

Suppose the researcher observes data $X^{(n)} \sim P_n \in \mathbf{P}_n$, where $\mathbf{P}_n$ is a set of distributions on a

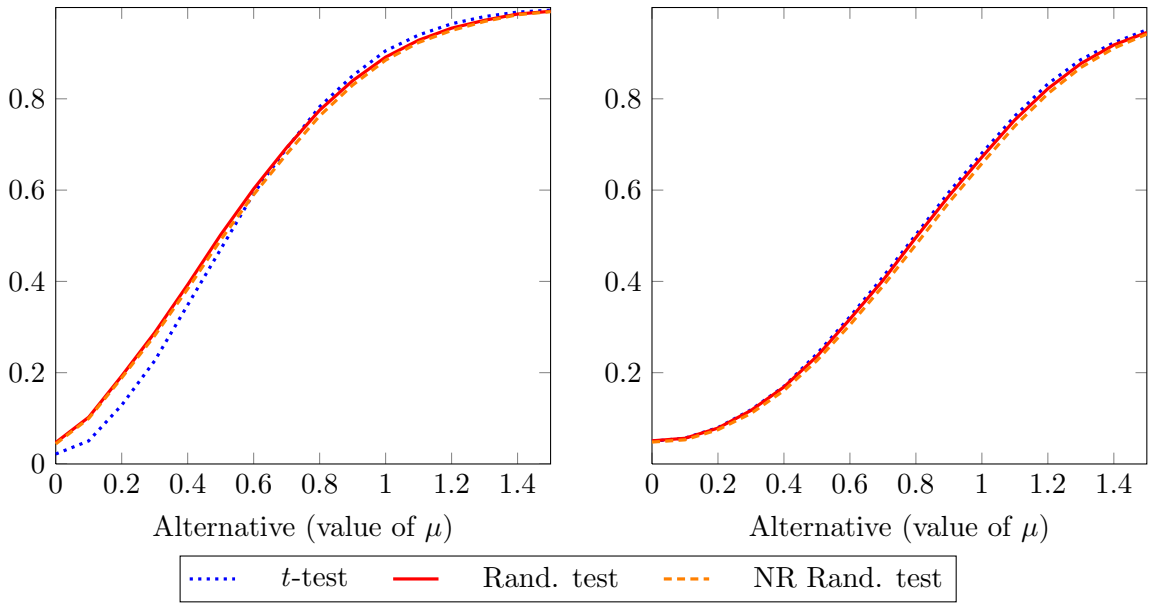


Figure 2: Rejection probabilities for $q = 8$ and different values of $\mu$ in the symmetric location example. Randomization test (randomized and non-randomized versions) versus $t$-test. $a = 0.1$ (left panel) and $a = 1$ (right panel).

sample space $\mathcal{X}_n$, and is interested in testing

$$H_0 : P_n \in \mathbf{P}_{n,0} \text{ versus } H_1 : P_n \in \mathbf{P}_n \setminus \mathbf{P}_{n,0} \ , \tag{12}$$

where $\mathbf{P}_{n,0} \subset \mathbf{P}_n$ , at level $\alpha \in (0,1)$. In contrast to Section 2, we no longer require that the distribution of $X^{(n)}$ exhibits symmetry whenever $P_n \in \mathbf{P}_{n,0}$. Instead, we require that $X^{(n)}$ exhibits approximate symmetry whenever $P_n \in \mathbf{P}_{n,0}$. In order to state this requirement more formally, we require some additional notation. Recall that $S_n$ denotes a function from $\mathcal{X}_n$ to a sample space $\mathcal{S}$. As before, let $T$ be a real-valued test statistic such that large values provide evidence against the null hypothesis, but we will assume that $T$ is a function from $\mathcal{S}$ to $\mathbf{R}$ as opposed to from $\mathcal{X}_n$ to $\mathbf{R}$. Finally, let $\mathbf{G}$ be a (finite) group of transformations from $\mathcal{S}$ to $\mathcal{S}$ and denote by $gs$ the action of $g \in \mathbf{G}$ on $s \in \mathcal{S}$. Using this notation, we have the following assumption:

**Assumption 3.1.** *If $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$, then*

*(i)* $S_n = S_n(X^{(n)}) \xrightarrow{d} S$ *under* $P_n$.

*(ii)* $gS \stackrel{d}{=} S$ *for all* $g \in \mathbf{G}$.

*(iii) For any two distinct elements* $g \in \mathbf{G}$ *and* $g' \in \mathbf{G}$,

$$\text{either } T(gs) = T(g's) \ \forall s \in \mathcal{S} \text{ or } P\{T(gS) \neq T(g'S)\} = 1 \ .$$

Assumption 3.1.(i)-(ii) formalizes what we mean by $X^{(n)}$ exhibiting approximate symmetry. Assumption 3.1.(iii) is a condition that controls the ties among the values of $T(gS)$ as $g$ varies over

$\mathbf{G}$. It requires that $T(gS)$ and $T(g'S)$ are distinct with probability one or deterministically equal to each other. For examples of $S$ that often arise in applications and typical choices of $T$, we verify Assumption 3.1.(iii) (see, in particular, Lemmas D.1-D.3 in the Appendix).

The construction of the randomization test in this setting parallels the one in Section 2 with $S_n$ replacing $X$. Let $M = |\mathbf{G}|$ and denote by

$$T^{(1)}(S_n) \leq T^{(2)}(S_n) \leq \cdots \leq T^{(M)}(S_n)$$

the ordered values of $\{T(gS_n) : g \in \mathbf{G}\}$. Let $k = \lceil M(1-\alpha) \rceil$ and define $M^+(S_n)$ and $M^0(S_n)$ as in (5) with $S_n$ replacing $X$. Using this notation, the proposed test is given by

$$\phi(S_n) = \begin{cases} 1 & T(S_n) > T^{(k)}(S_n) \\ a(S_n) & T(S_n) = T^{(k)}(S_n) \\ 0 & T(S_n) < T^{(k)}(S_n) \end{cases}, \tag{13}$$

where

$$a(S_n) = \frac{M\alpha - M^+(S_n)}{M^0(S_n)} .$$

The following theorem shows that this construction leads to a test that is asymptotically level $\alpha$ whenever Assumption 3.1 holds. In fact, the proposed test is asymptotically similar, i.e., has limiting rejection probability equal to $\alpha$ if $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$.

**Theorem 3.1.** *Suppose $X^{(n)} \sim P_n \in \mathbf{P}_n$ and consider the problem of testing (12). Let $S_n : \mathcal{X}_n \to \mathcal{S}$, $T : \mathcal{S} \to \mathbf{R}$ and $\mathbf{G} : \mathcal{S} \to \mathcal{S}$ be such that Assumption 3.1 holds. Assume further that $T : \mathcal{S} \to \mathbf{R}$ is continuous and that $g : \mathcal{S} \to \mathcal{S}$ is continuous for all $g \in \mathbf{G}$. Then, for any $\alpha \in (0,1)$, $\phi(S_n)$ defined in (13) satisfies*

$$E_{P_n}[\phi(S_n)] \to \alpha \tag{14}$$

*as $n \to \infty$ whenever $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$.*

**Remark 3.1.** If for every sequence $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$ there exists a subsequence $\{P_{n_k} \in \mathbf{P}_{n_k,0} : n_k \geq 1\}$ for which the statements in Assumption 3.1(i)-(iii) are satisfied with $P_{n_k}$ in place of $P_n$, then the conclusion of Theorem 3.1 can be strengthened as follows: for any $\alpha \in (0,1)$, $\phi(S_n)$ defined in (13) satisfies

$$\sup_{P_n \in \mathbf{P}_{n,0}} |E_{P_n}[\phi(S_n)] - \alpha| \to 0$$

as $n \to \infty$. ■

**Remark 3.2.** As described in Remark 2.1, the validity of the randomization test in finite samples is tightly related to fact that the conditional distribution of $X$ given $X \in \mathbf{G}^x$ is uniform on $\mathbf{G}^x$. While this property holds for the limiting random variable $S$ in our framework, it may not hold even approximately for $S_n$ for large $n$. The proof of Theorem 3.1 instead uses a novel argument that exploits the Almost Sure Representation Theorem (see, e.g., Theorem 2.19 in van der Vaart, 1998). ■

**Remark 3.3.** Earlier work on the asymptotic behavior of randomization tests includes Hoeffding (1952), Romano (1989) and Romano (1990). The arguments in these papers involve showing that the "randomization distribution" (see, e.g., Chapter 15 of Lehmann and Romano, 2005) settles down to a fixed distribution as $|\mathbf{G}| \to \infty$. In our framework, $|\mathbf{G}|$ is fixed and the "randomization distribution" will generally not settle down at all. For this reason, the analysis in these papers is not useful in our setting. ■

**Remark 3.4.** Comments analogous to those made in Remarks 2.2-2.4 after Theorem 2.1 apply to Theorem 3.1. In particular, Theorem 3.1 still holds when $\mathbf{G}$ is replaced by $\hat{\mathbf{G}}$ defined in (8), asymptotically valid $p$-values can be computed using (9), and the non-randomized test that rejects if $T(S_n) > T^{(k)}(S_n)$ is also asymptotically level $\alpha$. ■

# 4 Applications

In this section we present three applications of Theorem 3.1 to settings where the data may be grouped into a fixed number of "clusters," $q$, with a large number of observations within each cluster: time series regression, differences-in-differences, and clustered regression. Before proceeding to these specific examples, we highlight a common structure found in all of the applications.

Suppose the researcher observes data $X^{(n)} \sim P_n \in \mathbf{P}_n$ and considers testing the hypotheses in (12) with

$$\mathbf{P}_{n,0} = \{P_n \in \mathbf{P}_n : \theta_n(P_n) = \theta_0\} ,$$

where $\theta_n(P_n) \in \Theta \subseteq \mathbf{R}^d$ is some parameter of interest. Further suppose that the data $X^{(n)}$ can be grouped into $q$ clusters, $X_1^{(n)}, \ldots, X_q^{(n)}$, where the clusters are allowed to have observations in common. Let $\hat{\theta}_{n,j} = \hat{\theta}_{n,j}(X_j^{(n)})$ be an estimator of $\theta_n(P_n)$ based on observations from the $j$th cluster such that whenever $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$,

$$S_n(X^{(n)}) = \sqrt{n}(\hat{\theta}_{n,1} - \theta_0, \ldots, \hat{\theta}_{n,q} - \theta_0) \xrightarrow{d} N(0, \Sigma) \tag{15}$$

as $n \to \infty$, where $\Sigma = \text{diag}\{\Sigma_1, \ldots, \Sigma_q\}$ and each $\Sigma_j$ is of dimension $d \times d$. In this setting, the conditions of Theorem 3.1 hold for $\mathbf{G} = \{-1, 1\}^q$ and $T(S_n) = T_{\text{Wald}}(S_n)$, where

$$T_{\text{Wald}}(S_n) = q\bar{S}_{n,q}'\bar{\Sigma}_{n,q}^{-1}\bar{S}_{n,q} \tag{16}$$

with

$$\bar{\Sigma}_{n,q} = \frac{1}{q}\sum_{j=1}^{q} S_{n,j}S_{n,j}' \ , \ \bar{S}_{n,q} = \frac{1}{q}\sum_{j=1}^{q} S_{n,j} \ , \text{ and } S_{n,j} = \sqrt{n}(\hat{\theta}_{n,j} - \theta_0) \ .$$

See Lemma D.3 in the Appendix for details. In the special case where $d = 1$, the conditions of Theorem 3.1 also hold for $T(S_n) = T_{|t\text{-stat}|}(S_n)$, where

$$T_{|t\text{-stat}|}(S_n) = \frac{|\bar{S}_{n,q}|}{\sqrt{\frac{1}{q-1}\sum_{j=1}^{q}(S_{n,j} - \bar{S}_{n,q})^2}} \ .$$

See Lemmas D.1-D.2 in the Appendix for details. Equivalently,

$$T_{|t\text{-stat}|}(S_n) = \frac{|\bar{\hat{\theta}}_{n,q} - \theta_0|}{s_{\hat{\theta}}/\sqrt{q}} \ , \tag{17}$$

with

$$\bar{\hat{\theta}}_{n,q} = \frac{1}{q}\sum_{j=1}^{q} \hat{\theta}_{n,j} \text{ and } s^2_{\hat{\theta}} = \frac{1}{q-1}\sum_{j=1}^{q}(\hat{\theta}_{n,j} - \bar{\hat{\theta}}_{n,q})^2 \ .$$

In each of the applications below, we will therefore simply specify $X_j^{(n)}$ and $\hat{\theta}_{n,j}$ and argue that the convergence (15) holds when $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$.

**Remark 4.1.** In the special case where $d = 1$, the idea of grouping the data in this way and constructing estimators satisfying (15) has been previously proposed by Ibragimov and Müller (2010). Using the result on the $t$-test described in Section 2.1.1, they go on to propose a test that rejects the null hypothesis when $T_{|t\text{-stat}|}(S_n)$ in (17) exceeds the $1 - \frac{\alpha}{2}$ quantile of a $t$-distribution with $q - 1$ degrees of freedom. Further comparisons with this approach are provided below. ■

**Remark 4.2.** The convergence (15) permits dependence within each cluster. It also permits some dependence across clusters. See, for example, Jenish and Prucha (2009) for some relevant central limit theorems. The convergence (15) further allows for heterogeneity in the distribution of the data across clusters in the sense that $\Sigma_j$ need not be independent of $j$ in $\Sigma = \text{diag}\{\Sigma_1, \ldots, \Sigma_q\}$. ■

**Remark 4.3.** The asymptotic normality in (15) arises frequently in applications, but is not necessary for the validity of the test described above. All that is required is that the $q$ estimators (after an appropriate re-centering and scaling) have a limiting distribution that is the product of $q$ distributions that are symmetric about zero. This may even hold in cases where the estimators have infinite variances or are inconsistent. See Remark 4.9 below. ■

**Remark 4.4.** The test statistics in (16) and (17) are both invariant under scalar multiplication. As a result, the $\sqrt{n}$ in the definition of $S_n$ in (15) may be omitted or replaced with another sequence without changing the results. ■

**Remark 4.5.** Note that the convergence (15) allows the number of observations within each cluster to differ across clusters provided that ratio of the number of observations in any pair of clusters tends to a finite, nonzero limit as $n \to \infty$. ■

## 4.1 Time Series Regression

Suppose

$$Y_t = Z_t'\theta + \epsilon_t \text{ with } E[\epsilon_t Z_t] = 0 \ . \tag{18}$$

Here, the observed data is given by $X^{(n)} = \{(Y_t, Z_t) : 1 \leq t \leq n\} \sim P_n$ taking values on a sample space $\mathcal{X}_n = \prod_{1 \leq t \leq n} \mathbf{R} \times \mathbf{R}^d$. The scalar random variable $\epsilon_t$ is unobserved and $\theta \in \Theta \subseteq \mathbf{R}^d$ is the

parameter of interest. We focus on the linear case here for ease of exposition, but the construction we describe below applies more generally.

In order to state the null and alternative hypotheses formally, it is useful to introduce some further notation. Let $W^{(\infty)} = \{(\epsilon_t, Z_t) : 1 \leq t < \infty\} \sim Q \in \mathbf{Q}$ taking values on a sample space $\mathcal{W}_\infty = \prod_{1 \leq t < \infty} \mathbf{R} \times \mathbf{R}^d$ and $A_{n,\theta} : \mathcal{W}_\infty \to \mathcal{X}_n$ be the mapping implied by (18). Our assumptions on $\mathbf{Q}$ are discussed below. Using this notation, define

$$\mathbf{P}_n = \bigcup_{\theta \in \Theta} \mathbf{P}_n(\theta) \text{ with } \mathbf{P}_n(\theta) = \{Q A_{n,\theta}^{-1} : Q \in \mathbf{Q}\} \ .$$

Here, $A_{n,\theta}^{-1}$ denotes the pre-image of $A_{n,\theta}$. The null and alternative hypotheses of interest are thus given by (12) with $\mathbf{P}_{n,0} = \mathbf{P}_n(\theta_0)$.

As mentioned previously, in order to apply our methodology, we must specify $X_j^{(n)}$ and $\hat{\theta}_{n,j}$ and argue that the convergence (15) holds when $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$. To this end, for a pre-specified value of $q$, define

$$X_j^{(n)} = \{(Y_t, Z_t) : t = (j-1)b_n + 1, \ldots, jb_n\} \ ,$$

where $b_n = \lfloor n/q \rfloor$, and let $\hat{\theta}_{n,j}$ be the ordinary least squares estimator of $\theta$ in (18) using the data $X_j^{(n)}$. In other words, we divide the data into $q$ consecutive blocks of data of size $b_n$ and estimate $\theta$ using ordinary least squares within each block of data. For this choice of $X_j^{(n)}$ and $\hat{\theta}_{n,j}$, the convergence (15) holds when $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$ under a wide range of assumptions on $\mathbf{Q}$. Extensive discussions of such conditions can be found in Ibragimov and Müller (2010, Section 3.1) and Bester et al. (2011, Lemma 1). We therefore omit further discussion of these conditions here.

**Remark 4.6.** Our methodology allows for considerable heterogeneity in the sense that both

$$E\left[\frac{1}{b_n} \sum_{(j-1)b_n \leq t \leq jb_n} Z_t Z_t'\right] \text{ and } E\left[\frac{1}{b_n} \sum_{(j-1)b_n \leq t \leq jb_n} Z_t Z_t' \epsilon_t^2\right] \tag{19}$$

may depend on $j$ even asymptotically. With the exception of the $t$-test approach developed in Ibragimov and Müller (2010), the competing approaches we discuss in Section 5.1 below do not share this feature. ■

**Remark 4.7.** By replacing the time index $t$ with a vector index, as in Bester et al. (2011), we can accommodate more complicated dependence structures, such as those found in spatially dependent data or in panel data. ■

**Remark 4.8.** When $\mathbf{Q}$ includes distributions that are heavy-tailed, the asymptotic normality in (15) may fail, but the $q$ esimators (after an appropriate re-centering and scaling) may still have a limiting distribution that is the product of $q$ distributions that are symmetric about zero. Note in particular that the rate of convergence in this case may depend on the tail index of the distribution. See, for example, McElroy and Politis (2002) and Ibragimov and Müller (2010). Following the discussion in Remarks 4.3 and 4.4, the test described above remains valid in such situations. ■

### 4.2 Differences-in-Differences

Suppose

$$Y_{j,t} = \theta D_{j,t} + \eta_j + \gamma_t + \epsilon_{j,t} \quad \text{with} \quad E[\epsilon_{j,t}] = 0 \ . \tag{20}$$

Here, the observed data is given by $X^{(n)} = \{(Y_{j,t}, D_{j,t}) : j \in J_0 \cup J_1, t \in T_0 \cup T_1\} \sim P_n$ taking values on a sample space $\mathcal{X}_n = \prod_{j \in J_0 \cup J_1, t \in T_0 \cup T_1} \mathbf{R} \times \{0,1\}$, where $Y_{j,t}$ is the outcome of unit $j$ at time $t$, $D_{j,t}$ is the (non-random) treatment status of unit $j$ at time $t$, $T_0$ is the set of pre-treatment time periods, $T_1$ is the set of post-treatment time periods, $J_0$ is the set of controls units, and $J_1$ is the set of treatment units. The scalar random variables $\eta_j$, $\gamma_t$ and $\epsilon_{j,t}$ are unobserved and $\theta \in \Theta \subseteq \mathbf{R}$ is the parameter of interest.

As before, in order to state the null and alternative hypotheses formally, it is useful to introduce some further notation. Let $W^{(n)} = \{(\epsilon_{j,t}, \eta_j, \gamma_t, D_{j,t}) : j \in J_0 \cup J_1, t \in T_0 \cup T_1\} \sim Q_n \in \mathbf{Q}_n$ taking values on a sample space $\mathcal{W}_n = \prod_{j \in J_0 \cup J_1, t \in T_0 \cup T_1} \mathbf{R} \times \mathbf{R} \times \mathbf{R} \times \{0,1\}$ and $A_{n,\theta} : \mathcal{W}_n \to \mathcal{X}_n$ be the mapping implied by (20). Our assumptions on $\mathbf{Q}_n$ are discussed below. Using this notation, define

$$\mathbf{P}_n = \bigcup_{\theta \in \Theta} \mathbf{P}_n(\theta) \text{ with } \mathbf{P}_n(\theta) = \{Q_n A_{n,\theta}^{-1} : Q_n \in \mathbf{Q}_n\} \ .$$

The null and alternative hypotheses of interest are thus given by (12) with $\mathbf{P}_{n,0} = \mathbf{P}_n(\theta_0)$.

In order to apply our methodology, we must again specify $X_j^{(n)}$ and $\hat{\theta}_{n,j}$ and argue that the convergence (15) holds when $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$. Different specifications may be appropriate for different asymptotic frameworks. We first consider an asymptotic framework similar to the one in Conley and Taber (2011), where $|J_1| = q$ is fixed, $|J_0| \to \infty$, and $\min\{|T_0|, |T_1|\} \to \infty$ with $\frac{|T_1|}{|T_0|} \to c \in (0, \infty)$. A modification for an alternative asymptotic framework in which $|J_0|$ is also fixed is discussed in Remark 4.14 below. For such an asymptotic framework, for each $j \in J_1$, define

$$X_j^{(n)} = \{(Y_{k,t}, D_{k,t}) : k \in \{j\} \cup J_0, t \in T_0 \cup T_1\}$$

and let $\hat{\theta}_{n,j}$ be the ordinary least squares estimator of $\theta$ in (20) using the data $X_j^{(n)}$, including indicator variables appropriately in order to account for $\eta_j$ and $\gamma_t$. Note that in this case the $X_j^{(n)}$ are not disjoint. We may also express $\hat{\theta}_{n,j}$ more simply as

$$\hat{\theta}_{n,j} = \Delta_{n,j} - \frac{1}{|J_0|} \sum_{k \in J_0} \Delta_{n,k} \ , \tag{21}$$

where

$$\Delta_{n,k} = \frac{1}{|T_1|} \sum_{t \in T_1} Y_{k,t} - \frac{1}{|T_0|} \sum_{t \in T_0} Y_{k,t} \ .$$

It follows that for $\theta$ as in (20),

$$\begin{aligned}\sqrt{|T_1|}(\hat{\theta}_{n,j} - \theta) &= \sqrt{|T_1|}\left(\frac{1}{|T_1|}\sum_{t\in T_1}\epsilon_{j,t} - \frac{1}{|T_0|}\sum_{t\in T_0}\epsilon_{j,t}\right) \\ &\quad -\sqrt{|T_1|}\frac{1}{|J_0|}\sum_{k\in J_0}\left(\frac{1}{|T_1|}\sum_{t\in T_1}\epsilon_{k,t} - \frac{1}{|T_0|}\sum_{t\in T_0}\epsilon_{k,t}\right) \ .\end{aligned}$$

For this choice of $X_j^{(n)}$ and $\hat{\theta}_{n,j}$, the convergence (15) (with $|T_1|$ in place of $n$) therefore holds when $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$ under a wide range of assumptions on $\mathbf{Q}_n$. In particular, it suffices to assume that $\epsilon_j = (\epsilon_{j,t} : t \in T_0 \cup T_1)$ are independent across $j$, that for $1 \leq \ell \leq 2$

$$\frac{1}{|J_0|^2}\sum_{k\in J_0}\left(\frac{1}{|T_\ell|}\sum_{t\in T_\ell}\sum_{s\in T_\ell}E[\epsilon_{k,t}\epsilon_{k,s}]\right) \to 0 \ , \tag{22}$$

and that

$$\left(\frac{1}{\sqrt{|T_1|}}\sum_{t\in T_1}\epsilon_{j,t} \ , \frac{1}{\sqrt{|T_0|}}\sum_{t\in T_0}\epsilon_{j,t} : j \in J_1\right) \tag{23}$$

satisfies a central limit theorem (see, e.g., Politis et al., 1999, Theorem B.0.1).

**Remark 4.9.** The construction described above relies on the fact that $\min\{|T_0|, |T_1|\} \to \infty$ in order to apply an appropriate central limit theorem to (23). The construction remains valid, however, even if $|T_0|$ and $|T_1|$ are small provided that

$$\frac{1}{|T_1|}\sum_{t\in T_1}\epsilon_{j,t} \text{ and } \frac{1}{|T_0|}\sum_{t\in T_0}\epsilon_{j,t}$$

are independent and identically distributed. This property will hold, for example, if $|T_0| = |T_1|$ (which may be enforced by ignoring some time periods if necessary) and the distribution of $\epsilon_j$ is exchangeable (across $t$) for all $j$. While these assumptions may be strong, this discussion illustrates that the estimators $\hat{\theta}_{n,j}$ of $\theta$ need not even be consistent in order to apply our methodology. ■

**Remark 4.10.** The construction described above applies equally well in the case where (20) includes covariates $Z_{j,t}$. The estimators $\hat{\theta}_{n,j}$ of $\theta$ can no longer be expressed as in (21), but they may still be obtained using ordinary least squares using the $j$th cluster of data. Under an appropriate modification of the assumptions to account for the $Z_{j,t}$, the convergence (15) holds when $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$. ■

**Remark 4.11.** The requirement that $\epsilon_j$ are independent across $j$ can be relaxed using mixing conditions as in Conley and Taber (2011). In order to do so, it must be the case that the $\epsilon_j$ can be ordered linearly. ■

**Remark 4.12.** The construction described above applies equally well in the case where there are multiple observations for each unit $j$. This situation may arise, for example, when $j$ indexes states and individual-level data within each state is available. ■

**Remark 4.13.** The construction above may also be used if $T_0$ and $T_1$ vary across $j \in J_1$. In this case, we simply define $X_j^{(n)} = \{(Y_{k,t}, D_{k,t}) : k \in J_0 \cup \{j\}, t \in T_{0,j} \cup T_{1,j}\}$. ■

**Remark 4.14.** The requirement that $|J_0| \to \infty$ can be relaxed by modifying our proposed test in the following way. Suppose $|J_0|$ is fixed and that $|J_1| \leq |J_0|$ (if this is not the case, simply relabel treatment and control). Denote by $\{\tilde{J}_{0,l} : 1 \leq l \leq q\}$ a partition of $J_0$. For each $j \in J_1$, define

$$X_j^{(n)} = \{(Y_{k,t}, D_{k,t}) : k \in \tilde{J}_{0,j} \cup \{j\}, t \in T_0 \cup T_1\}$$

and let $\hat{\theta}_{n,j}$ be computed as before using the data $X_j^{(n)}$. For this choice of $X_j^{(n)}$ and $\hat{\theta}_{n,j}$, the convergence (15) continues to hold when $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$ under appropriate modifications of the assumptions described above. ■

### 4.3 Clustered Regression

Suppose

$$Y_{i,j} = \theta D_j + Z_{i,j}'\gamma + \epsilon_{i,j} \quad \text{with } E[\epsilon_{i,j}|D_j, Z_{i,j}] = 0 \ . \tag{24}$$

Here, the observed data is given by $X^{(n)} = \{(Y_{i,j}, Z_{i,j}, D_j) : i \in I_j, j \in J_0 \cup J_1\} \sim P_n$ taking values on a sample space $\mathcal{X}_n = \prod_{i \in I_j, j \in J_0 \cup J_1} \mathbf{R} \times \mathbf{R}^d \times \{0,1\}$, where $Y_{i,j}$ is the outcome of unit $i$ in area $j$, $Z_{i,j}$ is a vector of covariates of unit $i$ in area $j$, $D_j$ is the treatment status of area $j$, $I_j$ is the set of units in area $j$, $J_1$ is the set of treated areas, and $J_0$ is the set of untreated areas. The scalar random variable $\epsilon_{i,j}$ is unobserved, $\gamma \in \Gamma \subseteq \mathbf{R}^d$ is a nuisance parameter, and $\theta \in \Theta \subseteq \mathbf{R}$ is the parameter of interest. The mean independence requirement is stronger than needed; indeed, all that is required is that the $\epsilon_{i,j}$ is uncorrelated with $D_j$ and $Z_{i,j}$. For simplicity, we assume below that $|J_0| = |J_1| = q$, but the arguments are easily adapted to the case where $|J_0| \neq |J_1|$.

As before, in order to state the null and alternative hypotheses formally, it is useful to introduce some further notation. Let $W^{(n)} = \{(\epsilon_{i,j}, D_j, Z_{i,j}) : i \in I_j, j \in J_0 \cup J_1\} \sim Q_n \in \mathbf{Q}_n$ taking values on a sample space $\mathcal{W}_n = \prod_{i \in I_j, j \in J_0 \cup J_1} \mathbf{R} \times \{0,1\} \times \mathbf{R}^d$ and $A_{n,\theta,\gamma} : \mathcal{W}_n \to \mathcal{X}_n$ be the mapping implied by (24). Our assumptions on $\mathbf{Q}_n$ are discussed below. Using this notation, define

$$\mathbf{P}_n = \bigcup_{\theta \in \Theta, \gamma \in \Gamma} \mathbf{P}_n(\theta, \gamma) \text{ with } \mathbf{P}_n(\theta, \gamma) = \{Q_n A_{n,\theta,\gamma}^{-1} : Q_n \in \mathbf{Q}_n\} \ ,$$

where, as before, $A_{n,\theta,\gamma}^{-1}$ denotes the pre-image of $A_{n,\theta,\gamma}$. The null and alternative hypotheses of interest are thus given by (12) with

$$\mathbf{P}_{n,0} = \bigcup_{\gamma \in \Gamma} \mathbf{P}_n(\theta_0, \gamma) \ .$$

In order to apply our methodology, we must again specify $X_j^{(n)}$ and $\hat{\theta}_{n,j}$ and argue that the convergence (15) holds when $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$. Note that the clusters cannot be defined

by areas themselves because $\theta$ is not identified within a single area. Indeed, $D_j$ is constant within a single area. We therefore define the clusters by forming pairs of treatment and control areas, i.e., by matching each area in $J_1$ with an area in $J_0$. In experimental settings, such pairs are often suggested by the way in which treatment status was determined (see, e.g., the empirical application in Section 6). More specifically, for each $j \in J_1$, let $k(j) \in J_0$ be the area in $J_0$ that is matched with $j$. For each $j \in J_1$, define

$$X_j^{(n)} = \{(Y_{i,l}, Z_{i,l}, D_l) : i \in I_l, l \in \{j, k(j)\}\}$$

and let $\hat{\theta}_{n,j}$ be the ordinary least squares estimator of $\theta$ in (24) using the data $X_j^{(n)}$. For this choice of $X_j^{(n)}$ and $\hat{\theta}_{n,j}$, the convergence (15) holds when $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$ under a wide range of assumptions on $\mathbf{Q}_n$. Some such conditions can be found in Bester et al. (2011, Lemma 1).

# 5 Monte Carlo Simulations

## 5.1 Time Series Regression

In this section, we examine the finite-sample performance of our methodology with a simulation study designed around (18). Following Bester et al. (2011), we set

$$Z_t = 1 + \rho Z_{t-1} + \nu_{1,t}$$
$$\epsilon_t = \rho \epsilon_{t-1} + \nu_{2,t}$$

with $\theta = 1$ and $\{(\nu_{1,t}, \nu_{2,t}) : 1 \leq t \leq n\}$ distributed in one of the following three ways:

**N**: (*Normal*) $(\nu_{1,t}, \nu_{2,t}), t = 1, \ldots, n$ i.i.d. with a bivariate normal distribution with mean zero and identity covariance matrix.

**H**: (*Heterogeneous*) $\nu_{1,t} = a_t u_{1,t}$ and $\nu_{2,t} = b_t u_{2,t}$, where $(u_{1,t}, u_{2,t}), t = 1, \ldots, n$ are i.i.d. with

$$u_{\ell,t} \sim \frac{1}{3}N(-1, \frac{1}{2}) + \frac{1}{3}N(0, \frac{1}{2}) + \frac{1}{3}N(1, \frac{1}{2})$$

for all $1 \leq \ell \leq 2$ and $u_{1,t} \perp\!\!\!\perp u_{2,t}$ and the constants $a_t$ and $b_t$ are given by

$$a_t = \frac{1}{\sqrt{6}} I\{t \leq n/2\} + I\{t > n/2\} \quad \text{and} \quad b_t = \frac{1}{\sqrt{6}} I\{t \leq n/2\} + 3I\{t > n/2\} \ .$$

**HT**: (*Heavy-Tailed*) $(\nu_{1,t}, \nu_{2,t}), t = 1, \ldots, n$ are i.i.d. with $\nu_{1,t} \perp\!\!\!\perp \nu_{2,t}$ and, for $1 \leq \ell \leq 2$, $\nu_{\ell,t}$ has a $t$-distribution with 2 degrees of freedom for $t \leq \frac{n}{2}$ and a Pareto distribution with shape parameter 1 and scale parameter 2 re-centered to have mean zero for $t > \frac{n}{2}$.

Design N captures a homogeneous setting in the sense that the quantities in (19) do not depend on $j$. In other words, the distribution of observed data in this case is stationary. This design is considered by Bester et al. (2011). Design H, on the other hand, captures a heterogeneous (i.e., non-stationary) setting in the sense that the quantities in (19) depend on $j$ even asymptotically. Finally, design HT is not only heterogeneous (i.e., non-stationary), but also features heavy-tailed disturbances.

In the simulation results presented below, we compare our test (denoted Rand), the non-randomized version of our test (denoted NR R), and the following three alternative tests:

**IM**: This test is the one proposed by Ibragimov and Müller (2010). It is based on the result about the $t$-test developed by Bakirov and Székely (2006) and discussed in Section 2.1.1.

**BCH**: This test is the one proposed by by Bester, Conley and Hansen (2011). It rejects the null hypothesis when

$$\frac{\sqrt{n}|\hat{\theta}_n^F - \theta_0|}{\sqrt{\hat{\Gamma}_n^{-1}\hat{V}_n\hat{\Gamma}_n^{-1}}} \tag{25}$$

exceeds the $1-\frac{\alpha}{2}$ quantile of a $t$-distribution with $q-1$ degrees of freedom, where $\hat{\theta}_n^F$ is the ordinary least squares estimator of $\theta$ in (18) based on the full sample of data, $\hat{\Gamma}_n = n^{-1}\sum_{t=1}^n Z_t Z_t'$ and $\hat{V}_n$ is a "cluster covariance matrix estimator" with $q$ clusters.

**BRL**: This test is the one proposed by Bell and McCaffrey (2002), who refer to it as "bias reduced linearization." It is used by Angrist and Lavy (2009), whose analysis we revisit in our empirical application in Section 6. This test replaces $\hat{V}_n$ in (25) with a "bias reduced" version of it and rejects when the resulting quantity exceeds the $1-\frac{\alpha}{2}$ quantile of a $t$-distribution with degrees of freedom no greater than $q$. See page 8 of Bell and McCaffrey (2002) for exact expressions for the "bias reduced" covariance matrix estimator and the degrees of freedom correction. Further discussion is provided by Imbens and Kolesar (2012).

Table 1 reports rejection probabilities under the null hypothesis for our tests, Rand and NR R, as well as IM, BCH and BRL. The parameter values we use for the simulations are $n = 100$, $\alpha = 5\%$, $\rho \in \{0, 0.5, 0.8, 0.95\}$, and $q \in \{4, 8, 12\}$. All results are based on $10,000$ Monte Carlo repetitions. The results in Table 1 are consistent with the theoretical properties of our test. Relative to IM, Rand has rejection probabilities closer to the nominal level across all heterogeneous specifications (designs H and HT), while in the homogeneous specifications (design N) both tests perform similarly. This is consistent with Theorem 3.1, which shows that Rand has asymptotic rejection probability under the null hypothesis equal to the nominal level, while IM may have asymptotic rejection probability under the null hypothesis substantially below the nominal level when the data exhibit heterogeneity. Relative to BCH, Rand performs better under both heterogeneity and high levels of dependence (i.e., $\rho > 0.5$). Indeed, BCH is only shown to be valid under homogeneity in the distribution of

| | | Design N | | | | Design H | | | | Design HT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\rho$ | | | | $\rho$ | | | | $\rho$ | | | |
| $q$ | | 0 | 0.5 | 0.8 | 0.95 | 0 | 0.5 | 0.8 | 0.95 | 0 | 0.5 | 0.8 | 0.95 |
| | Rand | 5.0 | 5.2 | 5.2 | 5.4 | 5.1 | 5.1 | 5.3 | 5.3 | 5.2 | 5.3 | 5.5 | 5.2 |
| | IM | 4.9 | 5.0 | 5.2 | 5.0 | 2.7 | 2.8 | 2.9 | 5.0 | 2.7 | 2.9 | 3.2 | 3.4 |
| 4 | BCH | 5.4 | 6.1 | 8.2 | 16.2 | 18.1 | 17.6 | 18.8 | 18.2 | 8.4 | 9.0 | 10.7 | 17.6 |
| | BRL | 4.8 | 4.9 | 5.2 | 7.4 | 4.9 | 4.9 | 5.0 | 8.2 | 2.1 | 2.3 | 2.9 | 6.3 |
| | Rand | 5.0 | 5.4 | 5.8 | 5.4 | 4.9 | 5.3 | 5.8 | 5.6 | 5.2 | 5.4 | 5.7 | 5.1 |
| | NR R | 4.7 | 5.1 | 5.5 | 5.1 | 4.5 | 5.0 | 5.5 | 5.3 | 4.9 | 5.1 | 5.4 | 4.8 |
| 8 | IM | 4.7 | 5.2 | 5.6 | 5.0 | 3.7 | 4.0 | 4.4 | 5.4 | 2.9 | 3.2 | 3.6 | 3.6 |
| | BCH | 5.4 | 6.9 | 11.3 | 24.7 | 11.1 | 13.0 | 17.9 | 29.9 | 8.6 | 10.0 | 13.8 | 26.2 |
| | BRL | 4.7 | 5.3 | 7.0 | 14.6 | 6.9 | 7.4 | 8.7 | 19.2 | 1.8 | 2.3 | 4.1 | 12.7 |
| | Rand | 5.1 | 5.6 | 5.9 | 5.5 | 5.2 | 5.6 | 5.9 | 5.6 | 5.2 | 5.7 | 5.5 | 5.3 |
| | IM | 4.7 | 5.1 | 5.5 | 5.0 | 4.4 | 4.7 | 4.9 | 5.2 | 3.0 | 3.5 | 3.7 | 3.9 |
| 12 | BCH | 5.7 | 7.4 | 13.1 | 30.3 | 9.1 | 11.6 | 18.4 | 35.4 | 8.8 | 10.5 | 15.6 | 32.1 |
| | BRL | 4.9 | 5.7 | 8.8 | 21.7 | 6.4 | 7.4 | 10.5 | 42.2 | 1.8 | 2.5 | 5.4 | 20.1 |

Table 1: Rejection probabilities (in %) under the null hypothesis for different designs in the time series regression example.

$Z_t Z_t'$, which is violated in the heterogeneous specifications (designs H and HT), while Rand does not require such homogeneity assumptions. Relative to BRL, Rand performs better in most cases, except in design N with low levels of dependence (i.e., $\rho \leq 0.5$), in which case both tests perform well. BRL performs poorly under heterogeneity and higher levels of dependence, exhibiting both under-rejection (1.8%) and over-rejection (42.2%).

Overall, across all specifications, the rejection rates of Rand under the null hypothesis are between 4.9% and 5.9%. We also report results for NR R for the case $q = 8$. Its performance is very similar to that of Rand. Indeed, for $q = 12$, both Rand and NR R are numerically identical, so we omit these results in Table 1. Note that for $q = 4$, NR R is the trivial test, i.e., the test that simply does not reject, so we do omit these results in Table 1. See also Remark 2.4.

Figure 3 reports size-adjusted power curves for NR R, IM, BCH and BRL. The results are for designs N and H with $q = 8$ and $\rho \in \{0.8, 0.95\}$. In all scenarios, the size-adjusted power of NR R and IM are quite similar, the size-adjusted power of BCH and BRL are quite similar, and NR R and IM significantly outperform BRL and BCH. The difference in power is smallest for design N with $\rho = 0.8$. In unreported results for design N with $\rho \in \{0, 0.5\}$, BCH and BRL have size-adjusted power similar to Rand and IM. Finally, the size-adjusted power of all four tests for design HT are

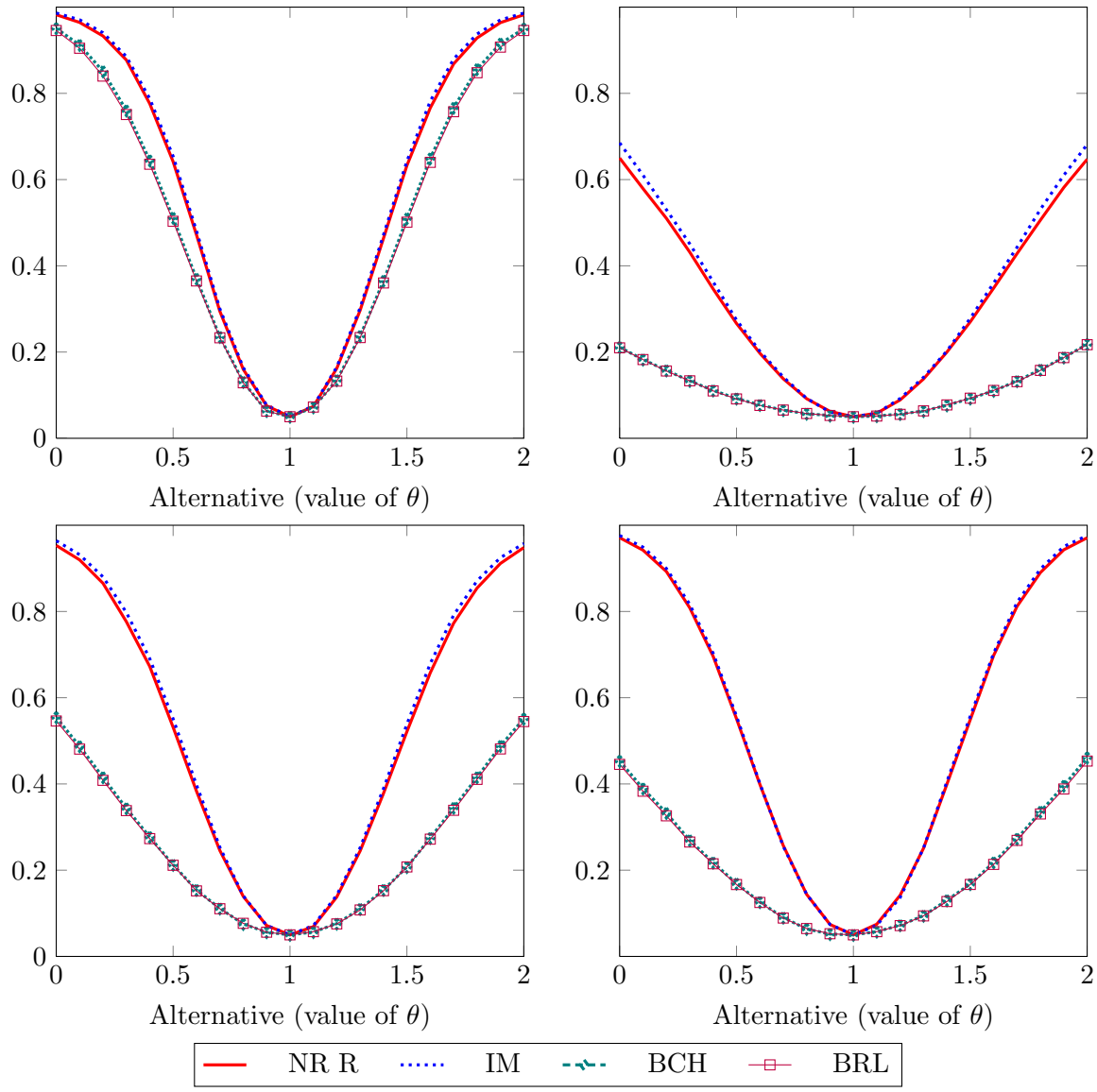


Figure 3: Size-adjusted power curves in the time series regression example with $q = 8$. Design N and $\rho = 0.8$ (upper left panel), Design H and $\rho = 0.8$ (upper right panel), Design N and $\rho = 0.95$ (lower left panel), and Design H and $\rho = 0.95$ (lower right panel).

very similar, so we do not report the results here.

It is important to emphasize that Rand and NR R have additional advantages over these competing tests that are not visible in the simulation study. First, they are available for any $\alpha \in (0, 1)$, which, as mentioned in Remark 2.3, allows the computation of $p$-values. IM and BCH, on the other hand, require $\alpha \leq 8.3\%$ and $q \geq 2$ or $\alpha \leq 10\%$ and $2 \leq q \leq 14$. Second, they allow for inference on vector-valued parameters, while both IM and BCH are restricted to scalar parameters. Third, the tests can be used with a variety of test statistics instead of only the $t$-statistic. Finally, our approximate symmetry requirement accommodates a broader range of situations, including, for

example, situations with an infinite variance.

**Remark 5.1.** Bester et al. (2011) and Ibragimov and Müller (2010) show in a simulation study that their respective tests outperform conventional tests that replace $\hat{V}_n$ in (25) with a heteroskedasticity-autocorrelation consistent covariance matrix estimator and reject when the resulting quantity exceeds the $1-\frac{\alpha}{2}$ quantile of the $N(0,1)$ distribution. These tests are justified by requiring $q \to \infty$. Bester et al. (2011) and Ibragimov and Müller (2010) also find that their tests outperform the test proposed by Kiefer and Vogelsang (2002, 2005), in which the $1-\frac{\alpha}{2}$ quantile of the $N(0,1)$ distribution is replaced with an alternative critical value that does not require $q \to \infty$. We therefore do not include these tests in our comparisons. ■

**Remark 5.2.** As mentioned previously, BRL involves a "bias reduced" covariance matrix estimator and a degrees of freedom correction for the $t$-distribution with which the test statistic is compared. The bias correction is highlighted by Angrist and Pischke (2008, page 320) and is used by Angrist and Lavy (2009) without the degrees of freedom adjustment. All our simulations, however, suggest that the good performance of BRL is largely driven by the degrees of freedom correction. For example, for $q = 8$ and $\rho = 0.8$, the rejection probabilities under the null hypothesis of a test that uses the $1-\frac{\alpha}{2}$ quantile of a standard normal distribution instead of the $1-\frac{\alpha}{2}$ quantile of the appropriate $t$-distribution would be 14.7%, 19.2%, and 15.4% for each of the three designs. The corresponding numbers using the degrees of freedom correction are 7.0%, 8.7%, and 4.1%, as reported in Table 1. ■

**Remark 5.3.** Imbens and Kolesar (2012) propose an alternative degrees of freedom correction for BRL. The results using this alternative correction are essentially the same as those using the correction by Bell and McCaffrey (2002). We therefore do not include them in Table 1. ■

### 5.2 Differences-in-Differences

In this section, we examine the finite-sample performance of our methodology with a simulation study designed around (20). Following Conley and Taber (2011), we set

$$\begin{aligned} Y_{j,t} &= \theta D_{j,t} + \beta Z_{j,t} + \epsilon_{j,t} \\ \epsilon_{j,t} &= \rho \epsilon_{j,t-1} + \nu_{1,j,t} \\ Z_{j,t} &= \gamma D_{j,t} + \nu_{2,j,t} \end{aligned} \tag{26}$$

with $\theta = 1$, $\beta = 1$, $\gamma = 0.5$. The distributions of $\nu_{1,j,t}$, $\nu_{2,j,t}$ and $D_{j,t}$ and the value of $\rho$ are specified below. The first specification is our baseline specification, and the other specifications only deviate from it in the specified ways.

**(a)**: We set $|J_1| = 8$, $|J_0| + |J_1| = 100$, $|T_0| + |T_1| = 10$, $\rho = 0.5$,

$$D_{j,t} = \begin{cases} 0 & \text{if } j \in J_0 \\ 0 & \text{if } j \in J_1 \text{ and } t < t_j^\star \\ 1 & \text{if } j \in J_1 \text{ and } t \geq t_j^\star \end{cases},$$

where $t_j^\star = \min\{2j, |T_0| + |T_1|\}$, and (independently of all other variables) $(\nu_{1,j,t}, \nu_{2,j,t}), j \in J_0 \cup J_1, t \in T_0 \cup T_1$ are i.i.d. $N(0, I_2)$, where $I_2$ is the two-dimensional identity matrix.

**(b)**: Everything as in (a), but $|J_0| + |J_1| = 50$.

**(c)**: Everything as in (a), but $|J_1| = 12$.

**(d)**: Everything as in (a), but $t_j^\star = \frac{|T_0|+|T_1|}{2}$.

**(e)**: Everything as in (a), but $\rho = 0.95$.

**(f)**: Everything as in (a), but $|T_0| + |T_1| = 3$.

**(g)**: Everything as in (a), but $\nu_{1,j,t}, j \in J_0, t \in T_0 \cup T_1$ are i.i.d. $\sim N(0, 1)$ and, independently, $\nu_{1,j,t}, j \in J_1, t \in T_0 \cup T_1$ are i.i.d. $\sim N(0, 4)$.

**(h)**: Everything as in (a), but $\nu_{1,j,t}, 1 \leq j \leq 4, t \in T_0 \cup T_1$ are i.i.d. $\sim N(0, 16)$ and, independently, $\nu_{1,j,t}, 4 < j \leq 100, t \in T_0 \cup T_1$ are i.i.d. $\sim N(0, 1)$.

In the simulation results presented below, we compare our tests, Rand and NR R, the IM and BRL tests described in the previous subsection, and the following three additional tests:

**CT**: This test is the one proposed by Conley and Taber (2011). It is based on $\hat{\theta}_n^F$, the ordinary least squares estimator of $\theta$ in (26) based on the full sample of data. In an asymptotic framework in which $|J_1|$ is fixed and $|J_0| \to \infty$, they show that $\hat{\theta}_n^F \xrightarrow{p} \theta + W$, where $W$ is a random variable defined in Conley and Taber (2011, Proposition 1). They then propose a novel approach to approximate the distribution of $W$ using simulation that is valid under the assumption that $(\epsilon_{j,t} : t \in T_0 \cup T_1)$ is i.i.d. across $j$ and independent of $(D_{j,t}, Z_{j,t} : t \in T_0 \cup T_1)$ (see Conley and Taber, 2011, Proposition 2).

**CCE**: This test is the one proposed by Bertrand et al. (2004). This test replaces $\hat{V}_n$ in (25) with a "cluster covariance matrix estimator" with $|J_0| + |J_1|$ clusters and rejects when the resulting quantity exceeds the $1 - \frac{\alpha}{2}$ quantile of a standard normal distribution.

**CGM**: This test is the one proposed by Cameron et al. (2008) based on the wild bootstrap. The authors argue that this test provides a higher-order asymptotic refinement over some other methods, such as CCE. See Cameron et al. (2008) for further details on implementation.

| Spec. | Rejection probabilities under $\theta = 1$ | | | | | | | Rejection probabilities under $\theta = 0$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rand | NR R | IM | CT | CCE | CGM | BRL | Rand | NR R | IM | CT | CCE | CGM | BRL |
| (a) | 5.58 | 5.26 | 5.51 | 5.85 | 10.37 | 5.88 | 4.16 | 66.49 | 65.21 | 67.70 | 80.37 | 81.11 | 66.42 | 64.13 |
| (b) | 6.39 | 6.01 | 6.32 | 7.21 | 9.36 | 5.61 | 3.93 | 64.69 | 63.35 | 65.60 | 78.74 | 79.51 | 65.15 | 63.89 |
| (c) | 6.26 | 6.26 | 6.10 | 6.42 | 8.52 | 5.35 | 4.41 | 85.37 | 85.37 | 85.58 | 91.28 | 89.76 | 84.54 | 81.36 |
| (d) | 5.56 | 5.32 | 5.57 | 6.75 | 9.50 | 5.41 | 4.79 | 69.44 | 68.20 | 70.40 | 82.58 | 81.61 | 69.43 | 69.26 |
| (e) | 6.06 | 5.67 | 5.89 | 6.39 | 9.92 | 5.53 | 4.23 | 32.29 | 31.14 | 32.91 | 41.92 | 29.20 | 32.90 | 17.08 |
| (f) | 5.41 | 5.15 | 5.44 | 6.66 | 9.50 | 5.69 | 4.79 | 59.06 | 57.58 | 59.93 | 73.45 | 73.61 | 58.85 | 58.99 |
| (g) | 4.78 | 4.58 | 4.86 | 62.02 | 11.14 | 5.55 | 4.97 | 9.54 | 8.99 | 9.69 | 70.12 | 18.36 | 10.40 | 9.15 |
| (h) | 5.52 | 5.24 | 3.84 | 51.81 | 11.08 | 6.10 | 2.93 | 20.11 | 19.51 | 16.61 | 66.49 | 27.55 | 20.73 | 14.59 |

Table 2: Rejection probabilities (in %) under the null and alternative hypotheses for different designs in the differences-in-differences example.

Note that with $|J_0|+|J_1| = 100$ clusters, the test proposed by Bester et al. (2011) performs similarly to CCE. We therefore do not include it in our comparisons.

Table 2 reports rejection probabilities under the null hypothesis (i.e., $\theta = 1$) for our tests, Rand and NR R, as well as IM, CT, CCE, and BRL. Table 2 also reports rejection probabilities for these tests when $\theta = 0$. The tests are all conducted with $\alpha = 5\%$. All results are based on $10,000$ Monte Carlo replications. We find that Rand and NR R perform well across all specifications. IM performs well, although, as expected, it is has rejection probability less than the nominal level when there is heterogenenity (specification (h)). CT, on the other hand, works very well when the conditions in Conley and Taber (2011) are met, but it severely over-rejects when $(\epsilon_{j,t} : t \in T_0 \cup T_1)$ is not i.i.d. across $j$ (specifications (g) and (h)). CCE over-rejects in all designs. CGM works remarkably well across all designs, though in unreported simulations involving high levels of heterogeneity we found that it could mildly over-reject. See also Ibragimov and Müller (2013), who find in a clustered regression setting that CGM can over-reject dramatically. Finally, BRL under-rejects in some specifications and typically delivers the lowest power across all specifications.

**Remark 5.4.** Conley and Taber (2011) show in a simulation study that their test outperforms the test proposed by Donald and Lang (2007). We therefore do not include the test proposed by Donald and Lang (2007) in our comparisons. ■

**Remark 5.5.** Tests Rand and CT are valid under non-nested assumptions. Unlike CT, Rand is valid in settings where $(\epsilon_{j,t} : t \in T_0 \cup T_1)$ is not i.i.d. across $j$, which might arise, for example, when there is heteroskedasticity conditional on treatment. The test by Conley and Taber (2011), on the other hand, is valid even when $q = 1$, whereas NR R may have poor power when $q$ is very small. See Remark 2.4. ■

**Remark 5.6.** The rejection probabilities under the null hypothesis of a version of BRL without the degrees of freedom correction are close to those of CCE across all designs. For example, in specification (a), such a test has rejection probability equal to 8.52% instead of 4.2%. ■

# 6 Empirical Application

In this section we revisit the analysis of Angrist and Lavy (2009, henceforth AL09), who study the effect of cash awards on Bagrut achievement – the high school matriculation certificate in Israel. This certificate is awarded after a sequence of tests in 10th–12th grades and is a formal prerequisite for university admission. Certification is largely determined by performance on a series of exams given in 10th–12th grades. AL09 find that the program was most successful for girls and that the impact on girls was driven by "marginal" students, i.e., students close to achieving certification based on their performance on tests given before the twelfth grade.

## 6.1 Program details and data

In December 2000, 40 nonvocational high schools with the lowest 1999 Bagrut rates in a national ranking were selected to participate in the Achievement Awards demonstration. These schools were matched into 20 pairs based on lagged values of the primary outcome of interest, the average 1999 Bagrut rate. Treatment status was then assigned randomly (i.e., with equal probability) within each pair. Treated schools were contacted shortly after random assignment and every student in a treated schools who received a Bagrut was eligible for a payment. Five treated schools are noncompliers in the sense that principals in these schools did not inform teachers about the program after the initial orientation or indicated that they did not wish to participate. Although the program was initially intended as a program that would provide cash awards to high school students in every grade, the actual implementation of the program focused on seniors. Thus, our analysis below, which follows AL09, is limited to high school seniors.

Baseline data were collected in January 2001, while the main Bagrut outcome comes from tests taken in June of 2001. One of the schools closed immediately after the start of the program, so the sample consists of 19 pairs of schools (the 6th matched pair is omitted). The data are publicly available at `http://economics.mit.edu/faculty/angrist/data1/data/angrist`. Below we index schools by $j \in J_0 \cup J_1$, where $J_0$ is the set of untreated schools and $J_1$ is the set of treated schools, and students in the $j$th school by $i \in I_j$. The data include the following variables: $Y_{i,j}$ is an indicator for Bagrut achievement; $D_j$ is an indicator for treatment; $W_j$ is a vector of school-level covariates, including an indicator for Arab school, an indicator for Jewish religious schools, and indicators for each of the matched pairs; $Z_{i,j}$ is a vector of covariates, including parental school, number of siblings, immigrants states, and credit-unit weighted averages of test scores prior to January 2001.

### 6.2 Model and empirical results

The model in this section fits into the framework described in Section 4.3 as follows,

$$Y_{i,j} = \Lambda[\theta D_j + Z_{i,j}'\gamma + W_j'\delta] + \epsilon_{i,j} \quad \text{with} \quad E[\epsilon_{i,j}|D_j, Z_{i,j}, W_j] = 0 \ , \tag{27}$$

where $\Lambda[\cdot]$ is the identity or logistic transformation. The parameter of interest is $\theta \in \Theta \subseteq \mathbf{R}$. While not discussed explicitly in Section 4.3, the logistic version of this model is handled in exactly the same way after replacing the ordinary least squares estimator of $\theta$ with the maximum likelihood estimator.

AL09 estimate the model in (27) by ordinary least squares and maximum likelihood using the full sample of schools. In order to circumvent the problem of having a small number of clusters (39 clusters at the school level), they estimate standard errors using the bias-reduced covariance matrix estimator proposed by Bell and McCaffrey (2002). AL09 do not report confidence intervals or $p$-values, so we do not know the exact critical values they used. A closer look at the paper (e.g., on page 1395, where $t$-statistics range from 1.7 to 2.1, the authors write "the 2001 estimates for girls are on the order of 0.10, and most are at least marginally significantly different from zero") suggests that they are using the $1 - \frac{\alpha}{2}$ quantile of standard normal distribution. We therefore use this approach to construct their confidence intervals in Tables 3-5. We note, however, that this is not equivalent to the BRL test we described in Sections 5.1 and 5.2. See also Remarks 5.2 and 5.6 for a discussion of the differences between these two methods.

In order to apply our methodology, we follow Section 4.3 and divide the data into $q$ clusters. We require that the parameter of interest, $\theta$, is identified within each cluster. With this in mind, it is natural to consider the 19 clusters defined by the 19 matched pairs of schools. Unfortunately, such an approach does not allow for certain school-level covariates in (27) because in some of the pairs $D_j$ and $W_j$ are perfectly collinear. We therefore form clusters by grouping the 19 matched pairs of schools in a way that guarantees that $D_j$ and $W_j$ are not perfectly collinear within each cluster. The total number of clusters resulting from this strategy depends on the particular subpopulation under consideration. In the sample of boys and girls, we form $q = 11$ clusters: {1,3}, {2,4}, {5,8}, {7}, {9,10}, {11}, {12,13}, {14,15}, {16,17}, {18,20}, {19}; in the sample of girls only, we form $q = 9$ clusters: {1,3}, {16,4}, {5,7}, {2,12}, {10,11}, {8,19}, {13}, {14,15}, {18,20}. Here, the notation $\{a, b\}$ means that the $a$th and $b$th matched pairs are grouped together. The median number of students per cluster is approximately 400 when boys and girls are included and approximately 200 when only girls are included.

Table 3 reports results for our test and the corresponding results from AL09 at the 5% and 10% significance levels for the sample of boys and girls. Table 4 reports the same results for the sample of girls only. These results correspond to those in Table 2 on page 1394 in AL09. We report the average of the $q$ estimators as our point estimate and compute our confidence intervals using

| | Treatment Effect: Boys & Girls | | | |
|---|---|---|---|---|
| | Randomization Test | | Angrist and Lavy (2009) | |
| | OLS | Logit | OLS | Logit |
| Sch. cov. only | 0.049 | -0.017 | 0.052 | 0.054 |
| 90% | [ -0.078 , 0.164 ] | [ -0.147 , 0.093 ] | [ -0.025 , 0.130 ] | [ -0.016 , 0.125 ] |
| 95% | [ -0.109 , 0.182 ] | [ -0.180 , 0.105 ] | [ -0.040 , 0.144 ] | [ -0.030 , 0.138 ] |
| Lagged score, micro. cov. | 0.075 | 0.022 | 0.067 | 0.055 |
| 90% | [ -0.034 , 0.178 ] | [ -0.058 , 0.102 ] | [ 0.008 , 0.126 ] | [ -0.004 , 0.114 ] |
| 95% | [ -0.059 , 0.198 ] | [ -0.077 , 0.117 ] | [ -0.003 , 0.138 ] | [ -0.015 , 0.125 ] |

Table 3: Results corresponding to boys and girls in Table 2 in AL09.

| | Treatment Effect: Girls only | | | |
|---|---|---|---|---|
| | Randomization Test | | Angrist and Lavy (2009) | |
| | OLS | Logit | OLS | Logit |
| Sch. cov. only | 0.036 | 0.037 | 0.105 | 0.093 |
| 90% | [ -0.132 , 0.195 ] | [ -0.099 , 0.165 ] | [ 0.005 , 0.205 ] | [ 0.006 , 0.179 ] |
| 95% | [ -0.182 , 0.234 ] | [ -0.144 , 0.183 ] | [ -0.014 , 0.224 ] | [ -0.010 , 0.197 ] |
| Lagged score, micro. cov. | 0.090 | 0.058 | 0.105 | 0.097 |
| 90% | [ -0.049 , 0.226 ] | [ -0.020 , 0.140 ] | [ 0.027 , 0.182 ] | [ 0.021 , 0.172 ] |
| 95% | [ -0.099 , 0.256 ] | [ -0.047 , 0.157 ] | [ 0.012 , 0.197 ] | [ 0.006 , 0.187 ] |

Table 4: Results corresponding to girls only in Table 2 in AL09.

test inversion. The row labeled "Sch. cov. only" includes the case where only school covariates are included. The row labeled "Lagged score, micro. cov." includes the individual covariates as well. Our results in Table 3 for the sample of boys and girls are consistent with those in AL09 and show that $\theta$ is not statistically significantly different from zero. The conclusions change for the sample of girls only in Table 4. While the confidence intervals for AL09 are consistent with the claim on page 1395 in AL09 of $\theta$ being "marginally significantly different from zero," our confidence intervals do not support this assertion.

AL09 re-estimate the logistic specification of (27) for the sample of "marginal" girls. The define "marginal" in two different ways. The first scheme splits students into approximately equal-sized groups according to the credit unit-weighted average test scores prior to January 2001. The second scheme splits students into approximately equal-sized groups using the fitted values obtained by estimating the logistic specification of (27) using the untreated sample only. We replicate AL09's results and apply our randomization test to the resulting samples in Table 5. The results show

| | Treatment Effect: Girls on top half of cohort | | | |
|---|---|---|---|---|
| | Randomization Test | | Angrist and Lavy (2009) | |
| | by lagged score | by pred. probability | by lagged score | by pred. probability |
| Sch. cov. only | 0.089 | 0.081 | 0.206 | 0.194 |
| 90% | [ -0.077 , 0.259 ] | [ -0.099 , 0.262 ] | [ 0.076 , 0.335 ] | [ 0.067 , 0.320 ] |
| 95% | [ -0.129 , 0.289 ] | [ -0.156 , 0.295 ] | [ 0.051 , 0.360 ] | [ 0.043 , 0.344 ] |
| Lagged score, micro. cov. | 0.091 | 0.076 | 0.213 | 0.207 |
| 90% | [ -0.064 , 0.252 ] | [ -0.095 , 0.249 ] | [ 0.083 , 0.342 ] | [ 0.079 , 0.334 ] |
| 95% | [ -0.113 , 0.286 ] | [ -0.150 , 0.279 ] | [ 0.058 , 0.367 ] | [ 0.054 , 0.359 ] |

Table 5: Results corresponding to "marginal" girls only in Table 4 in AL09.

again that our test does not support AL09's claim that $\theta$ is statistically significantly different from zero for this subsample.

Overall, the results using our test do not support the finding in AL09 that cash awards appeared to have generated substantial increases in the matriculation rates of "marginal" girls, though, as in AL09, we found no evidence of negative or perverse effects of the program either.

# A Proof of Theorem 2.1

The proof of this result is not new to this paper and can be found in Hoeffding (1952) and Lehmann and Romano (2005, Chapter 15). We include it here for completeness.

Let $P \in \mathbf{P}_0$ be given. Since for every $x \in \mathcal{X}$, $T^{(j)}(x) = T^{(j)}(gx)$ for all $g \in \mathbf{G}$ and $1 \leq j \leq M$,

$$\sum_{g \in \mathbf{G}} \phi(gx) = M^+(x) + a(x)M^0(x) = M\alpha \ .$$

In addition, since $X \stackrel{d}{=} gX$ under $P$ for any $P \in \mathbf{P}_0$ and $g \in \mathbf{G}$, we have

$$M\alpha = E_P\left[\sum_{g \in \mathbf{G}} \phi(gX)\right] = \sum_{g \in \mathbf{G}} E_P[\phi(X)] = ME_P[\phi(X)] \ ,$$

and the result follows. ■

# B Optimality of Randomization Test

Define

$$\begin{aligned} \mathbf{P} &= \{\otimes_{j=1}^q P_{j,\mu} : P_{j,\mu} = N(\mu, \sigma_j^2) \text{ with } \mu \geq 0 \text{ and } \sigma_j^2 \geq 0\} \\ \mathbf{P}_0 &= \{\otimes_{j=1}^q P_{j,\mu} : P_{j,\mu} = N(\mu, \sigma_j^2) \text{ with } \mu = 0 \text{ and } \sigma_j^2 \geq 0\} \ . \end{aligned}$$

Let $X = (X_1, \ldots, X_q) \sim P \in \mathbf{P}$ consider testing (3) at level $\alpha \in (0,1)$. Below we argue that the randomization test with $T(X) = T_{t-\text{stat}}(X)$ and $\mathbf{G} = \{-1,1\}^q$ is the uniformly most powerful unbiased level $\alpha$ test against the restricted class of of alternatives with $\sigma_j^2 = \sigma^2 > 0$ for all $1 \leq j \leq q$. A similar argument can be used to establish the corresponding two-sided result for the randomization test with $T(X) = T_{|t-\text{stat}|}(X)$ and $\mathbf{G} = \{-1,1\}^q$ when $\mathbf{P}$ and $\mathbf{P}_0$ according to (10)-(11). Related results have been obtained previously in Lehmann and Stein (1949).

Consider a test $\tilde{\phi}(X) = \tilde{\phi}(X_1, \ldots, X_q)$. Since the test is unbiased, it must be the case that $E_P[\tilde{\phi}(X)] \leq \alpha$ for all $P \in \mathbf{P}_0$ and $E_P[\tilde{\phi}(X)] \geq \alpha$ for all $P \in \mathbf{P}_1$. Using the dominated convergence theorem, it is straightforward to show that the requirement of unbiasedness therefore implies that the test is similar, i.e., $E_P[\tilde{\phi}(X)] = \alpha$ for all $P \in \mathbf{P}_0$.

Next, note that $U = (|X_1|, \ldots, |X_n|)$ is sufficient for $\mathbf{P}_0$. Indeed, the distribution of $X|U$ under any $P \in \mathbf{P}_0$ is uniform over the $2^n$ points of the form $(\pm|X_1|, \ldots, \pm|X_n|)$. Furthermore, $\mathbf{P}_0^U$, the family of distributions for $U$ under $P$ as $P$ varies over $\mathbf{P}_0$, is complete. To see this, for $\gamma \in \mathbf{R}^n$, define $P_\gamma$ to be the distribution with density

$$C(\gamma) \exp\left(-\sum_{j=1}^n \gamma_j x_j^2\right) \ ,$$

where $C(\gamma)$ is an appropriate constant. By construction, $P_\gamma \in \mathbf{P}_0$, so the desired result follows from Theorem 4.3.1 in Lehmann and Romano (2005). Therefore, by Theorem 4.3.2 in Lehmann and Romano (2005), we see that all similar tests have Neyman structure, i.e., $E_P[\tilde{\phi}(X)|U = u] = \alpha$ for all $P \in \mathbf{P}_0$ and all $u$ except those in a set $N$ such that $\sup_{P\in\mathbf{P}_0} P\{U \in N\} = 0$.

To find an optimal test, we therefore maximize the power of the test under $P = \otimes_{j=1}^{q} N(\mu, \sigma^2)$ where $\mu > 0$ and $\sigma^2 > 0$. Under the null, the distribution of $X|U$ is uniform, as described above. Under this alternative, the conditional probability mass function is proportional to

$$\prod_{1\le i\le n} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{1\le i\le n} x_i^2 - 2\mu \sum_{1\le i\le n} x_i + \mu^2\right)\right) .$$

Since $\sum_{1\le i\le n} X_i^2$ is constant conditional on $U = u$, the Neyman-Pearson Lemma implies that the optimal (conditional) test rejects when $\sum_{1\le i\le n} X_i > c(u)$ and rejects with probability $\gamma(u)$ when $\sum_{1\le i\le n} X_i = c(u)$, where the constants $c(u)$ and $\gamma(u)$ are chosen so that the test has (conditional) rejection probability equal to $\alpha$. Such tests are, of course, randomization tests with underlying choice of test statistic equal to $\sum_{1\le i\le n} X_i$, and this test is identical to the randomization test with underlying choice of test statistic equal to $T_{t-\text{stat}}(X)$ (see Example 15.2.4 in Lehmann and Romano (2005) for details). Denote this test by $\phi(X)$.

It remains to show that $\phi(X)$ is indeed unbiased. By construction, it is similar and therefore has rejection probability $= \alpha$ for all $P \in \mathbf{P}_0$. To see that the rejection probability is $\ge \alpha$ under any $P \in \mathbf{P}_1$, note that $\phi(X)$ is weakly increasing in each of its arguments. We therefore have that $E_P[\phi(X_1+\mu, \ldots, X_n+\mu)] \ge \alpha$ for all $\mu > 0$ and any $P \in \mathbf{P}_0$, from which the desired result follows.

**Remark B.1.** It is important to emphasize that this optimality result, like the one in Ibragimov and Müller (2010), is only for a restricted class of alternatives. On the other hand, it can readily be shown that the specified randomization test is in fact admissible whenever the set of alternatives contains this class and $\alpha$ is a multiple of $\frac{1}{2^q}$. The argument hinges on the fact that the above argument using the Neyman-Pearson lemma together with Lemma D.1 below guarantees that the optimal test is non-randomized for these values of $\alpha$. ■

**Remark B.2.** The argument presented above in fact shows that the specified randomization test remains uniformly most powerful unbiased against the same class of alternatives even if $\mathbf{P}_0$ is enlarged so that each $P_{j,\mu}$ is only required to be symmetric about zero. ■

# C Proof of Theorem 3.1

Let $\{P_n \in \mathbf{P}_{n,0} : n \ge 1\}$ be given and define $M = |\mathbf{G}|$. By Assumption 3.1(i) and the Almost Sure Representation Theorem (c.f van der Vaart, 1998, Theorem 2.19), there exists $\tilde{S}_n$, $\tilde{S}$, and

$U \sim U(0,1)$, defined on a common probability space $(\Omega, \mathcal{A}, P)$, such that

$$\tilde{S}_n \to \tilde{S} \text{ w.p.1 } ,$$

$\tilde{S}_n \stackrel{d}{=} S_n$, $\tilde{S} \stackrel{d}{=} S$, and $U \perp (\tilde{S}_n, \tilde{S})$. Consider the randomization test based on $\tilde{S}_n$, this is,

$$\tilde{\phi}(\tilde{S}_n, U) \equiv \begin{cases} 1 & T(\tilde{S}_n) > T^{(k)}(\tilde{S}_n) \text{ or } T(\tilde{S}_n) = T^{(k)}(\tilde{S}_n) \text{ and } U < a(\tilde{S}_n) \\ 0 & T(\tilde{S}_n) < T^{(k)}(\tilde{S}_n) \end{cases} .$$

Denote the randomization test based on $\tilde{S}$ by $\tilde{\phi}(\tilde{S}, U)$, where the same uniform variable $U$ is used in $\tilde{\phi}(\tilde{S}_n, U)$ and $\tilde{\phi}(\tilde{S}, U)$.

Since $\tilde{S}_n \stackrel{d}{=} S_n$, it follows immediately that $E_{P_n}[\phi(S_n)] = E_P[\tilde{\phi}(\tilde{S}_n, U)]$. In addition, since $\tilde{S} \stackrel{d}{=} S$, Assumption 3.1(ii) and Theorem 2.1 imply that $E_P[\tilde{\phi}(\tilde{S}, U)] = \alpha$. It therefore suffices to show

$$E_P[\tilde{\phi}(\tilde{S}_n, U)] \to E_P[\tilde{\phi}(\tilde{S}, U)] \ . \tag{28}$$

In order to show (28), let $E_n$ be the event where the orderings of $\{T(g\tilde{S}) : g \in \mathbf{G}\}$ and $\{T(g\tilde{S}_n) : g \in \mathbf{G}\}$ correspond to the same transformations $g_{(1)}, \ldots, g_{(M)}$. We first claim that $I\{E_n\} \to 1$ w.p.1. To see this, note that by Assumption 3.1(iii) and $\tilde{S} \stackrel{d}{=} S$, any two $g, g' \in \mathbf{G}$ are such that either

$$T(gs) = T(g's) \ \forall s \in \mathcal{S} \ , \tag{29}$$

or

$$T(g\tilde{S}) \neq T(g'\tilde{S}) \text{ w.p.1 under } P \ . \tag{30}$$

It follows that there exists a set with probability one under $P$ such that for all $\omega \in \Omega$ in this set, $\tilde{S}_n(\omega) \to \tilde{S}(\omega)$ and $T(g\tilde{S}(\omega)) \neq T(g'\tilde{S}(\omega))$ for any two $g, g' \in \mathbf{G}$ not satisfying (29). For any $\omega$ in this set, let $g_{(1)}(\omega), \ldots, g_{(M)}(\omega)$ be the transformations such that

$$T(g_{(1)}(\omega)\tilde{S}(\omega)) \leq T(g_{(2)}(\omega)\tilde{S}(\omega)) \leq \cdots \leq T(g_{(M)}(\omega)\tilde{S}(\omega)) \ .$$

For any two consecutive elements $g_{(j)}(\omega)$ and $g_{(j+1)}(\omega)$ with $1 \leq j \leq M-1$, there are only two possible cases: either $T(g_{(j)}(\omega)\tilde{S}(\omega)) = T(g_{(j+1)}(\omega)\tilde{S}(\omega))$ or $T(g_{(j)}(\omega)\tilde{S}(\omega)) < T(g_{(j+1)}(\omega)\tilde{S}(\omega))$. If $T(g_{(j)}(\omega)\tilde{S}(\omega)) = T(g_{(j+1)}(\omega)\tilde{S}(\omega))$ then by (29) it follows that

$$T(g_{(j)}(\omega)\tilde{S}_n(\omega)) = T(g_{(j+1)}(\omega)\tilde{S}_n(\omega)) \ \forall n \geq 1 \ .$$

If $T(g_{(j)}(\omega)\tilde{S}(\omega)) < T(g_{(j+1)}(\omega)\tilde{S}(\omega))$, then

$$T(g_{(j)}(\omega)\tilde{S}_n(\omega)) < T(g_{(j+1)}(\omega)\tilde{S}_n(\omega)) \ \text{ for } n \text{ sufficiently large } ,$$

as $\tilde{S}_n(\omega) \to \tilde{S}(\omega)$ and the continuity of $T : \mathcal{S} \to \mathbf{R}$ and $g : \mathcal{S} \to \mathcal{S}$ imply that $T(g_{(j)}(\omega)\tilde{S}_n(\omega)) \to T(g_{(j)}(\omega)\tilde{S}(\omega))$ and $T(g_{(j+1)}(\omega)\tilde{S}_n(\omega)) \to T(g_{(j+1)}(\omega)\tilde{S}(\omega))$. We can therefore conclude that

$$I\{E_n\} \to 1 \ w.p.1 \ ,$$

which proves the first claim.

We now prove (28) in two steps. First, we note that

$$E_P[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] = E_P[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] \ . \tag{31}$$

This is true because, on the event $E_n$, if the transformation $g = g_{(m)}$ corresponds to the $m$th largest value of $\{T(g\tilde{S}) : g \in \mathbf{G}\}$, then this same transformation corresponds to the $m$th largest value of $\{T(g\tilde{S}_n) : g \in \mathbf{G}\}$. In other words, $\tilde{\phi}(\tilde{S}_n, U) = \tilde{\phi}(\tilde{S}, U)$ on $E_n$. Second, since $I\{E_n\} \to 1$ *w.p.*1 it follows that $\tilde{\phi}(\tilde{S}, U)I\{E_n\} \to \tilde{\phi}(\tilde{S}, U)$ w.p.1 and $\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\} \to 0$ w.p.1. We can therefore use (31) and invoke the dominated convergence theorem to conclude that,

$$\begin{aligned} E_P[\tilde{\phi}(\tilde{S}_n, U)] &= E_P[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] + E_P[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &= E_P[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] + E_P[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &\to E_P[\tilde{\phi}(\tilde{S}, U)] \ . \end{aligned}$$

This completes the proof. ■

# D Auxiliary Lemmas

**Lemma D.1.** *Let $S = (S_1, \ldots, S_q)$ where $S_j \perp\!\!\!\perp S_{j'}$ for all $j \neq j'$ and each $S_j$ is symmetrically distributed about 0. Let $\mathbf{W} = \{w = (w_1, \ldots, w_q) \in \mathbf{R}^q : w_j \neq 0$ for at least one $0 \leq j \leq q\}$. If for every $w \in \mathbf{W}$ and $w_0 \in \mathbf{R}$*

$$w_0 + \sum_{j=1}^{q} w_j S_j \neq 0 \ \textit{w.p.1} \ , \tag{32}$$

*then Assumption 3.1(iii) is satisfied for $T(S) = T_{t\text{-}stat}(S)$, where*

$$T_{t\text{-}stat}(S) = \frac{\bar{S}_q}{\sqrt{\frac{1}{q-1}\sum_{j=1}^{q}(S_j - \bar{S}_q)^2}} \quad \textit{with} \quad \bar{S}_q = \frac{1}{q}\sum_{j=1}^{q} S_j \ ,$$

*and $\mathbf{G} = \{-1, 1\}^q$. In particular, if the distribution of $S_j$ is absolutely continuous with respect to Lebesgue measure for all $1 \leq j \leq q$, then the requirement in (32) holds.*

Proof: We prove the result by contradiction. Suppose there exist two distinct elements $g, g' \in \mathbf{G}$ such that $T(gS) = T(g'S)$ with positive probability, where

$$T(gS) = \frac{\frac{1}{q}\sum_{j=1}^{q} g_j S_j}{\sqrt{\frac{1}{q-1}\sum_{j=1}^{q} S_j^2 - \frac{q}{q-1}(\sum_{j=1}^{q} g_j S_j)^2}} \ . \tag{33}$$

We first claim that the denominator in (33) is nonzero w.p.1 for all $g \in \mathbf{G}$. Let $\tilde{\sigma}_S^2 = \frac{1}{q-1}\sum_{j=1}^{q} S_j^2$, $\tilde{w}_0 = \sqrt{\frac{q-1}{q}\tilde{\sigma}_S^2}$, and note that $\tilde{\sigma}_S^2 - \frac{q}{q-1}(\sum_{j=1}^{q} g_j S_j)^2 = 0$ with positive probability if and only if

$$\tilde{w}_0 + \sum_{j=1}^{q} g_j S_j = 0 \quad \text{or} \quad -\tilde{w}_0 + \sum_{j=1}^{q} g_j S_j = 0$$

with positive probability. Since $g_j \neq 0$ for all $1 \leq j \leq q$, $(g_1, \ldots, g_q) \in \mathbf{W}$ and (32) implies this cannot happen.

We next note that $T(gS) = T(g'S)$ implies that

$$\frac{1}{q}\sum_{j=1}^{q} g_j S_j \left\{ \tilde{\sigma}_S^2 - \frac{q}{q-1}\left(\sum_{j=1}^{q} g_j' S_j\right)^2 \right\}^{1/2} = \frac{1}{q}\sum_{j=1}^{q} g_j' S_j \left\{ \tilde{\sigma}_S^2 - \frac{q}{q-1}\left(\sum_{j=1}^{q} g_j S_j\right)^2 \right\}^{1/2} .$$

Additional algebra using this last expression implies that $T(gS) = T(g'S)$ with positive probability if and only if

$$\sum_{j=1}^{q} \Delta g_j S_j = 0 \quad \text{or} \quad \sum_{j=1}^{q} (g_j + g_j') S_j = 0 \ , \tag{34}$$

where $\Delta g_j = g_j - g_j'$. Since $g$ and $g'$ are distinct, it follows that $\Delta g_j \neq 0$ for at least one $1 \leq j \leq q$ and so $(\Delta g_1, \ldots, \Delta g_q) \in \mathbf{W}$. By (32), $\sum_{j=1}^{q} \Delta g_j S_j \neq 0$ w.p.1. In addition, since $g \neq g'$, it follows that $g_j + g_j' \neq 0$ for at least one $1 \leq j \leq q$ and so $(g_1 + g_1', \ldots, g_q + g_q') \in \mathbf{W}$. By (32), $\sum_{j=1}^{q} (g_j + g_j') S_j \neq 0$ w.p.1. We conclude that (34) cannot hold with positive probability and this completes the first part of the proof.

To prove the last claim of the Lemma, let $Z(w) = \sum_{j=1}^{q} w_j S_j$ and suppose by way of contradiction that the requirement in (32) fails. Then, there exists $w_0 \in \mathbf{R}$ and $w \in \mathbf{W}$ such that $Z(w) = -w_0$ holds with positive probability. However, since $w_j \neq 0$ for at least one $0 \leq j \leq q$ and $S_j$ is continuously distributed for all $1 \leq j \leq q$, it follows that $Z(w)$ is continuously distributed for all $w \in \mathbf{W}$, which leads to a contradiction. ■

**Lemma D.2.** *Let $S = (S_1, \ldots, S_q)$ where $S_j \perp\!\!\!\perp S_{j'}$ for all $j \neq j'$ and each $S_j$ is symmetrically distributed about 0. Let $\mathbf{W} = \{w = (w_1, \ldots, w_q) \in \mathbf{R}^q : w_j \neq 0$ for at least one $0 \leq j \leq q\}$. If for every $w \in \mathbf{W}$ and $w_0 \in \mathbf{R}$,*

$$w_0 + \sum_{j=1}^{q} w_j S_j \neq 0 \text{ w.p.1} \ , \tag{35}$$

*then Assumption 3.1(iii) is satisfied for $T(S) = T_{|t\text{-stat}|}(S)$ defined in (17) and $\mathbf{G} = \{-1, 1\}^q$. In particular, if the distribution of $S_j$ is absolutely continuous with respect to Lebesgue measure for all $1 \leq j \leq q$, then the requirement in (35) holds.*

Proof: Let $T(S) = T_{|t\text{-stat}|}(S)$ as defined in (17). Take any two distinct elements $g, g' \in \mathbf{G}$ and consider the following two cases. If $g \neq -g'$, then the same arguments as those in the proof of Lemma D.1 show that $T(gS) \neq T(g'S)$ w.p.1. On the other hand, if $g' = -g$, then it follows that for any $s \in \mathcal{S}$,

$$T(gs) = \left| \frac{\frac{1}{q}\sum_{j=1}^{q} g_j s_j}{\frac{1}{q-1}\sum_{j=1}^{q} s_j^2 - \frac{q}{q-1}(\sum_{j=1}^{q} g_j s_j)^2} \right| = \left| -\frac{\frac{1}{q}\sum_{j=1}^{q} (-g_j) s_j}{\frac{1}{q-1}\sum_{j=1}^{q} s_j^2 - \frac{q}{q-1}(-\sum_{j=1}^{q} g_j s_j)^2} \right| = T(g's) \ .$$

The result follows. Finally, the proof of the last claim follows from the proof of Lemma D.1. ■

**Lemma D.3.** *Let $S = (S_1, \ldots, S_q)$ where $S_j \perp\!\!\!\perp S_{j'}$ for all $j \neq j'$ and each $S_j \in \mathbf{R}^d$ is symmetrically distributed about 0. Let $\mathbf{W} = \{w = (w_1, \ldots, w_q) \in \mathbf{R}^q : w_j \neq 0$ for at least one $0 \leq j \leq q\}$. If for every $w \in \mathbf{W}$ and $w_0 \in \mathbf{R}^d$*

$$w_0 + \sum_{j=1}^{q} w_j S_j \neq 0 \text{ w.p.1} , \tag{36}$$

*then Assumption 3.1(iii) is satisfied for $T(S) = T_{\text{Wald}}(S)$ defined in (16) and $\mathbf{G} = \{-1, 1\}^q$. In particular, if the distribution of $S_j$ is absolutely continuous with respect to Lebesgue measure on $\mathbf{R}^d$ for all $1 \leq j \leq q$, then the requirement in (36) holds.*

PROOF: Let $T(gS) = q\bar{S}_q(g)'\bar{\Sigma}_q^{-1}\bar{S}_q(g)$, where $\bar{\Sigma}_q = q^{-1}\sum_{j=1}^q g_j^2 S_j S_j'$ and $\bar{S}_q(g) = q^{-1}\sum_{j=1}^q g_j S_j$, noting that $\bar{\Sigma}_q$ is invariant to sign changes since $g_j^2 = 1$ for $1 \leq j \leq q$. Take two distinct elements $g, g' \in \mathbf{G} = \{-1, 1\}^q$ and consider the following two cases: either $g' = -g$ or $g \neq -g'$. If $g' = -g$, then for any $s \in \mathcal{S}$, $q\bar{s}_q(g) = \sum_{j=1}^q g_j s_j = -\sum_{j=1}^q -g_j s_j = -q\bar{s}_q(g')$. It follows immediately that $T(gs) = q\bar{s}_q(g)'\bar{\Sigma}_q^{-1}\bar{s}_q(g) = q\bar{s}_q(g')'\bar{\Sigma}_q^{-1}\bar{s}_q(g') = T(g's)$. If $g \neq -g'$, then we claim that $T(gS) \neq T(g'S)$ w.p.1. To this end, note that $\bar{\Sigma}_q$ is symmetric by definition and positive definite w.p.1 by (36). We can then write

$$T(gS) - T(g'S) = q(\bar{S}_q(g) - \bar{S}_q(g'))'\bar{\Sigma}_q^{-1}(\bar{S}_q(g) + \bar{S}_q(g')) .$$

Since $\bar{\Sigma}_q$ is positive definite w.p.1, it follows that $T(gS) = T(g'S)$ with positive probability if and only if

$$\bar{S}_q(g) - \bar{S}_q(g') = 0 \quad \text{or} \quad \bar{S}_q(g) + \bar{S}_q(g') = 0 , \tag{37}$$

with positive probability. First, note that $\bar{S}_q(g) - \bar{S}_q(g') = q^{-1}\sum_{j=1}^q \Delta g_j S_j$. Since $g$ and $g'$ are distinct, it follows that $\Delta g_j \neq 0$ for at least one $1 \leq j \leq q$ and so $(\Delta g_1, \ldots, \Delta g_q) \in \mathbf{W}$. By (36), $\bar{S}_q(g) - \bar{S}_q(g') \neq 0$ w.p.1. Second, note that $\bar{S}_q(g) + \bar{S}_q(g') = q^{-1}\sum_{j=1}^q (g_j + g_j')S_j$. Since $g + g' \neq 0$, it follows that $g_j + g_j' \neq 0$ for at least one $1 \leq j \leq q$ and so $(g_1 + g_1', \ldots, g_q + g_q') \in \mathbf{W}$. By (36), $\bar{S}_q(g) + \bar{S}_q(g') \neq 0$ w.p.1. We conclude that (37) cannot hold with positive probability and this completes the proof.

The proof of the last claim follows from arguments similar to those used in the proof of Lemma D.1. ■

## References


Angrist, J. D. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review* 1384–1414.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Bakirov, N. K. and Székely, G. (2006). Students $t$-test for Gaussian scale mixtures. *Journal of Mathematical Sciences*, **139** 6497–6505.

Bell, R. M. and McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, **28** 169–182.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, **119** 249–275.

Bester, C. A., Conley, T. G. and Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, **165** 137–151.

Cameron, A. C., Gelbach, J. B. and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, **90** 414–427.

Conley, T. G. and Taber, C. R. (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics*, **93** 113–125.

Donald, S. G. and Lang, K. (2007). Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, **89** 221–233.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, **23** 169–192.

Ibragimov, R. and Müller, U. K. (2010). $t$-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, **28** 453–468.

Ibragimov, R. and Müller, U. K. (2013). Inference with few heterogenous clusters. *Manuscript*.

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

Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of econometrics*, **150** 86–98.

Kiefer, N. M. and Vogelsang, T. J. (2002). Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. *Econometrica* 2093–2095.

KIEFER, N. M. and VOGELSANG, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, **21** 1130–1164.

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

LEHMANN, E. L. and STEIN, C. (1949). On the theory of some non-parametric hypotheses. *The Annals of Mathematical Statistics* 28–45.

MCELROY, T. and POLITIS, D. N. (2002). Robust inference for the mean in the presence of serial correlation and heavy-tailed distributions. *Econometric Theory*, **18** 1019–1039.

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. Springer, New York.

ROMANO, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, **17** 141–159.

ROMANO, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, **85** 686–692.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.