Supplement to "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

November 16, 2016

Abstract

This document provides additional results for the authors' paper "Randomization Tests under an Approximate Symmetry Assumption". It includes an application to time series regression, Monte Carlo simulations, an empirical application revisiting the analysis of Angrist and Lavy (2009), the proof of Theorem 2.1, and three auxiliary lemmas.

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. This research was supported in part through the computational resources and staff contributions provided for the Social Sciences Computing cluster (SSCC) at Northwestern University. Sergey Gitlin provided excellent research assistance.

†Research supported by NSF Grant SES-1530534.

‡Research supported by NSF Grant DMS-1307973.

§Research supported by NSF Grants DMS-1308260, SES-1227091, and SES-1530661.

S.1 Application: Time Series Regression

Suppose

$$Y_t = Z_t'\theta + \epsilon_t \text{ with } E[\epsilon_t Z_t] = 0 \ . \tag{S.1}$$

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 ϵ_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 (S.1). Our assumptions on $\mathbf{Q}$ are discussed below. Using this notation, define

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

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 in Section 4, in order to apply our methodology, we must specify $X_j^{(n)}$ and $\hat{\theta}_{n,j}$ and argue that the convergence (15) holds under weak assumptions on the sequence $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$. 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 θ in (S.1) 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 θ 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 under $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$ with $P_n = Q_n A_{n,\theta_0}^{-1}$ under weak assumptions on $\{Q_n \in \mathbf{Q}_n : n \geq 1\}$. 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 S.1.1. Our methodology allows for considerable heterogeneity in the sense that both

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

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 S.2.1 below do not share this feature. Note, however, that even this approach is only available for $d = 1$. ■

Remark S.1.2. 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 S.1.3. When **Q** includes distributions that are heavy-tailed, the asymptotic normality in (15) may fail, but the q estimators (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. ■

S.2 Monte Carlo Simulations

S.2.1 Time Series Regression

In this section, we examine the finite-sample performance of our methodology with a simulation study designed around (S.1). 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 (S.2) 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 (S.2) 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

$$\frac{\sqrt{n}|\hat{\theta}_n^F - \theta_0|}{\sqrt{\hat{\Gamma}_n^{-1}\hat{V}_n\hat{\Gamma}_n^{-1}}} \tag{S.3}$$

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 θ in (S.1) 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 S.3. This test replaces $\hat{V}_n$ in (S.3) 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 S.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 S.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 $Z_t Z_t'$, 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

		Design N				Design H				Design HT			
		ρ				ρ				ρ			
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 S.1: Rejection probabilities (in %) under the null hypothesis for different designs in the time series regression example.

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 S.1. Note that for $q = 4$, NR R is the trivial test, i.e., the test that simply does not reject, so we omit these results in Table S.1. See also Remark 2.4.

Figure S.1 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 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

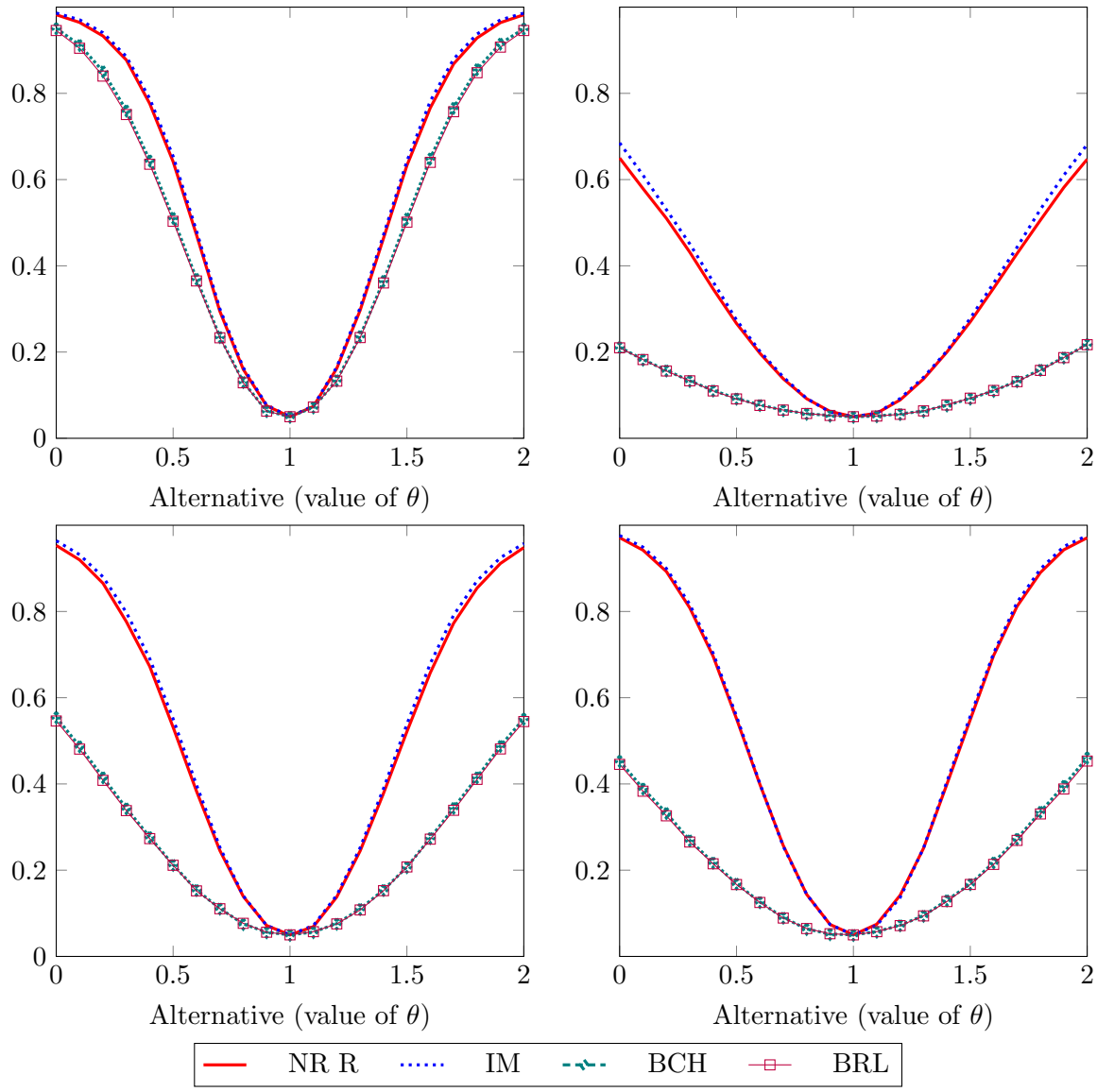

Figure S.1: Size-adjusted power curves in the time series regression example with $q = 8$. Design N and $\rho = 0.8$ (upper left panel), Design H and $\rho = 0.8$ (upper right panel), Design N and $\rho = 0.95$ (lower left panel), and Design H and $\rho = 0.95$ (lower right panel).

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.

Remark S.2.1. 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 (S.3) 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 S.2.2. 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 S.1. ■

Remark S.2.3. 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 S.1. ■

S.2.2 Differences-in-Differences

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

$$\begin{aligned} Y_{j,t} &= \theta D_{j,t} + \beta Z_{j,t} + \epsilon_{j,t} \\ \epsilon_{j,t} &= \rho \epsilon_{j,t-1} + \nu_{1,j,t} \\ Z_{j,t} &= \gamma D_{j,t} + \nu_{2,j,t} \end{aligned} \tag{S.4}$$

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 ρ 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 θ in (S.4) 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 (S.3) 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.

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 S.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 S.2 also reports rejection probabilities for these

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
(i)	7.00	6.65	5.81	7.55	8.92	5.50	2.93	25.32	24.36	22.77	18.81	30.48	23.46	17.14

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

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 (2016), 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 S.2.4. 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 S.2.5. 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 S.2.6. 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%. ■

S.3 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.

S.3.1 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 jth 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.

S.3.2 Model and empirical results

The model in this section fits into the framework described in Section 4.2 as follows,

$$Y_{i,j} = \Lambda[\theta D_j + Z_{i,j}'\gamma + W_j'\delta] + \epsilon_{i,j} \quad \text{with} \quad E[\epsilon_{i,j}|D_j, Z_{i,j}, W_j] = 0 \ , \tag{S.5}$$

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 4.2, the logistic version of this model is handled in exactly the same way after replacing the ordinary least squares estimator of θ with the maximum likelihood estimator.

AL09 estimate the model in (S.5) 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 S.3-S.5. We note, however, that this is not equivalent to the BRL test we described in Sections S.2.1 and S.2.2. See also Remarks S.2.2 and S.2.6 for a discussion of the differences between these two methods.

In order to apply our methodology, we follow Section 4.2 and divide the data into q clusters. We require that the parameter of interest, θ, 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 (S.5) 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 ath and bth 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 S.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 S.4 reports the same results for the sample of girls only. These results correspond to those in Table 2 on page 1394 in AL09. For comparison, we report the average of the q estimators as our point estimate, though there is no

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

reason one could not report a different estimator, such as the full-sample estimator used in AL09. We compute our confidence intervals using 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 S.3 for the sample of boys and girls are consistent with those in AL09 and show that θ is not statistically significantly different from zero. The conclusions change for the sample of girls only in Table S.4. While the confidence intervals for AL09 are consistent with the claim on page 1395 in AL09 of θ being "marginally significantly different from zero," our confidence intervals do not support this assertion.

AL09 re-estimate the logistic specification of (S.5) 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 (S.5) using the untreated sample only. We replicate AL09's

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

results and apply our randomization test to the resulting samples in Table S.5. The results show again that our test does not support AL09's claim that θ 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.

S.4 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. ■

S.5 Auxiliary Lemmas

Lemma S.5.1. *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 $w_0 \in \mathbf{R}$*

$$w_0 + \sum_{j=1}^{q} w_j S_j \neq 0 \text{ w.p.1} , \tag{S.6}$$

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

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

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 (S.6) 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

$$T(gS) = \frac{\frac{1}{q}\sum_{j=1}^{q} g_j S_j}{\sqrt{\frac{1}{q-1}\sum_{j=1}^{q} S_j^2 - \frac{q}{q-1}(\sum_{j=1}^{q} g_j S_j)^2}} . \tag{S.7}$$

We first claim that the denominator in (S.7) 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 (S.6) 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

$$\sum_{j=1}^{q} \Delta g_j S_j = 0 \quad \text{or} \quad \sum_{j=1}^{q} (g_j + g_j') S_j = 0 , \tag{S.8}$$

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 (S.6), $\sum_{j=1}^{q} \Delta g_j S_j \neq 0$ 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 (S.6), $\sum_{j=1}^q (g_j + g_j') S_j \neq 0$ w.p.1. We conclude that (S.8) 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 (S.6) 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 S.5.2. *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 $w_0 \in \mathbf{R}$,*

$$w_0 + \sum_{j=1}^q w_j S_j \neq 0 \text{ w.p.1} , \tag{S.9}$$

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 (S.9) 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 S.5.1 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 S.5.1. ■

Lemma S.5.3. *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 $w_0 \in \mathbf{R}^d$*

$$w_0 + \sum_{j=1}^q w_j S_j \neq 0 \text{ w.p.1} , \tag{S.10}$$

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 (S.10) 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 (S.10). 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

$$\bar{S}_q(g) - \bar{S}_q(g') = 0 \quad \text{or} \quad \bar{S}_q(g) + \bar{S}_q(g') = 0 \ , \tag{S.11}$$

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 (S.10), $\bar{S}_q(g) - \bar{S}_q(g') \neq 0$ 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 (S.10), $\bar{S}_q(g) + \bar{S}_q(g') \neq 0$ w.p.1. We conclude that (S.11) 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 S.5.1. ■

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). *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. (2016). Inference with few heterogenous clusters. *The Review of Economics and Statistics*, **98** 83–96.

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

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.

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.