Queen's Economics Department Working Paper No. 1355

Randomization Inference for Difference-in-Differences with Few Treated Clusters

James G. MacKinnon
Queen's University

Matthew D. Webb
Carleton University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

3-2018

Randomization Inference for Difference-in-Differences with Few Treated Clusters *

James G. MacKinnon
Queen's University
jgm@econ.queensu.ca

Matthew D. Webb
Carleton University
matt.webb@carleton.ca

March 24, 2018

Abstract

Inference using difference-in-differences with clustered data requires care. Previous research has shown that, when there are few treated clusters, t tests based on cluster-robust variance estimators (CRVEs) severely over-reject, different variants of the wild cluster bootstrap can either over-reject or under-reject dramatically, and procedures based on randomization inference show promise. We study two randomization inference (RI) procedures. A procedure based on estimated coefficients, which is essentially the one proposed by Conley and Taber (2011), has excellent power but may not perform well when the treated clusters are atypical. We therefore propose a new RI procedure based on t statistics. It typically performs better under the null, except when there is just one treated cluster, but at the cost of some power loss. Two empirical examples demonstrate that alternative procedures can yield dramatically different inferences.

Keywords: CRVE, grouped data, clustered data, panel data, wild cluster bootstrap, difference-in-differences, DiD, randomization inference

*This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are grateful to Taylor Jaworski for directing our attention to what became one of the empirical examples, and to three referees and an editor for helpful comments. We would also like to thank Chris Taber and participants at the University of Calgary Empirical Microeconomics Conference (2015), the Canadian Econometric Study Group (2015), McMaster University, New York Camp Econometrics (2016), NYU-Shanghai, the CIREQ Econometrics Conference in Honor of Jean-Marie Dufour, University of Copenhagen, the Society of Labor Economists (2016), the Canadian Economics Association Meetings (2016), and Dalhousie University for helpful comments on preliminary versions. Code for the proposed procedure is available from the authors.

1 Introduction

Inference for estimators that use clustered data, which in practice are very often difference-in-differences estimators, has received considerable attention in the past decade. Cameron and Miller (2015) provides a recent and comprehensive survey. While much progress has been made, there are still situations in which reliable inference is a challenge. It is particularly challenging when there are very few treated clusters. Past research, including Conley and Taber (2011), has shown that inference based on cluster-robust t statistics greatly over-rejects in this case. MacKinnon and Webb (2017b) explains why this happens and why the wild cluster bootstrap of Cameron, Gelbach and Miller (2008) does not solve the problem. In fact, the wild cluster bootstrap either greatly under-rejects or greatly over-rejects, depending on whether or not the null hypothesis is imposed on the bootstrap DGP.

Several authors have considered randomization inference (RI) as a way to obtain tests with accurate size when there are few treated groups (Barrios, Diamond, Imbens and Kolesár, 2012; Conley and Taber, 2011; Ferman and Pinto, 2015; Canay, Romano and Shaikh, 2017). We focus on procedures like the one proposed by Conley and Taber which use OLS estimates and are designed for samples with very few treated clusters, with many control clusters, and clustering at the 'state' level.

RI procedures necessarily rely on strong assumptions about the comparability of the control and treated groups. We show that, for any procedure based on estimated coefficients, such as the Conley-Taber procedure, these assumptions necessarily fail to hold in certain commonly-encountered cases. In particular, they fail to hold when the treated groups have either more or fewer observations than the control groups. As a consequence, the procedure can over-reject or under-reject quite severely if the treated groups are substantially smaller or larger than the controls.

We are motivated by the many studies that use individual data, in which there is variation in treatment across both groups and time periods. Such models are often expressed as follows. If i indexes individuals, g indexes groups, and t indexes time periods, then a classic "difference-in-differences" (or "DiD") regression can be written as

$$\begin{aligned} y_{igt} &= \alpha + \boldsymbol{T}_{igt}\boldsymbol{\gamma} + \boldsymbol{D}_{igt}\boldsymbol{\eta} + \beta\,\mathrm{TREAT}_{igt} + \epsilon_{igt}, \\ & i = 1,\ldots,N_g, \quad g = 1,\ldots,G, \quad t = 1,\ldots,T. \end{aligned} \tag{1}$$

Here $\boldsymbol{T}_{igt}$ is a row vector of time dummies, $\boldsymbol{D}_{igt}$ is a row vector of group dummies, and TREAT_{igt} is equal to 1 for observations that were in a treated group during a treated period and zero otherwise.[1] There may of course be other regressors as well. The coefficient of interest is β, which shows the effect on treated groups in periods when there is treatment. Following the literature, we divide the G groups into G_1 treated groups and G_0 control groups in which no observations are treated, so that $G = G_0 + G_1$. We are concerned with cases in which G_1 is small and G_0 is not too small. For example, the procedures we discuss might be viable for $G_1 = 2$ and $G_0 = 21$, but not for $G_1 = 3$ and $G_0 = 3$. Why this is so will become apparent in Subsection 3.1.

Section 2 discusses cluster-robust variance estimation, and Subsection 2.1 shows why it fails when there are few treated clusters. Section 3 introduces randomization inference.

[1] Since there is a constant term, one group dummy and one time dummy must be omitted.

Subsection 3.1 describes the coefficient-based (Conley-Taber) approach to RI, Subsection 3.2 discusses the design of our Monte Carlo experiments, and Subsection 3.3 explores the performance of coefficient-based RI with and without cluster heterogeneity. Subsection 3.4 then proposes an alternative RI procedure based on t statistics and examines theoretically how its properties compare with those of the coefficient-based one. Unfortunately, neither procedure can be expected to perform well when G_1 is extremely small and the treated clusters differ systematically from the controls.

The remainder of Section 3 presents a variety of simulation results. In Subsection 3.5, we find, as the theory of Subsection 3.4 suggests, that none of the existing procedures yields reliable inferences when groups vary in size and only one group is treated. However, the new RI procedure based on t statistics always performs reasonably well when two or more groups are treated. In Subsection 3.7, we find that fairly moderate differences between the error variances for treated and control clusters can have severe effects on rejection frequencies, especially for coefficient-based RI. However, we find in Subsection 3.8 that coefficient-based RI can have substantially more power than t-based RI, or than existing bootstrap procedures. Section 4 briefly discusses some alternative (non-RI) procedures for which we present simulation results in Appendix A. Section 5 presents results for two empirical examples, one based on Bailey (2010) and one based on Conley and Taber (2011), and Section 6 concludes.

2 Inference with Few Treated Clusters

A linear regression model with clustered errors may be written as

$$\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_G \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_G \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_G \end{bmatrix}, \quad \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \boldsymbol{\Omega}, \tag{2}$$

where each of the G clusters, indexed by g, has N_g observations. The matrix $\boldsymbol{X}$ and the vectors $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ have $N = \sum_{g=1}^{G} N_g$ rows, $\boldsymbol{X}$ has k columns, and the parameter vector $\boldsymbol{\beta}$ has k elements. OLS estimation of equation (2) yields estimates $\hat{\boldsymbol{\beta}}$ and residuals $\hat{\boldsymbol{\epsilon}}$. As usual in the literature on cluster-robust inference, we assume that

$$\mathrm{E}(\boldsymbol{\epsilon}_g \boldsymbol{\epsilon}_g') = \boldsymbol{\Omega}_g \quad \text{and} \quad \mathrm{E}(\boldsymbol{\epsilon}_g \boldsymbol{\epsilon}_h') = \mathbf{0} \ \text{ for } g \neq h,$$

where the $\boldsymbol{\epsilon}_g$ are vectors with typical elements ϵ_{ig}, and the $\boldsymbol{\Omega}_g$ are $N_g \times N_g$ positive definite covariance matrices. The $N \times N$ covariance matrix $\boldsymbol{\Omega}$ is then

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_1 & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\Omega}_2 & \dots & \mathbf{O} \\ \vdots & \vdots & & \vdots \\ \mathbf{O} & \mathbf{O} & \dots & \boldsymbol{\Omega}_G \end{bmatrix}.$$

Because the elements of the $\boldsymbol{\epsilon}_g$ are in general neither independent nor identically distributed, both classical OLS and heteroskedasticity-robust standard errors for $\hat{\boldsymbol{\beta}}$ are invalid.

As a result, conventional inference can be seriously unreliable. It is therefore customary to use a cluster-robust variance estimator, or CRVE. There are several of these, of which the earliest may be the one proposed in Liang and Zeger (1986). The CRVE we investigate, which we call CV_1, is defined as:

$$\frac{G(N-1)}{(G-1)(N-k)}(\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{3}$$

where $\hat{\boldsymbol{\epsilon}}_g$ is the subvector of $\hat{\boldsymbol{\epsilon}}$ that corresponds to cluster g. It yields reliable inferences when the number of clusters is large (Cameron, Gelbach and Miller, 2008) and the number of observations per cluster does not vary too much (Carter, Schnepel and Steigerwald, 2017; MacKinnon and Webb, 2017b). This is the estimator that is used when the `cluster` command is invoked in Stata.[2] However, Conley and Taber (2011) and MacKinnon and Webb (2017b) show that t statistics based on (3) over-reject severely when the parameter of interest is the coefficient on a treatment dummy and there are very few treated clusters. Rejection frequencies can be over 80% when only one cluster is treated, even when the t statistics are assumed to follow a $t(G-1)$ distribution, as is now commonly done based on the results of Donald and Lang (2007) and Bester, Conley and Hansen (2011).

2.1 Cluster-Robust Variance Estimation

It is important to understand precisely why inference based on the CRVE (3) fails when there are few treated clusters. The analysis in this subsection extends the one in MacKinnon and Webb (2017b, Section 6), which applies to the pure treatment case, by allowing only some observations in the treated clusters to be treated. Consider the following simplified version of regression (2), in which the fixed effects have been omitted and the t subscript suppressed:

$$y_{ig} = \alpha + \beta d_{ig} + \epsilon_{ig}. \tag{4}$$

Here the treatment dummy d_{ig} equals 1 for at least some of the observations in the first G_1 clusters, 0 for the remaining observations in those clusters, and 0 for all observations in the last $G_0 = G - G_1$ clusters. Making equation (4) more complicated by adding fixed effects or other additional regressors would not change anything important. The fundamental problem, as we will see shortly, is that the residuals sum to zero over all treated observations.

Equation (4) may be rewritten in vector notation as $\boldsymbol{y} = \alpha\boldsymbol{\iota} + \beta\boldsymbol{d} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$, $\boldsymbol{\iota}$, $\boldsymbol{d}$, and $\boldsymbol{\epsilon}$ are N-vectors with typical elements y_{ig}, 1, d_{ig}, and ϵ_{ig}, respectively, and i is assumed to vary more rapidly than g. Then the OLS estimate of β is

$$\hat{\beta} = \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{y}}{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})} = \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}}{N(\bar{d} - \bar{d}^2)}, \tag{5}$$

where the second equality holds under the null hypothesis that $\beta = 0$, and $\bar{d}$ denotes the sample mean of the d_{ig}, that is, the proportion of treated observations.

[2] Expression (3) is not the only CRVE; see Bell and McCaffrey (2002), Imbens and Kolesár (2016), and Young (2016). The best-known alternative to (3) will be discussed in Section 4. However, we do not focus on it because it is not widely used, does not solve the problem discussed in this subsection, and is computationally demanding to the point of being infeasible when any of the N_g is too large.

The variance of $\hat{\beta}$ is evidently

$$\text{Var}(\hat{\beta}) = \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\Omega}(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})}{\big((\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})\big)^2} = \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\Omega}(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})}{N^2\bar{d}^2(1-\bar{d})^2}, \tag{6}$$

where $\boldsymbol{\Omega}$ is an $N \times N$ block diagonal matrix with the $N_g \times N_g$ covariance matrices $\boldsymbol{\Omega}_g$ forming the diagonal blocks. From expression (3), the corresponding CRVE is

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{G(N-1)}{(G-1)(N-k)N^2\bar{d}^2(1-\bar{d})^2} \sum_{g=1}^{G} (\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g), \tag{7}$$

where $\boldsymbol{d}_g$ is the subvector of $\boldsymbol{d}$ that corresponds to cluster g, and $\boldsymbol{\iota}_g$ is an N_g-vector of 1s. Thus expression (7) should provide a good estimate of $\text{Var}(\hat{\beta})$ if the summation provides a good estimate of the quadratic form in (6). Unfortunately, this is not the case when there are few treated clusters.

The summation in expression (7) can be written as the sum of two summations, one for the treated clusters and one for the controls. In both cases, a typical term is

$$\bar{d}^2(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2 + (\boldsymbol{d}_g'\hat{\boldsymbol{\epsilon}}_g)^2 - 2\bar{d}\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'\boldsymbol{d}_g. \tag{8}$$

However, since $\boldsymbol{d}_g = \boldsymbol{0}$ for the control clusters, the second and third terms in (8) must vanish for those clusters. Thus the summation in (7) becomes

$$\sum_{g=1}^{G_1}\big(\bar{d}^2(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2 + (\boldsymbol{d}_g'\hat{\boldsymbol{\epsilon}}_g)^2 - 2\bar{d}\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'\boldsymbol{d}_g\big) + \bar{d}^2\sum_{g=G_1+1}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2. \tag{9}$$

This expression is supposed to estimate the quadratic form in expression (6), which can be written as

$$\sum_{g=1}^{G_1}\big(\bar{d}^2\boldsymbol{\iota}_g'\boldsymbol{\Omega}_g\boldsymbol{\iota}_g + \boldsymbol{d}_g'\boldsymbol{\Omega}\boldsymbol{d}_g - 2\bar{d}\boldsymbol{\iota}_g'\boldsymbol{\Omega}\boldsymbol{d}_g\big) + \bar{d}^2\sum_{g=G_1+1}^{G}\boldsymbol{\iota}_g'\boldsymbol{\Omega}_g\boldsymbol{\iota}_g. \tag{10}$$

Unfortunately, expression (9) estimates expression (10) very badly when G_1 is small, because the first summation in the former severely underestimates the first summation in the latter. Consider the extreme case in which $G_1 = 1$. The treatment dummy must be orthogonal to the residuals. Since $d_{ig} = 0$ for $g > 1$, this implies that $\hat{\boldsymbol{\epsilon}}_1'\boldsymbol{d}_1 = 0$. Therefore, the second and third terms in the first summation in expression (9) vanish. All that remains is $\bar{d}^2(\boldsymbol{\iota}_1'\hat{\boldsymbol{\epsilon}}_1)^2$. The terms that are supposed to estimate the last two terms in the first summation in expression (10) are missing.

This might not matter much if the last two terms in the first summation in (10) were small. But, in most cases, the opposite must be the case. Both the remaining term in the first summation and the entire second summation involve factors of $\bar{d}^2$, that is, the square of the proportion of treated observations. Unless the first cluster is much larger than any of the other clusters, and most of the observations in it are treated, $\bar{d}$ will typically be much less than one half when $G_1 = 1$, and these terms will tend to be quite small. This analysis explains why the CRVE (3) often produces standard errors that are too small by a factor of five or more when there is just one treated cluster.

When two or more clusters are treated, the residuals for the treated observations will not sum to zero for each treated cluster, but they must sum to zero over all the treated clusters.[3] In consequence, the first summation in expression (9) must underestimate the corresponding summation in (10). The first two terms in the former do not actually vanish, but they are often much too small when G_1 is small. The problem evidently goes away as G_1 increases, provided the sizes of the treated and control clusters are not changing systematically, and simulation results in MacKinnon and Webb (2017b) suggest that it does so quite quickly.

In this discussion, we have ignored the presence of fixed effects and other regressors in the regression of interest. Taking these into account would greatly complicate the analysis. However, it clearly would not change the basic result. The standard error of $\hat{\beta}$ is severely underestimated because the residuals sum to zero over all the treated observations, and that must be the case no matter how many other regressors there may be.

As we discuss in Section 4, expression (3) is not the only available CRVE, and other procedures may well work better. However, there appears to be no way to avoid severely underestimating the standard error of $\hat{\beta}$ when G_1 is small. This provides the motivation to consider alternative approaches, including randomization inference.

3 Randomization Inference

Randomization inference (RI) was first proposed by Fisher (1935) as a procedure for performing exact tests in the context of experiments. Rosenbaum (1996) mentions the possibility of using randomization inference for group-level interventions. Monte Carlo tests are closely related to randomization inference; see Dufour (2006). A formal theoretical treatment of RI may be found in Lehmann and Romano (2008, Chapter 15). A more accessible discussion focused on individual-level data is provided in Imbens and Rubin (2015, Chapter 5).

Imagine that we are interested in testing the sharp null hypothesis

$$H_0\colon \quad \mathrm{E}(Y_i|D_i = 1) - \mathrm{E}(Y_i|D_i = 0) = 0,$$

where Y_i is the outcome for individual i, $D_i = 1$ indicates treatment, and $D_i = 0$ indicates no treatment. Under this null hypothesis, the missing *potential* outcomes are equal to the observed outcome for each individual. That is, if there were no treatment effect, each individual would have the same outcome with or without treatment. We could calculate a test statistic for our original sample as

$$\tau = |\bar{Y}_{D_i=1} - \bar{Y}_{D_i=0}|, \tag{11}$$

where $\bar{Y}_{D_i=1}$ and $\bar{Y}_{D_i=0}$ are the average outcomes for treated and untreated individuals, respectively.

We can also calculate the test statistic (11) for any other random assignment of treatments to individuals. For any re-randomization r, the test statistic would be

$$\tau_r^* = |\bar{Y}_{D_i^r=1} - \bar{Y}_{D_i^r=0}|, \tag{12}$$

[3] See equation (A.2) and surrounding discussion in the online appendix of MacKinnon and Webb (2017b), which studies a pure treatment model rather than a DiD model.

where D_i^r denotes the re-randomized treatment assignment. We can repeat this process for all possible re-randomizations, or for a subset of them. If it is reasonable to believe that treatment was assigned at random, then it makes sense to compare τ with the τ_r^*. If the null hypothesis of no treatment effect is true, then τ and the τ_r^* must be drawn from the same distribution. A randomization test simply compares τ with the empirical distribution of the τ_r^*. If $\hat{\tau}$ is in one of the tails of that empirical distribution, then this is evidence against the null hypothesis of no treatment effect.

In the context of cluster-robust inference, to be discussed in Subsections 3.1 and 3.4 below, we will randomize at the group level rather than the individual level. We therefore let G denote the number of units and G_1 the number of treated units. The number of possible re-randomizations is then ${}_G\mathrm{C}_{G_1}$, that is, the number of ways to choose G_1 out of G units without replacement. One of these randomizations corresponds to the original sample. If we omit it, we have $S = {}_G\mathrm{C}_{G_1} - 1$ re-randomizations that can be used to compute the τ_r^*.

Suppose that we wish to reject when τ is large in absolute value. Then we must compare $|\tau|$ with the $|\tau_r^*|$. It is natural to sort the latter from smallest to largest and see how extreme $|\tau|$ is relative to the sorted list. Equivalently, we can calculate a P value based on the empirical distribution of the $|\tau_r^*|$:

$$p^* = \frac{1}{S}\sum_{r=1}^{S} \mathbb{I}\big(|\tau_r^*| > |\tau|\big), \tag{13}$$

which is the proportion of re-randomizations for which τ_r^* is more extreme in absolute value than τ. The test rejects at level α whenever $p^* \leq \alpha$. Of course, we could use a one-tailed test instead if the null hypothesis were that the treatment does not have a positive effect.

An alternative approach that is more common in the RI literature is not to omit the actual sample from the set of re-randomizations. The sorted list contains $S + 1 = {}_G\mathrm{C}_{G_1}$ elements, one of which is equal to $|\tau|$. If $c \equiv S + 1 - [\alpha(S+1)]$, where $[\cdot]$ denotes the largest integer no larger than its argument, then element number c of the sorted list, say $|\tau_r^{(c)}|$, can be thought of as a critical value. The RI test is then defined as

$$\phi(\boldsymbol{Y}) = \begin{cases} 0 \text{ if } |\tau| < |\tau_r^{(c)}| \\ a = \alpha(S+1) - \sum_{r=1}^{S+1} \mathbb{I}\big(|\tau_r^*| > |\tau|\big) \text{ if } |\tau| = |\tau_r^{(c)}| \\ 1 \text{ if } |\tau| > |\tau_r^{(c)}|, \end{cases} \tag{14}$$

where $\boldsymbol{Y}$ denotes the sample, $\phi(\boldsymbol{Y}) = 1$ denotes rejection, and $\phi(\boldsymbol{Y}) = 0$ denotes non-rejection. Theorem 15.2.1 of Lehmann and Romano (2008, Section 15.2) proves that, when $|\tau|$ and the $|\tau_r^*|$ follow the same distribution, $\mathrm{E}(\phi(\boldsymbol{Y})) = \alpha$. In other words, the expectation of the RI test $\phi(\boldsymbol{Y})$ defined in (14) across all randomizations is equal to the level of the test. The test is therefore exact.

Since $0 < a \leq 1$, the middle outcome in (14), which occurs whenever $|\tau|$ and $|\tau_r^{(c)}|$ coincide, can be interpreted as a probability. It is included because, otherwise, $\mathrm{E}(\phi(Y)) \neq \alpha$ unless we make further assumptions about S. This outcome does not directly tell us either to reject or not to reject, which seems unsatisfactory. However, we can decide whether or not to reject by drawing a random number η from the $\mathrm{U}(0,1)$ distribution. If we reject

whenever $\eta \leq a$, the test always gives an answer and is still exact, but now it depends on the random value of η, which is also not entirely satisfactory. Because the middle outcome occurs with probability $1/(S+1)$, it can safely be ignored when S is large.[4]

There is also an important special case in which the middle outcome does yield a definitive result. Suppose that $\alpha(S+1)$ is an integer. Then $c = (1-\alpha)(S+1)$, and the summation in the middle outcome equals $\alpha(S+1) - 1$, because that is the number of $|\tau_r^*|$ that exceed number $(1-\alpha)(S+1)$ in the sorted list. This implies that the middle outcome is simply equal to 1 in this case. Thus, when $\alpha(S+1)$ is an integer, the test in (14) simplifies to

$$\phi(\boldsymbol{Y}) = \begin{cases} 0 \text{ if } |\tau| < |\tau_r^{(c)}| \\ 1 \text{ if } |\tau| \geq |\tau_r^{(c)}|. \end{cases} \tag{15}$$

It is easy to see that this test must yield exactly the same result as a test based on (13), because $|\tau_r^*| > |\tau|$ if and only if $|\tau| \geq |\tau_r^{(c)}|$.

Writing the RI test as (15) suggests another way to compute the P value:

$$p^{*\prime} = \frac{1}{S+1}\left(1 + \sum_{r=1}^{S} \mathbb{I}\big(|\tau_r^*| > |\tau|\big)\right). \tag{16}$$

When $(1-\alpha)(S+1)$ is an integer, rejecting whenever $p^{*\prime} \leq \alpha$ must yield exactly the same outcome as rejecting whenever $p^* < \alpha$. However, when $(1-\alpha)(S+1)$ is not an integer, the two tests will yield different results. The one based on p^* will over-reject, and the one based on $p^{*\prime}$ will under-reject. If rejection frequencies are plotted as a function of S for, say, $\alpha = 0.05$, they will form two sawtooth patterns, which meet at 0.05 for $S = 19$, $S = 39$, and so on. The test based on $p^{*\prime}$ never rejects for $S < 19$ and never rejects more than 5% of the time for any S, while the test based on P^* never rejects less than 5% of the time. See Racine and MacKinnon (2007, Figure 1). A modified version of the wild bootstrap to overcome this problem is proposed in MacKinnon and Webb (2018b).

RI procedures are valid only when the distribution of the test statistic is invariant to the realization of the re-randomizations across permutations of assigned treatments (Lehmann and Romano, 2008, Section 15.2). It is therefore important to incorporate all available information about treatment assignment in conducting the re-randomization (Yates, 1984). For example, if the investigator knows that treatment was only assigned to units with particular characteristics, then any re-randomization should also assign treatment only to units with those characteristics. Of course, that may or may not feasible, depending on how many such units there are and how much information about unit characteristics is available.

3.1 Randomization Inference based on Coefficients

Classic RI procedures were designed for treatment assigned randomly at the observation level, as in the case of agricultural experiments. Extending them to DiD models was first proposed in Conley and Taber (2011), which suggests two procedures for inference with few treated groups. Both of them involve using information about the control groups to

[4] In writing (14), we have implicitly assumed that there can never be more than one value of $|\tau_r^*|$ that equals $|\tau_r^{(c)}|$. The expression for the middle outcome would be more complicated without that assumption.

learn about the distribution of a test statistic. The Γ^* procedure is a form of randomization inference. It involves constructing an empirical distribution by randomizing the assignment of groups to "treatment" and "control" and using this distribution to conduct inference. Of the two procedures, Γ^* is more attractive, because it can be used whether or not $G_0 > G_1$ and because it often has better size properties in the Monte Carlo experiments reported in the paper. The procedure we now discuss is very similar to Γ^*.

Up to this point, we have not said much about the test statistic τ on which randomization inference is based. One approach, which is the one taken in Conley and Taber (2011), is to use a coefficient estimate as the test statistic. We now propose a simple coefficient-based RI procedure, which we call RI-β, for the DiD model (1). It is not identical to the Γ^* procedure of Conley and Taber (2011), but it is much simpler to describe, and it seems to yield extremely similar results. The principal difference between the RI-β and Γ^* procedures is that the former explicitly uses the OLS estimate $\hat{\beta}$ from equation (1), while the latter uses a quantity that is not identical to $\hat{\beta}$ but is extremely highly correlated with it. For the Georgia Hope example of Merit Scholarships analyzed in Subsection 5.2, the correlation is greater than 0.9999.

The RI-β procedure works as follows:

1. Estimate the DiD regression model (1) to calculate $\hat{\beta}$, the coefficient of interest
2. Generate a (preferably large) number of β_r^* statistics to compare $\hat{\beta}$ with.
   - When $G_1 = 1$, assign a group from the G_0 control groups as the "treated" group g^* for each repetition, re-estimate the model using the observations from all G groups, and calculate a new coefficient, β_r^*, indicating randomized treatment. Repeat this process for all G_0 control groups. Thus the empirical distribution of the β_r^* will have G_0 elements.
   - When $G_1 > 1$, sequentially treat every set of G_1 groups except the set actually treated, re-estimate equation (1), and calculate a new β_r^*. There are potentially ${}_G\mathrm{C}_{G_1} - 1$ sets of groups to compare with. When this number is not too large, obtain all of the β_r^* by enumeration.[5] When it exceeds an upper limit B, picked on the basis of computational cost and with the property that $\alpha(B+1)$ is an integer so as to ensure that $p^{*\prime} = p^*$, choose the comparators randomly, without replacement, from the set of potential comparators. Thus the empirical distribution will have $\min({}_G\mathrm{C}_{G_1} - 1, B)$ elements.
3. Compute either the P value p^* defined in (13) or the P value $p^{*\prime}$ defined in (16). Recall that $p^{*\prime} \geq p^*$.

In the context of the DiD model (1), one important practical issue is how to assign treatment years for the re-randomizations. The treated clusters are numbers 1 through G_1,

[5] The number of comparators can easily be too large. For example, if $G = 50$ and $G_1 = 4$, there are 230,299 possible re-randomizations.

for which treatment begins in periods t_1^1, t_2^1,..., $t_{G_1}^1$, respectively.[6] Let the clusters chosen for treatment in each re-randomization be numbered 1^*, 2^*, ..., G_1^*. For example, 1^* might denote cluster 11, 2^* might denote cluster 8, and so on. It is natural to assign starting year t_j^1 to cluster j^*. However, since both orderings are arbitrary, there is more than one way to do this. We considered two of them.

In the first procedure, the original clusters are ordered from smallest to largest, so that $N_1 \leq N_2 \ldots \leq N_{G_1}$, and the clusters chosen for each re-randomization are ordered in the same way, so that $N_{1^*} \leq N_{2^*} \ldots \leq N_{G_1^*}$. Thus the smallest cluster for each re-randomization is "treated" for the same years as the smallest actual treated cluster, the second-smallest for the same years as the second-smallest actual treated cluster, and so on. In the second procedure, the re-randomized clusters are not ordered in any way, so the assignment of years of treatment is random. In several experiments, we found very little to choose between the two procedures. All the results we report below are for the first procedure, because it is slightly easier to implement.

When treatment is randomly assigned at the individual level, the invariance of the distribution of $\hat{\beta}$ to re-randomization follows naturally. However, if treatment is assigned instead at the group level, which is almost always the case for difference-in-differences, it may be hard to argue that assignment was random. Unless the units are indistinguishable, randomization inference is not appropriate when assignment is not random. Moreover, the extent to which clusters are heterogeneous can affect how close the distribution is to being invariant.

The RI-β procedure evidently depends on the strong assumption that $\hat{\beta}$ and the β_r^* all follow the same distribution. But that cannot be the case if the coefficients for some clusters are estimated more efficiently than for others. This may occur whenever the clusters are heterogeneous. As we will demonstrate in Subsection 3.3, several types of heterogeneity can substantially affect the reliability of inference based on RI-β. The most readily observable type of heterogeneity, which is very likely to occur with individual data in a wide variety of contexts, is variation in cluster sizes, and we therefore focus on it.

It is possible to interpret RI-β and the procedures proposed in Conley and Taber (2011) as testing a joint null hypothesis of no treatment effect and random assignment. However, since the treated clusters are observed, we want to make inferences conditional on them. Even if treatment actually were assigned at random (which seems unlikely in many contexts where DiD is used, such as assessing the effects of policy changes at the jurisdiction level), the RI-β procedure is potentially either over-sized or under-sized conditional on which clusters were actually treated. In Subsection 3.3, we attempt to see just how serious these size distortions are likely to be.

3.2 Design of the Monte Carlo Experiments

In the next subsection, and in several later ones, we report results of a number of Monte Carlo experiments that study the performance of various inferential procedures, including ones not based on randomization inference, when the number of treated clusters is small and clusters are heterogeneous. In this subsection, before we report any results, we describe

[6] Here we implicitly assume that, for all treated clusters, treatment begins at some point in time and never ends. This is also what we assume in our simulation experiments. However, it is easy to extend the procedures we discuss to handle situations in which treatment has an end date as well as a start date.

the model and experimental design.

The model we use is a simplified version of the DiD model (1) with no group fixed effects. In the data generating process, the ϵ_{igt} are normally distributed and generated by a random effects model at the group level. The correlation between any two error terms that belong to the same cluster is ρ.[7] Each observation is assigned to one of 20 "years", and the starting year of "treatment" is randomly assigned to years between 6 and 16. The null hypothesis, which was maintained in most of the experiments, is that $\beta = 0$.

In most experiments, we assign N total observations unevenly among G clusters using the following formula:

$$N_g = \left[N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad g = 1, \ldots, G-1, \tag{17}$$

where $[x]$ means the integer part of x. The value of N_G is then set to $N - \sum_{g=1}^{G_1} N_g$. The key parameter here is γ, which determines how uneven the cluster sizes are. When $\gamma = 0$ and N/G is an integer, equation (17) implies that $N_g = N/G$ for all g. As γ increases, however, cluster sizes vary more and more.[8]

Most experiments have $G = 40$, although some have $G = 20$. These numbers were chosen so that, when $G_1 = 1$, the number of re-randomizations is either 39 or 19, which ensures that the two P values (13) and (16) are identical. For randomization inference procedures with $G = 40$, the number of randomizations is as follows: 39 for $G_1 = 1$; ${}_{40}\mathrm{C}_2 - 1 = 779$ for $G_1 = 2$; and 999 for $G_1 \geq 3$. The number of Monte Carlo replications in most experiments is 100,000. Rejection frequencies are calculated at the 1%, 5%, and 10% levels, although only the 5% rejection frequencies are reported.

3.3 Performance of RI-β when Cluster Sizes Vary

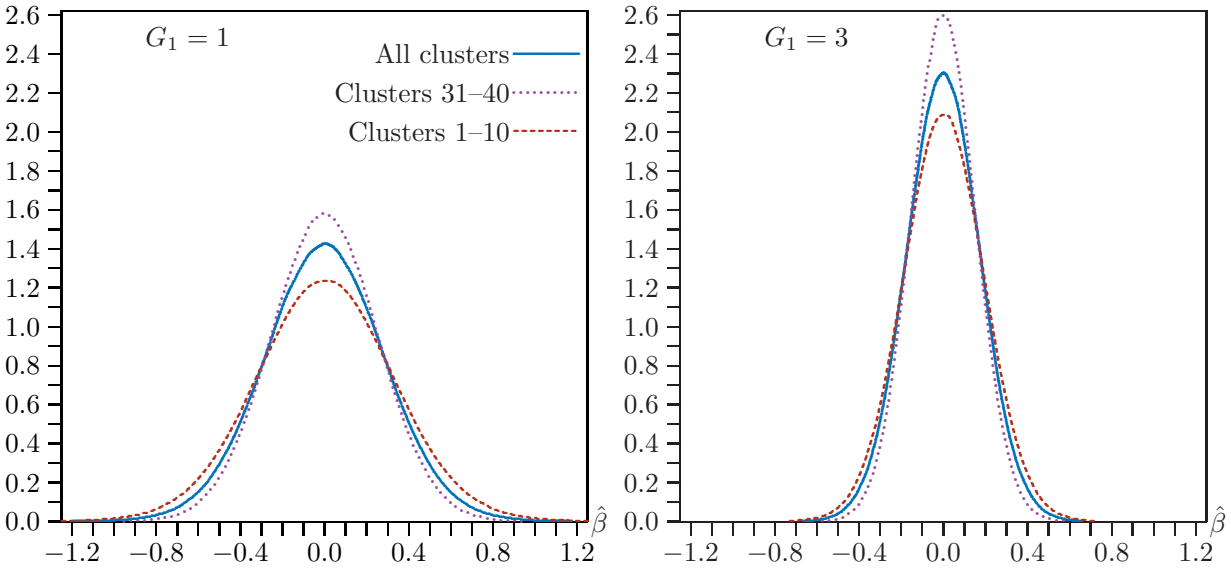
As we saw in Subsection 3.1, the RI-β procedure cannot be expected to work perfectly if the treated clusters do not have the same characteristics as the control clusters. We focus initially on what happens when cluster sizes differ systematically. Specifically, we treat either 1 or 3 clusters from a set of 40 unbalanced clusters, with $N = 4000$ and cluster sizes determined by equation (17) with $\gamma = 2$. In each case, we plot three distributions of $\hat{\beta}$, which were obtained by kernel density estimation using 1,999,999 replications. One of these is the unconditional distribution, for which the treated clusters are selected at random from all 40 clusters. The other two are conditional distributions, for which the treated clusters are selected at random either from clusters 1-10 (the smallest clusters) or from clusters 31-40 (the largest clusters).

Figure 1 presents these results. The left panel of the figure shows densities for $G_1 = 1$, and the right panel shows densities for $G_1 = 3$. It is evident that, in each case, the two conditional distributions differ from the unconditional one. When small clusters are treated,

[7] We did not include group fixed effects in the model partly to save computer time and partly because, if they were included, they would completely explain the random effects, effectively eliminating intra-cluster correlation. Because the model did include time fixed effects, the DGP did not include time random effects.

[8] Many of our experiments have 4000 observations and $\gamma = 2$. In these experiments, the sizes of the 40 clusters are: 32, 33, 35, 37, 39, 41, 43, 45, 47, 50, 52, 55, 58, 61, 64, 67, 71, 75, 78, 82, 87, 91, 96, 101, 106, 112, 117, 123, 130, 136, 143, 151, 158, 167, 175, 184, 194, 204, 214, and 246.

Figure 1: Conditional and Unconditional Distributions of $\hat{\beta}$



Notes: Based on 1,999,999 samples with $G = 40$, $G_1 = 1$ or 3, $\gamma = 2$, and $\rho = 0.05$

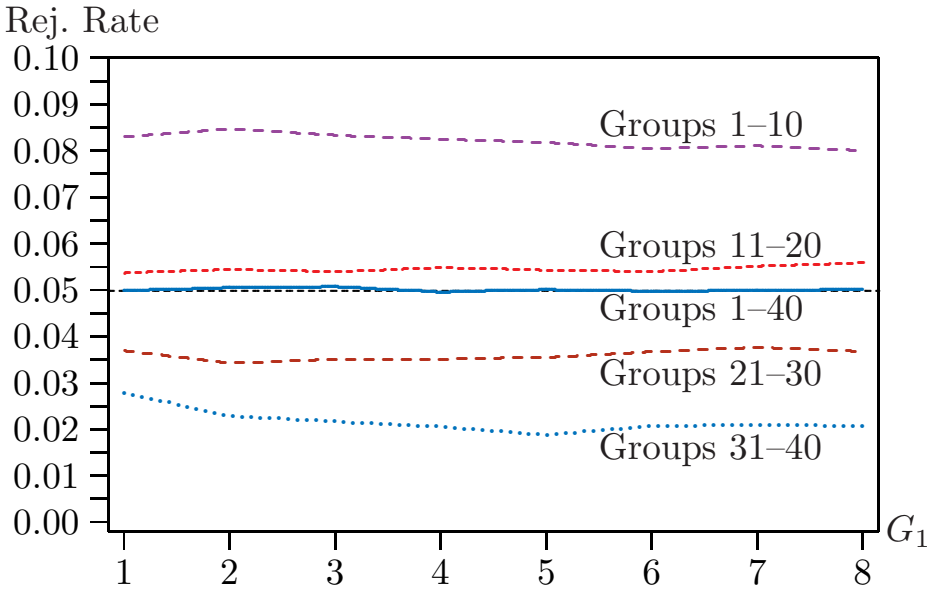
the distribution has a lower peak and is more spread out. When large clusters are treated, it has a higher peak and is less spread out. Notice that, although all distributions are less spread out when $G_1 = 3$ than when $G_1 = 1$, the differences between the conditional and unconditional ones are essentially the same in both cases.

Figure 1 highlights a subtle difference between ex-post and ex-ante analysis. Imagine that you were to conduct an experiment in which treatment was randomly assigned to a single cluster. This is the setting of the left panel in Figure 1. Imagine also that you were given only the values of $\hat{\beta}$ and the β_r^*, but did not know which cluster was actually treated. In this case, even if the clusters were quite heterogeneous, the expected rejection frequency of the RI-β test for the null of no treatment effect would be α, because any particular cluster has an equal chance of being treated. Without knowledge of which cluster is treated, both $\hat{\beta}$ and all the β_r^* are drawn from the "all clusters" distribution in the figure.

Unfortunately, it is unrealistic except from an ex-ante perspective to assume that the investigator knows nothing about which cluster was treated. From an ex-post perspective, we almost always do know which cluster is treated. With heterogeneous clusters, the expected rejection frequency conditional on the cluster that is actually treated will not be α. In the experiment considered in the left panel of the figure, imagine that the treated cluster happens to be one of the 10 smallest ones. In this case, $\hat{\beta}$ will be drawn from the corresponding conditional distribution (the red dashed line in the figure). However, the 39 β_r^* coefficients will be drawn from the unconditional distribution (the solid blue line in the figure), but with the treated cluster omitted. These distributions are clearly not the same. In practice, the difference between the distributions of $\hat{\beta}$ and β_r^* can be even more severe, as will be discussed below.

Even if an experiment is designed in such a way that treatment is random, anyone hoping

Figure 2: Rejection Frequencies for RI-β Tests



Notes: Based on 100,000 replications with $G = 40$, $\gamma = 2$, and $\rho = 0.05$

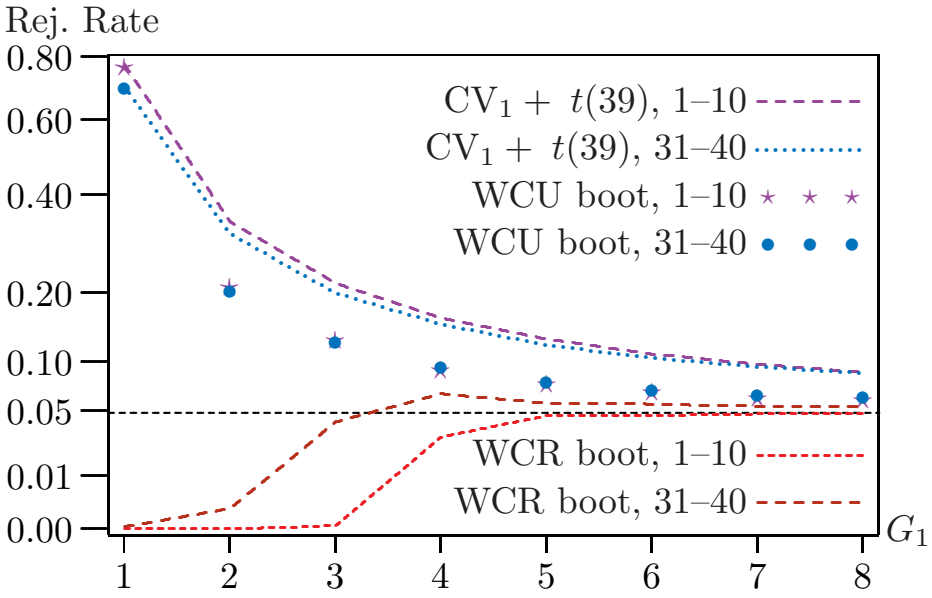
to do ex-post analysis *should* condition on the treated cluster. Even if the treatment were randomly assigned initially, both panels of Figure 1 strongly suggest that, conditional on the assignment of cluster(s) to treatment, the RI-β procedure will not yield a test that rejects α% of the time. We will see shortly see that this is indeed the case.

Since researchers always know the sizes, and usually the identities, of the clusters that are treated, it generally makes no sense to pretend that treatment assignment is unknown. Failing to condition on what the researcher knows about the treated and control clusters inevitably results in unreliable inference. Ferman and Pinto (2015) eloquently makes this point in the context of aggregate (panel) data.

For the RI-β procedure, the β_r^* are always drawn from the unconditional distribution. However, unless treatment really is assigned at random and nothing is known about the treated clusters, $\hat{\beta}$ may actually be drawn from a conditional distribution like the ones in Figure 1. This suggests that RI-β will over-reject when the treated clusters tend to be small and under-reject when they tend to be large. To investigate this phenomenon, we perform 40 experiments, with $G = 40$, $N = 4000$, and $\gamma = 2$. In eight of the experiments, the treated clusters are drawn at random from all 40 clusters. In the other 32 experiments, they are drawn from clusters 1-10, 11-20, 21-30, or 31-40. In each case, the number of treated clusters G_1 varies from 1 to 8.

Figure 2 shows rejection frequencies for tests at the .05 level based on the RI-β procedure for these 40 experiments. As expected, the procedure works perfectly in the unconditional case where the treated clusters are chosen at random from the entire set of clusters, subject to the small experimental errors to be expected with 100,000 replications. However, it over-rejects noticeably when the treated clusters are chosen from numbers 1-10, and slightly

Figure 3: Rejection Frequencies for Asymptotic and Bootstrap Tests



Notes: Based on 100,000 replications with $G = 40$, $\gamma = 2$, $\rho = 0.05$, and $B = 399$

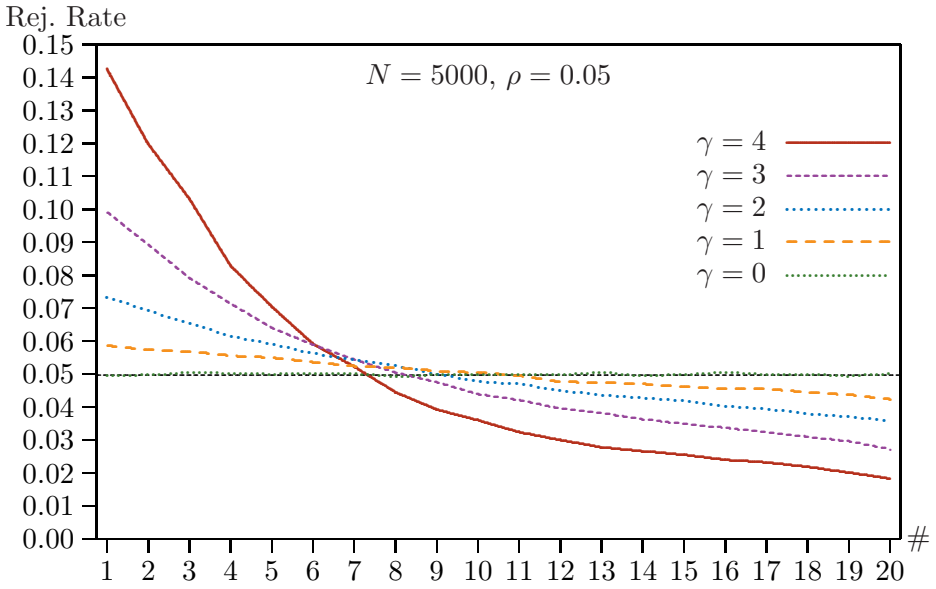
for numbers 11-20. In contrast, it under-rejects moderately for numbers 21-30, and quite noticeably for numbers 31-40. Thus the smaller/larger the treated clusters are relative to the entire set, the more prone is the procedure to over-/under-reject.

One interesting feature of these experiments is that the performance of RI-β varies only a little with G_1. This contrasts sharply with both bootstrap and asymptotic procedures for cluster-robust inference, which typically perform extremely badly when $G_1 = 1$ but then improve rapidly as G_1 increases. However, these procedures are much less sensitive to the relative sizes of the treated and control clusters.

Several existing procedures that are not based on randomization inference are discussed in Section 4, and simulation results for them are presented in Appendix A. We present a few results for the same experiments as Figure 2 here to highlight the contrast between the RI-β test, on the one hand, and existing asymptotic and bootstrap tests, on the other. Figure 3 shows rejection frequencies for what is still by far the most commonly used testing procedure, which is simply to compare a t statistic based on the CRVE (3) with the $t(G-1)$ distribution. Note that the vertical axis has been subjected to a nonlinear transformation in order to show all the results on the same graph. As the analysis in Subsection 2.1 implies, the conventional procedure over-rejects very severely when $G_1 = 1$, because the CRVE standard error for $\hat{\beta}$ is much too small. The over-rejection becomes less severe as G_1 increases, but the test rejects at least 8.5% of the time even when $G_1 = 8$.

The figure also shows rejection frequencies for two forms of wild cluster bootstrap test, which will be discussed in Section 4. One of these (WCR, where the bootstrap DGP is based on restricted estimates) was proposed in Cameron, Gelbach and Miller (2008), and both were shown to be asymptotically valid in Djogbenou, MacKinnon and Nielsen (2017).

Figure 4: Rejection Frequencies for RI-β tests when $G_1 = 1$



Notes: Based on 400,000 replications with $N = 5000$, $G = 20$, and $\rho = 0.05$

However, MacKinnon and Webb (2017b) showed theoretically that WCR will under-reject very severely when G_1 is small and WCU (where the bootstrap DGP is based on unrestricted estimates) will over-reject very severely. That is exactly what is observed in Figure 3.

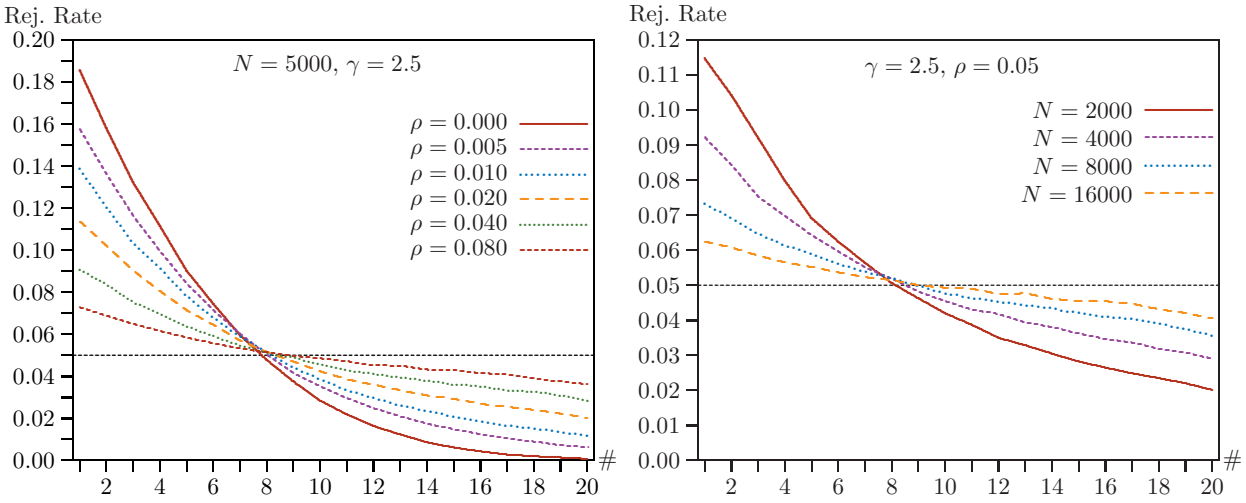
For clarity, Figure 3 only shows results for the relatively extreme cases in which the treated clusters are chosen from numbers 1-10 and numbers 31-40. Except for WCR, where for $2 \leq G_1 \leq 4$ the omitted results lie between the ones shown in the figure, the omitted results are very similar to the ones that are shown.

It is evident from Figures 2 and 3 that RI-β always works very much better than any of the other procedures when $G_1 \leq 2$. This is true even for the extreme cases in which the treated clusters are drawn from numbers 1-10 or 31-40 and RI-β does not work particularly well. For $G_1 \geq 4$, however, WCR typically works better than RI-β, except of course for the case in which all clusters are potentially treated, where RI-β works perfectly. For $G_1 \geq 5$, even WCU works better for the extreme cases than RI-β does.

The error terms in the DGP used to generate all the simulation results presented in this section are normally distributed. Additional results for a DGP with lognormally distributed error terms, which display a great deal of positive skewness, are presented in Appendix B. It will be seen that the distribution of the error terms matters, but none of the principal findings is overturned.

The experimental results presented so far necessarily depend on various features of the DGP. One important feature is the distribution of cluster sizes and the position of the treated cluster(s) within it. The next set of experiments focuses on the case of $G_1 = 1$ (which is by far the cheapest to study, allowing us to use 400,000 replications). In it, the distribution of cluster sizes is changed by varying γ. Figure 4 shows rejection frequencies

Figure 5: Rejection Frequencies for RI-β tests when $G_1 = 1$



Notes: Based on 400,000 replications with $N = 5000$ and $G = 20$

for the RI-β procedure when $G_1 = 1$ and $G = 20$. The horizontal axis shows the rank of the treated cluster, ordered from smallest to largest. There are five curves, which correspond to five values of γ. Each point on a curve represents a rejection frequency for a different treated cluster.

When $\gamma = 0$, the RI-β tests work perfectly, except for simulation error. This must be the case, because the clusters are homogeneous when $\gamma = 0$, so that the distributions of $\hat{\beta}$ and β_r^* must be the same. When $\gamma > 0$, however, the rejection frequencies depend on which cluster is treated. As expected, the tests over-reject when the treated cluster is small and under-reject when it is large. Both over-rejection and under-rejection become more severe as γ increases.

Figure 5 performs two similar exercises with $\gamma = 2.5$, so that the ratio of the largest to smallest cluster sizes is nearly 11. In the left panel, N is fixed at 5000, and ρ varies across the curves. In this case, the problem of over-rejection and under-rejection becomes less severe as ρ becomes larger. In the right panel, $\rho = 0.05$, and the sample size N varies across the curves. In this case, the problem of over-rejection and under-rejection becomes less severe as N increases. Both of these results could have been predicted from the fact that, for the model we are using, the ratio of the information provided by two clusters of different sizes is decreasing in both ρ and N. Thus the distributions of $\hat{\beta}$ vary less across the rank of the treated cluster as either ρ or N increases.

The key message from Figures 4 and 5 is that RI-β can overreject or underreject much more severely than it does in Figure 2. This is most likely to happen when the treated cluster(s) are very much smaller or larger than the average cluster, when there is not much intra-cluster correlation, and when the sample size is fairly small.

3.4 Randomization Inference based on t Statistics

Randomization inference does not have to be based on coefficient estimates. It can instead be based on any sort of test statistic, as Imbens and Rubin (2015, Chapter 5) points out. An obvious alternative to RI-β is an RI procedure based on cluster-robust t statistics. Instead of comparing $\hat{\beta}$ to the empirical distribution of the β_r^*, we compare the actual t statistic t_β, which equals $\hat{\beta}$ divided by the square root of the appropriate diagonal element of the CRVE (3), to the empirical distribution of the t_r^*, which are computed in the same way for each of the β_r^*. This is similar to one of the procedures studied in Young (2015). We will refer to this procedure as "t statistic randomization inference," or RI-t for short.

It seems plausible that randomization inference should perform better under the null hypothesis when it is based on t statistics than when it is based on coefficients, because the former are asymptotically pivotal (that is, invariant to any unknown parameters or other unknown features of the DGP) and the latter are not. Djogbenou, MacKinnon and Nielsen (2017) proves formally that t_β is asymptotically distributed as $\mathrm{N}(0,1)$, and therefore asymptotically pivotal, under certain regularity conditions.

For the DiD model, however, these conditions imply that G and G_1 must both tend to infinity together. This suggests that the distribution of t_β may be far from standard normal when G_1 is small. In fact, as we now demonstrate, $\hat{\beta}$ and t_β do not behave in the usual way as G becomes large when $G_1 = 1$. This suggests that RI-β and RI-t are likely to yield similar (and not always accurate) inferences when $G_1 = 1$, but that inferences based on the latter should improve as G_1 increases.

To see what happens when $G_1 = 1$ as $G \to \infty$, we need to make a few assumptions. In particular, we assume that N is proportional to G, and that N_g/N is $O(1/G)$ for all g. These assumptions rule out extreme cases in which the first cluster either never becomes small relative to the entire sample as $G \to \infty$ or becomes small faster than other clusters do. We also assume that, for treated clusters, the number of treated observations is, on average, proportional to N_g. If G_1 were allowed to vary with G instead of being held fixed, these assumptions would satisfy the more formal regularity conditions of Djogbenou, MacKinnon and Nielsen (2017).

Consider again the simplified DiD model (4). Under the null hypothesis, the parameter estimate $\hat{\beta}$ is given in equation (5), which can be rewritten as

$$\hat{\beta} = \frac{1}{N\bar{d}(1-\bar{d})} \left(\sum_{g=1}^{G_1} \boldsymbol{d}_g' \boldsymbol{\epsilon}_g - \bar{d} \sum_{g=1}^{G} \boldsymbol{\iota}_g' \boldsymbol{\epsilon}_g \right). \tag{18}$$

When $G_1 = 1$, the first summation here reduces to $\boldsymbol{d}_1' \boldsymbol{\epsilon}_1$, which is $O_p(N_1^{1/2})$. The second summation is $O(N_1/N)O_p(N^{1/2}) = O_p(N_1/N^{1/2})$. Since $N_1 < N$, it must be the case that $N_1^{1/2} > N_1/N^{1/2}$. Therefore, taking account of the first factor in (18), which is $O(1/N_1)$, we conclude that $\hat{\beta} = O_p(N_1^{-1/2})$.

Now consider the denominator of the t statistic when $G_1 = 1$. From (9), this is simply the square root of

$$\frac{1}{N^2(1-\bar{d})^2} \left((\boldsymbol{\iota}_1' \hat{\boldsymbol{\epsilon}}_1)^2 + \sum_{g=2}^{G} \boldsymbol{\iota}_g' \hat{\boldsymbol{\epsilon}}_g \hat{\boldsymbol{\epsilon}}_g' \boldsymbol{\iota}_g \right) = O(N^{-2})O_p(N) = O_p(N^{-1}), \tag{19}$$

because the second term inside the parenthesis is the leading-order one. The t statistic itself is therefore $O_p(N^{1/2}N_1^{-1/2})$. Thus, for any given sample with N fixed, the t statistic itself, like $\hat{\beta}$, is $O_p(N_1^{-1/2})$.

We conclude from this analysis that both RI procedures will tend to over-reject when N_1 is small and under-reject when N_1 is large. In the former case, both $\hat{\beta}$ and t_β will tend to be more variable than the β_r^* and t_r^* with which they are being compared, because $N_1^{-1/2}$ is larger than $N_r^{*-1/2}$ for most of the other clusters. In the latter case, by the same argument in reverse, both $\hat{\beta}$ and t_β will tend to be less variable than the β_r^* and t_r^* with which they are being compared. Thus neither RI procedure can possibly provide valid inferences when $G_1 = 1$ and the treated cluster is larger or smaller than the controls.

The case of $G_1 = 1$ is the most extreme one. As G_1 increases, we would expect the distribution of t_β eventually to lose any dependence on the sizes of the treated clusters, because the statistic is asymptotically pivotal. In contrast, the distribution of $\hat{\beta}$ will continue to depend on the sizes of the treated clusters. Thus we would expect the behavior of the two RI procedures to become less and less similar as G_1 increases in cases with unbalanced clusters where neither of them yields valid inferences when $G_1 = 1$.

The failure of both RI-β and RI-t when G_1 is small and cluster sizes vary, and of the former even when G_1 is not small, is a consequence of the fact that $\hat{\beta}$ and t_β depend on $\bar{d}$, which is not invariant across re-randomizations. As such, it is not surprising that both randomization inference procedures fail with unbalanced clusters, as the simulation results in Subsections 3.3 (above) and 3.5 (below) demonstrate. The more treated observations the treated clusters have (at least up to about half the sample size), the more efficiently β should be estimated.[9] Thus, except perhaps when $G_1 = 1$, we would expect randomization inference based on coefficient estimates to perform less well than randomization inference based on t statistics when the treated clusters are unusually large or small.

Conley and Taber (2011) originally suggested their Γ^* procedure, which is similar to RI-β, for use either with aggregate data or with individual data that have been aggregated into time-cluster cells. It seems to be a weaker assumption that $\hat{\beta}$ and the β_r^* follow the same distribution in those cases than in the case of individual data. Nevertheless, this assumption is still a very strong one. Variations across clusters in the number of underlying observations per cell, in the values of other regressors, or in the variances of the error terms may all invalidate this crucial assumption.[10] In contrast, t_β and the t_r^* can be expected to follow approximately the same distribution whenever G_1 and G are not too small.

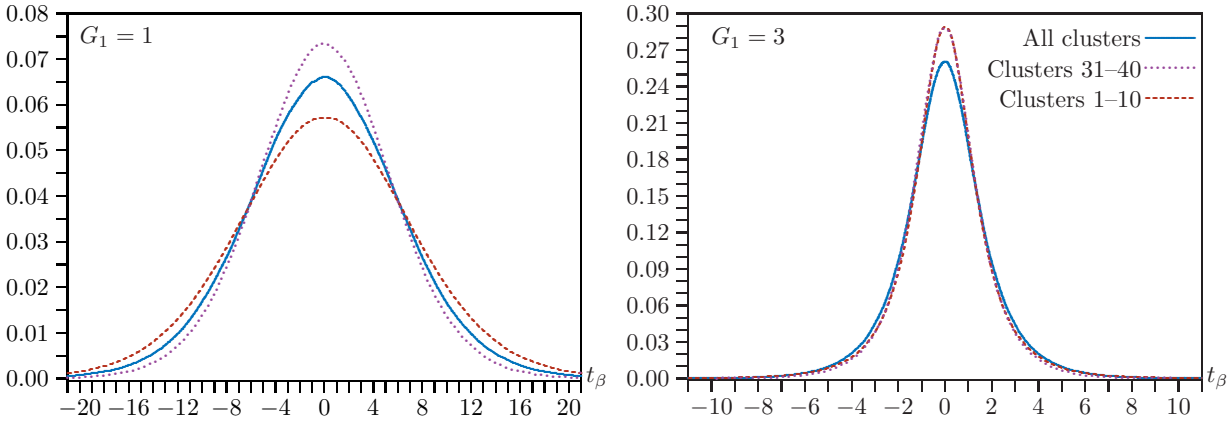
3.5 Performance of RI-t when Cluster Sizes Vary

As in Figure 1, and using results from the same simulations, we first plot conditional and unconditional distributions of t_β in Figure 6. Results for $G_1 = 1$ are again shown in the left panel and results for $G_1 = 3$ in the right panel. When $G_1 = 1$, the two conditional distributions are once again quite different from the unconditional distribution. This will

[9] It is difficult to be precise about this, because efficiency will also depend on intra-cluster correlations and the values of other regressors.

[10] Ferman and Pinto (2015) shows that aggregation of unbalanced clusters introduces heteroskedasticity in the aggregate data. When either large or small clusters are treated, this causes problems for randomization inference that are very similar to the ones with individual data.

Figure 6: Conditional and Unconditional Distributions of t_β



Notes: Based on 1,999,999 samples with $G = 40$, $G_1 = 1$ or 3, $\gamma = 2$, and $\rho = 0.05$

inevitably lead to the same inference problems as before, with RI-t over-rejecting when small clusters are treated and under-rejecting when large clusters are treated.

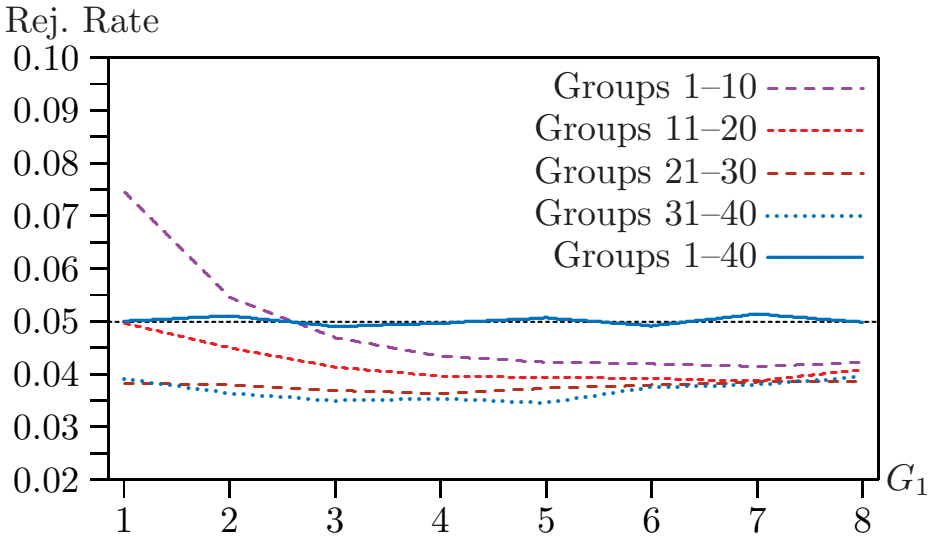
In contrast, the right panel of Figure 6, in which $G_1 = 3$, does not look much like the left panel. There are still some differences between the conditional distribution and the unconditional ones, but they are much less evident than they were in Figure 1. Moreover, and this is somewhat surprising, both of the conditional distributions are less spread out than the unconditional one. This suggests that RI-t may tend to under-reject for $G_1 > 1$ even when the treated clusters are relatively large.

To compare the performance of RI-t and RI-β, the experiments of Figure 2 are repeated using the former procedure instead of the latter. The results are shown in Figure 7. As must be the case, the RI-t procedure works perfectly (except for experimental error) when all clusters are potentially treated. As the analysis of Subsection 3.4 implies, it overrejects somewhat when the smallest clusters are treated and $G_1 = 1$, but not as much as RI-β. It also overrejects slightly in that case when $G_1 = 2$. However, as Figure 6 suggests, it actually underrejects in every other case. RI-t clearly outperforms RI-β when either the smallest or the largest clusters are treated, but there is not much to choose between the two procedures when intermediate clusters (numbers 11–20 or 21–30) are treated.

3.6 Varying the Number of Clusters

Our findings in Subsections 3.3 and 3.5 are not dependent on the total number of clusters being fairly small. Figure 8 shows what happens when G increases from 20 to 100 by increments of 10, with $N = 100G$ and $G_1 = G/10$ in each case. As in many of our experiments, $\gamma = 2$, which implies that the largest cluster is between about 7.2 (when $G = 20$) and 8.9 (when $G = 100$) times as large as the smallest one. In the left panel, the treated clusters are drawn from the lower half of the distribution. In the right panel, they are drawn from the upper half. The lackluster performance of RI-β is seen to be almost unrelated to the number of clusters in both cases. In the left panel, RI-β always over-rejects moderately, and in the right panel it always under-rejects quite substantially. In both panels,

Figure 7: Rejection Frequencies for RI-t Tests



Notes: Based on 100,000 replications with $G = 40$, $\gamma = 2$, and $\rho = 0.05$

RI-t works noticeably better than RI-β, but it too does not improve as G increases.

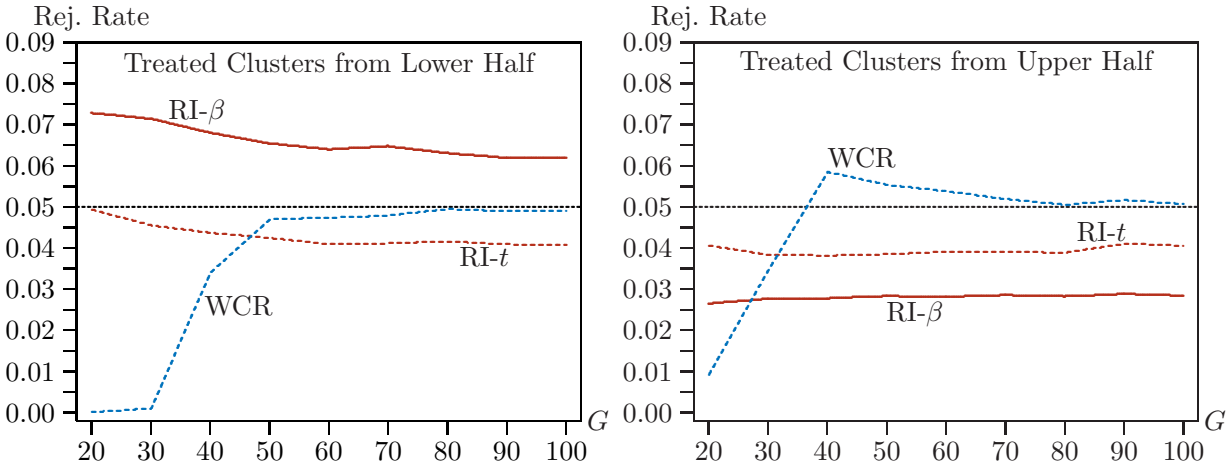
The figure also shows rejection frequencies for the WCR bootstrap. As in Figure 3, it under-rejects extremely severely when G, and hence also G_1, is small. The problem goes away more rapidly as G increases when the treated clusters are relatively large, as the theory in MacKinnon and Webb (2017b, Section 6) implies. In both panels, WCR works very well indeed for the largest values of G and G_1. Thus the figure illustrates a major difference between the wild cluster bootstrap and randomization inference. The bootstrap may work badly in small samples, but it is asymptotically valid under quite weak conditions. In contrast, if randomization inference is not valid in finite samples, it is typically not valid asymptotically either.

There is one feature of Figure 8 that requires comment. For $G = 20$ with $G_1 = 2$, the value of S for the two RI procedures is 189. This number does *not* satisfy the condition that $(B + 1) \times .05$ be an integer. Therefore, the two RI P values (13) and (16) yield different results. The rejection frequencies we report in the figure for $G = 20$ are the average of the ones for the two P values. They are essentially what would have been obtained if we had used the procedure involving a random number η discussed between equations (14) and (15) to determine the outcome of the test whenever (13) and (16) yield conflicting results.

3.7 RI Procedures with Heteroskedasticity across Clusters

In the experiments reported so far, the distributions of $\hat{\beta}$ and the corresponding β_r^*, and of t_β and the corresponding t_r^*, can only differ across clusters when cluster sizes vary. However, this is not the only possible reason for those distributions to differ. Another possibility is that the error terms of the treated clusters may have larger or smaller variances than those of the controls. For simplicity, suppose there are just two variances, with the ratio of the

Figure 8: Rejection Frequencies when G and G_1 Vary Together



Notes: Based on 100,000 replications with $\gamma = 2$, $\rho = 0.05$, $N = 100G$, $G_1 = G/10$.

ones for the treated and control clusters equal to λ^2.

Consider the extreme case in which $G_1 = 1$. From equation (18), it is evident that the larger is the variance of the error terms for the treated cluster, the larger will be the variance of $\hat{\beta}$. This follows from the fact that the first summation within the large parentheses, which depends on those error terms, is proportional to λ. The variance of t_β must also be increasing in λ in this case, because the second summation within the large parentheses in (19) is the leading-order one. Therefore, the cluster-robust standard error underestimates the true standard error of $\hat{\beta}$ more and more severely as λ becomes larger.

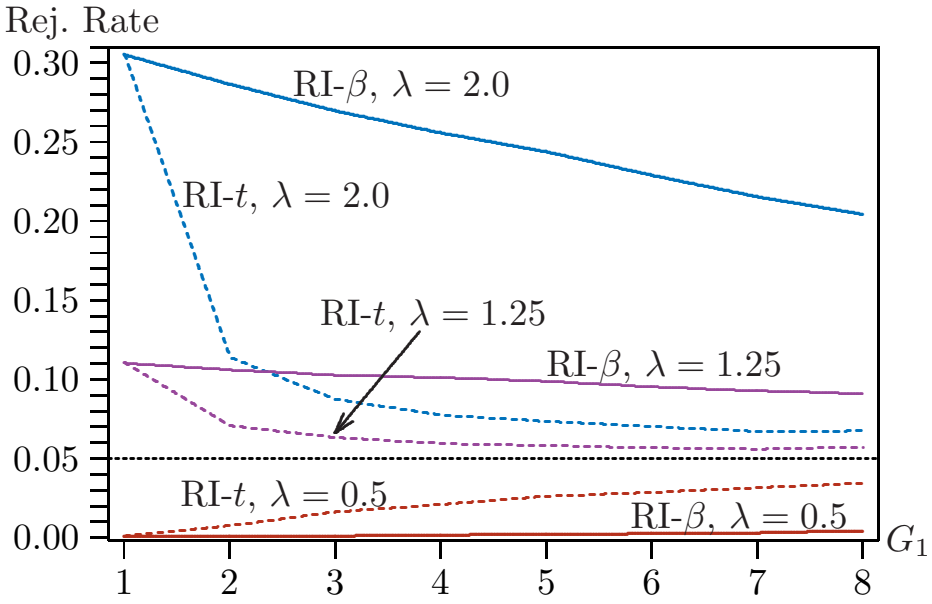
To investigate this phenomenon, we perform an additional set of experiments in which the standard error for the treated clusters is λ times the standard error for the controls. We would expect over-rejection when $\lambda > 1$ and under-rejection when $\lambda < 1$. As G_1 increases, we expect the problem to go away for RI-t but not for RI-β.

Figure 9 shows rejection frequencies for the RI-β and RI-t procedures for a DiD model with 40 equal-sized clusters and 4000 observations. Results are shown for three values of λ, namely, $\lambda = 2.0$, $\lambda = 1.25$, and $\lambda = 0.5$. As expected, both procedures over-reject when $\lambda > 1$ and under-reject when $\lambda < 1$. When $G_1 = 1$, both the over-rejection for $\lambda = 2.0$ and the under-rejection for $\lambda = 0.5$ are very severe. In this case, they are identical for RI-β and RI-t for all three values of λ.

As G_1 increases, the performance of RI-t initially improves quite quickly, while that of RI-β improves very slowly. However, the rate of improvement for RI-t slows down greatly as G_1 increases. It still over-rejects noticeably for $G_1 = 8$ when $\lambda = 2.0$ and it under-rejects noticeably when $\lambda = 0.5$.[11] The size distortions in Figure 9 are much more severe

[11] Using a different experimental design, Canay, Romano and Shaikh (2017, Appendix) also studies the performance of a Conley and Taber (2011) procedure (and several other tests) when the treated clusters have greater variance than the untreated ones. They find even more severe over-rejection for $\lambda = 2.0$ than we do.

Figure 9: Rejection Frequencies with Heteroskedasticity



Notes: Based on 100,000 replications with $G = 40$, $\gamma = 2$, $N = 4000$, and $\rho = 0.05$

than in previous figures, which suggests that heteroskedasticity associated with treatment status may be a serious impediment to valid randomization inference. This might occur, for example, if treatment caused individual outcomes to become more or less variable.

3.8 Power of Alternative Procedures

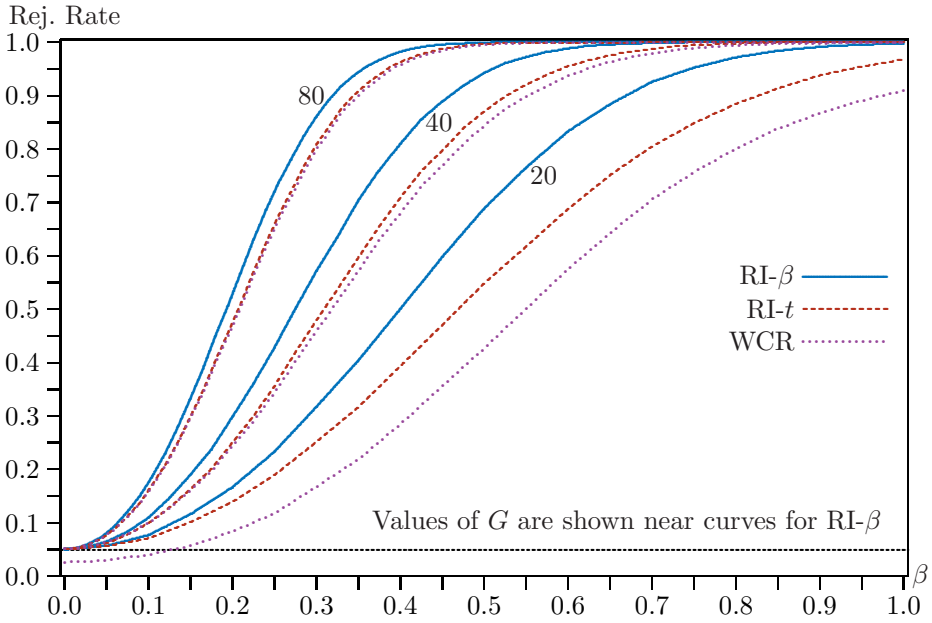
One possible drawback of RI-t, and of every other procedure based on asymptotic t statistics, is that the denominator of the t statistic adds noise, and noise inevitably reduces power. Because the CRVE (3) can be a rather inefficient estimator when G is small, the loss of power is potentially substantial. In this subsection, we investigate this issue by conducting a set of Monte Carlo experiments in which $G = 20, 40,$ or 80 with $G_1 = 3$, 6, or 12 treated clusters, respectively. All clusters are the same size, with $N_g = 50$ for all g. This ensures that the RI procedures have the correct size under the null.

We vary the true value of β between 0 and 1 and plot the power functions of RI-β, RI-t, and the WCR bootstrap in Figure 10. As expected, the power of all procedures increases with the number of clusters, and the differences between them diminish. However, RI-β evidently has substantially higher power than RI-t. Its power advantage is clearly evident even when $G = 80$, which is a relatively large number of clusters.

The WCR bootstrap underrejects quite severely when $G = 20$ and $G_1 = 3$, so it is not surprising that it has substantially less power than RI-t in that case. However, it performs very well under the null in the other two cases, and it still has slightly less power than RI-t. This suggests that it may be desirable to use RI-t rather than WCR when G_1 is not particularly small, whether or not all clusters are the same size.

Of course, in cases where RI-β over-rejects under the null, it will appear to have an

Figure 10: Power of Alternative Tests with Equal-Sized Clusters



Notes: Based on 100,000 replications with $N = 50G$, $G_1/G = 0.15$, $\rho = 0.05$, and $B = 999$

even greater power advantage than it does in Figure 10. However, even in cases where RI-β under-rejects under the null, it may have more power than RI-t for large enough values of β. We found this in some experiments that we do not report.

4 Other Inferential Procedures

It has been known for some time that detecting treatment effects reliably when very few clusters are treated is extremely difficult unless one is willing to make uncomfortably strong assumptions about the error terms (for example, that they are uncorrelated within each cluster). Many procedures for tackling this difficult problem have therefore been proposed. In this section, we briefly discuss a number of these procedures. Simulation evidence for some of them is presented in Appendices A and B.

4.1 Bootstrap Methods

The wild cluster bootstrap was proposed in Cameron, Gelbach and Miller (2008). The key feature of this bootstrap method is that there is one drawing of an auxiliary random variable for each cluster, instead of one per observation as for the ordinary wild bootstrap. Every residual in cluster g is multiplied by the same auxiliary random variable, say v_g^*, when generating each bootstrap sample. In all our experiments, we draw the v_g^* from the Rademacher distribution, which takes the values -1 and $+1$ with equal probability.

MacKinnon and Webb (2017b, Section 6) explains why the wild cluster bootstrap fails when the number of treated clusters is small. The WCR bootstrap, which imposes the null hypothesis on the bootstrap DGP, leads to severe under-rejection, as was seen in Figures 3

and 8. In contrast, the WCU bootstrap, which does not impose the null hypothesis, leads to severe over-rejection, as can be seen in Figure 3. When just one cluster is treated, WCU over-rejects almost as much as using CRVE t statistics with the $t(G-1)$ distribution. This is unfortunate, because it is easy to use WCU to form studentized bootstrap confidence intervals, but they tend to under-cover severely when there are few treated clusters.

Recently, MacKinnon and Webb (2018a) suggested using the ordinary wild bootstrap together with cluster-robust standard errors, and Djogbenou, MacKinnon and Nielsen (2017) proved that doing so is asymptotically valid. The WR (for restricted) and WU (for unrestricted) versions of this procedure can work remarkably well when cluster sizes are equal. In addition, they are essentially unaffected by heteroskedasticity at the cluster level. However, like RI-β, they are very sensitive to variable cluster sizes. Moreover, unlike RI-β, they are also sensitive to variation in the number of treated observations per cluster, which often occurs in a DiD context. Some evidence about the performance of WR is presented in Appendix A; WU performs very similarly in most cases.

A very different bootstrap procedure, usually called the pairs cluster bootstrap, was suggested by Bertrand, Duflo and Mullainathan (2004). In this procedure, each bootstrap sample is obtained by resampling all of the data at the cluster level. Thus each bootstrap sample contains G clusters, some of them repeats, and the sample size varies across bootstrap samples unless all clusters are the same size. The number of treated clusters also varies across bootstrap samples and may even be zero for some of them when G_1 is small for the actual sample. Simulation results for this procedure are presented in MacKinnon and Webb (2017a). When $G_1 = 1$, the pairs cluster bootstrap over-rejects extremely severely, about the same as WCU, but it can perform quite well when neither G nor G_1 is too small.

4.2 Bias Correction and Degrees-of-Freedom Methods

Carter, Schnepel and Steigerwald (2017) discusses the asymptotic properties of the CRVE (3) when the number of observations per cluster is not constant. It shows that, when clusters are unbalanced, a sample typically has an effective number of clusters, G^*, which is less than G (sometimes very much less). Simulations in MacKinnon and Webb (2017b) show that using critical values based on G^* can work fairly well when intermediate numbers of clusters are treated. However, when very few clusters are treated in the DiD context, it can either over-reject or under-reject. We consider the performance of what we call the $t(G^*)$ procedure in some of our simulation experiments.

Alternative degrees-of-freedom corrections, in some cases based on alternative CRVEs, have also been proposed in Bell and McCaffrey (2002), Imbens and Kolesár (2016), and Young (2016). The first two of these papers propose procedures that use an alternative CRVE, which we call CV_2, that is analogous to the HC_2 HCCME discussed in MacKinnon and White (1985). It requires finding the inverse symmetric square root matrix $\boldsymbol{M}_{gg}^{-1/2}$ for each of the $N_g \times N_g$ diagonal blocks $\boldsymbol{M}_{gg}$ of the $N \times N$ matrix $\boldsymbol{M}_{\boldsymbol{X}} \equiv \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. The N_g-vector of residuals for each cluster is then premultiplied by $\boldsymbol{M}_{gg}^{-1/2}$. Doing so has the effect of inflating the residuals, thereby increasing the cluster-robust standard errors. However, this is computational demanding. For the cases we study, simply computing CV_2 is much more costly than using either randomization inference or the wild cluster bootstrap.

As we illustrate in Appendix A, using CV_2 rather than CV_1 leads to substantially less

over-rejection when G_1 is small. However, it still yields rejection rates that are much too high. The procedures of Bell and McCaffrey (2002) and Imbens and Kolesár (2016) combine CV_2 with estimated degrees-of-freedom parameters, the computation of which can be extremely demanding.[12] When G_1 is small, these parameters also tend to be small. In consequence, the critical values can be very much larger than the ones from the $t(G-1)$ distribution that are conventionally used.

A much less computationally demanding procedure was proposed in Young (2016). It starts with the CV_1 CRVE (3), then inflates each diagonal element by a factor (which is different for every coefficient) that is designed to offset its downward bias, and finally computes an alternative degrees-of-freedom parameter that is conceptually similar to the one in Bell and McCaffrey (2002). In MacKinnon and Webb (2018a), we found that Young's procedure tends to yield rejection frequencies that are quite similar to the ones from the Imbens-Kolesár procedure. We present a number of results for it in Appendix A.

4.3 Methods that Use Different Estimates of β

We consider a large number of inferential procedures in this paper. In order to keep the results manageable, we restrict our experiments to methods based on OLS estimation of equation (2). However, several methods that use other estimates have also been proposed.

Building off results in Donald and Lang (2007), Ibragimov and Müller (2016) studies the generalized Behrens-Fisher problem of comparing the means of two groups with different unknown covariance matrices. The paper focuses on differences in means for treated and control groups and proves that appropriately constructed t tests for these differences follow asymptotic distributions with degrees of freedom equal to $\min(G_0, G_1) - 1$. When $G_1 = 1$, this number is 0, which implies that the Ibragimov-Müller procedure is inapplicable when there is only one treated group. The procedure is primarily designed for the pure treatment case, but the paper also discusses how to extend it to a DiD model with a common treatment start date. However, it does not explain how to deal with models in which treatment starts at different dates, as in all of our experiments. We therefore do not attempt to study the performance of this procedure.

Canay, Romano and Shaikh (2017) proposes a related procedure which requires a matching of treated clusters to control clusters. In their general framework, G_1 is small and G_0 is large. When both G_0 and G_1 are small, the required matching is not easily accomplished, and the paper recommends the procedure of Ibragimov and Müller (2016). The former procedure has power at the 5% level that is always strictly less than one when the minimum of G_0 and G_1 is less than 5, because there are too few re-randomizations. Since the RI-β and RI-t procedures are most attractive for cases with $G_1 \leq 4$, and do provide consistent tests even when $G_1 = 1$, it is not interesting to compare them with the procedure of Canay, Romano and Shaikh (2017).

A very different procedure is proposed by Abadie, Diamond and Hainmueller (2010). Like

[12] We ran into computational difficulties when we attempted to compute these parameters for $G_1 = 1$. We were able to compute them for $G_1 > 1$, but at great computational cost. Even for the rather modest sample sizes in our experiments (often just 4000), the procedure of Imbens and Kolesár (2016) was many times more expensive than any of the randomization inference or bootstrap procedures. MacKinnon and Webb (2018a) provides evidence on how the cost of this procedure, and others, varies with the sample size.

the RI procedures, it bases inference on an empirical distribution generated by perturbing the assignment of treatment. However, the procedure differs substantially from the ones considered in this paper, because it constructs a "synthetic control" as a weighted average of potential control groups, based on the characteristics of the explanatory variables for these groups. This results in both a different estimate of the "treatment effect" and a different P value. For this reason, we do not study the synthetic controls approach in this paper.

Two other procedures that we do not investigate require much stronger assumptions about the error terms than the assumptions in (2). Ferman and Pinto (2015) proposes a form of RI procedure, which requires users to estimate a pattern of cross-cluster heteroskedasticity. Brewer, Crossley and Joyce (2018) proposes a feasible GLS procedure, which requires users to estimate autocorrelation.

5 Empirical Examples

In this section, we consider two empirical examples. In the first of them, $G_1 = 2$, so that randomization inference may be expected to work well if the treated clusters are not atypical, but other methods can be expected to work poorly. In the second, $G_1 = 10$, so that many methods should work well. We include the second example because it was used in Conley and Taber (2011).

5.1 Birth Control Pills

Bailey (2010) examines the relationship between the introduction of the birth control pill and the decrease in fertility in the United States since about 1957. The paper uses state-by-state variation in "Comstock laws," which prohibited, among other things, the advertising and sale of the birth control pill. The practice of using these laws to restrict the sale of birth control pills was essentially ended by the U.S. Supreme Court's 1965 Griswold v. Connecticut decision. Part of the analysis in Bailey (2010) shows that women in states with sales restrictions on the birth control pill were indeed less likely to have taken the pill by 1965. The analysis employs a DiD regression using data on married, white women from the *National Fertility Surveys* for the years 1965 and 1970. The women come from 47 states, and clustering is done at the state level.

Bailey estimates a probit regression in which the dependent variable is an indicator variable that equals 1 in 1965 or 1970 if the respondent had ever taken the birth control pill by that year. The key regressors are an indicator variable `Salesban` that equals 1 if the state had a sales ban on the birth control pill in 1960, and `Salesban` interacted with a dummy variable `D1970` for observations from 1970. Estimated coefficients and standard errors for these two regressors are presented in her Table 2, Column 1. Other regressors include `D1970`, three regional dummies, an indicator variable equal to 1 if the state had a physician exemption to the sales ban, and each of these variables interacted with `D1970`.

There is no real need to use a probit model in this case. Because all regressors are indicator variables, and the mean of the dependent variable (which is 0.515) is far from the limits of 0 and 1, using OLS inevitably produces results almost identical to the probit ones. In fact, the probit t statistics for `Salesban` and `Salesban`×`D1970` are −2.76 and 1.46, and the OLS ones are −2.71 and 1.37; these are all based on cluster-robust standard errors.[13]

[13] Although we attempted to use the same sample as Bailey, our sample has 6929 observations, and hers

Prior to the "Griswold" decision, several states repealed their previously existing sales bans. In particular, Illinois and Colorado repealed their Comstock laws in 1961. It is of interest to ask whether women in these early-repeal states were more or less likely to use the pill than women in other states with a sales ban. We therefore created an indicator variable `rep61` equal to 1 for those two states and added `rep61`×`D1965` and `rep61`×`D1970` to the base specification. Results for the four coefficients of interest are shown in Table 1.

Table 1: Effects of Sales Ban and Early Repeal, Full Sample

	Coef.	Std. Err.	CR t-stat	RI-β p^*	RI-t p^*	
`Salesban`	−0.042	0.016	−2.677			
`Salesban`×`D1970`	0.029	0.027	1.059			
`rep61`×`D1965`	−0.125	0.023	−5.432	0.063	0.056	
`rep61`×`D1970`	−0.043	0.029	−1.488	0.615	0.445	
	Young p	CSS p	IM Coef.	IM p	WR p^*	WCR p^*
`Salesban`	0.019	0.034			0.035	0.028
`Salesban`×`D1970`	0.184	0.318			0.366	0.320
`rep61`×`D1965`	0.028	0.059	−0.338	0.221	0.021	0.546
`rep61`×`D1970`	0.315	0.322	0.338	0.221	0.638	0.458

Notes: Outcome variable is whether respondent had ever taken the birth control pill. The sample is women from 47 states, 23 of which had a sales ban. `rep61` = 1 for individuals in Illinois and Colorado. Standard errors are clustered at the state level.

Taken at face value, the cluster-robust t statistic for `rep61`×`D1965` in column 3 of Table 1 appears to be telling us that living in an early-repeal state very significantly lowered the probability of using the pill in 1965. However, because there are only two such states, the analysis of Section 2.1 suggests that this t statistic is probably much too large. In contrast, the WCR bootstrap (based on $B = 99{,}999$) yields a P value of about 0.55, which the analysis of MacKinnon and Webb (2017b) suggests is probably much too conservative. Thus the cluster-robust t statistic and the bootstrap P value yield wildly contradictory results, which could have been expected before even computing them, and are therefore of no real use in this case.

We also compute two randomization inference P values for each regressor involving `rep61`. Because $G = 47$, the value of S is $(47 \cdot 46)/2 - 1 = 1080$. We report RI P values computed using equation (16), because they are slightly more conservative than ones based on equation (13). The two RI procedures yield results that are very similar to each other, with P values just a little greater than .05. Although the RI P values do not entirely resolve the uncertainty about whether the coefficient on `rep61`×`D1965` is significant, they at least yield sensible results that could not have been predicted in advance.

P values for other procedures discussed in Section 4 are also reported. The ones for the procedure of Young (2016) and the ones based on the effective degrees of freedom proposed

has 6950. We are unable to explain this minor discrepancy. Bailey does not explicitly report t statistics. Calculating them from coefficients and standard errors reported to only two decimal places, her t statistics are similar enough to our probit ones that they could actually be equal.

in Carter et al. (2017) give broadly similar results. Moreover, these P values are quite similar to the WR P values and the RI P values (where they exist). Specifically, all of the P values for `rep61`×`D1965` are below 0.10 (except for WCR), while all of the P values for `rep61`×`D1970` are well above 0.10. We also calculate P values and coefficient estimates using the procedure in Ibragimov and Müller (2016). One limitation of this procedure is that, although standard difference-in-differences analysis allows for year-specific treatment effects, the IM procedure always estimates these coefficients to be the negative of one another when there are only two periods.

One way to investigate the robustness of these results is to limit attention to the 23 states that had sales bans in 1960. This reduces the sample size to 3780 observations and requires us to drop the variables `Salesban` and `Salesban`×`D1970`. Results for the two coefficients of interest are shown in Table 2.

Table 2: Effects of Early Repeal, Sales Ban Sample

	Coef.	Std. Err.	CR t-stat	RI-β p^*	RI-t p^*	
`rep61`×`D1965`	−0.120	0.024	−4.917	0.040	0.079	
`rep61`×`D1970`	−0.046	0.032	−1.445	0.316	0.289	
	Young p	CSS p	IM Coef.	IM p	WR p^*	WCR p^*
`rep61`×`D1965`	0.032	0.062	−0.336	0.225	0.029	0.439
`rep61`×`D1970`	0.316	0.326	0.336	0.225	0.641	0.460

Notes: Outcome variable is whether respondent had ever taken the birth control pill. The sample is women from 23 states which had a sales ban. `rep61` = 1 for individuals in Illinois and Colorado. Standard errors are clustered at the state level.

The results in Table 2 are very similar to the ones in Table 1. The most noticeable difference is that the RI-β P value is now just 0.040, while the RI-t P value is almost exactly twice as large. Since these are based on $S = (23 \cdot 22)/2 - 1 = 252$, they would have been noticeably smaller (0.036 and 0.075, respectively) if we had used equation (13) rather than equation (16) to compute them. It is difficult to understand why early repeal would have reduced pill use in 1965. The RI-β P value suggests that there is stronger evidence against the null than does the RI-t P value. However, the results in Figure 2 suggest that RI-t should be more reliable than RI-β.

The Young P and CSS P values again give results that are quite similar both to one another and to the RI P values. The WR and WCR bootstrap P values exhibit the same pattern as in Table 1. The WCR P value is quite large because there are only two treated clusters, while the WR one rejects the null at the 5% level for the `D1965` coefficient, but it provides little evidence against the null for the `D1970` coefficient. We again calculate IM coefficients and P values, but the coefficients are negatives of one another by construction and thus not meaningful.

Although the results in Tables 1 and 2 are not entirely definitive, randomization inference certainly yields results that are much more plausible, and much less predictable, than using either cluster-robust t statistics or the wild cluster bootstrap. Moreover, these P values are consistent with those from the Young and CSS procedures, and the ordinary wild bootstrap.

5.2 Merit Scholarships

In this subsection, we consider an empirical example studied in Conley and Taber (2011). It deals with the impact of state-level merit scholarships initiated during the 1989-2000 period. These programs generally offered scholarships for students to attend college in their home state conditional on being above some academic threshold. The details differ state by state, but they are not important for our purposes.

Conley and Taber (2011) attempts to determine whether the 10 merit scholarships that were in operation by the end of 2000 had any impact on college enrollment by estimating the following DiD regression using data from 1989-2000:

$$\begin{aligned}\text{college}_{ist} = \beta_0 + \beta_1 \text{merit}_{ist} + \beta_2 \text{male}_{ist} + \beta_3 \text{black}_{ist} + \beta_4 \text{asian}_{ist} \\ + \sum_{j=2}^{51} \gamma_j \text{state}_{ist}^j + \sum_{k=2}^{12} \delta_k \text{year}_{ist}^k + \epsilon_{ist}.\end{aligned}$$

Here college_{ist} is the outcome of interest, a binary indicator for whether individual i in state s and year t was enrolled in college, and the treatment variable merit_{ist} equals 1 if state s offered a merit scholarship in year t. The remaining variables are all binary indicator variables. The dataset has $N = 42{,}161$ observations taken from all states, including the District of Columbia, so that $G = 51$.

Table 3: Effect of Merit Scholarships on College Enrollment

	Coef.	Std. Err	CR t-stat	RI β p^*	RI t p^*
merit	0.034	0.013	2.654	0.117	0.034
	Young p	CSS p	WR p^*	WCR p^*	
merit	0.018	0.071	.030	0.021	

Notes: Outcome variable is whether individual had ever enrolled in college. Sample is 42,161 individuals from all 50 states and DC. `Merit` = 1 for individuals in the 10 states with merit scholarships in the relevant treatment years. Standard errors are clustered at the state level.

The original paper presents estimates of β_1, along with several different confidence intervals, in Column C of Table II. The table reports that $\hat{\beta}_1 = 0.034$, along with a 95% CRVE confidence interval of [0.008, 0.059]. We calculate several alternative P values and present the results in Table 3. Although it is not explicitly reported, we calculated the P value for the test of $\beta_1 = 0$ based on the $t(50)$ distribution to be 0.010. However, using a method that essentially inverts RI-β P values, the paper estimates a 95% confidence interval for β_1 of $[-0.003, 0.093]$.[14] Thus, unlike the conventional CRVE confidence interval, the Conley-Taber 95% confidence interval contains 0.

We use the Conley-Taber data and modify their Stata code to conduct inference on β_1 using both RI-β and RI-t.[15] With 9999 randomizations and symmetric P values, we obtain

[14] The procedure searches separately for both the upper and lower limit of the confidence interval, by re-randomizing treatment amongst 10 of the 51 states.

[15] We thank the authors for making their code and data easily available.

an RI-t P value of 0.032 and an RI-β P value of 0.117. Like the Conley-Taber confidence interval, the RI-β P value fails to reject the null at the 5% level. In contrast, our RI-t P value of 0.032 suggests that there is a statistically significant effect at the 5% level.

We also calculate the WCR P value for $\beta_1 = 0$, based on $B = 99,999$ bootstraps. It is 0.021, which is quite similar to the RI-t P value. With 10 treated states, the WCR P value should be quite reliable. The WR P value, which should also be reliable, is very similar. We also calculate Young and CSS P values. The former rejects the null at the 5% level, and the latter rejects it at the 10% level.[16] In view of these results and the fact that, in all our Monte Carlo experiments, the RI-t procedure tended to be slightly under-sized, but quite close to 5%, we conclude that the merit scholarship programs did have a statistically significant impact.

6 Conclusion

We compare several new and existing procedures for inference with few treated clusters in the context of difference-in-differences, focusing on two methods that are based on randomization inference (RI). There are four main findings, some of which were obtained theoretically for a simple model in Subsection 3.4, and all of which were confirmed by simulation results in various subsections of Section 3 and in Appendices A and B.

The first result is that none of the procedures we study works well when there are very few treated clusters and those clusters are atypical in terms of either the number of observations or the variance of the error terms. The second is that both RI procedures appear to work well when the treated clusters are typical. In an ex ante sense, they work perfectly when the treated clusters are chosen at random. The third is that, when the number of treated clusters (G_1) equals 1, RI-t, the RI procedure based on t statistics, performs very similarly to RI-β, the one based on coefficient estimates. However, the performance of RI-t tends to improve as G_1 increases, at least up to a point, while that of RI-β may or may not improve. We never encountered a case where RI-t performs really badly for $G_1 \geq 3$. The fourth result, which makes sense theoretically, is that RI-β tends to have substantially more power than RI-t or other procedures based on cluster-robust standard errors.

The performance of all the procedures we study depends in a complicated way on G, G_1, the sizes of the treated and control clusters, and the number of treated observations within the treated clusters. This suggests that the best procedure to use will depend on the specific dataset under analysis. Accordingly, prudent empirical researchers would benefit from conducting their own small-scale Monte Carlo experiments using the values of G, G_1, and the N_g found in their dataset, in addition to plausible values of ρ.

[16] We are unable to calculate IM P values here as treatment starts at different years.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010) 'Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program.' *Journal of the American Statistical Association* 105(490), 493–505

Bailey, Martha A. (2010) '"Momma's got the pill": How Anthony Comstock and Griswold v. Connecticut shaped US childbearing.' *American Economic Review* 100(1), 98–129

Barrios, Thomas, Rebecca Diamond, Guido W. Imbens, and Michal Kolesár (2012) 'Clustering, spatial correlations, and randomization inference.' *Journal of the American Statistical Association* 107(498), 578–591

Bell, Robert M., and Daniel F. McCaffrey (2002) 'Bias reduction in standard errors for linear regression with multi-stage samples.' *Survey Methodology* 28(2), 169–181

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) 'How much should we trust differences-in-differences estimates?' *The Quarterly Journal of Economics* 119(1), pp. 249–275

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen (2011) 'Inference with dependent data using cluster covariance estimators.' *Journal of Econometrics* 165(2), 137–151

Brewer, Mike, Thomas F. Crossley, and Robert Joyce (2018) 'Inference with difference-in-differences revisited.' *Journal of Econometric Methods*

Cameron, A. Colin, and Douglas L. Miller (2015) 'A practitioner's guide to cluster robust inference.' *Journal of Human Resources* 50, 317–372

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2008) 'Bootstrap-based improvements for inference with clustered errors.' *The Review of Economics and Statistics* 90(3), 414–427

Canay, Ivan A, Joseph P Romano, and Azeem M Shaikh (2017) 'Randomization tests under an approximate symmetry assumption.' *Econometrica* 85(3), 1013–1030

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (2017) 'Asymptotic behavior of a t test robust to cluster heterogeneity.' *Review of Economics and Statistics* 99(4), 698–709

Conley, Timothy G., and Christopher R. Taber (2011) 'Inference with "Difference in Differences" with a small number of policy changes.' *The Review of Economics and Statistics* 93(1), 113–125

Djogbenou, Antoine, James G. MacKinnon, and Morten Ø. Nielsen (2017) 'Validity of wild bootstrap inference with clustered errors.' Working Paper 1383, Queen's University, Department of Economics

Donald, Stephen G, and Kevin Lang (2007) 'Inference with difference-in-differences and other panel data.' *The Review of Economics and Statistics* 89(2), 221–233

Dufour, Jean-Marie (2006) 'Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics.' *Journal of Econometrics* 133(2), 443–477

Ferman, Bruno, and Christine Pinto (2015) 'Inference in differences-in-differences with few treated groups and heteroskedasticity.' Technical Report, Sao Paulo School of Economics

Fisher, R.A. (1935) *The Design of Experiments* (Oliver and Boyd)

Ibragimov, Rustam, and Ulrich K. Müller (2016) 'Inference with few heterogeneous clusters.' *Review of Economics & Statistics* 98(1), 83–96

Imbens, Guido W., and Donald B. Rubin (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press)

Imbens, Guido W., and Michal Kolesár (2016) 'Robust standard errors in small samples: Some practical advice.' *Review of Economics and Statistics* 98(4), 701–712

Lehmann, E. L., and Joseph P. Romano (2008) *Testing Statistical Hypotheses* Springer Texts in Statistics (Springer New York)

Liang, Kung-Yee, and Scott L. Zeger (1986) 'Longitudinal data analysis using generalized linear models.' *Biometrika* 73(1), 13–22

MacKinnon, James G., and Halbert White (1985) 'Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties.' *Journal of Econometrics* 29(3), 305–325

MacKinnon, James G., and Matthew D. Webb (2017a) 'Pitfalls when estimating treatment effects using clustered data.' *The Political Methodologist* 24(2), 20–31

MacKinnon, James G., and Matthew D. Webb (2017b) 'Wild bootstrap inference for wildly different cluster sizes.' *Journal of Applied Econometrics* 32, 233–254

MacKinnon, James G., and Matthew D. Webb (2018a) 'The wild bootstrap for few (treated) clusters.' *Econometrics Journal* 21, to appear

MacKinnon, James G., and Matthew D. Webb (2018b) 'Wild bootstrap randomization inference for few treated clusters.' Working Paper 1404, Queen's University, Department of Economics

Racine, Jeffrey S., and James G. MacKinnon (2007) 'Simulation-based tests that can use any number of simulations.' *Communications in Statistics: Simulation and Computation* 36(2), 357–365

Rosenbaum, Paul R (1996) 'Observational studies and nonrandomized experiments.' *Handbook of Statistics* 13, 181–197

Yates, F. (1984) 'Tests of significance for 2 × 2 contingency tables.' *Journal of the Royal Statistical Society. Series A (General)* 147(3), 426–463

Young, Alwyn (2015) 'Channelling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results.' Technical Report, London School of Economics

Young, Alwyn (2016) 'Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections.' Technical Report, London School of Economics

Appendix A: Simulation Results for Additional Methods

In this appendix, we present simulation results for several of the alternative procedures discussed in Section 4. Figure 11 reports additional results for three of the five experiments initially reported in Figures 2 and 3. To keep the figure readable, rejection frequencies are shown only for the case in which all groups are treated and for the two extreme cases in which groups 1-10 (the smallest ones) and groups 31-40 (the largest ones) are treated.

In the left panel, it is evident that the restricted wild cluster bootstrap (WCR) almost never rejects when $G_1 \leq 2$ and under-rejects severely for $G_1 = 3$, except when the largest clusters are treated. These results are explained in MacKinnon and Webb (2017b, Section 6). They are caused by dependence between the actual t statistic and the bootstrap t statistics. The left panel of Figure 11 also shows rejection frequencies for the ordinary restricted wild bootstrap (WR). For $G_1 \leq 4$, WR works considerably better than WCR, except when $G_1 = 1$ and the smallest clusters are treated, although its performance is far from perfect. For $G_1 \geq 5$, however, WCR works a bit better than WR. For larger values of G_1, the results for WR appear to be much more sensitive to the size of the treated clusters than the results for WCR. Broadly similar results are reported in MacKinnon and Webb (2018a).

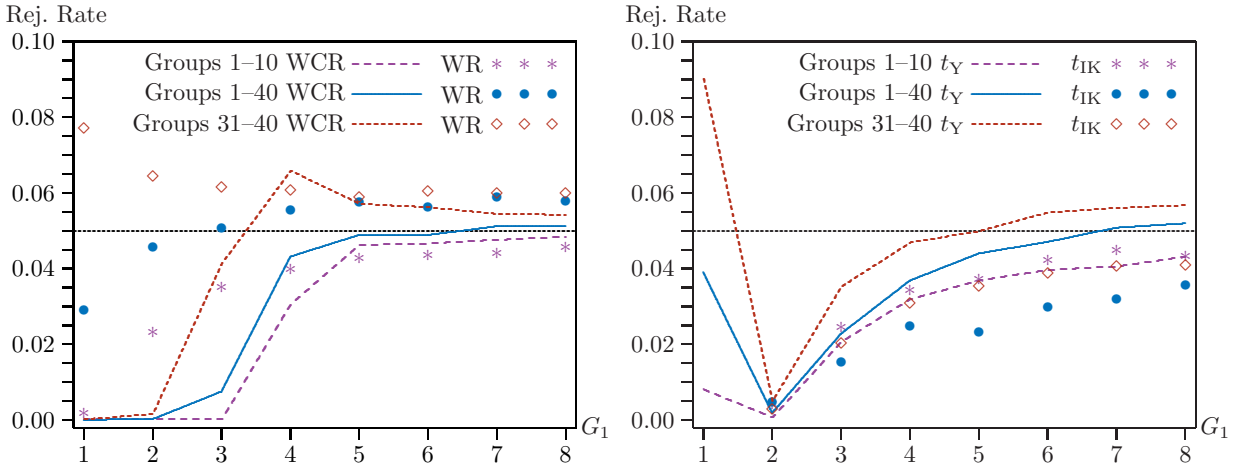
The right panel of Figure 11 reports rejection frequencies for two procedures that use standard errors which differ from the usual ones based on CV_1 and also use critical values based on calculated degrees of freedom that are smaller (often very much smaller) than $G - 1$. The procedure called t_{Y} in the figure is due to Young (2016), and the one called t_{IK} is due to Imbens and Kolesár (2016); see Subsection 4.2. The former procedure is very inexpensive to compute, but the latter is extremely expensive. Results for it (which are not available for $G_1 = 1$) are therefore based on only 20,000 replications. In the figure, the performance of t_{Y} is usually a bit better than that of t_{IK}. For $G_1 \geq 4$, the t_{Y} procedure generally works quite well.

If we compare the results in Figure 11 with those in Figures 2 and 3, we see that, for $G_1 \geq 4$, all of the alternative procedures outperform RI-β when either the smallest or largest groups are treated. Several of them also outperform RI-t for some or all of the same cases. Of course, both RI procedures work perfectly when all groups are treated, and they typically work better than most of the alternative procedures when $G_1 \leq 2$.

The procedures considered in the right panel of Figure 11 both differ in two ways from the usual one. Two procedures that each differ in only one way are to use standard errors based on CV_2 together with the usual $t(G-1)$ critical values, and to use ordinary CV_1 standard errors together with critical values based on the effective degrees of freedom G^* suggested by Carter, Schnepel and Steigerwald (2017). Figure 12 shows rejection frequencies for these two procedures for the same three cases as Figure 11. Using CV_2 works quite a bit better than using CV_1, but it still over-rejects substantially even for the largest values of G_1. In contrast, using $t(G^*)$ critical values works remarkably well, especially when all groups or the largest ones are treated and $G_1 \geq 4$.

In Figure 9, we showed that heteroskedasticity at the cluster level severely impacts the performance of RI-β. RI-t also performs very badly when $G_1 = 1$, but it improves rapidly as G_1 increases. In Figure 13, we report results of the same experiments for the two restricted bootstrap tests. The ordinary wild bootstrap (WR) works very much better than the wild

Figure 11: Rejection Frequencies for Alternative Procedures



Notes: Based on 100,000 or 20,000 replications with $G = 40$, $\gamma = 0$, and $\rho = 0.05$

cluster bootstrap (WCR) in these simulations. Moreover, WR always performs very much better than the two RI procedures. Its only defect is that it underrejects moderately when $G_1 = 1$, as the theory of MacKinnon and Webb (2018a) predicts.

In Figure 14, we report results of the same experiments for t_Y and $t(G^*)$. These procedures perform much less well with heteroskedasticity and constant cluster sizes than they did in Figures 11 and 12 with homoskedasticity and variable cluster sizes. Note the nonlinear scale of the vertical axis. We do not report results for t_{IK} because it is extremely expensive to compute.[17] When G_1 is small, there are considerable differences between the performance of t_Y and $t(G^*)$. With $G_1 = 1$ and $\lambda = 2.0$, both procedures severely over-reject. With $G_1 = 1$ and $\lambda = 0.5$, both procedures severely under-reject. Interestingly, when $G_1 = 2$ and $\lambda = 1.25$, the $t(G^*)$ procedure rejects nearly 11% of the time, while the t_Y procedure rejects only 0.4% of the time. Neither of these procedures offers an improvement over RI-t for $G_1 \leq 2$.

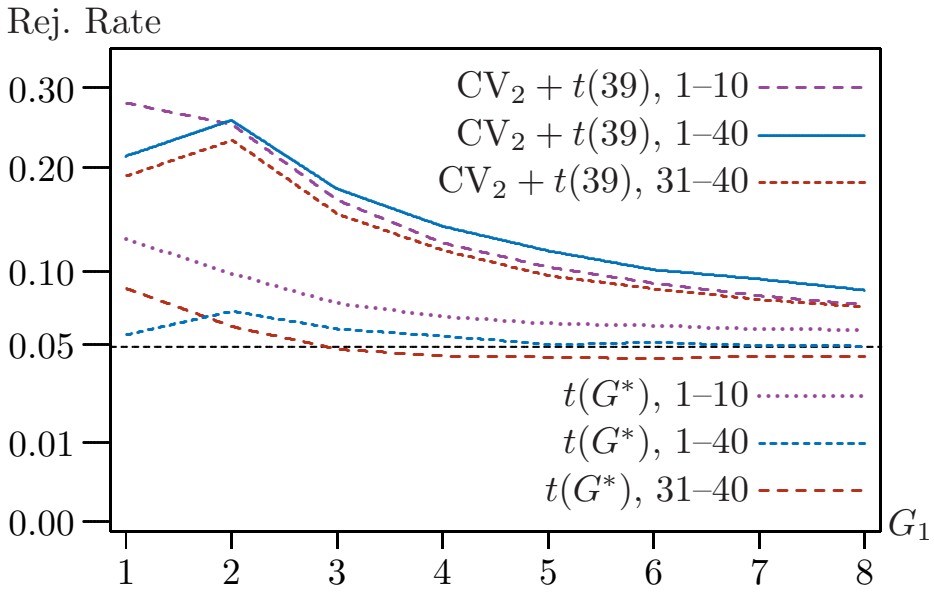
Appendix B: Simulation Results with Lognormal Errors

For all of our experiments up to this point, the error terms have been normally distributed. In this appendix, we report some additional results in which the error terms are instead lognormal, rescaled to have mean 0 and variance 1. These errors are strongly skewed to the right. Not surprisingly, this affects the performance of all the procedures.

Figure 15 shows rejection frequencies for RI-β and RI-t for the extreme cases in which either groups 1-10 or groups 31-40 are treated. RI-t performs more or less the same as it did in Figure 7, but RI-β performs noticeably worse that it did in Figure 2, at least for larger values of G_1. Of course, when all groups are treated, both procedures continue to work perfectly, and we do not show those results.

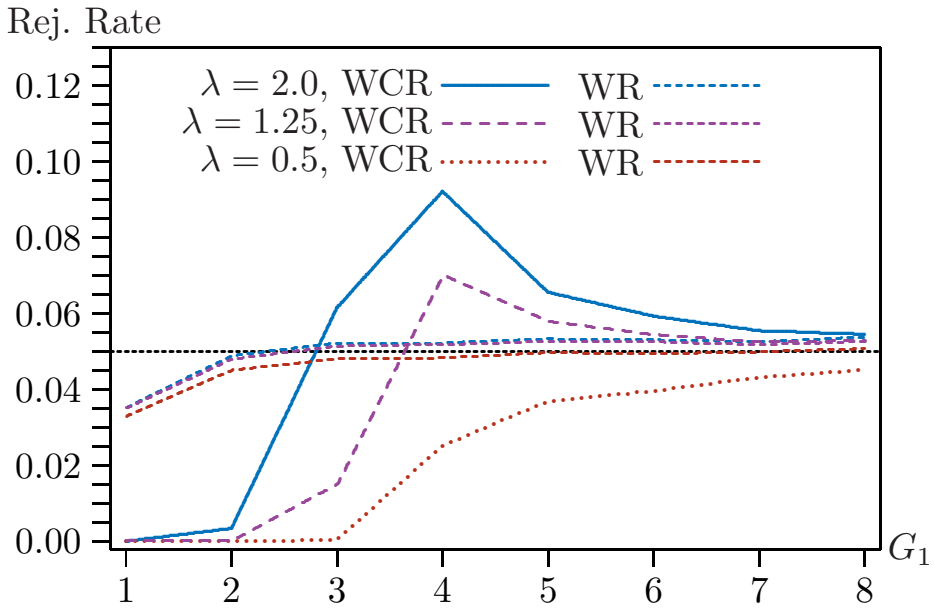
[17] For a different DGP that also involves heteroskedasticity, MacKinnon and Webb (2018a, Figure 13) reports results for both t_Y and t_{IK}, and they are quite similar.

Figure 12: Rejection Frequencies for Alternative Procedures



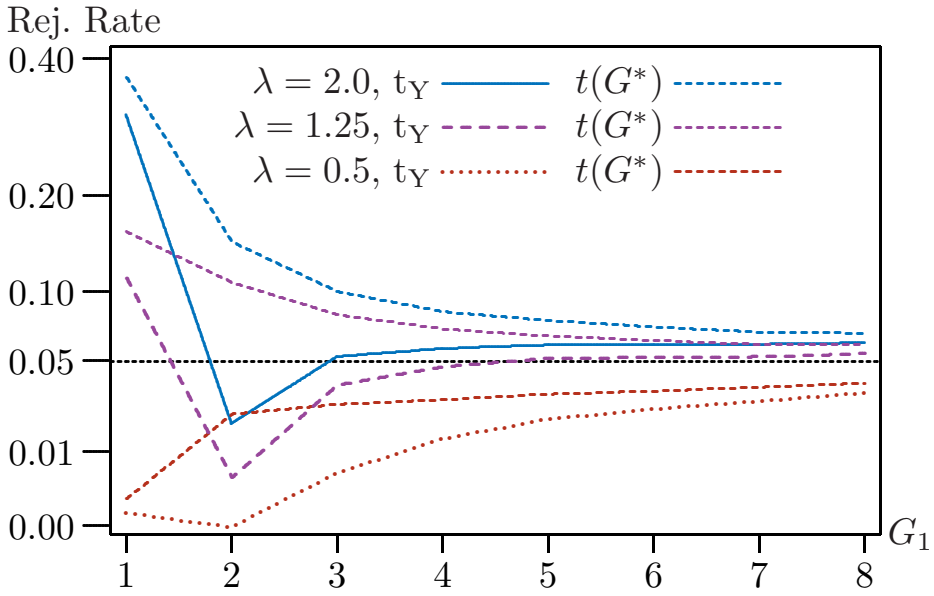
Notes: Based on 100,000 replications with $G = 40$, $\gamma = 0$, and $\rho = 0.05$

Figure 13: Rejection Frequencies for Wild Bootstrap Procedures with Heteroskedasticity



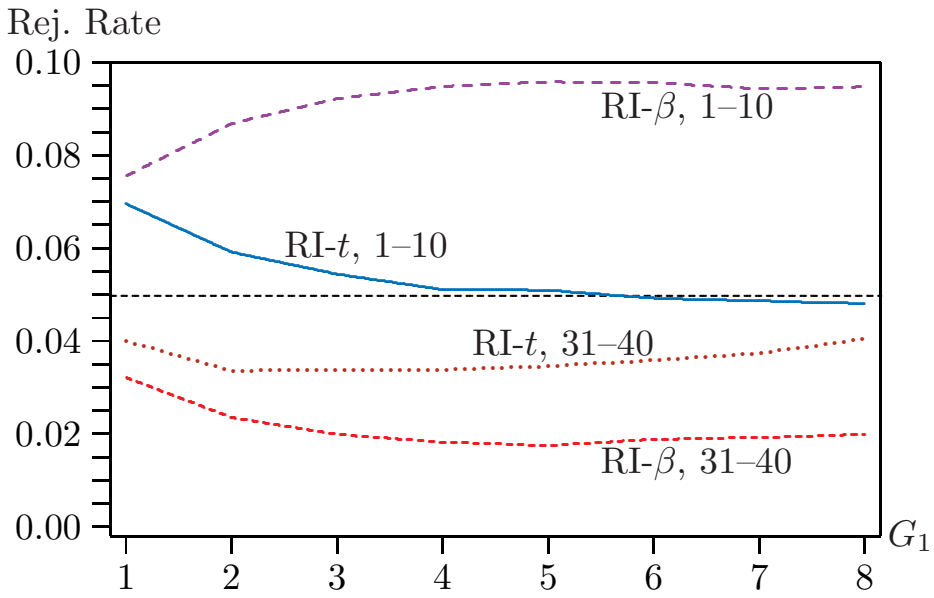
Notes: Based on 100,000 replications with $G = 40$, $\gamma = 0$, and $\rho = 0.05$

Figure 14: Rejection Frequencies for Alternative Procedures With Heteroskedasticity



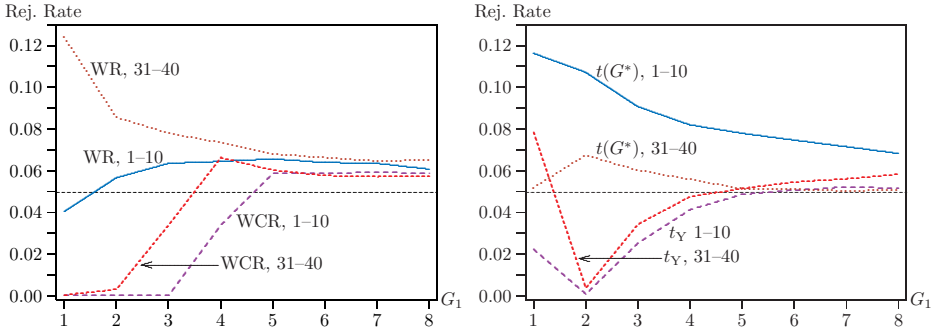
Notes: Based on 100,000 replications with $G = 40$, $\gamma = 0$, and $\rho = 0.05$

Figure 15: Rejection Frequencies for RI Procedures With Lognormal Errors



Notes: Based on 100,000 replications with $G = 40$, $\gamma = 2$, and $\rho = 0.05$

Figure 16: Rejection Frequencies for Alternative Procedures With Lognormal Errors



Notes: Based on 100,000 replications with $G = 40$, $\gamma = 2$, and $\rho = 0.05$

Figure 16 shows rejection frequencies for the two restricted wild bootstrap tests in the left panel and for t_Y and $t(G^*)$ in the right panel. These may be compared with the results in Figure 11. There are a number of differences between the two figures. Notably, WCR now rejects between about 5.7% and 5.9% of the time even for the largest values of G_1, and WR rejects noticeably more than that. However, the overall shapes of the rejection frequency curves as functions of G_1 are quite similar in the two figures.

All of the tests that we examine in this paper are two-tailed. If we had studied one-tailed tests, we would have found the effect of skewed error terms to be much greater. When the error terms are heavily right-skewed, upper-tail tests tend to reject much less often than symmetric tests, and lower-tail tests much more often.