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

1-2019

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

January 4, 2019

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 overreject, and different variants of the wild cluster bootstrap can either overreject or underreject dramatically. We study two randomization inference (RI) procedures. A procedure based on estimated coefficients may be unreliable when clusters are heterogeneous. A procedure based on t-statistics typically performs better (although by no means perfectly) under the null, but at the cost of some power loss. An empirical example demonstrates 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 the empirical example, to Azim Shaikh for a discussion that led is to find Romano (1990), 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 earlier versions.

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 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 overrejects in this case. MacKinnon and Webb (2017b, 2018) explain why this happens and why the wild cluster bootstrap of Cameron, Gelbach and Miller (2008), which yields reliable inferences in many other cases, does not solve the problem.

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 2019; Canay, Romano and Shaikh 2017). These are designed to handle various situations. We focus on procedures like the one proposed by Conley and Taber that use OLS estimates and are designed for samples with very few treated clusters, many control clusters, and clustering at the "state" level.

We are motivated by the many studies that use individual data, in which there is variation in treatment across both groups and time periods. This variation may arise either from randomization induced by the research design or from quasi-experimental policy changes. Such studies often use a classic "difference-in-differences" (or "DiD") regression that can be written as

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

Here i indexes individuals, g indexes groups, t indexes time periods, $\boldsymbol{T}_{igt}$ 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. Since there is a constant term, one group dummy and one time dummy must be omitted, and there may of course be other regressors as well.

The coefficient of interest in (1) is β, which shows the effect on treated groups in periods when there is treatment. The G groups are divided 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.

RI procedures necessarily rely on strong assumptions about the comparability of the control and treated groups. We show that, for procedures based on estimated coefficients, these assumptions 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, such procedures can overreject or underreject quite severely if the treated groups are substantially smaller or larger than the controls.

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.

Subsection 3.1 describes the coefficient-based approach to RI, and Subsection 3.2 discusses some of the underlying assumptions. Subsection 3.3 discusses the design of our Monte Carlo experiments, and Subsection 3.4 explores the performance of coefficient-based RI with and without cluster heterogeneity. Subsection 3.5 then proposes an alternative RI procedure based on cluster-robust t-statistics. Simulations suggest that the t-based RI procedure does not provide reliable inferences when only one group is treated and groups vary in size. When two or more groups are treated and the errors are homoskedastic, however, it tends to underreject, thus yielding conservative tests. Not surprisingly, there is a cost to basing inference on t-statistics. Subsection 3.6 shows that coefficient-based RI can have substantially more power than t-based RI, or than existing bootstrap procedures. Section 4 presents results for an empirical example based on Bailey (2010), and Section 5 concludes. There are also five online appendices.

2 Inference with Few Treated Clusters

A linear regression model with clustered errors may be written as

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

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. However, Conley and Taber (2011) and MacKinnon and Webb (2017b) show that t-statistics based on (3) overreject 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, which is commonly done (it is the default for Stata) and can be justified by results in 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 only regressor is a treatment dummy:

$$\boldsymbol{y} = \alpha\boldsymbol{\iota} + \beta\boldsymbol{d} + \boldsymbol{\epsilon}, \tag{4}$$

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. 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. 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.

The variance of $\hat{\beta}$, conditional on $\boldsymbol{d}$, is evidently

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

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

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

where $c \equiv G(N-1)/\big((G-1)(N-k)\big)$, $\boldsymbol{d}_g$ 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 $\operatorname{Var}(\hat{\beta})$ if the summation provides a good estimate of the quadratic form in (6). But 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 $\boldsymbol{d}_g = \mathbf{0}$ for the control clusters, the second and third terms in (8) must vanish for those clusters. Thus the summation in (7) becomes

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

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

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

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.[1] 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 tend to be 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.

[1] 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.

As we discuss in Appendix B, expression (3) is not the only available CRVE, and other procedures may well work better; see Bell and McCaffrey (2002), Imbens and Kolesár (2016), and Young (2016). 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).

Using the potential outcomes framework, let D_i represent the treatment assignment status of an individual, where $D_i = 1$ indicates treatment, and $D_i = 0$ indicates no treatment. Each individual then has two potential outcomes for Y_i, one when they receive treatment, $Y_i(D_i = 1)$, and one when they do not, $Y_i(D_i = 0)$. Of course, only one of these outcomes is observed. Imagine that we are interested in testing the sharp null hypothesis

$$H_0\colon\;\; \mathrm{E}\big(Y_i(D_i = 1) - Y_i(D_i = 0)\big) = 0 \quad \forall\, i.$$

Under this null hypothesis, the missing *potential* outcome is 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: $Y_i(D_i = 1) = Y_i(D_i = 0) \;\; \forall\, i$. We could then calculate a test statistic for our original sample as

$$\tau = \bar{Y}_i(D_i = 1) - \bar{Y}_i(D_i = 0), \tag{11}$$

where $\bar{Y}_i(D_i = 1)$ and $\bar{Y}_i(D_i = 0)$ 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. The outcomes in such a re-randomization have not actually changed, but we pretend that the treatment assignments were different. For any re-randomization r, the test statistic is

$$\tau^*_r = \bar{Y}_i(D^r_i = 1) - \bar{Y}_i(D^r_i = 0), \tag{12}$$

where D^r_i 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 for DiD models, to be discussed in Subsections 3.1 and 3.5 below, we will randomize at the group-period level rather than the individual level, because treatment affects all individuals for certain groups and certain time periods.

As above, we let G denote the number of groups and G_1 the number of treated groups. 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 groups 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 the null hypothesis 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 $p^* \le \alpha$. Of course, if the null hypothesis were that the treatment does not have a positive effect, we could instead use a one-tailed test. There is another way to compute p^*, discussed in Appendix A, in which the original sample is included in the set of $|\tau_r^*|$. This usually yields slightly more conservative inferences than (13), and it never leads to less conservative ones.

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 with few treated groups was first proposed in Conley and Taber (2011), which suggests two procedures called Γ and Γ^*. These procedures use $\hat{\beta}$ from the regression of interest as the test statistic. It is compared with an estimate of the distribution of the treatment parameter based on residuals from the control groups. 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.

Up to this point, we have not said much about the test statistic τ on which randomization inference is based. 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 very similar results. The principal difference between the RI-β and Γ^* procedures is that, instead of using residuals to estimate the distribution of the treatment parameter, the former explicitly uses OLS estimates of $\hat{\beta}$ from equation (1) based on re-randomized samples to obtain the τ_r^*.

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 re-randomization, re-estimate the model using the observations from all G groups, and calculate a new coefficient, β_r^*. 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.[2] When it exceeds an upper limit B with the property that $\alpha(B+1)$, picked on the basis of computational cost, choose the comparators randomly, without replacement, from the set of potential ones. 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 one discussed in Appendix A.

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, for which treatment begins in periods t_1^1, t_2^1,. . ., $t_{G_1}^1$, respectively.[3] Let the clusters chosen for treatment in each re-randomization be numbered 1^*, $2^*, \ldots, G_1^*$. 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 find 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.

3.2 Assumptions Underlying Randomization Inference

When considering whether to perform a hypothesis test using randomization inference, it is important to think about three things:

1. Whether there was random assignment.

[2] 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.

[3] 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.

2. Whether the groups are homogeneous.

3. Whether the investigator knows which groups were assigned to treatment.

Whether there was indeed random assignment is knowable for an actual experiment. It is perhaps a maintained assumption for a DiD analysis. When there is not random assignment, then it is not appropriate to use randomization inference.

RI procedures evidently depend on the strong assumption that τ and the τ_r^* follow the same distribution. Recall that, under the sharp null, there is no treatment effect for any individual. When treatment is randomly assigned at the individual level, there will seldom be any differences in inference based on the choice of test statistic. Specifically, using either the conventional test statistic $\tau = \bar{Y}_i(D_i = 1) - \bar{Y}_i(D_i = 0)$ or its regression analog $\hat{\beta}$, the invariance of the distributions under re-randomization follows naturally. However, if treatment is instead assigned at the group level, which is almost always the case for difference-in-differences, the choice of test statistic can matter.

When treatment is assigned at the individual level, the homogeneity assumption is rather weak. When it assigned at the group level, however, the homogeneity assumption is much stronger. Several types of heterogeneity can substantially affect the reliability of inference based on RI-β. The most readily observable type of heterogeneity is variation in cluster sizes. Since this is very likely to occur with individual data in a wide variety of contexts, we focus on it.[4] With heterogeneous cluster sizes, the coefficients for some clusters are estimated more efficiently than for others. This means that $\hat{\beta}$ and the β_r^* can follow different distributions.

Knowledge of which groups are treated, even with the maintained assumption of random assignment, is particularly important when the groups are heterogeneous. When all groups are homogeneous, the RI procedures yield a rejection frequency equal to α regardless of treatment assignment.[5] Things are more complicated when groups are heterogeneous, however. With heterogeneous groups, the expected rejection frequency is still equal to α before treatment is assigned. But once the outcome of the random assignment is realized, the rejection frequency can be either larger or smaller than α, even with random assignment. When assignment is random but the probabilities of treatment are not equal for all groups, the rejection frequencies will also differ from α when the groups are heterogeneous.

As an analogy, consider the following two-stage game. Stage 1 randomly determines, with equal probability, whether to play game A or game B in the second stage. Game A results in a loss of \$$X$ or a loss/gain of \$0 with equal probability. Game B results in a gain of \$$X$ or a loss/gain of \$0 with equal probability. Clearly, a risk-neutral person would be willing to pay \$0 to play this game before the start of stage 1, because at this point the game has an expected value of \$0. However, once stage 1 has concluded, the expected values are quite different.

Random assignment can be thought of as equivalent to stage 1 and the randomization inference procedure as equivalent to stage 2. Before treatment is assigned, the P values from the experiment (across all possible treatment assignments) are uniformly distributed between

[4] Another damaging type of heterogeneity is heteroskedasticity at the cluster level; see Appendix C.

[5] Here and elsewhere in this section, we are implicitly assuming that $\alpha(S + 1)$ is an integer. When S is small and that is not the case, the rejection frequency may noticeably exceed α if (13) is used to obtain p^*. Ways to avoid this are discussed in Appendix A.

0 and 1. Therefore, the *expected* rejection frequency is equal to α. However, the distribution of P values for any given treatment assignment (stage 2) can be quite different from uniform. Therefore, the expected rejection frequency conditional on treatment assignment can differ from α. Just as a gambler would be wise to revise their expectation of the value of the game after stage 1, a researcher would be wise to revise their expectation of rejecting the null hypothesis after treatment has been assigned. This distinction is often referred to as conditional versus unconditional inference. In this case, failing to condition on the *outcome* of the random assignment can lead to poor test size. A nice discussion of the issue of conditional inference in an RI setting can be found in Ferman and Pinto (2015).[6]

One could perhaps interpret the P value resulting from the RI-β procedure described in Subsection 3.1 as testing the joint null hypothesis of no treatment effect and random assignment. However, since treatment status is observed, it seems more natural to make conditional inferences about the effect of treatment. Even when treatment is randomly assigned, the RI-β procedure is potentially either oversized or undersized conditional on the clusters that were actually treated. In Subsection 3.4, we provide some evidence about just how serious these size distortions are likely to be.

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. Ferman and Pinto (2019) 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. In Appendix C, we study the performance of RI procedures when there is heteroskedasticity.

3.3 Design of the Monte Carlo Experiments

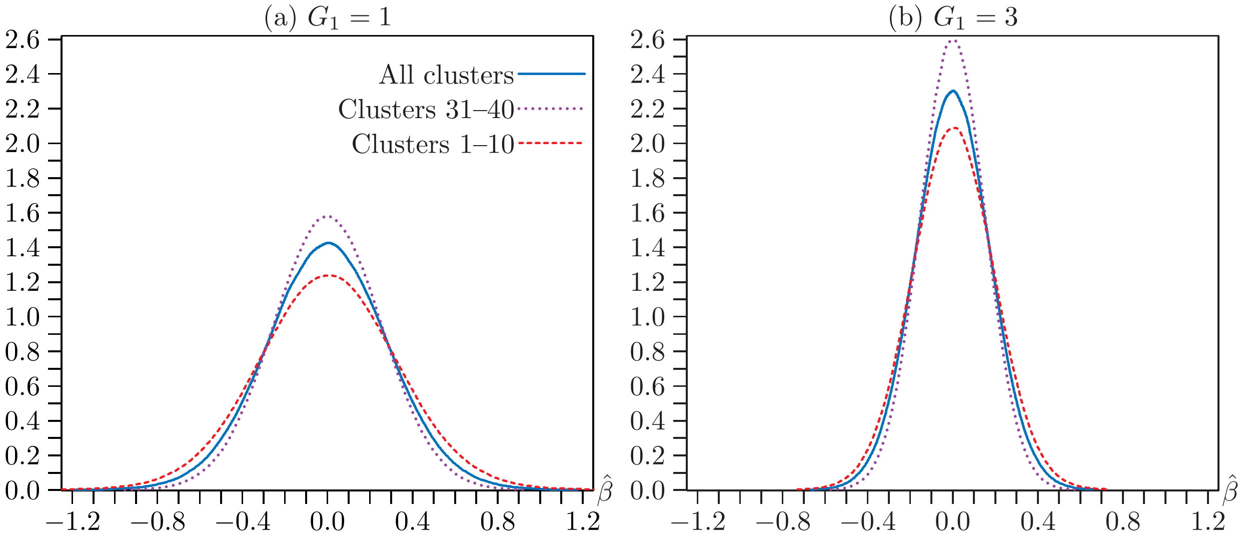
In the next subsection, several later ones, and Appendices B through D, we report results of a number of Monte Carlo experiments that study the performance of various inferential procedures, including ones not based on randomization inference, when the number of treated clusters is small and clusters are heterogeneous. In this subsection, before we report any results, we describe 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 $\beta = 0$.

[6] Note that this discussion is not in Ferman and Pinto (2019), a later version of the same paper.

[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.

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 $\rho = 0.05$

We assign N total observations unevenly among G clusters using the following formula:

$$N_g = \left[N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad g = 1, \ldots, G-1, \tag{14}$$

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 (14) implies that $N_g = N/G$ for all g. As γ increases, however, cluster sizes vary more and more. Many of our experiments have 4000 observations, 40 clusters, and $\gamma = 2$. In these experiments, the cluster sizes range from 32 to 246. For randomization inference procedures with $G = 40$, the number of randomizations is 39 for $G_1 = 1$, ${}_{40}\mathrm{C}_2 - 1 = 779$ for $G_1 = 2$, and 999 for $G_1 \geq 3$. For most experiments, the number of Monte Carlo replications is 100,000, and we report rejection frequencies at the 5% level.

3.4 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 and control clusters have different characteristics. 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 (14) 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).

Panel (a) of Figure 1 shows densities for $G_1 = 1$, and panel (b) shows densities for $G_1 = 3$. In both cases, the two conditional distributions differ from the unconditional one.

When small clusters are treated, 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. 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 the importance of whether we know which clusters were treated; see the discussion of point 3 in Subsection 3.2. Imagine conducting an experiment in which treatment was randomly assigned to a single cluster. This is the setting of panel (a) in Figure 1. Imagine also that you were given the values of $\hat{\beta}$ and the β_r^*, but did not know which cluster was actually treated. In this case, both $\hat{\beta}$ and all the β_r^* are drawn from the "all clusters" distribution in the figure. Therefore, 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.

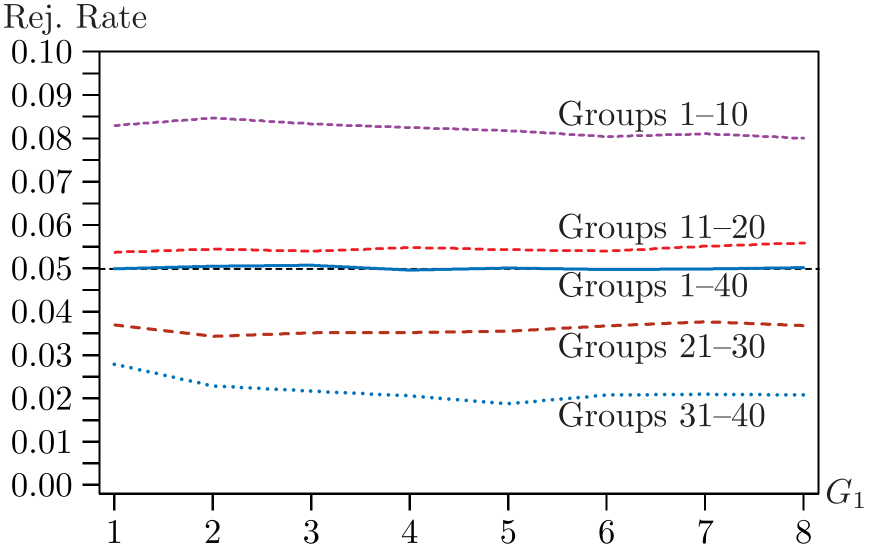
The thought experiment of the previous paragraph is not realistic. Even if an experiment is designed so that treatment is genuinely random, researchers always know the sizes, and usually the identities, of the clusters that are treated. With heterogeneous clusters, the expected rejection frequency conditional on the cluster that is actually treated will not be α. In the experiment considered in panel (a) of Figure 1, 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.

Both panels of Figure 1 strongly suggest that, conditional on the assignment of cluster(s) to treatment, the RI-β procedure will generally not yield a test that rejects $\alpha\%$ of the time when cluster sizes vary. 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 overreject when the treated clusters tend to be small and underreject 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. This is equivalent to assigning treatment randomly across all 40 clusters and then performing RI only for cases in which the treated clusters happen to be drawn from one of the four bins. In each case, the number of treated clusters, G_1, varies from 1 to 8. Of course, under random assignment, it would be increasingly unlikely for all treated clusters to fall into one bin as G_1 increases. The point of the experiments is to show that the problems of inference with heterogeneous clusters are due to the heterogeneity and are not limited to very small values of G_1.

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 overrejects noticeably conditional on the treated clusters being in the range of 1-10, and slightly for the range 11-20. In contrast, it underrejects moderately for the range 21-30,

Figure 2: Rejection Frequencies for RI-β Tests



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

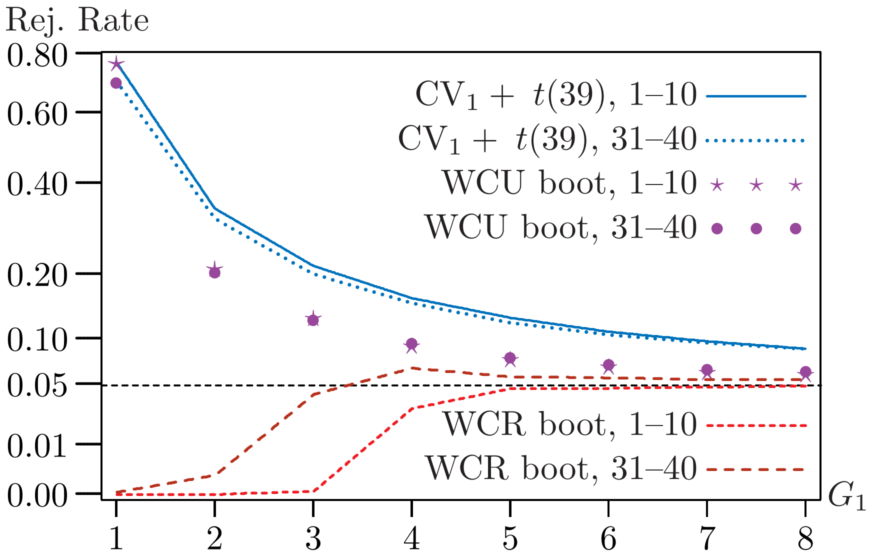
and quite noticeably for the range 31-40. Thus the smaller/larger the treated clusters are relative to the entire set, the more prone is the procedure to overreject/underreject.

One interesting feature of these experiments is that the performance of RI-β varies only a little with G_1. This contrasts sharply with asymptotic and bootstrap procedures for cluster-robust inference, which typically perform extremely badly when $G_1 = 1$ but then improve rapidly as G_1 increases, whether or not cluster sizes vary. Figure 3 presents a few results for the same experiments as Figure 2 to highlight the differences between the RI-β test, on the one hand, and existing tests, on the other. The vertical axis has been subjected to a nonlinear transformation in order to accommodate rejection frequencies that vary greatly.

The most commonly used procedure for testing whether $\beta_2 = 0$ is to compare a t-statistic based on the CRVE (3) with the $t(G-1)$ distribution. As the analysis in Subsection 2.1 implies, this procedure overrejects very severely when $G_1 = 1$, because the CRVE standard error for $\hat{\beta}$ is much too small. The overrejection becomes less severe as G_1 increases, but the test rejects at least 8.5% of the time even when $G_1 = 8$. In sharp contrast to RI-β, the rejection frequencies are a little larger when the smallest clusters are treated than when the largest clusters are treated, but the difference is always fairly small.

Figure 3 also shows rejection frequencies for two forms of wild cluster bootstrap test, which are discussed in Appendix B. One of these (WCR, where the bootstrap DGP is based on restricted estimates) was proposed in Cameron, Gelbach and Miller (2008). The other (WCU, where the bootstrap DGP is based on unrestricted estimates) is less widely used. Both WCR and WCU are shown to be asymptotically valid in Djogbenou, MacKinnon and Nielsen (2018). However, MacKinnon and Webb (2017b) shows theoretically that WCR will underreject very severely when G_1 is small and that WCU will overreject very severely. That is exactly what is observed in Figure 3. Cluster sizes have only a small effect on the rejection frequencies for WCU, but they have a large effect on WCR when G_1 is small.

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 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.

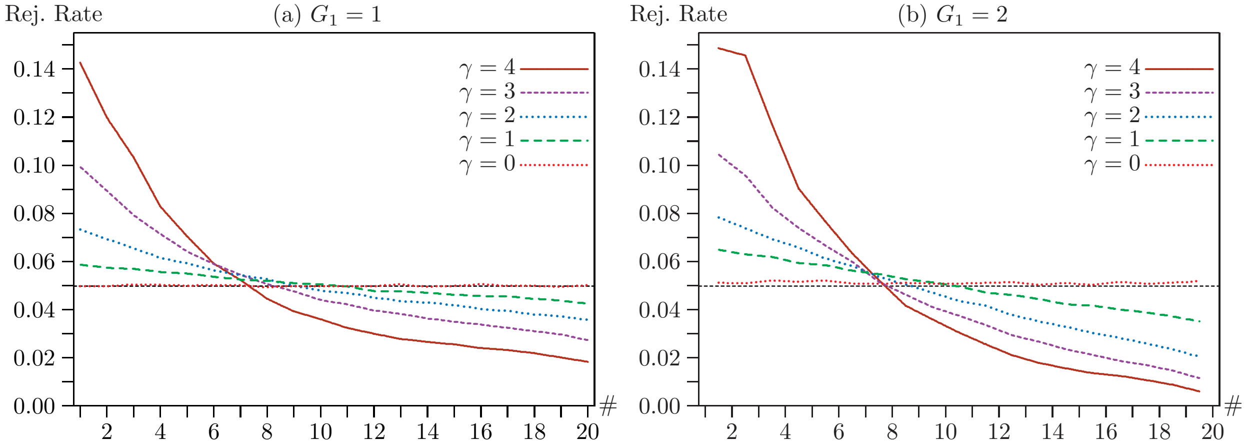
It is evident from Figures 2 and 3 that RI-β always works very much better than any of the other procedures when $G_1 \leq 2$. This is true even for the extreme cases in which the treated clusters are drawn from numbers 1-10 or 31-40 and RI-β does not work particularly well. For $G_1 \geq 4$, however, 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 DGP used in this section has normally distributed error terms. Additional results for a DGP with lognormally distributed error terms, which display a great deal of positive skewness, are presented in Appendix D. It will be seen that the distribution of the error terms matters, but none of the principal findings is overturned.

Two important features of the DGP are the distribution of cluster sizes and the positions of the treated cluster(s) within that distribution. The next set of experiments deals with these issues. It focuses on the cases of $G_1 = 1$ and $G_1 = 2$, where methods other than RI work badly. Because computation is relatively cheap in these cases, we use 400,000 replications. Figure 4 shows rejection frequencies for the RI-β procedure when $N = 5000$ and $G = 20$, *conditional* on the clusters that are actually treated. The five curves in each panel correspond to five values of γ from 0 to 4. When $\gamma = 4$, the variation in cluster sizes is quite extreme: The smallest cluster has 20 observations, and the largest has 935. Smaller values of γ are probably more realistic.

In panel (a), just one cluster is treated, and the horizontal axis shows its rank, ordered from smallest to largest. Since $S = 19$, the condition that $.05(S+1)$ be an integer is satisfied.

Figure 4: Rejection Frequencies for RI-β tests when $G_1 = 1$ and $G_1 = 2$



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

In panel (b), pairs of adjacent clusters are treated. There are 19 such pairs (1 & 2, 2 & 3, 3 & 4, and so on). The horizontal axis now shows the average rank of each pair (1.5, 2.5, and so on). Since $G = 20$ and $G_1 = 2$, the value of S for these experiments is 189, so that $.05(S+1)$ is not an integer. We therefore report the average of two rejection frequencies, one based on equation (13) and one based on equation (A.3). The former is always somewhat larger than the latter; see Appendix A.

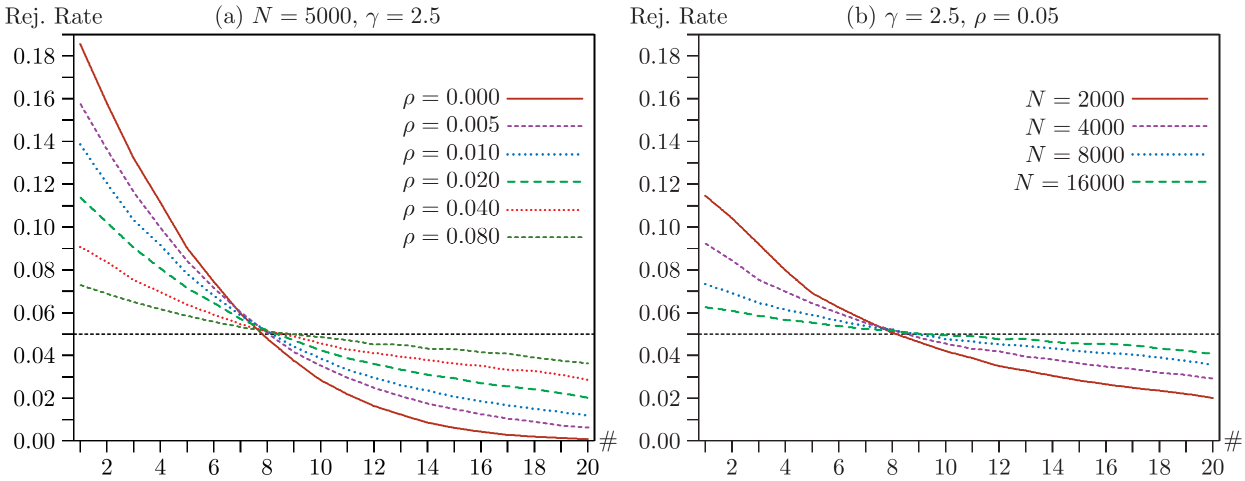
When $\gamma = 0$, the RI-β tests work perfectly, except for simulation error.[8] This must be the case, because the clusters are homogeneous when $\gamma = 0$, so that the distributions of $\hat{\beta}$ and the β_r^* must be the same. When $\gamma > 0$, however, the rejection frequencies depend on the treated cluster(s). As expected, the tests overreject for small treated clusters and underreject for large ones. Both overrejection and underrejection become more severe as γ increases. Interesting, the results are actually somewhat worse for $G_1 = 2$ than for $G_1 = 1$.

Figure 5 performs two similar exercises with $G_1 = 1$ and $\gamma = 2.5$. The ratio of the largest to the smallest cluster sizes is nearly 11. As in Figure 4, the horizontal axis shows the rank of the treated cluster. In panel (a), N is fixed at 5000, and ρ varies across the curves. In this case, the problem of poor test size becomes less severe as ρ becomes larger. In panel (b), $\rho = 0.05$, and N varies across the curves. In this case, the problem of poor test size becomes less severe as N increases. Both of these results reflect the facts that, for the model we are using, the standard deviation of $\hat{\beta}$ varies less across the rank of the treated cluster as either ρ or N increases. It is illuminating to see why this is the case.

For simplicity, we study the simplified DiD model (4) with $G_1 = 1$. We assume that N is proportional to G, that N_g/N is $O(1/G)$ for all g, and that N_T, the number of treated observations, is proportional to N_1. Thus N_T is either fixed, if cluster sizes are constant,

[8] Sharp-eyed readers may notice that this is not quite true in panel (b). That is because $.05 \times 190$ is not an integer. The test based on equation (13) overrejects, and the one based on equation (A.3) underrejects. The average of the two rejection frequencies, which we report, is apparently biased upwards very slightly. The two RI tests yield identical rejection frequencies at the .10 level, because $.10 \times 190 = 19$, and these appear to differ from .10 only because of simulation error when $\gamma = 0$.

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



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

or growing more slowly than N at the same rate as all the clusters, if cluster sizes are increasing. Under the null hypothesis, the parameter estimate $\hat{\beta}$ is given in equation (5), which, when $G_1 = 1$, can be rewritten as

$$\hat{\beta} = \frac{1}{N\bar{d}(1-\bar{d})}\left(\boldsymbol{d}_1'\boldsymbol{\epsilon}_1 - \bar{d}\sum_{g=1}^{G}\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g\right). \tag{15}$$

The scalar $\boldsymbol{d}_1'\boldsymbol{\epsilon}_1$ is the sum of N_T random variables ϵ_{1i}, each with mean 0. If they were uncorrelated, the variance of this sum would be $O_p(N_\mathrm{T})$. When they are correlated, however, the sum involves N_T variances and $N_\mathrm{T}(N_\mathrm{T}-1)/2$ covariances. Thus there are two terms, one that is $O_p(N_\mathrm{T})$ and one that is $O_p(N_\mathrm{T}^2)$. If, as is typically the case, the correlations are fairly small, the first term will be larger than the second when N_T is small. However, because the second term is of higher order, it will eventually become the dominant one as N_T becomes large; see Djogbenou, MacKinnon and Nielsen (2018). Thus, for finite N, the standard deviation of $\boldsymbol{d}_1'\boldsymbol{\epsilon}_1$ is bounded by $O_p(N_\mathrm{T}^{1/2})$ and $O_p(N_\mathrm{T})$.

The second term inside the large parentheses in (15) is $\bar{d} = N_\mathrm{T}/N$ times a summation of G uncorrelated quantities. By exactly the same argument as for the first term, the variance of each of these quantities is between $O_p(N/G)$ and $O_p\big((N/G)^2\big)$. To find the standard deviation of the second term, we multiply these variances by G, take the square root, and then multiply by N_T/N. When there is no intra-cluster correlation, the standard deviation of this summation is $O_p(N_\mathrm{T}N^{-1/2})$, and, when there is, it is $O_p(N_\mathrm{T}G^{-1/2})$. In both cases, the leading-order term within the large parentheses in (15) is the first one, because, for N and G large and N_T small relative to N, $N_\mathrm{T}^{1/2} > N_\mathrm{T}N^{-1/2}$, and $N_\mathrm{T} > N_\mathrm{T}G^{-1/2}$.

Since $1-\bar{d} = O(1)$, the leading factor in (15) must be $O(N_\mathrm{T}^{-1})$. Multiplying the first term by this leading factor, and recalling that N_T is proportional to N_1, we find that the standard deviation of $\hat{\beta}$ must be between $O_p(N_1^{-1/2})$, when there is no intra-cluster correlation, and $O_p(1)$, when the treated cluster is large and there is a lot of intra-cluster correlation.

This result helps to explain both panels of Figure 5. For the simulations reported in that figure, there is correlation among all the elements within each $\boldsymbol{\epsilon}_g$ vector. Nevertheless, when ρ and N/G are both small, the standard deviation of $\hat{\beta}$ is fairly well approximated by $O_p(N_1^{-1/2})$. As either or both of them become larger, however, this standard deviation becomes closer to being $O_p(1)$. Thus we expect to see the standard deviations depend more on the size of the treated cluster when ρ and N/G are both small than when either or both of them is large. In consequence, as the figure shows, the rejection frequencies for RI-β tests become less variable across the rank of the treated cluster as either ρ or N/G increases.

The argument above can be extended to the more general case in which $G_1 > 1$. The expression for $\hat{\beta}$ is almost the same as the one in equation (15), except that $\boldsymbol{d}_1'\boldsymbol{\epsilon}_1$ is replaced by $\sum_{g=1}^{G_1} \boldsymbol{d}_g'\boldsymbol{\epsilon}_g$. Provided $\bar{d}$ is small, which requires that not too many observations are treated, the leading-order term is still the first one. Its variance is the sum of G_1 variances, which will depend on the numbers of treated observations in each of the G_1 clusters in the same way that the variance of $\boldsymbol{d}_1'\boldsymbol{\epsilon}_1$ depends on N_{T} when $G_1 = 1$.

The key message from this analysis, and 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.

3.5 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, including conventional t-statistics, as Imbens and Rubin (2015, Chapter 5) points out. A natural 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 t_β, which equals $\hat{\beta}$ divided by the square root of the appropriate diagonal element of the CRVE in Eq. (3), to the empirical distribution of the corresponding t_r^*. This is similar to one of the procedures studied in Young (2019). We will refer to this procedure as "cluster-robust t-statistic randomization inference," or RI-t for short.

When testing equality of means in a two-sample problem (sometimes called the Behrens-Fisher problem), randomization inference based on coefficients does not yield tests with the correct size unless either the sample sizes are the same or the variances of the two populations are the same (Lehmann and Romano 2008, pp. 642–643). However, randomization inference based on (ordinary) t-statistics does yield asymptotically valid tests even when neither of these conditions holds (Romano 1990). Since testing for $\beta = 0$ in Eq. (1) can be thought of as a generalization of the problem of testing the equality of two means, it seems plausible that randomization inference should perform better under the null hypothesis when it is based on cluster-robust t-statistics than when it is based on coefficients.

Djogbenou, MacKinnon and Nielsen (2018) proves formally that t_β is asymptotically distributed as $\text{N}(0, 1)$ under the null hypothesis. The proof requires regularity conditions that allow cluster sizes to vary substantially, but not too much, as $G \to \infty$. Since the asymptotic distribution of the t_r^* must likewise be $\text{N}(0, 1)$, this implies that RI-t yields asymptotically valid tests under these conditions. This argument does not imply that there is any sort of asymptotic refinement, however. Indeed, there is no reason to believe that inferences based on RI-t improve any faster than inferences based on other procedures.

The key regularity conditions in Djogbenou, MacKinnon and Nielsen (2018) concern what happens as the sample becomes large. For the DiD model, these conditions imply that G and G_1 must both tend to infinity together. Thus the asymptotic theory does not apply when G_1 is small and fixed. In fact, as we saw in Subsection 2.1, the CRVE (7) estimates the variance of $\hat{\beta}$ very poorly when G_1 is small, because the second and third terms in the first summation in expression (9) vanish. This suggests that the distribution of t_β may be far from standard normal in that case. However, the performance of the variance estimator (7) apparently improves quite quickly as G_1 increases; see MacKinnon and Webb (2017b). Thus we would expect the distributions of t_β and the t^*_r to become closer as G_1 increases for given G, at least up to a point, depending on the sizes of the treated and control clusters.

In Subsection 3.4, we studied the relationship between $\hat{\beta}$ and the number of treated observations, N_T, when $G_1 = 1$. This analysis can easily be extended to t_β. From (7) and (9), the denominator of t_β is simply

$$\frac{1}{N\bar{d}(1-\bar{d})}\left(\sum_{g=1}^{G}(\boldsymbol{\iota}'_g\hat{\boldsymbol{\epsilon}}_g)^2\right)^{1/2}. \tag{16}$$

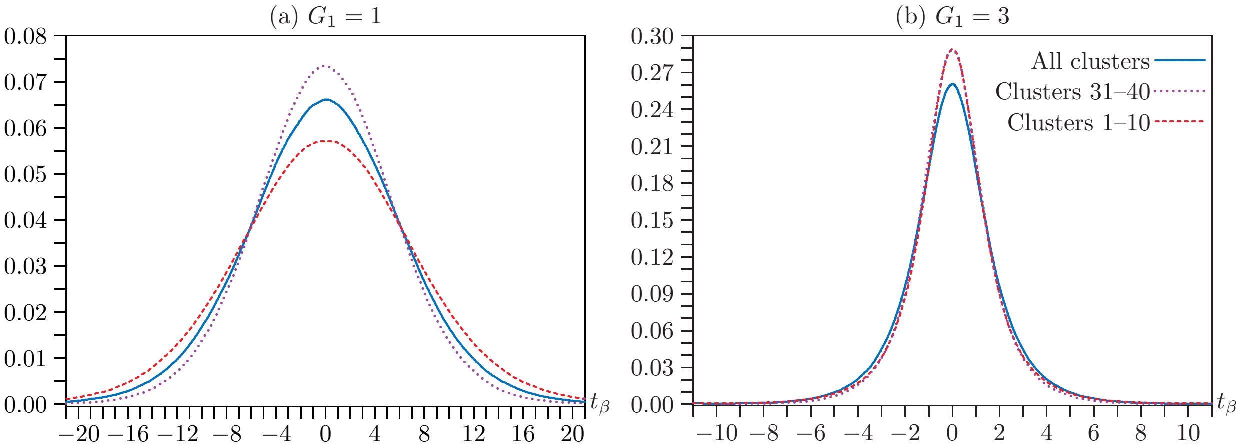
The first factor here is relatively large when $\bar{d}$ is small and decreases with $\bar{d}$ up to $\bar{d} = 1/2$. In contrast, the second factor in (16) can be expected to vary little across re-randomizations. The residual vectors $\hat{\boldsymbol{\epsilon}}_g$ change when the treated cluster does, but only because the parameter estimates change. The underlying error vectors $\boldsymbol{\epsilon}_g$ are identical across re-randomizations. The residuals for the treated observations may differ sharply when different clusters are treated, but the ones for other observations change only because the estimated constant term changes. With $\bar{d}$ small, there are far more untreated residuals than treated residuals.

We saw in Subsection 3.4 that the standard deviation of $\hat{\beta}$ decreases as N_T increases, albeit at a rate that is difficult to pin down. We have just seen that expression (16) tends to decrease with N_T. By itself, this would cause the standard deviation of t_β to increase with N_T. Thus there are two opposing forces. Which of them is dominant will depend on the values of $\bar{d} = N_\text{T}/N$ and on the rate at which $\hat{\beta}$ is decreasing, which, as we have seen, depends on the cluster sizes and intra-cluster correlations. It is quite possible that the standard deviation of t_β may decrease with N_T over some range of cluster sizes and increase with it over another range. We conclude that the standard deviations of the t^*_r may vary across re-randomizations quite differently from the way in which those of the β^*_r vary.

Figure 6 plots conditional and unconditional distributions of t_β using results from the same simulations as Figure 1. Results for $G_1 = 1$ are again shown in panel (a) and results for $G_1 = 3$ in panel (b). When $G_1 = 1$, the two conditional distributions are once again quite different from the unconditional distribution. Indeed, apart from scale, panel (a) of Figure 1 and panel (a) of Figure 6 look remarkably similar. This suggests that we will encounter the same inference problems as before, with RI-t overrejecting when small clusters are treated and underrejecting when large clusters are treated.

In contrast to Figure 1, however, panel (b) of Figure 6, in which $G_1 = 3$, does not look much like panel (a). There are still some differences between the unconditional distribution and the conditional ones, but they are much less evident than they were in panel (b) of Figure 1. Moreover, and this is somewhat surprising, both of the conditional distributions are

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 $\rho = 0.05$

less spread out than the unconditional one. This suggests that RI-t may tend to underreject for $G_1 > 1$ even when the treated clusters are relatively large.

To compare the performance of RI-t and RI-β, we perform two sets of experiments. First, the experiments of Figure 2 are repeated for RI-t in Figure 7. As must be the case, RI-t works perfectly (except for experimental error) when all clusters are potentially treated. 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 conditional on either the smallest or the largest clusters being treated, but there is not much to choose between the two procedures when intermediate clusters (numbers 11–20 or 21–30) are treated.

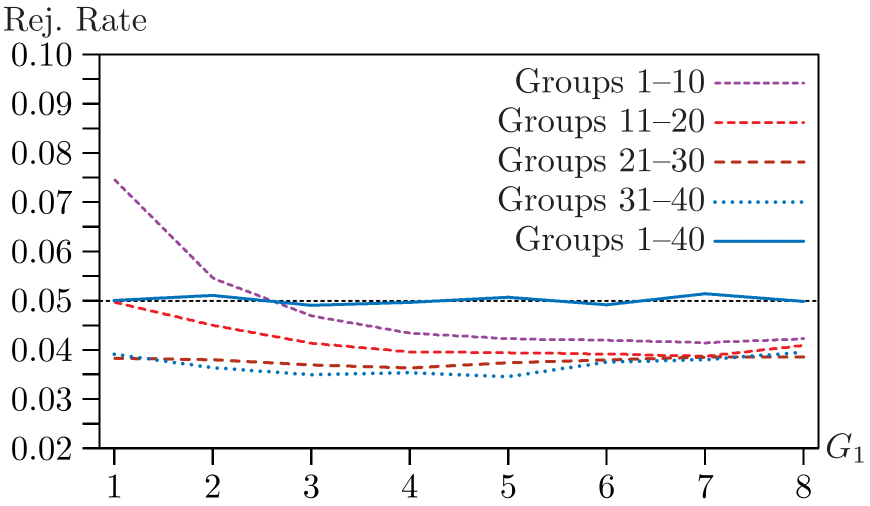
Figure 4 reported rejection frequencies for RI-β when either one or two clusters are treated. Figure 8 reports them for RI-t for the same experiments. When $G_1 = 1$, the RI-t procedure works perfectly when all clusters are the same size ($\gamma = 0$). When cluster sizes differ, however, the rejection frequencies in panel (a) are U-shaped and look very different from the ones for RI-β in panel (a) of Figure 4. Depending on the value of γ, the RI-t test tends to overreject when the treated cluster is very small or very large, and to underreject for some intermediate values. This is consistent with the analysis following expression (16).

In panel (b) of Figure 8, we report rejection frequencies for RI-t when pairs of adjacent clusters are treated. Somewhat surprisingly, RI-t always underrejects when $\gamma > 0$, although only to a limited extent when the very smallest clusters are treated. Even with just two treated clusters, RI-t generally works much better than RI-β; compare panel (b) of Figure 4. Its principal defect is that it can underreject quite noticeably when γ is large and the treated clusters are not among the smallest ones.

3.6 Power of Alternative Procedures

One possible drawback of RI-t, and of every other procedure based on t-statistics, is that the denominator of the t-statistic adds noise, and noise inevitably reduces power. Because

Figure 7: Rejection Frequencies for RI-t Tests



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

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 9. As expected, the power of all procedures increases with the number of clusters, and the differences among 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. In cases where RI-β overrejects under the null, it will appear to have an even greater power advantage than it does in Figure 9. However, even when RI-β underrejects 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.

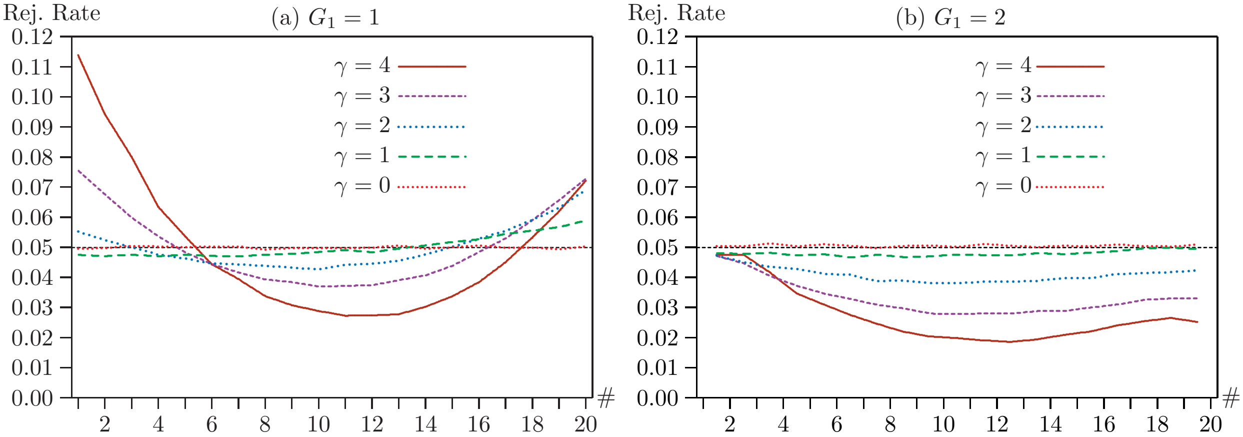
The WCR bootstrap underrejects quite severely when $G = 20$ and $G_1 = 3$, so it is not surprising that it has substantially less power than RI-t in that case. However, it performs very well under the null in the other two cases, and it still has slightly less power than RI-t. This suggests that there may be an advantage to using RI-t rather than WCR even when G_1 is not particularly small.

4 Empirical Example

In this section, we consider an empirical example for which $G_1 = 2$, 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.

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

Figure 8: Rejection Frequencies for RI-t Tests when $G_1 = 1$ and $G_1 = 2$



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

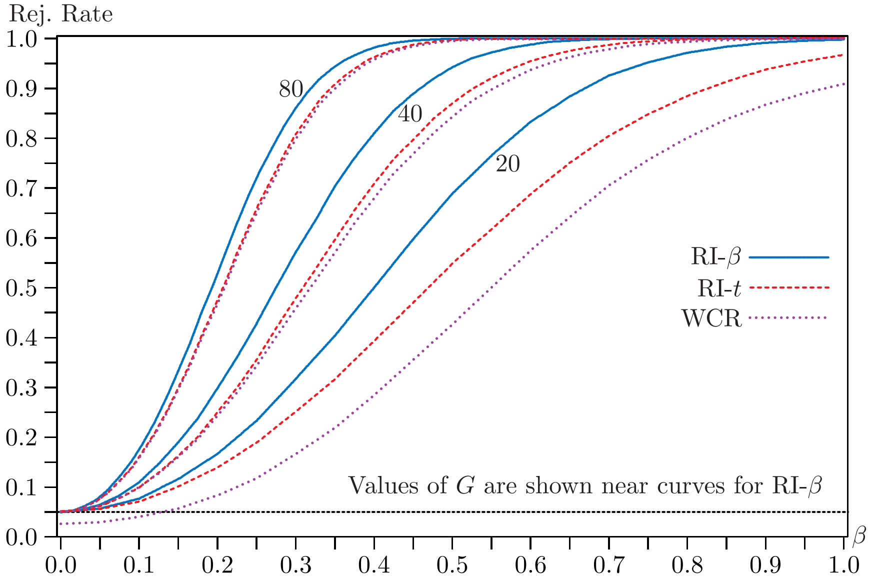
and sale of the birth control pill. The practice of using these laws to restrict the sale of birth control pills was essentially ended by the U.S. Supreme Court's 1965 Griswold v. Connecticut decision. Part of the analysis in Bailey (2010) shows that women in states with sales restrictions on the birth 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. Although we attempted to use the same sample as Bailey, our sample has 6929 observations, and hers 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.

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

Figure 9: 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 $B = 999$

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.

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 (A.3), because they are slightly more conservative than ones based on equation (13). For `rep61`×`D1965`, 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 Appendix B are also reported. The ones for the procedure of Young (2016) and the ones based on the effective degrees of freedom proposed in Carter et al. (2017) give broadly similar results. Moreover, these P values are quite similar to the ordinary wild bootstrap (WR) P values and to the RI P values (where they exist). Specifically, all of the P values for `rep61`×`D1965` are below 0.10 (except for WCR), while

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.

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.

Although the results in Table 1 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, the RI P values are reasonably consistent with those from the Young and CSS procedures, and from the ordinary wild bootstrap.

In Appendix E, we present results for another empirical example taken from Conley and Taber (2011). In this case, the RI-β and RI-t procedures yield different conclusions, with the latter providing somewhat stronger evidence against the null hypothesis.

5 Conclusion

We study two methods based on randomization inference (RI) for difference-in-differences estimation with few treated clusters, and we compare them with other methods. With random assignment, *unconditional* inference based on any form of RI would always be valid. This is true even with heterogeneous clusters. In practice, however, empirical economists observe their samples after treatment has been assigned and outcomes realized. Because the characteristics of the treated and control clusters often differ, they generally wish to make inferences *conditional* on the clusters that were actually treated. In particular, the numbers of observations may differ between treated and control clusters. As we have seen, this causes the distributions of coefficient estimates and, to a lesser extent, of cluster-robust t-statistics to depend on which clusters are treated. In consequence, with heterogeneous clusters, randomization inference conditional on the sample may not be valid.

There are five main findings. Some of these were obtained theoretically for a simple model in Subsections 3.4 and 3.5. Others were obtained by Monte Carlo simulation methods in Subsections 3.4, 3.5, and 3.6 and in Appendices B, C, and D.

The first result is that none of the procedures we study works well when there is just one treated cluster and it is atypical in terms of either the number of observations or the variance of the error terms. In particular, the procedure based on coefficient estimates, RI-β, can overreject severely when the treated cluster is unusually small and underreject severely when it is unusually large. This can also happen when there are several treated clusters and they are all atypical in the same way.

The second result is that both RI procedures actually work quite well when the clusters are approximately homogeneous, even when G_1 is extremely small. They tend to work far better than the wild cluster bootstrap when $G_1 \leq 2$.

The third result is that, as the number of treated clusters G_1 increases, holding the total number G constant, the performance of the procedure based on cluster-robust t-statistics, RI-t, initially improves quite rapidly. In contrast, the performance of RI-β may or may not improve as G_1 increases. RI-t asymptotically yields valid inferences under suitable regularity conditions when G_1 and G increase together, but RI-β does not. However, G may have to be quite large for RI-t to perform really well.

The fourth result is that both the sample size and the extent of intra-cluster correlation matter when clusters are heterogeneous. The theory in Subsection 3.4 and the simulation results in Figure 5 both suggest that the performance of RI-β for atypical treated clusters improves as the number of observations per cluster increases and as the extent of intra-cluster correlation increases. These will also affect the performance of RI-t, but the relationship is more complicated; see Subsection 3.5.

The final result is that RI-β tends to have substantially more power than RI-t or other procedures based on cluster-robust standard errors. This is predictable, but the extent of the power gain may be surprisingly large.

The performance of all the procedures we study depends in a complicated way on the numbers and sizes of the treated and control clusters, the cluster-level covariance matrices of the error terms, and the numbers 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 for their dataset, in addition to the actual exogenous variables, if any, and plausible values of the intra-cluster correlation coefficient ρ.

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* 7(1), 1–16

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 (2018) 'Asymptotic theory and wild bootstrap inference with clustered errors.' Working Paper 1399, 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

Ferman, Bruno, and Christine Pinto (2019) 'Inference in differences-in-differences with few treated groups and heteroskedasticity.' *Review of Economics and Statistics* 101, to appear

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 and Statistics* 98(1), 83–96

Imbens, Guido W., and Donald B. Rubin (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (New York: 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* (New York: Springer)

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 (2018) 'The wild bootstrap for few (treated) clusters.' *Econometrics Journal* 21, 114–135

MacKinnon, James G., and Matthew D. Webb (2019) 'Wild bootstrap randomization inference for few treated clusters.' In *The Econometrics of Complex Survey Data: Theory and Applications,* ed. Kim P. Huynh, David Tomás Jacho-Chávez, and Gautam Tripathi, vol. 39 of *Advances in Econometrics* (Emerald Group) chapter 3, pp. 61–85

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

Romano, Joseph P. (1990) 'On the behavior of randomization tests without a group invariance assumption.' *Journal of the American Statistical Association* 85(411), 686–692

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

Young, Alwyn (2019) 'Channelling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results.' *Quarterly Journal of Economics* 134, to appear

Appendix A: Alternative Approaches to Randomization Inference

Calculating a P value as in (13) is not the only way to perform a test based on a statistic $|\tau|$ and S re-randomizations $|\tau_r^*|$. In fact, an alternative approach is more common in the theoretical literature. In this approach, the actual sample is included in the set of re-randomizations. Thus the total number of test statistics is $R \equiv S+1 = {}_G\mathrm{C}_{G_1}$. When these are sorted from smallest to largest, they may be denoted $|\tau_r^{(j)}|, j = 1, \ldots, R$.

Now define c as $R-[\alpha R]$, where $[\cdot]$ denotes the largest integer no larger than its argument. Then $|\tau_r^{(c)}|$, which is element number c of the sorted list, can be thought of as a critical value. The RI test is then defined as

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

where $\boldsymbol{Y}$ denotes the sample, $\phi(\boldsymbol{Y}) = 1$ denotes rejection, and $\phi(\boldsymbol{Y}) = 0$ denotes non-rejection. The expectation of the RI test $\phi(\boldsymbol{Y})$ defined in (A.1) 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 (A.1), 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 R. 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 $1/R$, it can safely be ignored when R is large.[1]

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

$$\phi(\boldsymbol{Y}) = \begin{cases} 0 \text{ if } |\tau| < |\tau_r^{(c)}| \\ 1 \text{ if } |\tau| \geq |\tau_r^{(c)}|. \end{cases} \tag{A.2}$$

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 (A.2) suggests another way to compute the P value:

$$p^{*\prime} = \frac{1}{R}\left(1 + \sum_{r=1}^{R} \mathbb{I}\big(|\tau_r^*| > |\tau|\big)\right). \tag{A.3}$$

[1] In writing (A.1), 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.

When $(1-\alpha)R$ is an integer, rejecting whenever $p^{*\prime} \leq \alpha$ must yield exactly the same outcome as rejecting whenever $p^* < \alpha$. However, when $(1-\alpha)R$ is not an integer, the two tests will yield different results. The one based on p^* will overreject, and the one based on $p^{*\prime}$ will underreject. If rejection frequencies are plotted as a function of S (or R) 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 variant of randomization inference that makes use of additional bootstrap samples was proposed in MacKinnon and Webb (2019) in part to increase S in settings where ${}_G\mathrm{C}_{G_1}$ is small.

Appendix B. 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, and then present some simulation evidence.

B.1 Bootstrap Methods

The wild cluster bootstrap was proposed in Cameron, Gelbach and Miller (2008) and shown to be asymptotically valid in Djogbenou, MacKinnon and Nielsen (2018). 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. The v_g^* are usually drawn 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 underrejection. In contrast, the WCU bootstrap, which does not impose the null hypothesis, leads to severe overrejection. For both cases, see Figure 3. When just one cluster is treated, WCU overrejects 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 (2018) suggested using the ordinary wild bootstrap together with cluster-robust standard errors, and Djogbenou, MacKinnon and Nielsen (2018) 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 and to the number of treated observations per cluster.

A very different bootstrap procedure, usually called the pairs cluster bootstrap, was suggested in 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 overrejects extremely severely, about the same as WCU, but it can perform quite well when neither G nor G_1 is too small.

B.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 overreject or underreject. We consider the performance of what we call the $t(G^*)$ procedure in some of the simulation experiments in Subsection B.4.

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 computationally demanding. Simply computing CV_2 can be much more costly than using either randomization inference or the wild cluster bootstrap.

As we illustrate in Subsection B.4, using CV_2 rather than CV_1 leads to substantially less overrejection 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.[2] 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 is 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 (2018), we find that Young's procedure tends

[2] 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 (2018) provides evidence on how the cost of this procedure, and others, varies with the sample size.

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 Subsection B.4.

B.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).

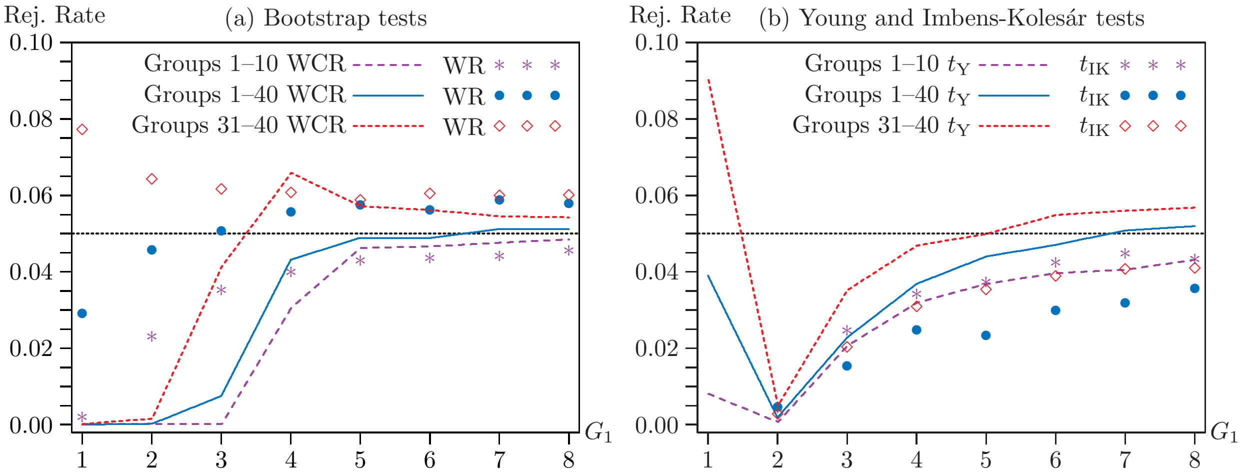
A very different procedure is proposed in Abadie, Diamond and Hainmueller (2010). Like the RI procedures, it bases inference on an empirical distribution generated by perturbing the assignment of treatment. However, the procedure differs substantially from the ones considered in this paper, because it 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 (2019) 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 the parameters of an autoregressive process.

B.4 Simulation Results for Additional Methods

In this section, we present simulation results for several of the procedures discussed above. Figure B.1 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

Figure B.1: Rejection Frequencies for Alternative Procedures



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

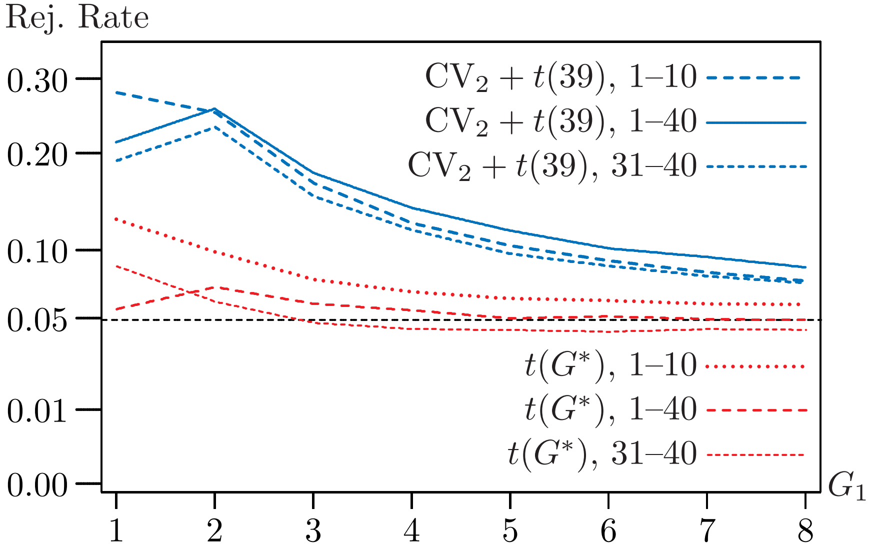
case in which all groups are treated and for the two extreme cases in which we condition on groups 1-10 (the smallest ones) and groups 31-40 (the largest ones) being treated.

The left panel of Figure B.1 deals with the restricted wild cluster bootstrap (WCR) and the ordinary restricted wild bootstrap (WR). It is evident that WCR almost never rejects when $G_1 \leq 2$ and underrejects 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. Because this dependence is very much less for the ordinary wild bootstrap, WR works considerably better than WCR for $G_1 \leq 4$, except when $G_1 = 1$ and the smallest clusters are treated. However, its performance is far from perfect, and for $G_1 \geq 5$, 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 (2018).

The right panel of Figure B.1 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 B.2. The former procedure is inexpensive to compute, but the latter is extremely expensive. Results for it (which are not available for $G_1 = 1$ because it cannot be computed) 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 B.1 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 $G_1 \leq 2$.

Figure B.2: Rejection Frequencies for Alternative Procedures



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

Figure B.2 shows rejection frequencies for two procedures that are less sophisticated than the ones in the right panel of Figure B.1. One of these uses standard errors based on CV_2 together with the usual $t(G - 1)$ critical values, and the other uses CV_1 standard errors together with critical values based on the effective degrees of freedom G^* suggested in Carter, Schnepel and Steigerwald (2017). Note that the vertical axis has been subjected to a square-root transformation. CV_2 does not overreject as severely as CV_1 (compare Figure 3), but it still overrejects 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 $G_1 \geq 4$.

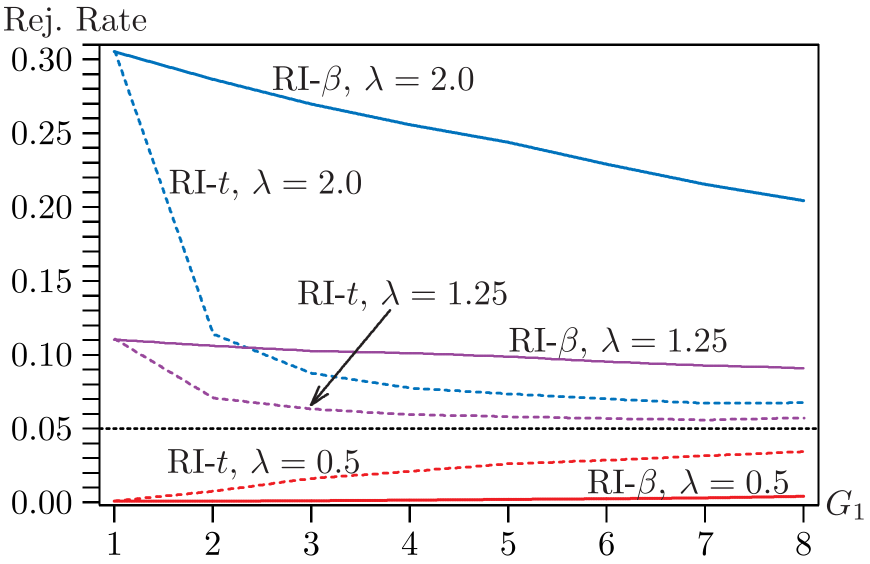
Appendix C. RI Procedures with Heteroskedasticity

In the experiments reported in the body of the paper, 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 those for the treated and control clusters equal to λ^2. Then, from equation (10), it is evident that the larger is the variance of the error terms for the treated clusters, the larger will be the variance of $\hat{\beta}$. This follows from the fact that the first summation, which depends on those error terms, is proportional to λ^2. The variance of t_β must also be increasing in λ in this case, at least when G_1 is small, because the first summation in (10) is precisely what the CRVE underestimates. Thus, the more λ differs from 1, the worse we expect both RI procedures to perform.

To investigate this phenomenon, we perform an additional set of experiments in which

Figure C.1: Rejection Frequencies with Heteroskedasticity



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

the standard deviation of the errors for the treated clusters is λ times the standard deviation of the errors for the controls. We would expect overrejection when $\lambda > 1$ and underrejection when $\lambda < 1$. As G_1 increases, we expect the problem to go away for RI-t but not for RI-β.

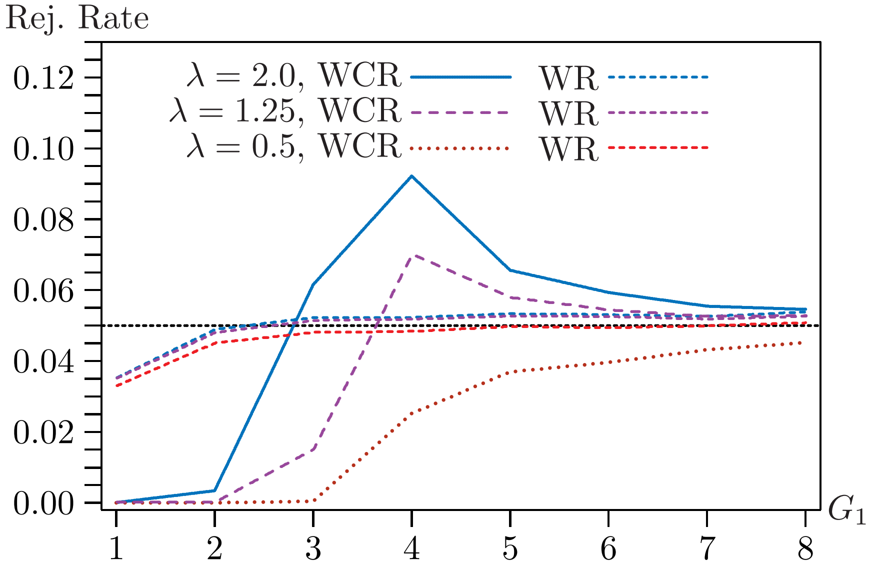
Figure C.1 shows rejection frequencies for the RI-β and RI-t procedures for a DiD model with 40 equal-sized clusters and 4000 observations. Results are shown for three values of λ, namely, $\lambda = 2.0$, $\lambda = 1.25$, and $\lambda = 0.5$. As expected, both procedures overreject when $\lambda > 1$ and underreject when $\lambda < 1$. When $G_1 = 1$, both the overrejection for $\lambda = 2.0$ and the underrejection 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 overrejects noticeably for $G_1 = 8$ when $\lambda = 2.0$, and it underrejects noticeably when $\lambda = 0.5$.[3] The size distortions in Figure C.1 are much more severe 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.

In Figure C.2, 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 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 (2018) predicts.

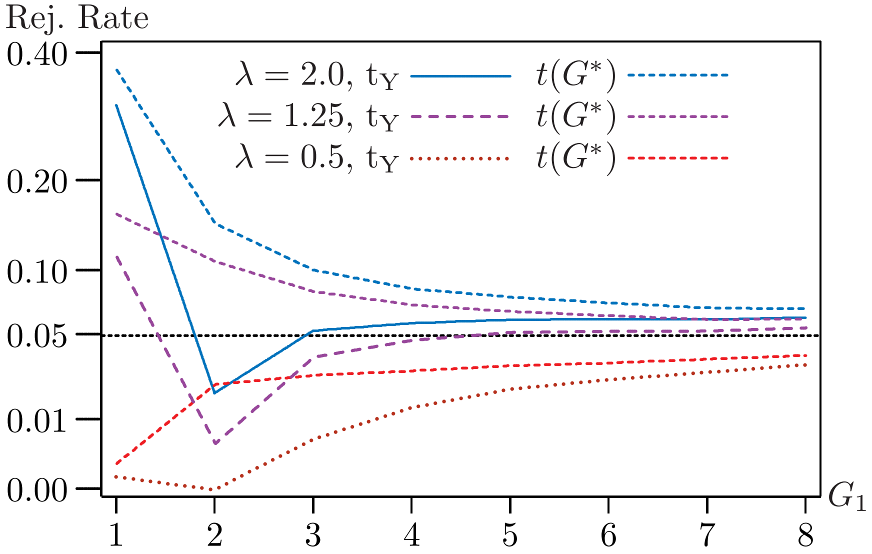
[3] Using a different experimental design, Canay, Romano and Shaikh (2017, Appendix) studies the performance of the Conley and Taber (2011) Γ procedure (and several other tests) when the treated clusters have greater variance than the untreated ones. They find even more severe overrejection for $\lambda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

Figure C.2: Rejection Frequencies for Wild Bootstrap Procedures with Heteroskedasticity



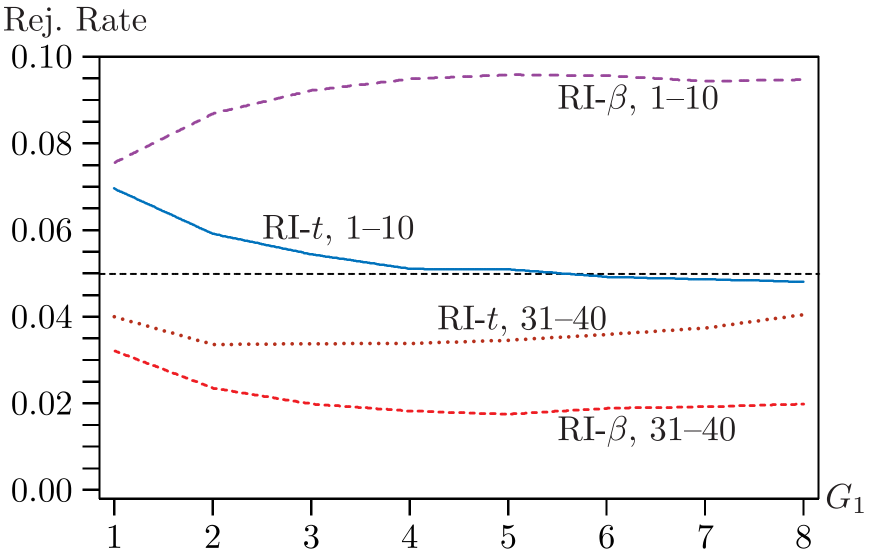
Notes: Based on 100,000 replications with $G = 40$, $\gamma = 0$, and $\rho = 0.05$

Figure C.3: Rejection Frequencies for Alternative Procedures With Heteroskedasticity



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

Figure D.1: Rejection Frequencies for RI Procedures With Lognormal Errors



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

In Figure C.3, 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 do in Figures B.1 and B.2 with homoskedasticity and variable cluster sizes. Note that the vertical axis has been subjected to a square-root transformation. We do not report results for t_{IK} because it is extremely expensive to compute.[4] 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 overreject. With $G_1 = 1$ and $\lambda = 0.5$, both procedures severely underreject. 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 D: Simulation Results with Lognormal Errors

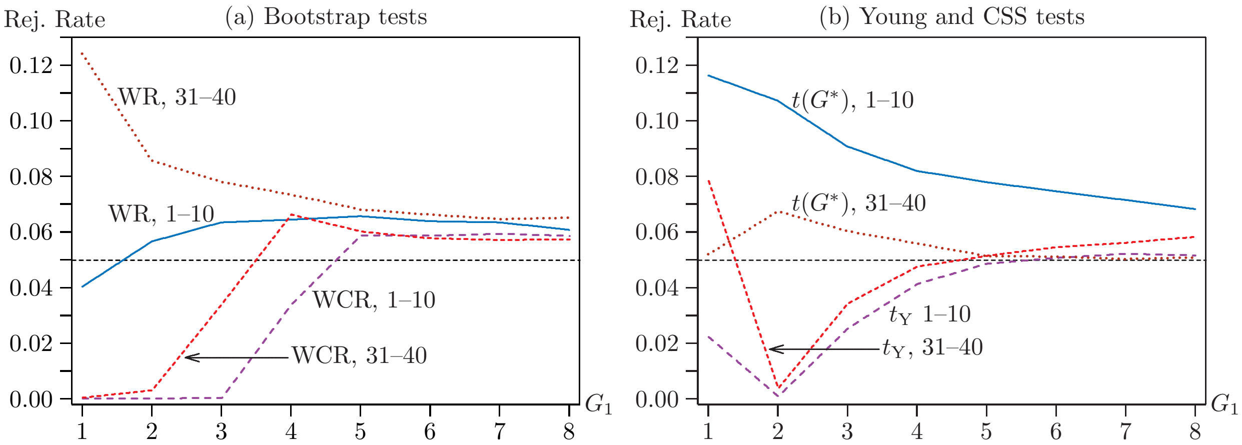
For all of our experiments up to this point, the error terms have been normally distributed. Here we report some additional results in which they are instead lognormal, rescaled to have mean 0 and variance 1. These errors are strongly skewed to the right. Not surprisingly, this affects the performance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

Figure D.1 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 potentially treated, both procedures continue to work perfectly, and we do not show those results.

Figure D.2 shows rejection frequencies for the two restricted wild bootstrap tests in the

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

Figure D.2: Rejection Frequencies for Alternative Procedures With Lognormal Errors



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

left panel and for $t_{\rm Y}$ and $t(G^*)$ in the right panel. These may be compared with the results in Figure B.1. 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.

Appendix E. Merit Scholarships

In this appendix, 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:

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

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_{st} equals 1 if state s offered a merit scholarship in year t. The remaining variables are all binary indicator variables. The state dummies equal 1 when $j = s$, and the year dummies equal 1 when $k = t$. The dataset has $N = 42{,}161$ observations taken from all states, including the District of Columbia, so that $G = 51$.

Conley and Taber (2011), hereafter referred to as CT, 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 $[0.008,\ 0.059]$. Using a method that essentially inverts RI-β P values, the paper estimates a 95% confidence interval for β_1 of $[-0.003,\ 0.093]$.[5] Thus, unlike the conventional CRVE confidence interval, the CT 95% confidence interval contains 0.

Table 2: 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
	$t(50)$ p	Young p	CSS p	WR p^*	WCR p^*
merit	0.010	0.018	0.071	0.030	0.021

Notes: The outcome variable is whether an individual had ever enrolled in college. The 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.

We calculate several alternative P values and present the results in Table 2. We calculate both RI-β and RI-t using the CT data and a modified version of their Stata code.[6] With 9999 randomizations and symmetric P values, we obtain an RI-t P value of 0.032 and an RI-β P value of 0.117. Like the CT 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.[7] In view of these results and the fact that, in all our Monte Carlo experiments, the RI-t procedure tended to be slightly undersized, but quite close to 5%, we conclude that the merit scholarship programs did have a statistically significant impact.

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

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

[7] We are unable to calculate IM P values here because treatment starts in different years.