Queen's Economics Department Working Paper No. 1404

Wild Bootstrap Randomization Inference for Few Treated Clusters

James G. MacKinnon
Queen's University

Matthew D. Webb
Carleton University

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

3-2018

Wild Bootstrap Randomization Inference for Few Treated Clusters*

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

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

March 24, 2018

Abstract

When there are few treated clusters in a pure treatment or difference-in-differences setting, t tests based on a cluster-robust variance estimator (CRVE) can severely over-reject. Although procedures based on the wild cluster bootstrap often work well when the number of treated clusters is not too small, they can either over-reject or under-reject seriously when it is. In a previous paper, we showed that procedures based on randomization inference (RI) can work well in such cases. However, RI can be impractical when the number of clusters is small. We propose a bootstrap-based alternative to randomization inference, which mitigates the discrete nature of RI P values in the few-clusters case.

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

*The procedure discussed in this paper was originally proposed in a working paper circulated as "Randomization Inference for Difference-in-Differences with Few Treated Clusters." A revised version of that paper no longer contains the WBRI procedure. This draft borrows heavily from the previous version of the working paper. We are grateful to Jeffrey Wooldridge, seminar participants at the Complex Survey Data conference on October 19-20, 2017, and two anonymous referees for helpful comments. This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. Joshua Roxborough and Oladapo Odumosu provided excellent research assistance.

1 Introduction

During the past decade or two, it has become common for empirical work in many areas of economics to involve models where the error terms are allowed to be correlated within clusters. Much of this work employs difference-in-differences (DiD) estimators, where the dataset has both a time and a cross-section dimension, and clustering is typically at the cross-section level (say, by state or province). Cameron and Miller (2015) provides a recent and comprehensive survey of econometric methods for cluster-robust inference.

Despite considerable progress in the development of suitable econometric methods over the past decade, it can still be a challenge to make reliable inferences. Doing so is particularly challenging in the DiD context when there are very few treated clusters. Past research, including Conley and Taber (2011), has shown that inference based on cluster-robust test statistics can greatly over-reject in this case. MacKinnon and Webb (2017b) explains why this happens and why the wild cluster bootstrap of Cameron, Gelbach and Miller (2008) does not solve the problem; see also MacKinnon and Webb (2017a). When there are very few treated clusters, the restricted wild cluster bootstrap often severely under-rejects, and the unrestricted wild cluster bootstrap often severely over-rejects.

One potentially attractive way to obtain tests with accurate size when there are few treated clusters is to use randomization inference (RI). This involves comparing estimates based on the clusters that were actually treated with ones based on control clusters that were not treated. Several authors have recently investigated this approach; see Conley and Taber (2011), Canay, Romano and Shaikh (2017), Ferman and Pinto (2015), and MacKinnon and Webb (2018a).

RI procedures necessarily rely on strong assumptions about how similar the control clusters are to the treated clusters. MacKinnon and Webb (2018a) shows that, for the Conley-Taber procedure, these assumptions almost always fail to hold when the treated clusters have either more or fewer observations than the control clusters. As a consequence, the procedure can over-reject or under-reject quite noticeably when the treated clusters are substantially smaller or larger than the controls. MacKinnon and Webb (2018a) suggests that more reliable inferences can often be obtained by basing randomization inference on t statistics rather than coefficient estimates.

In Section 2, we briefly discuss some existing procedures for inference with clustered errors. Subsection 2.1 explains how the wild cluster bootstrap works. Subsection 2.2 then introduces randomization inference and discusses two variants of it, the one based on coefficient estimates proposed in Conley and Taber (2011) and the one based on t statistics proposed in MacKinnon and Webb (2018a). All RI procedures encounter a serious practical problem when the number of controls is small. Since there are not many ways to compare the treated clusters with the control clusters, the RI P value can take on only a small number of values in such cases. We discuss this problem in Section 3.

Section 4 then introduces a modified RI procedure closely related to the wild cluster bootstrap that solves this problem in some cases. This procedure is the main contribution of this paper. In Subsection 4.1, we show that it can substantially improve inference in cases where the only problem is an insufficient number of control clusters. An empirical example from Decarolis (2014) is presented in Section 5. Section 6 concludes.

2 Cluster-Robust Inference

A linear regression model with clustered errors may be written as

$$\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_G \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_G \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_G \end{bmatrix}, \tag{1}$$

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 rows. Each subvector $\boldsymbol{\epsilon}_g$ is assumed to have covariance matrix $\boldsymbol{\Omega}_g$ and to be uncorrelated with every other subvector. The covariance matrix $\boldsymbol{\Omega}$ of the entire error vector is block diagonal with diagonal blocks the $\boldsymbol{\Omega}_g$. OLS estimation of equation 1 yields estimates $\hat{\boldsymbol{\beta}}$ and residuals $\hat{\boldsymbol{\epsilon}}$.

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 severely unreliable. The true covariance matrix for the model (1) is

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{\Omega}_g \boldsymbol{X}_g \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2}$$

This can be estimated by using a cluster-robust variance estimator, or CRVE. The most popular CRVE is:

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

where $\hat{\boldsymbol{\epsilon}}_g$ is the subvector of $\hat{\boldsymbol{\epsilon}}$ that corresponds to cluster g. This is the estimator that is used when the `cluster` command is invoked in Stata.[1] Based on the results of Donald and Lang (2007) and Bester, Conley and Hansen (2011), it is common to assume that the t statistics follow a $t(G-1)$ distribution; this is what Stata does by default.

It is not obvious that using t statistics based on the CRVE (3) is valid asymptotically. The proof requires technical assumptions about the distributions of the errors and the regressors and how the number of clusters and their sizes change as the sample size tends to infinity. See Djogbenou, MacKinnon and Nielsen (2018). Nevertheless, test statistics based on (3) seem to yield reliable inferences when the number of clusters is large and there is not too much heterogeneity across clusters. In particular, the number of observations per cluster must not vary too much; see Carter, Schnepel and Steigerwald (2017) and MacKinnon and Webb (2017b). However, t statistics based on (3) tend to over-reject severely when the parameter of interest is the coefficient on a treatment dummy and there are very few treated clusters; see Conley and Taber (2011) and MacKinnon and Webb (2017b). Rejection frequencies can be over 75% when all the treated observations belong to the same cluster.

[1] One of the earliest CRVEs was suggested in Liang and Zeger (1986). Alternatives to (3) have been proposed in Bell and McCaffrey (2002) and Imbens and Kolesár (2016), among others.

In this paper, we are primarily concerned with the difference-in-differences (DiD) model, which is often appropriate for studies that use individual data in which there is variation in treatment across both clusters (or groups) and time periods. We can write such a model as

$$\begin{aligned} y_{igt} &= \beta_1 + \beta_2 \mathrm{GT}_{ig} + \beta_3 \mathrm{PT}_{it} + \beta_4 \mathrm{TREAT}_{igt} + \epsilon_{igt}, \\ & i = 1, \ldots, N_g, \quad g = 1, \ldots, G, \quad t = 1, \ldots, T, \end{aligned} \tag{4}$$

where i indexes individuals, g indexes groups, and t indexes time periods. Here GT_{ig} is a "group treated" dummy that equals 1 if group g is treated in any time period, PT_{it} is a "period treated" dummy that equals 1 if any group is treated in time period t, and TREAT_{igt} is a dummy that equals 1 if an observation is actually treated.

The coefficient of most interest, on which we focus in this paper, is β_4, which measures the effect on treated groups in periods when there is treatment. In many cases, of course, regression (4) would also contain additional regressors, such as group and/or time dummies, which might make it necessary to drop GT_{ig}, PT_{it}, or both. Following the literature, we divide the G groups into G_0 control groups, for which $\mathrm{GT}_{ig} = 0$, and G_1 treated groups, for which $\mathrm{GT}_{ig} = 1$, so that $G = G_0 + G_1$.

We are concerned with the case in which G_1 is small. In this case, as previously noted, CRVE-based inference fails. It also fails when G_0 is small if every cluster is either treated or not treated. However, in a DiD model where treatment only takes place in some time periods, it is possible for CRVE-based inference to perform well even when $G_0 = 0$; see MacKinnon and Webb (2017a,b). In the remainder of the paper, since we are focusing on the DiD case, we assume that only G_1 may be small.

The reason for the failure of CRVE-based inference when G_1 is small is explained in detail in MacKinnon and Webb (2017b, Section 6). Essentially, the problem is that the least squares residuals must be orthogonal to the treatment dummy variable. This implies that they sum to zero over all the treated observations. When those treated observations are spread over many clusters, there is no problem. But when they are concentrated in just a few clusters, some of the terms that are summed in the middle matrix of (3) severely underestimate the corresponding quantities in the matrices $\boldsymbol{X}'\boldsymbol{\Omega}_g\boldsymbol{X}$.[2] This causes the standard error of $\hat{\beta}_4$ to be seriously underestimated.

2.1 The Wild Cluster Bootstrap

The wild cluster bootstrap (WCB) was proposed in Cameron, Gelbach and Miller (2008) as a method for reliable inference in cases with a small number of clusters, and its asymptotic validity is proved in Djogbenou, MacKinnon and Nielsen (2018).[3] The WCB was studied extensively in MacKinnon and Webb (2017b) for the cases of unbalanced clusters and/or few treated clusters. Because we will be proposing a new procedure that is closely related to the wild cluster bootstrap in Section 4, we review how the latter works.

[2] Of course, even when G_1 is not small, the matrices $N_g^{-1}\boldsymbol{X}_g'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'\boldsymbol{X}_g$ in (3) do not estimate the corresponding matrices $N_g^{-1}\boldsymbol{X}'\boldsymbol{\Omega}_g\boldsymbol{X}$ in (2) consistently, because the former matrices necessarily have rank 1. But the summation in the middle of expression (3), appropriately normalized, does consistently estimate the matrix $\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}$, appropriately normalized. See Djogbenou, MacKinnon and Nielsen (2018) for details.

[3] A different, but less effective, bootstrap procedure for cluster-robust inference was previously suggested in Bertrand, Duflo and Mullainathan (2004); see MacKinnon and Webb (2017a).

Without loss of generality, we consider how to test the hypothesis that β_k, the last coefficient of $\boldsymbol{\beta}$ in equation (1), is zero. Then the (restricted) wild cluster bootstrap works as follows:

1. Estimate equation (1) by OLS.

2. Calculate $\hat{t}_k$, the t statistic for $\beta_k = 0$, using the square root of the k^{th} diagonal element of (3) as a cluster-robust standard error.

3. Re-estimate the model (1) subject to the restriction that $\beta_k = 0$, so as to obtain restricted residuals $\tilde{\boldsymbol{\epsilon}}$ and restricted estimates $\tilde{\boldsymbol{\beta}}$.

4. For each of B bootstrap replications, indexed by b, generate a new set of bootstrap dependent variables y_{ig}^{*b} using the bootstrap DGP

$$y_{igt}^{*b} = \boldsymbol{X}_{igt}\tilde{\boldsymbol{\beta}} + \tilde{\epsilon}_{igt} v_g^{*b}. \tag{5}$$

   Here y_{igt}^{*b} is an element of the vector $\boldsymbol{y}^{*b}$ of observations on the bootstrap dependent variable, $\boldsymbol{X}_{igt}$ is the corresponding row of $\boldsymbol{X}$, and v_g^{*b} is an auxiliary random variable that follows the Rademacher distribution; see Davidson and Flachaire (2008). It takes the values 1 and -1 with equal probability.[4]

5. For each bootstrap replication, estimate regression (1) using $\boldsymbol{y}^{*b}$ as the regressand. Calculate t_k^{*b}, the bootstrap t statistic for $\beta_k = 0$, using the square root of the k^{th} diagonal element of (3), with bootstrap residuals replacing the OLS residuals, as the standard error.

6. Calculate the bootstrap P value as

$$\hat{p}_{\text{s}}^{*} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\big(|t_k^{*b}| > |\hat{t}_k|\big), \tag{6}$$

   where $\mathbb{I}(\cdot)$ denotes the indicator function. Equation (6) assumes that the distribution of t_k is symmetric. Alternatively, one can use a slightly more complicated formula to calculate an equal-tail bootstrap P value.

The procedure just described is known as the restricted wild cluster, or WCR, bootstrap, because the bootstrap DGP (5) uses restricted parameter estimates and restricted residuals. We could instead use unrestricted estimates and residuals in step 4 and calculate bootstrap t statistics for the hypothesis that $\beta_k = \hat{\beta}_k$ in step 5. This yields the unrestricted wild cluster, or WCU, bootstrap.

MacKinnon and Webb (2017b) explains why the wild cluster bootstrap fails when the number of treated clusters is small. The WCR bootstrap, which imposes the null hypothesis, leads to severe under-rejection. In contrast, the WCU bootstrap, which does not impose

[4] Because v_g^{*b} takes the same value for all observations within each group, we would not want to use the Rademacher distribution if G were smaller than about 12; see Webb (2014), which proposes an alternative for such cases.

the null hypothesis, leads to severe over-rejection. When just one cluster is treated, it over-rejects at almost the same rate as using CRVE t statistics with the $t(G-1)$ distribution.

MacKinnon and Webb (2018b) proposes modified versions of the WCR and WCU bootstraps for inference with few treated clusters. Rather than the auxiliary random variables being drawn at the level at which the errors are believed to be clustered, they are drawn at a lower level corresponding to 'subclusters' of the original clusters. For example, in the DiD model (4), we might draw GT independent values of v_{gt}^{*b}. It is up to the investigator to decide on the appropriate level of subclustering. In the limit, there is only one observation per subcluster, and the subcluster wild bootstrap is simply the ordinary wild bootstrap.

The asymptotic validity of the ordinary wild bootstrap for the model (1) combined with the CRVE (3) is proved in Djogbenou, MacKinnon and Nielsen (2018). Theoretical and simulation results in MacKinnon and Webb (2018b) and simulation results in Djogbenou, MacKinnon and Nielsen (2018) suggest that the ordinary wild bootstrap works well for equal-sized clusters and for cluster-specific heteroskedasticity, even when there are few treated clusters. However, it works less well for unequal cluster sizes and other types of heterogeneity across clusters. In general, the restricted version works better than the unrestricted version.

2.2 Randomization Inference

Randomization inference, first proposed in Fisher (1935), is a procedure for performing exact tests in the context of experiments. The idea is to compare an observed test statistic $\hat{\tau}$ with an empirical distribution of test statistics τ_j^* for $j = 1, \ldots, S$ generated by re-randomizing the assignment of treatment across experimental units. To compute each of the τ_j^*, we use the actual outcomes while pretending that certain non-treated experimental units were treated. If $\hat{\tau}$ is in the tails of the empirical distribution of the τ_j^*, then this is evidence against the null hypothesis of no treatment effect.

Randomization tests 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). Whether this key assumption is true in the context of policy changes such as those typically studied in the DiD literature is debatable. Any endogeneity in the way policies are implemented over jurisdictions and time would presumably cast doubt on the assumption.

When treatment is randomly assigned at the individual level, the invariance of the distribution of the test statistic to re-randomization follows naturally. However, if treatment assignment is instead at the group level, as is always the case for DiD models like (4), then the extent of unbalancedness can determine how close the distribution is to being invariant.

It is obvious that the proportion of treated observations matters for $\hat{\beta}_4$ in (4) and its cluster-robust standard error. Let $\bar{d} = \left(\sum_{g=1}^{G_1} N_g\right)/N$ denote this proportion. When clusters are balanced, the value of $\bar{d}$ will be constant across re-randomizations. However, when clusters are unbalanced, $\bar{d}$ may vary considerably across re-randomizations. This implies that the distributions of $\hat{\beta}_4$ may also vary substantially. Randomization inference may not work well in such cases.

MacKinnon and Webb (2018a) studies two types of RI procedure. One uses $\hat{\beta}_4$ in (4) as $\hat{\tau}$, and the other uses the cluster-robust t statistic that corresponds to $\hat{\beta}_4$. The former procedure, which we refer to as RI-β, is quite similar to a procedure proposed in Conley and

Taber (2011). It is only valid, even in large samples, if re-randomizing does not change the distribution of the $\hat{\beta}_{4j}^*$. The latter procedure, which we refer to as RI-t, is evidently valid in large samples whenever the cluster-robust t statistics follow an asymptotic distribution that is invariant to $\bar{d}$ and to any other features of the individual clusters. However, as MacKinnon and Webb (2018a) shows, it is generally not valid in finite samples when $\bar{d}$ varies across re-randomizations, especially when G_1 is small. Nevertheless, the RI-t procedure typically works much better than the RI-β procedure, especially when G_1 is not too small.

When there is just one treated group, it is natural to compare $\hat{\tau}$ to the empirical distribution of G_0 different τ_j^* statistics. However, when there are two or more treated groups and G_0 is not quite small, the number of potential τ_j^* to compare with can be very large. In such cases, we may pick S of them at random. To avoid ties, we never include the actual $\hat{\tau}$ among the τ_j^*. Some RI procedures do in fact include $\hat{\tau}$, however. Provided S is large, this is inconsequential.

The randomization inference procedures discussed in MacKinnon and Webb (2018a) for the model (4) work as follows. Here $\hat{\tau}$ denotes either $\hat{\beta}_4$ or its cluster-robust t statistic, and τ_j^* denotes the corresponding quantity for the j^{th} re-randomization.

1. Estimate the regression model and calculate $\hat{\tau}$.

2. Generate a number of τ_j^* statistics, S, to compare $\hat{\tau}$ with.

   - When $G_1 = 1$, assign a group from the G_0 control groups as the "treated" group g^* for each repetition, re-estimate the model using the observations from all G groups, and calculate a new statistic, τ_j^*, indicating randomized treatment. Repeat this process for all G_0 control groups. Thus the empirical distribution of the τ_j^* 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 (4), and calculate a new τ_j^*. There are potentially ${}_G\mathrm{C}_{G_1} - 1$ sets of groups to compare with, where ${}_n\mathrm{C}_k$ denotes "n choose k." When this number is not too large, obtain all of the τ_j^* by enumeration. When it exceeds B (picked on the basis of computational cost), choose the comparators randomly, without replacement, from the set of potential comparators. Thus the empirical distribution will have $S = \min({}_G\mathrm{C}_{G_1} - 1, B)$ elements.

3. Sort the vector of τ_j^* statistics.

4. Determine the location of $\hat{\tau}$ within the sorted vector of the τ_j^*, and compute a P value. This may be done in more than one way, as we discuss in the next section.

In the above procedures, we need to assign a starting period for "treatment" in each re-randomization if we are dealing with a DiD model like (4). The method used in the simulation experiments in MacKinnon and Webb (2018a) and in Subsection 4.1 below is to make the treatment period(s) the same for each re-randomization as for the actual sample. Thus if, for example, $G_1 = 1$ and treatment began in 1978, the single "treated" group in all re-randomizations would start treatment in 1978. If $G_1 = 2$ and treatment began in 1978

and 1982, then, for each re-randomization, one group would begin treatment in 1978 and the other in 1982. In our simulations, we ordered both the actually treated groups and the controls by size. Thus if, for example, treatment began in 1978 for group 3 and in 1982 for group 11, and $N_3 > N_{11}$, then treatment would begin in 1978 for the larger control group and in 1982 for the smaller one. We also experimented with assigning treatment years at random and found that doing so made very little difference.

3 Randomization Inference and Interval P Values

The most natural way to calculate an RI P value is probably to use the equivalent of equation (6). As before, S denotes the number of repetitions, which would be G_0 when $G_1 = 1$ and the minimum of ${}_G\mathrm{C}_{G_1} - 1$ and B when $G_1 > 1$, where B is a user-specified target number of replications. Then the analog of (6) is

$$\hat{p}_1^* = \frac{1}{S}\sum_{j=1}^{S} \mathbb{I}\big(|\tau_j^*| > |\hat{\tau}|\big). \tag{7}$$

This makes sense if we are testing the null hypothesis that $\beta_4 = 0$ and expect the τ_j^* to be symmetrically distributed around zero. If we were instead testing the one-sided null hypothesis that $\beta_4 \leq 0$, we would want to remove the absolute value signs.

Equation (7) is not the only way to compute an RI P value for a point null hypothesis. A widely-used alternative is

$$\hat{p}_2^* = \frac{1}{S+1}\left(1 + \sum_{j=1}^{S} \mathbb{I}\big(|\tau_j^*| > |\hat{\tau}|\big)\right). \tag{8}$$

Both procedures are valid, as would be any procedure that yields a number between $\hat{p}_1^*$ and $\hat{p}_2^*$, because P values based on a finite number of simulations are interval-identified rather than point-identified.[5]

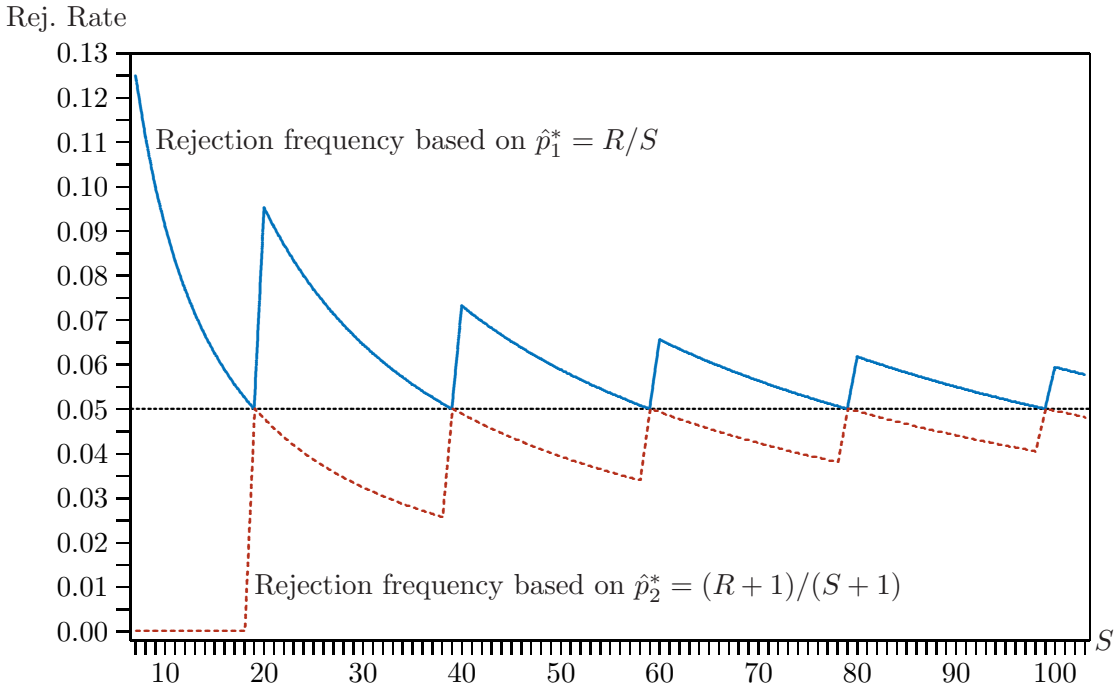
It is easy to see that the difference between $\hat{p}_1^*$ and $\hat{p}_2^*$ is $O(1/S)$, so that they tend to the same value as $S \to \infty$. There is evidently no problem if S is large, but the two P values can yield quite different inferences when S is small. The analogous issue should not arise for bootstrap tests, because the investigator can always choose B (the number of bootstrap samples, which plays the same role as S here) in such a way that equations (7) and (8) yield the same inferences. This will happen whenever $\alpha(B+1)$ is an integer, where α is the level of the test. That is why it is common to see $B = 99$, $B = 999$, and so on.

For small values of S, the conflict between inferences based on $\hat{p}_1^*$ and $\hat{p}_2^*$ can be substantial. Figure 1 shows analytical rejection frequencies for tests at the .05 level based on equations (7) and (8), respectively. The tests would reject exactly 5% of the time if S were infinite, but the figure is drawn for values of S between 7 and 103. In the figure, R denotes the number of times that $\hat{t}$ is more extreme than t_j^*, so that $\hat{p}_1^* = R/S$ and $\hat{p}_2^* = (R+1)/(S+1)$.

It is evident that $\hat{p}_1^*$ always rejects more often than $\hat{p}_2^*$, except when $S = 19$, 39, 59, and so on. Even for fairly large values of S, the difference between the two rejection frequencies

[5] The problem with P values not being point-identified is discussed at length in Webb (2014).

Figure 1: Rejection Frequencies and Number of Simulations



can be substantial. The figure is drawn under the assumption that we reject whenever either P value is equal to or less than .05. This is the only correct procedure for $\hat{p}_2^*$. However, for $\hat{p}_1^*$ it might be more natural to reject only when $\hat{p}_1^* < .05$. If that were done, the results for $\hat{p}_1^*$ with $S = 20$, 40, 60, and so on would be identical to the results for $\hat{p}_2^*$ with those values of S. The remainder of the figure would be unchanged.

Suppose the data come from Canada, which has just ten provinces. If one province is treated, then $G_1 = 1$, $G_0 = 9$, and the P value can lie in only one of nine intervals: 0 to 1/10, 1/9 to 2/10, 2/9 to 3/10, and so on. Even if $R = 0$, it would never be reasonable to reject at the .01 or .05 levels.

It is possible to eliminate the interval and obtain a single P value by using a draw from the U[0, 1] distribution. The procedure proposed in Racine and MacKinnon (2007b) simply replaces the 1 after the large left parenthesis in (8) with such a draw. A similar procedure, which allows for ties, is used in Young (2015). However, these procedures have the unfortunate property that the outcome of the test depends on the realization of a single random variable drawn by the investigator. The gap between $\hat{p}_1$ and $\hat{p}_2$ still remains. We have simply chosen a number between the two by, in effect, flipping a coin. This means that two different researchers using the same dataset will randomly obtain different P values.

4 Wild Bootstrap Randomization Inference

In this section, we suggest a novel way to overcome the problem of interval P values. We propose a procedure that we refer to as wild bootstrap randomization inference, or WBRI. The WBRI procedure essentially combines the wild cluster bootstrap of Subsection 2.1 with

the RI-t procedure of Subsection 2.2. We only consider RI-t, because, at least under the null hypothesis, it seems to be better to use t statistics rather than coefficients for randomization inference.[6]

Recall the example of Canadian provinces given in the previous section, and suppose that $R = 0$, so that the treated province has a more extreme outcome than any of the others. In the strict context of randomization inference, all we can say is that the P value is between 0, according to equation (7), and 0.10, according to equation (8). In saying this, however, we have made no use of the actual values of $\hat{t}$ and the t_j^*. Only the location of $|\hat{t}|$ in the sorted list affects either P value. If the outcome for the treated province differed a lot from the outcomes for the other nine provinces, that is, if $|\hat{t}|$ were much larger than any of the $|t_j^*|$, then the evidence against the null hypothesis would seem to be quite strong. On the other hand, if $|\hat{t}|$ were just slightly larger than the largest of the $|t_j^*|$, the evidence against the null would seem to be rather weak. But neither of the RI P values takes this into account.

The key idea of the WBRI procedure is to replace the small number (S) of test statistics obtained by randomization by a much larger number generated by a restricted wild cluster bootstrap DGP like (5). However, instead of simply imposing the null hypothesis that we are actually interested in when we generate the bootstrap samples, we impose $S+1$ different nulls, corresponding to the actual treatment and the S re-randomized ones.

Why should this procedure work? Provided the clusters are reasonably homogeneous and S is not too small, the RI-t procedure seems to work very well; see MacKinnon and Webb (2018a). But when S is small, it encounters the interval P value problem. The idea of WBRI is to keep the good properties of RI-t for large S even when S is not large by generating a large number of bootstrap statistics that resemble the t_j^* obtained by re-randomization.

Of course, we could obtain as many bootstrap statistics t_b^* as we desire simply by using the wild cluster bootstrap. But, when G_1 is small, the $|t_b^*|$ tend to be positively correlated with $|\hat{t}|$. This is the reason for the failure of the WCR bootstrap with few treated clusters; see MacKinnon and Webb (2017b). When $G_1 = 1$, the correlation tends to be very high, and this often leads to extreme under-rejection.

With the WBRI procedure, the bootstrap statistics $|t_{bj}^*|$ that correspond to the j^{th} re-randomization will undoubtedly be correlated with $|\hat{t}_j|$ when G_1 is small. But only the ones that correspond to the actual null hypothesis should be strongly correlated with $|\hat{t}|$. Thus WBRI should not encounter anything like the sort of extreme failure that WCR routinely does when G_1 is small. Of course, we do not expect that WBRI will ever work perfectly, especially when the number of clusters is very small. But it seems plausible that it should yield P values which are reasonably accurate and much more precise than the interval $[\hat{p}_1, \hat{p}_2]$. We provide evidence on this point in Section 4.1.

Formally, the WBRI procedure for generating the t_b^* and t_{bj}^* statistics is as follows:

1. Estimate equation (4) by OLS and calculate $\hat{t}$ for the coefficient of interest using CRVE standard errors.

[6] A very different approach was proposed, in the context of bootstrap tests, in Racine and MacKinnon (2007a); we plan to investigate it in a future version of this paper.

2. Estimate a restricted version of equation (4) with $\beta_4 = 0$, and retain the restricted estimates $\tilde{\boldsymbol{\beta}}$ and residuals $\tilde{\boldsymbol{\epsilon}}$.

3. Construct B bootstrap samples indexed by b, say $\boldsymbol{y}_b^*$, using the restricted wild cluster bootstrap procedure discussed in Subsection 2.1. For each b, estimate equation (4) using $\boldsymbol{y}_b^*$ and calculate a bootstrap t statistic t_b^* based on CRVE standard errors.

4. For each of the S possible re-randomizations, indexed by j, construct B more bootstrap samples in exactly the same way.[7] For each b, estimate the version of equation (4) appropriate for whatever set of groups is "treated," and calculate a bootstrap t statistic t_{bj}^*. When $G_1 = 1$, each of the re-randomizations corresponds to "treating" one of the G_0 control groups.

5. Use equation (7) to calculate a P value for $\hat{t}$ based on the $B \times {}_G\mathrm{C}_{G_1}$ bootstrap statistics. These include B bootstrap statistics t_b^* that correspond to the G_1 actually treated groups and are drawn from exactly the same distribution as the t statistics in the restricted wild cluster bootstrap procedure, along with $B({}_G\mathrm{C}_{G_1} - 1)$ bootstrap statistics t_{bj}^* in which each set of groups other than the actual one is "treated" in turn.

Since every possible set of G_1 clusters is "treated" in the bootstrap samples, the number of test statistics is $B \times {}_G\mathrm{C}_{G_1}$.[8] Unless G is quite small, this will be a large number for $G_1 \geq 2$ even when B is small. We suggest choosing B so that $B \times {}_G\mathrm{C}_{G_1}$ is at least 1000.

The number of possible bootstrap DGPs is only 2^G if one uses the Rademacher distribution. Therefore, when G is small, one may want to enumerate the DGPs (that is, pick every possible value from the Rademacher distribution) or use an alternative bootstrap distribution such as the 6-point distribution suggested in Webb (2014).

In general, it makes sense to use the WBRI procedure only when the RI-t procedure does not provide enough t_j^* for the interval P value problem to be negligible. As a rule of thumb, we suggest using WBRI when $G_1 = 1$ and $G < 500$, or $G_1 = 2$ and $G < 45$, or $G_1 = 3$ and $G < 20$. Code for this procedure is available from the authors.

4.1 Monte Carlo Results for WBRI

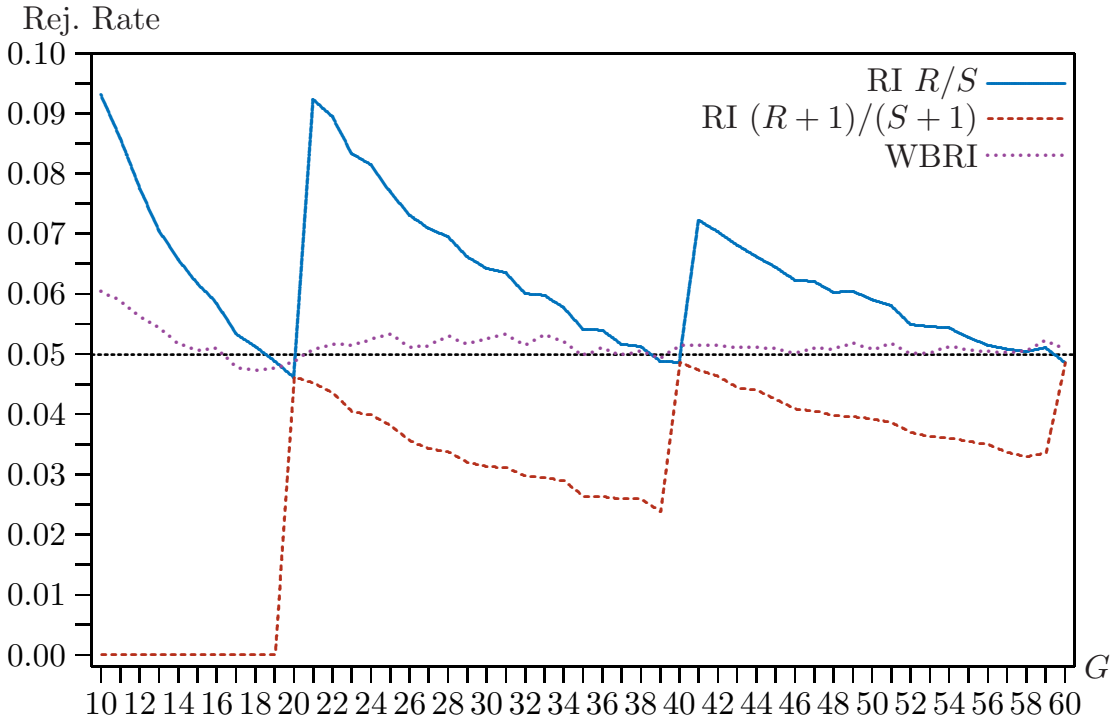
The WBRI procedure described above is designed to avoid the problem of interval P values. Based on Figure 2, it seems to be quite effective at doing so. The figure shows rejection frequencies for three procedures (RI-t using $\hat{p}_1$, RI-t using $\hat{p}_2$, and WBRI) for 51 different experiments, each with 100,000 replications. It deals with the case in which $G_1 = 1$, which is when the interval P value problem is most severe.

Every cluster has 100 observations, and the number of clusters varies from 10 to 60, which implies that the number of controls varies from 9 to 59. When $G = 20$, 40, and 60, the two RI P values must yield the same outcomes. In every other case, however, $\hat{p}_1^* = R/S$

[7] Since the null hypothesis does not depend on which observations are being treated, we could simply use the same B bootstrap samples for every re-randomization. However, this would create dependence among the S different test statistics that would then depend on each bootstrap sample. It is surely much safer to use $B \times S$ different bootstrap samples, as we do.

[8] Not surprisingly, the estimated P value is the same regardless of whether the randomization occurs within a bootstrap replication or a bootstrap occurs within a randomization.

Figure 2: WBRI Rejection Frequencies and RI Intervals



must reject more often than $\hat{p}_2^* = (R+1)/(S+1)$. As expected, the observed rejection frequencies for the two RI tests look very similar to the theoretical ones in Figure 1.

In Figure 2, the WBRI rejection frequencies are almost always between the two RI rejection frequencies and are always quite close to 5% except when G is very small. This is what we would like to see. However, it must be remembered that the figure deals with a very special case in which all clusters are the same size and the error terms are homoskedastic. The WBRI procedure cannot be expected to work any better than the RI-t procedure when the treated clusters are smaller or larger than the untreated clusters, or when their error terms have different variances.

5 Empirical Example

In this section, we consider an empirical example from Decarolis (2014). Part of the analysis deals with how the introduction of first price auctions (FPA) in Italy affected winning discounts in public works procurement. From January 2000 to June 2006, the use of average bid auctions (ABA) was required for all contracts with reserve prices below €5 million. However, after a case of collusion in ABAs was discovered, the Municipality of Turin and the County of Turin switched from ABAs to FPAs in early 2003. The central government mounted a legal challenge against these reforms that essentially prevented all other public administrations (PA) from making a similar switch.

The timing and exclusivity of the switch in Turin is exploited to estimate a regression analogous to difference-in-differences. Each of the two treated PAs (the county and the

municipality) is considered separately in the following model:

$$\text{W.Discount}_{ist} = a_s + b_t + cX_{ist} + \beta \text{FPA}_{st} + \epsilon_{ist}. \tag{9}$$

The outcome of interest, W.Discount_{ist}, is the winning discount offered in auction i of PA s in year t. FPA is a binary indicator equal to 1 for an FPA and 0 otherwise. The coefficient of interest, β, is the effect of using an FPA on the winning discount conditional on fixed effects for PA (a_s), time (b_t), and other covariates (X_{ist}). Analysis is restricted to public works auctions with reserve prices between €300,000 and €5 million, consisting of simple work types such as roadwork construction and repair jobs.

Table 5 presents our results. We first recreate the first two columns of Table 5 in Decarolis (2014). That paper implements a matching strategy, based on similarities in total number of auctions held in each PA during the sample period, to define control groups from other jurisdictions for each of the two treated PAs. This results in 14 control groups for the Municipality of Turin and 17 control groups for the County of Turin. Thus, $G = 15$ for the Municipality of Turin, and $G = 18$ for the County of Turin, with $G_1 = 1$ in both cases. In the municipality regression, Turin is the largest cluster with 200 observations out of 1,262, and the smallest cluster has 28. In the county regression, Turin is again the largest cluster with 147 observations out of 1,355, and the smallest cluster has 27. Results in MacKinnon and Webb (2018a) suggest that the RI tests should be conservative when the largest clusters are treated, as is the case in both our samples.

The model above is used to estimate 95% confidence intervals for β under two specifications. Both specifications control for year, PA, a municipality dummy, type of public work dummies, and reserve price. The first specification, which we call Model 1 and is called "W. Discount (1)" in the paper, controls for fiscal efficiency, the ratio of total yearly realized revenue to estimated revenue of the PA. The second specification, which we call Model 2, and is called "W. Discount (2)" in the paper, controls for time trends and PA-specific time trends, but not fiscal efficiency. For each panel, the first and second rows provide estimates when standard errors are clustered at the PA-Year and PA levels, while the third row uses the method of constructing confidence intervals proposed in Conley and Taber (2011). Following the original paper, confidence intervals are rounded to the nearest integer value.

In addition to reproducing the original results, we compute RI-β, RI-t, and WBRI P values using the same two samples and two models. We do this clustering only by PA. As expected, the RI-β P values are identical to the RI-t P values because there is only one treated cluster; see MacKinnon and Webb (2018a) for details. The four RI P value intervals for Model 1 contain .05, while the four RI P value intervals for Model 2 contain .10. In the former case, this makes it impossible either to reject or not reject at the .05 level. In the latter case, we evidently cannot reject at the .05 level, but it is impossible either to reject or not reject at the .10 level.

The WBRI P values shown in the table are obtained with $B = 700$ for Panel A and $B = 600$ for Panel B. This means that there are $700 \times {}_{15}C_1 = 10{,}500$ and $600 \times {}_{18}C_1 = 10{,}800$ bootstrap t statistics, respectively. Under Model 1, we find WBRI P values that are very close to $\hat{p}_1^*$ and highly significant. Under Model 2, we again find WBRI P values that are very close to $\hat{p}_1^*$. However, they are greater than .05, even though the Conley-Taber confidence intervals do not contain 0.

Table 1: 95% Confidence Intervals and P values for FPA coefficient

	Model 1	Model 2
Panel A: Municipality of Turin		
$\hat{\beta}$	12.18	6.14
PA-Year Clustering (CI)	(10, 15)	(4, 9)
PA Clustering (CI)	(10, 14)	(4, 8)
Conley-Taber (CI)	(10, 16)	(5, 8)
RI-β (P values)	(0.000, 0.067)	(0.071, 0.133)
RI-t (P value)	(0.000, 0.067)	(0.071, 0.133)
WBRI (P value)	0.0002	0.0769
N	1,262	1,262
G	15	15
Panel B: County of Turin		
$\hat{\beta}$	8.71	5.69
PA-Year Clustering (CI)	(7, 11)	(3, 8)
PA Clustering (CI)	(8, 10)	(4, 7)
Conley-Taber (CI)	(7, 14)	(4, 8)
RI-β (P values)	(0.000, 0.056)	(0.058, 0.111)
RI-t (P value)	(0.000, 0.056)	(0.058, 0.111)
WBRI (P value)	0.0006	0.0653
N	1,355	1,355
G	18	18
Regressors		
Fiscal Efficiency	Yes	No
PA Specific Time Trends	No	Yes

Notes: Entries of the form (0.000, 0.056) represent the P value pairs ($\hat{p}_1^*$, $\hat{p}_2^*$). WBRI P values are obtained with $B = 700$ for Panel A and $B = 600$ for Panel B, ensuring that $B \times_G C_1 > 10,000$ for both panels.

The evidence against the null hypothesis is probably even stronger than the WBRI results suggest. In MacKinnon and Webb (2018a), we showed that RI procedures tend to under-reject when the treated clusters are unusually large. Since the only treated cluster is either the Municipality or the County of Turin, and each of those is the largest cluster in its sample, we would expect the WBRI P values to be biased upwards. Thus the fact that the WBRI test rejects at the .001 level for Model 1 and at the .10 level for Model 2 suggests that there is quite strong evidence against the null hypothesis.

6 Conclusion

We introduce a bootstrap-based modification of randomization inference which appears to solve the problem of interval P values when there are few control groups. This procedure, which we call WBRI for "wild bootstrap randomization inference," is easiest to understand as

a modified version of the wild cluster bootstrap. Like the WCB, it generates a large number of bootstrap samples and uses them to compute bootstrap test statistics. However, unlike the WCB, only some of the bootstrap test statistics are testing the actual null hypothesis. Most of them are testing fictional null hypotheses obtained by re-randomizing the treatment. If there are S possible re-randomizations (when only one group is treated, S would equal $G_0 = G - 1$), then $B/(S+1)$ of the bootstrap test statistics are testing the actual null hypothesis and $BS/(S+1)$ of them are testing fictional null hypotheses.

The WBRI procedure can be used to generate as many t^* statistics as desired by making B large enough. Thus it can solve the problem of interval P values. However, it shares some of the properties of the RI-t procedure, which performs conventional randomization inference based on cluster-robust t statistics; see MacKinnon and Webb (2018a). In particular, like RI-t, WBRI can be expected to over-reject (or under-reject) when the treated clusters are smaller (or larger) than the control clusters and G_1 is very small. Thus we cannot expect it to yield reliable inferences in every case.

References

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

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

Davidson, Russell, and Emmanuel Flachaire (2008) 'The wild bootstrap, tamed at last.' *Journal of Econometrics* 146(1), 162 – 169

Decarolis, Francesco (2014) 'Awarding Price, Contract Performance, and Bids Screening: Evidence from Procurement Auctions.' *American Economic Journal: Applied Economics* 6(1), 108–132

Djogbenou, Antoine, James G. MacKinnon, and Morten Ø. Nielsen (2018) 'Asymptotic 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

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

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

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 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(2), 233–254

MacKinnon, James G., and Matthew D. Webb (2018a) 'Randomization inference for difference-in-differences with few treated clusters.' Working Paper 1355, Queen's University, Department of Economics

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

Racine, Jeffrey S., and James G. MacKinnon (2007a) 'Inference via kernel smoothing of bootstrap P values.' *Computational Statistics & Data Analysis* 51(12), 5949–5957

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

Webb, Matthew D. (2014) 'Reworking wild bootstrap based inference for clustered errors.' Working Paper 1315, Queen's University, Department of Economics, August

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