An Exact Hypothesis Test For Samples With Few Effective Clusters

Akiva Yonah Meiselman[*]

January 2025

Abstract

I propose a hypothesis test for clustered samples. This test is exact in samples with few clusters, few ever-treated clusters, cluster size outliers, or treatment intensity outliers; these features cause previous tests to over- or under-reject true hypotheses. I derive my test by inverting the distribution of the test statistic under a standard assumption about the errors, so that critical values can be selected from a distribution that matches the test statistic. I use Monte Carlo simulations to demonstrate where this adjustment is most impactful in achieving exact tests compared to previous hypothesis tests, and I apply my test to an empirical setting from the literature.

1 Introduction

Researchers often find that their samples include observations that are not independent. Rather, the observations are grouped into independent clusters. Common research designs based on fixed effects exploit the dependence among clustered observations to avoid bias due to omitted unobservable characteristics. Statistical inference in such samples must account for clustering. Failing to do so can lead to dramatically inaccurate standard errors, confidence intervals, and p-values.

Since White (1984, theorem 6.3), there have been tools that allow researchers to perform cluster-robust inference. The tools are consistent; they are valid in asymptotically large samples. However, those tools are not guaranteed to work in samples with a small number of especially important clusters. Conventional cluster-robust hypothesis tests will over-reject or under-reject true hypotheses in four environments: (1) when the number of clusters is small; (2) when the number of clusters with treatment variation is small; (3) when there are cluster

[*] (email: meiselman@utexas.edu) I am deeply grateful to Leigh Linden, Brendan Kline, Richard Murphy, and Sandy Black for their guidance and support. I would also like to thank Dean Spears, Stephen Trejo, Gerald Oettinger, Manuela Angelucci, Marika Cabral, Michael Geruso, Haiqing Xu, Julie Cullen, Eric Shulman, Jung Hyub Lee, Eric Chyn, and Ben Meiselman for providing valuable feedback. All errors are my own.

size outliers; and (4) when there are treatment intensity outliers. I refer to samples with one or more of these features as having few "effective clusters," since the asymmetry between clusters in (2), (3), and (4) causes the test statistic to behave similarly to (1).

These environments with few effective clusters often arise in empirical work using common research designs. Karaivanov et al. (2021), studied COVID-related mask mandates using a panel of 13 Canadian provinces and territories (few clusters). Myers (2017) examined an abortion-access policy among US states, only 4 of which implemented the policy (few clusters with treatment variation). Kuziemko et al. (2018) study managed care and mortality in Texas, with over 40% of the population located in 5 of the 254 county clusters (cluster size outliers). Bound et al. (2020) analyze the impact of state appropriations to public colleges on foreign public college enrollees, and some states had much larger year-to-year changes in state appropriations (treatment intensity outliers). In all of these cases, some adjustment to the conventional cluster-robust hypothesis test would be necessary. Moreover, previous tests, including those recommended by the highly-cited Cameron and Miller (2015), are all vulnerable to one or more of these features.

In this paper, I propose an exact hypothesis test for clustered samples with cluster-level fixed effects that rejects true hypotheses at the correct rate even with few effective clusters. I develop my test by calculating the inverse of the distribution of the conventional cluster-robust test statistic under the assumption that the errors are normal and homoskedastic. This is a common assumption made in the existing literature. Unlike previous hypothesis tests, my test uses the exact distribution of the test statistic under these assumptions. If the errors are not normal or homoskedastic, then my test performs similar to those in the literature; it is asymptotically consistent, in the sense that the rejection rate converges to the nominal size of the test. Thus, my test improves on the existing tests by performing as well as existing tests in large samples while also being an exact test in small samples under the required error conditions. The main advantage of my test over other asymptotically consistent tests is that my test performs better than those tests when there are few effective clusters.

I demonstrate using Monte Carlo simulations that my test is exact and that previous tests are not exact in samples with few effective clusters. I calculate the rejection rates of my test and the other tests in randomly-generated samples where I vary the number of clusters, the number of clusters with treatment variation, the size of a cluster size outlier, and the treatment intensity of a treatment intensity outlier. I also calculate rejection rates when the errors are non-normal, heteroskedastic, or serially correlated. My test performs as well as any other test when the assumption of normal, homoskedastic errors is violated.

Additionally, I show that the applied economics literature is replete with examples of samples with few effective clusters, underscoring the need for a hypothesis test that does not rely so heavily on asymptotic validity. I compile a list of studies that match a popular research design based on treatment variation across U.S. states, and I find that many of these studies have small effective sample sizes. I then illustrate that my test changes inference in Ang (2019), a study of the impact of the Voting Rights Act.

This is the first paper to explicitly target exact inference in a finite, clustered sample. There have been other studies which make adjustments to improve on the conventional cluster-robust hypothesis test, including C. B. Hansen (2007), Bell and McCaffrey (2002), Carter et al. (2017), and Cameron, Gelbach, et al. (2008), but those previous tests fail to reject at the correct rate when there are few effective clusters. Two of those studies, Bell and McCaffrey (2002) and Carter et al. (2017), make an assumption that is similar to the one I use to derive my test (requiring normal, homoskedastic errors), but neither use the exact distribution of the test statistic, instead approximating the test-statistic with a t-distribution. My test does not use an approximation. As a result, my test performs better in samples with few effective clusters.

Section 2 gives context for my contribution to the literature on cluster-robust inference. Section 3 describes the model with clustering and cluster-level fixed effects that is the setting for this paper. Section 4 introduces my test and shows that it is exact. Section 5 presents evidence from Monte Carlo simulations that my test rejects true hypotheses at the correct rate even when other tests fail to do so. Section 6 discusses empirical settings where my test would be especially useful and applies my test to the setting in Ang, 2019. Section 7 concludes.

2 Literature

This paper contributes to an existing literature that addresses cluster-robust inference with a particular focus on applied research designs. Consistent cluster-robust variance estimators (CRVEs) were developed by White (1984), Liang and Zeger (1986), and Arellano (1987). During the credibility revolution of the 1990s, research designs that included fixed effects came into much broader use. In the wake of this, Bertrand et al. (2004) elevated awareness of CRVEs among applied economists.

I build on previous work that focuses on attaining asymptotically consistent inference in clustered samples. White (1984) shows that, when clusters are equally sized and the errors are homoskedastic, the basic CRVE (henceforth $\hat{V}_{CR0}$) can consistently estimate the variance of the OLS estimator. C. B. Hansen (2007) relaxes the homoskedasticity assumption, showing that equal-sized clusters alone allow the CRVE to be consistent and that it converges at a rate determined by the number of clusters G. He recommends critical values be drawn from a t-distribution with $G - 1$ degrees of freedom.

My proposed test is asymptotically consistent based on the same logic as in Carter et al. (2017). They show that $\hat{V}_{CR0}$ is consistent even when cluster sizes vary, but the rate of convergence is instead determined by G^*, what Carter et al. (2017) call the “effective number of clusters.” They recommend calculating G^* as a diagnostic tool. If G^* is large, inference can rely upon the asymptotic properties of the test statistic to determine its behavior, so critical values can reasonably be drawn from the standard normal distribution $N(0, 1)$. My test does not rely on calculating G^*, but I do use the phrase “few effective clusters” to refer to samples with small G^*, and in Section 6, I measure G^* in a number of empirical settings from the applied literature.

I will compare my test to the hypothesis tests recommended by Cameron and Miller (2015), a paper that many applied economists use as a guide for how to handle cluster-robust inference. Those tests are derived from C. B. Hansen (2007), Carter et al. (2017), Bell and McCaffrey (2002), and Cameron, Gelbach, et al. (2008). Section 4.2 describes these tests in more detail; here, I give a brief overview.

C. B. Hansen (2007), Carter et al. (2017), and Bell and McCaffrey (2002) select critical values for the test statistic from a t-distribution. C. B. Hansen (2007) sets the degrees of freedom to $G-1$ and Carter et al. (2017) sets the degrees of freedom to the effective number of clusters G^*.

Bell and McCaffrey (2002) address finite-sample cluster-robust inference by developing two additional CRVEs ($\hat{V}_{CR2}$ and $\hat{V}_{CR3}$), aimed at reducing bias in the variance estimation step. I build on their framework, making a similar assumption and discarding the approximation embedded in their test.

Taking a different approach, Cameron, Gelbach, et al. (2008) generate a reference distribution for the test statistic through a resampling method, the wild cluster bootstrap with restricted residuals (henceforth WCR).[1] Djogbenou et al. (2019) show that, in addition to performing well in simulations, WCR is a formal asymptotic refinement of the conventional test based on $\hat{V}_{CR0}$ and Hansen's $G-1$ degrees of freedom.

3 Model

In this section, I describe a linear model with clustering in a single dimension and cluster-level fixed effects. I include fixed effects because they are a common feature of models where there may also be concerns about clustering. Consider the model:

$$y_{ig} = x_{ig}\beta + \gamma_g + \epsilon_{ig} \tag{1}$$

where x_{ig} is a $(1 \times K)$ vector of K covariates and γ_g is a cluster-level fixed effect. Clusters are indexed by g, and individual observations are indexed by i. Let N_g be the (deterministic) number of observations in cluster g, and then $N = \sum_g N_g$ is the total number of observations. Additionally, for ease of notation:

$$Y_g = \begin{bmatrix} y_{1,g} \\ y_{2,g} \\ \cdots \\ y_{N_g,g} \end{bmatrix}, \quad X_g = \begin{bmatrix} x_{1,g} \\ x_{2,g} \\ \cdots \\ x_{N_g,g} \end{bmatrix}, \quad \epsilon_g = \begin{bmatrix} \epsilon_{1,g} \\ \epsilon_{2,g} \\ \cdots \\ \epsilon_{N_g,g} \end{bmatrix}$$

And similarly let Y (an $(N \times 1)$ matrix), X (an $(N \times k)$ matrix), and ϵ (an $(N \times 1)$ matrix) stack up the outcomes, covariates, and errors of all the clusters, so that X_g contains the rows of X corresponding to cluster g.

[1] The "restricted residuals" used in WCR are the residuals from the model estimated subject to the restriction that the null hypothesis is true.

For standard fixed-effects estimation of β, the fixed effects γ_g are absorbed. That is, rather than estimate γ_g, I transform the sample using cluster-level averages[2]. Let $\ddot{Y}_g = Y_g - \frac{1}{N_g}\sum_{i=1}^{N_g} y_{ig}$, $\ddot{X}_g = X_g - \frac{1}{N_g}\sum_{i=1}^{X_g} x_{ig}$, and $\ddot{\epsilon}_g = \epsilon_g - \frac{1}{N_g}\sum_{i=1}^{N_g} \epsilon_{ig}$. Assuming that $\mathbb{E}(\epsilon_g \mid X_g) = 0$, the fixed effects estimator can consistently estimate β:

$$\hat{\beta} = (\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\ddot{Y} \tag{2}$$

For inference on β, I examine two-sided tests of hypotheses with the form $H_0 : c_0^T\beta = a_0$, where c_0 is a $(K \times 1)$ vector and a_0 is a scalar. Without loss of generality, I normalize[3] c_0 so that $c_0^Tc_0 = 1$. Inference involves calculating a test statistic t_0 and comparing it to a critical value q^*. An exact test will reject a true hypothesis with some probability α, which is the "size" of the test.

Calculating t_0 begins with estimating $\hat{V}(\hat{\beta})$. The true variance of $\hat{\beta}$ is given by:

$$V(\hat{\beta}) = (\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\mathbb{E}(\ddot{\epsilon}\ddot{\epsilon}^T)\ddot{X}(\ddot{X}^T\ddot{X})^{-1}$$

In a sample of independent, identically distributed observations, inference could rely on the assumption that the errors are all mutually independent. However, in this clustered setting, I make only the (standard) weaker assumption that the errors are uncorrelated across clusters:

$$\mathbb{E}(\epsilon_g\epsilon_{g'}^T) = 0, \forall g \neq g'$$

Let $\hat{\epsilon}_g = \ddot{Y}_g - \ddot{X}_g\hat{\beta}$ be the residuals for cluster g. The simplest cluster-robust variance estimator, $\hat{V}_{CR0}$ from White (1984), takes the form:

$$\hat{V}(\hat{\beta}) = (\ddot{X}^T\ddot{X})^{-1}\left(\sum_g \ddot{X}_g^T\hat{\epsilon}_g\hat{\epsilon}_g^T\ddot{X}_g\right)(\ddot{X}^T\ddot{X})^{-1}$$

Finally, a test statistic can be generated, using the parameter estimator $\hat{\beta}$ and the variance estimator $\hat{V}(\hat{\beta})$:

$$t_0 = \frac{c_0^T\hat{\beta} - a_0}{\sqrt{c_0^T\hat{V}(\hat{\beta})c_0}}$$

If the hypothesis H_0 is true, then generating a test statistic with a large magnitude is relatively unlikely. So t_0 can be compared to some critical value q^*, and H_0 is rejected if $|t_0| > q^*$.

[2] For an extensive treatment of fixed effects models and fixed effects absorption (also called the "within transformation"), including the consistency and unbiasedness of the fixed effects estimator, see B. E. Hansen (2021), sections 17.8, 17.9, and 17.20.

[3] Suppose there is some hypothesis such that $\tilde{c}_0^T\tilde{c}_0 \neq 1$. Note that $\tilde{c}_0^T\tilde{c}_0 > 0$, since $\tilde{c}_0^T\tilde{c}_0 = 0$ would not actually be testing a linear hypothesis. Then let $c_0 = \frac{\tilde{c}_0}{\sqrt{\tilde{c}_0^T\tilde{c}_0}}$, so that $c_0^Tc_0 = \frac{\tilde{c}_0^T\tilde{c}_0}{\tilde{c}_0^T\tilde{c}_0} = 1$.

4 Hypothesis Tests

4.1 My Proposed Test

I propose a new method of testing linear hypotheses, which I develop here. I prove that my test is valid in two cases: normal, homoskedastic errors; and asymptotically large samples.

When the errors are normal and homoskedastic (see Assumption 1 below), my test rejects true hypotheses with probability equal to the nominal test size α. This assumption is standard among finite-sample adjustments to cluster-robust inference. Furthermore, under the (weaker) assumptions that allow all cluster-robust hypothesis tests to be consistent, my test is also consistent in the sense that its rejection rate converges in probability to α. In other words, my test maintains the good asymptotic properties of previous cluster-robust hypothesis tests.

In order to perform an exact hypothesis test, I would like to find a critical value $q^*(H_0, \alpha)$ such that, if H_0 is true, then:

$$P(|t_0|> q^*(H_0, \alpha)) = \alpha$$

The optimal method for selecting critical values would be some $q^*(\cdot, \cdot)$ that gives an exact test for any hypothesis H_0 and any test size α.

If the CDF of t_0^2 were known, I could invert that CDF to find the critical value. Using $F_{t_0^2}(\cdot)$ to denote the CDF of t_0^2, that critical value $q^*(\cdot, \cdot)$ is given by:

$$F_{t_0^2}((q^*(H_0, \alpha)^2) = 1 - \alpha$$
$$q^*(H_0, \alpha) = \sqrt{F_{t_0^2}^{-1}(1 - \alpha)}$$

The intuition for my test is that I make an assumption that is strong enough to determine $F_{t_0^2}(\cdot)$.

Assumption 1 *The errors are normal and homoskedastic:*

$$\epsilon_g \sim N(0, \sigma^2 I_g)$$

Note that σ may be unknown and that any random effects structure, with equicorrelated errors within a cluster, would be asborbed by the cluster-level fixed effects. I refer to this assumption as “standard” because it is the same assumption that several other papers use to adjust cluster-robust inference in finite samples (Carter et al., 2017; Bell and McCaffrey, 2002). With this assumption, $F_{t_0^2}(\cdot)$ can be derived in Theorem 1 below.

I introduce some notation in preparation for the test definition. Let I_g be an identity matrix of size N_g, and let ι_g be a column vector of length N_g whose elements are all 1. So then let $M_g = I_g - \frac{1}{N_g}\iota_g\iota_g^T$. Additionally, let $H = \ddot{X}(\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T$, let I be an identity matrix of size N, and let $(I - H)_g$ be the rows of $(I-H)$ corresponding to cluster g. Furthermore, I define a function $L(\cdot; X, H_0)$ that I will show is equal to the CDF $F_{t_0^2}(\cdot)$ in Theorem 1:

$$\text{Let } L(q; X, H_0) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{\sin\left(\frac{1}{2}\sum_{j=1}^{G+1}\tan^{-1}(\lambda_j u)\right)}{u\prod_{j=1}^{G+1}\left(1+\lambda_j^2 u^2\right)^{\frac{1}{4}}} du \tag{3}$$

$$\text{where } d_0 = \ddot{X}(\ddot{X}^T\ddot{X})^{-1}c_0, \quad d_g = (I-H)_g^T\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0,$$

$$D_+ = [d_0 \quad d_1 \ldots d_g \ldots d_G], \quad D_- = [\frac{1}{q}d_0 \quad -d_1 \cdots -d_g \cdots -d_G],$$

M block-diagonal, with g-th block M_g, and

(λ_j) are the eigenvalues of $D_-^T M D_+$

To implement my test[4]:

1. Calculate the test statistic $t_0 = \frac{c_0^T\hat{\beta}-a_0}{\sqrt{c_0^T\hat{V}(\hat{\beta})c_0}}$
2. Find $L(q; X, H_0)$
3. Determine the critical value q^* such that $L((q^*)^2; X, H_0) = 1-\alpha$
4. Reject H_0 if $|t_0| > q^*$

Note that, since $L(q; X, H_0)$ is increasing in q, $L(.; X, H_0)$ can easily be inverted numerically. Finally, let $\alpha^*(X, H_0) = P\left(|t_0| > q^*(\alpha; X, H_0)\right)$ be the rejection rate of my test – the rate at which my test rejects a true hypothesis H_0.

Theorem 1 *If Assumption 1 holds and H_0 is true, then:*

$$F_{t_0^2}(q) = L(q; X, H_0)$$
$$\textit{and } \alpha^* = \alpha$$

where $L(\cdot; \cdot, \cdot)$ is defined in Equation 3.

I prove Theorem 1 by deriving $L(\cdot; \cdot, \cdot)$ in Appendix A.

In the simplest version of the test, I calculate the test statistic t_0 using the variance estimator $\hat{V}_{CR0}$. In Section 5, I use two additional variants of my test that are based on different variance estimators. These variants use $\hat{V}_{CR2}$ and $\hat{V}_{CR3}$, the estimators given by Bell and McCaffrey (2002). The proof of Theorem 1 in Appendix A holds for both of these variants; they are exact tests under Assumption 1.

My test using $\hat{V}_{CR0}$ is also asymptotically consistent under the relatively weak assumptions described by Carter et al. (2017). They demonstrate that, when using $\hat{V}_{CR0}$, the test statistic converges to a standard normal distribution: $t_0 \xrightarrow{d} N(0,1)$. I build on this result, using their two main assumptions. The first ensures that the errors have finite fourth moments. The second ensures that the observations aren't too concentrated in a small number of clusters.

[4] My test is available in R as the function "p.value.meis()" in the package "clubsoda", available through github.

Assumption 2 *For each cluster g, there is some positive scalar B and some $(N_g \times N_g)$ matrix Ω_g such that $\epsilon_g = \Omega_g^{\frac{1}{2}} \eta_g$, where the η_g is a vector of length N_g whose elements are uncorrelated random variables and where:*

$$\begin{aligned}
&\mathbb{E}(\eta_{ig}\eta_{jg}\eta_{kg}\eta_{lg}) = 0, \quad \mathbb{E}(\eta_{ig}\eta_{jg}\eta_{kg}^2) = 0 \\
&\mathbb{E}(\eta_{ig}\eta_{jg}^3) = 0, \quad \mathbb{E}(\eta_{ig}^2\eta_{jg}^2) = 0 \\
&\mathbb{E}(\eta_{ig}^4) \leq B
\end{aligned}$$

Assumption 3 *Let $\lambda_g^{CSS} = c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}_g^T\mathbb{E}(\ddot{\epsilon}_g\ddot{\epsilon}_g^T)\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0$, let $V_0 = c_0^T V(\hat{\beta})c_0$, and let $P_g = (\ddot{X}^T\ddot{X})^{-1}\ddot{X}_g^T\ddot{X}_g$. As $n \to \infty$:*

$$\begin{aligned}
&G \to \infty \\
&\mathbb{E}\left(\frac{\sum_g (\lambda_g^{CSS})^2}{\left(\sum_g \lambda_g^{CSS}\right)^2}\right) \to 0 \\
&\frac{1}{V_0} c_0^T \left(\sum_g (P_g - \frac{1}{G}I)V(\hat{\beta})(P_g - \frac{1}{G}I)^T\right) c_0 \xrightarrow{p} 0
\end{aligned}$$

Recall that α^* is the rejection rate of my test.

Theorem 2 *If Assumptions 2 and 3 hold and H_0 is true, then:*

$$\alpha^* \xrightarrow{p} \alpha$$

I prove Theorem 2 in Appendix B. The intuition for the proof is that the test statistic t_0 and the reference distribution defined by my test both converge in distribution to $N(0, 1)$.

In large samples, my test will reject a true hypothesis with probability α. Proving that the variants of my test which use $\hat{V}_{CR2}$ and $\hat{V}_{CR3}$ are also asymptotically consistent remains an area of future work.

I have shown that my test's rejection rate α^* is equal to the nominal test size α in finite samples under Assumption 1 (normal, homoskedastic errors). I have shown separately, without Assumption 1, that α^* converges to α in an asymptotically large sample under Assumptions 2 and 3, so that my test is approximately exact when there are many effective clusters. Section 5 uses Monte Carlo simulations to demonstrate that this test performs well (rejecting at roughly the correct rate) even when dealing with few effective clusters and even when Assumption 1 does not hold.

4.2 Other Tests

In this section, I briefly discuss how previous cluster-robust hypothesis tests from the literature work and how they relate to my tests. Specifically, I will look at the tests recommended in Cameron and Miller (2015). These tests can be roughly divided into analytic tests, which select a critical value for t_0 from a known distribution, and resampling-based tests, which generate a simulated distribution of test statistics from which critical values are drawn. In Section 5, I compare my test's performance with the performance of these other tests.

In a test I refer to as "Hansen", derived from C. B. Hansen (2007), it is recommended to estimate $V(\hat{\beta})$ with $\hat{V}_{CR3}$ and to select critical values for the test statistic t_0 from $T(G-1)$, a t-distribution with $G-1$ degrees of freedom.[5]

The test from Bell and McCaffrey (2002), henceforth "BM", involves estimating $V(\hat{\beta})$ with $\hat{V}_{CR2}$ and selecting critical values for t_0 from $T(m)$, where m is calculated according to a "Satterthwaite approximation" of t_0. For an explanation of how the Satterthwaite approximation works and what assumptions it relies on, see Appendix C. Imbens and Kolesár (2016) innovate on BM by allowing a random effects structure on the errors specifically for the purpose of calculating m. However, because my setting includes cluster-level fixed effects, any random effects would be absorbed, so I use BM's formulation.

Cameron and Miller (2015) also recommend a test, henceforth "CSS", derived from Carter et al. (2017). In this test, $V(\hat{\beta})$ is estimated with $\hat{V}_{CR0}$, and critical values for t_0 are selected from $T(G^*)$, where G^* is called the "effective number of clusters". In Appendix C, I show how G^* is a simplified version of the Satterthwaite approximation.

Hansen, BM, and CSS all approximate the test statistic t_0 as a t-distribution. By contrast, the last method recommended by Cameron and Miller (2015) is a resampling method, the wild cluster bootstrap from Cameron, Gelbach, et al. (2008). Using this method, $V(\hat{\beta})$ is estimated with $\hat{V}_{CR0}$, and then over many bootstrap iterations, the residuals are resampled by multiplying them by values drawn from an auxiliary distribution with mean 0 and variance 1. A critical value q^* is then selected from the bootstrapped distribution of test statistics.

For choosing among the various specifications of the wild cluster bootstrap, I follow Djogbenou et al. (2019), a more recent study that tested many variants in simulations. They recommend resampling the restricted residuals (the residuals from the restricted model, subject to H_0), with the auxiliary distribution being either the Rademacher distribution or the Mammen distribution. I refer to these tests as "WCR-R" and "WCR-M", respectively.

There are some parallels between my test and previous analytic tests in the literature. However, these other methods all approximate the distribution of the test statistic. By contrast, I have made an assumption that is strong enough to fully determine the distribution of the test statistic. In the next section, I will demonstrate that this approach makes my test perform better in many samples; my test rejects true hypotheses at the correct rate even when other tests fail to do so.

[5] The method of selecting critical values from $T(G-1)$ is from C. B. Hansen (2007), while Cameron and Miller (2015) recommend this paired with $\hat{V}_{CR3}$ as the variance estimator.

5 Simulations

In this section, I show the results of Monte Carlo simulations that demonstrate that that my test is exact and that previous tests fail to reject true hypotheses at the correct rate under various conditions. Specifically, I focus on samples with few effective clusters: few clusters, few clusters with treatment variation, cluster size outliers, and treatment intensity outliers. I borrow certain features of the data-generating process from Djogbenou et al. (2019), another simulation study of cluster-robust inference.

Depending on the specification, I vary certain features of the design matrix X:

- $G =$ number of clusters
- $J =$ number of clusters with treatment variation
- $N_1 =$ size of first cluster
- $\phi =$ treatment intensity of first cluster

In my simulation experiments, I use the following data-generating process:

$$\begin{aligned} x_{1ig} &= \mathbb{1}(g \leq J) \times \phi^{\mathbb{1}(g=1)} \times \frac{x^*_{1ig} - 8}{4}, \quad x^*_{1ig} \sim \chi^2_8 \\ x_{2ig} &= \frac{x^*_{2ig} - 8}{4}, \quad x^*_{2ig} \sim \chi^2_8 \\ y_{ig} &= \beta_0 + \beta_1 x_{1ig} + \beta_2 x_{2ig} + \gamma_g + \epsilon_{ig} \end{aligned} \tag{4}$$

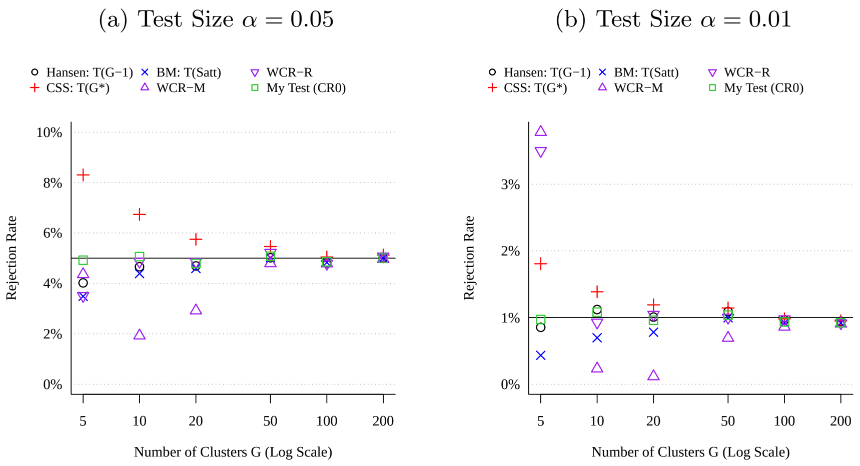
Following Djogbenou et al. (2019), the covariates x_{1ig} and x_{2ig} are generated with distributions that are both skewed and leptokurtic. This highlights the fact that my main result does not require normally-distributed covariates.

For each cluster beyond the first, there are 5 observations in that cluster — that is, for $g > 1$, $N_g = 5$. I set $\beta_0 = 1$, $\beta_1 = 2$, and $\beta_2 = 3$. Since fixed effects are absorbed before any estimation, the values of γ_g do not affect estimation or inference, so for simplicity I set $\gamma_g = 0$ for all g. Unless otherwise noted, the specification parameters have default values $G = 200$, $J = 200$, $N_1 = 5$, and $\phi = 1$.

In this section, I generate $\epsilon_{ig} \sim N(0, 1)$; this data generating process meets the conditions of Assumption 1. In Section 5.1, I will alter this specification with several violations of Assumption 1. Besides simply confirming what I showed in Theorem 1, this set of simulations serves to demonstrate the conditions where previous tests tend to over- or under-reject true hypotheses.

In each simulation, I generate a sample according to the data generating process, and I estimate $\hat{\beta}$ with the standard fixed effects estimator. Then, I test the true hypothesis $H_0 : \beta_1 = 2$ using my test as well as each of the tests recommended by Cameron and Miller (2015). Different tests require estimating $\hat{V}(\hat{\beta})$ using different variance estimators and comparing test statistics to critical

Figure 1: Comparison of Hypothesis Tests With Varying G


(a) Test Size $\alpha = 0.05$

(b) Test Size $\alpha = 0.01$

Notes: In these simulations, $J = 200$, $N_1 = 5$, and $\phi = 1$. The errors are i.i.d. standard normal. Each specification has 30,000 simulations.

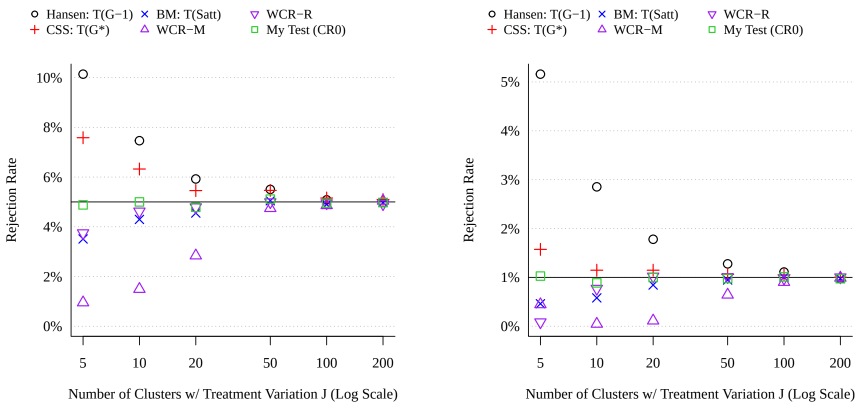
values selected according to different methods. I do this for each of the different tests discussed in Section 4.

First, I present results for simulations in which the number of clusters G takes on the following values: $G = 5, 10, 20, 50, 100, 200$. In Figure 1, I plot the rejection rates in this DGP for each of the tests discussed in Section 4. In general, the degree of a given test's over- or under-rejection depends on the size of the test α. I therefore show rejection rates for 5% tests in Panel 1a and 1% tests in Panel 1b. As the number of clusters gets small, CSS tends to overreject and WCR-M tends to underreject fairly dramatically. Hansen and BM look more reasonable but also tend to underreject for small G. WCR-R rejects at the correct rate except when $G = 5$, where it seems to fail completely. When a sample has a small number of clusters, my test is the only exact test.

Next, in Figure 2, I show rejection rates from simulations where I vary J, the number of clusters with treatment variation. As seen in (4), for clusters beyond the J-th cluster, I simply multiply the value of x_{1ig} by 0. Hansen and CSS both overreject and WCR-M underrejects at $J \leq 20$. BM and WCR-R both struggle at $J = 5$. When a sample has a small number of clusters with treatment variation, my test is the only exact test.

Figure 3 plots rejection rates for different degrees of cluster size heterogeneity, where $N_1 = 20, 100, 200, 500$. In each specification, there are 200 clusters, and for $g > 1$, $N_g = 5$. So when $N_1 = 500$, the first cluster contains about a third of the observations in the sample. In that most extreme case, BM and CSS underreject, Hansen overrejects, and WCR-M overrejects for a 5% test only. My test is exact here, and WCR-R seems to at least be resilient to this form of cluster size heterogeneity.

Figure 2: Comparison of Hypothesis Tests With Varying J


(a) Test Size $\alpha = 0.05$

(b) Test Size $\alpha = 0.01$

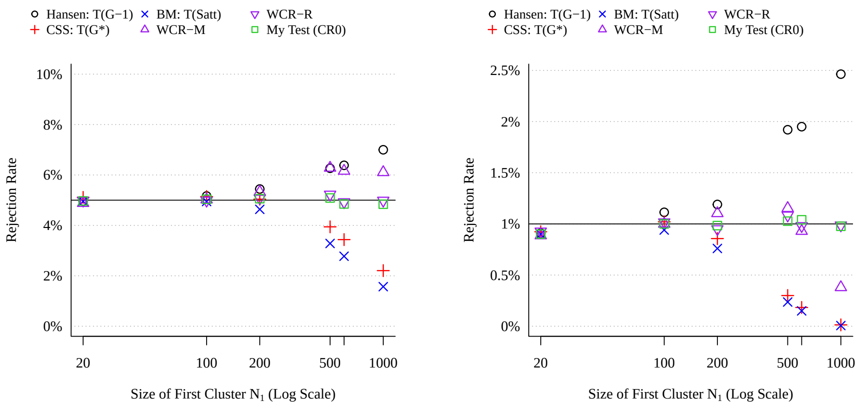
Notes: In these simulations, $G = 200$, $N_1 = 5$, and $\phi = 1$. The treatment variable x_{1ig} is distributed as $x_{1ig} = \mathbb{1}(g \leq J) \times \phi^{\mathbb{1}(g=1)} \times \frac{x^*_{1ig} - 8}{4}$, where $x^*_{1ig} \sim \chi^2_8$. The errors are i.i.d. standard normal. Each specification has 30,000 simulations.

In Figure 4, I show results for several values of the treatment intensity outlier parameter, so $\phi = 1, 5, 9, 13, 18, 24, 30$. Recall that the value of x_{1ig} is multiplied by ϕ for $g = 1$ only, so that a large value of ϕ creates a cluster that is an outlier in terms of variance in the treatment variable. Hansen overrejects for $\phi \geq 9$ and BM and CSS both underreject for $\phi \geq 9$. Around $\phi = 18$, both WCR-R and WCR-M begin to substantially underreject. In the presence of a large treatment intensity outlier, my test is the only exact test.

As discussed in Section 4, several of the methods (Hansen, CSS, and BM) approximate the test statistic t_0 with a t-distribution. Of those, Hansen and CSS can underreject or overreject, depending on the specification. This happens because variance estimation and critical value selection lead to different biases. The approximation to a t-distribution depends implicitly on having an unbiased variance estimator[6]. However, Hansen uses $\hat{V}_{CR3}$, which is biased up in this DGP (corresponding to underrejection), and CSS uses $\hat{V}_{CR0}$, which is biased down in this DGP (corresponding to overrejection). It is also true that both of these methods can select inappropriate critical values due to the approximation itself. The reason that Hansen and CSS underreject in some specifications and overreject in other specifications is that the bias in the variance estimator causes the rejection rate to move in one direction and misspecified critical value selection causes the rejection rate to move in the opposite direction. For a deeper discussion of the approximation of the test statistic t_0 with a t-distribution, see

[6] Since my test does not approximate the test statistic as a t-distribution, it does not require unbiased variance estimation

Figure 3: Comparison of Hypothesis Tests With Varying N_1


(a) Test Size $\alpha = 0.05$

(b) Test Size $\alpha = 0.01$

Notes: In these simulations, $G = 200$, $J = 200$, and $\phi = 1$. The errors are i.i.d. standard normal. Each specification has 30,000 simulations.

Section C.

5.1 Robustness

So far, the Monte Carlo simulations have met the conditions of Assumption 1. Here, I present additional simulation evidence regarding the robustness of my test to violations of Assumption 1.

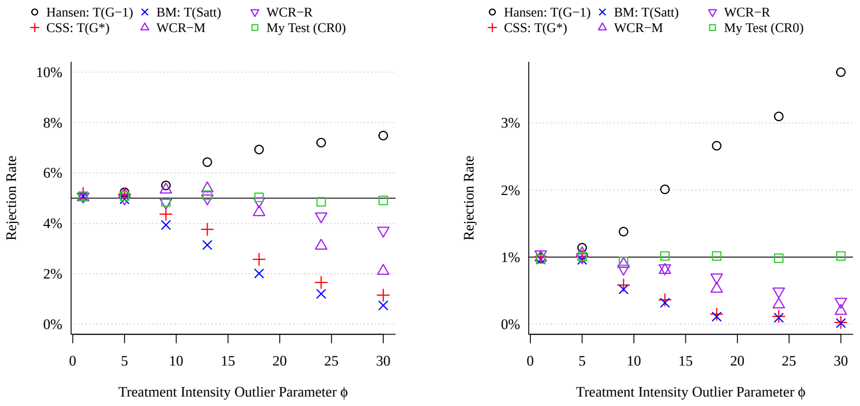
Since my test is asymptotically consistent, violations of Assumption 1 can only affect the performance of my test in samples with few effective clusters.

I focus on three violations of Assumption 1:

- Non-normal errors
- Serial correlation, where ϵ_{ig} is an AR(1) process
- Heteroskedasticity

These three violations correspond roughly to different parts of Assumption 1: normality, constant intracluster correlation, and homoskedasticity. It may be that normality of the errors can be relaxed when N_g, the number of observations per cluster, is large, and proving this is an area for future work. Still, I test robustness to non-normal errors here. Bertrand et al. (2004) highlight serial correlation as an important potential problem in differences-in-differences analyses of panel data, and CRVEs are powerful tools for addressing serial correlation. For that reason, it seems natural to check test performance when ϵ_{ig} is serially correlated as an AR(1) process. MacKinnon and Webb (2018) find that a simple analytic test using a CRVE is less reliable when the errors are heteroskedastic.

Figure 4: Comparison of Hypothesis Tests With Varying ϕ


(a) Test Size $\alpha = 0.05$

(b) Test Size $\alpha = 0.01$

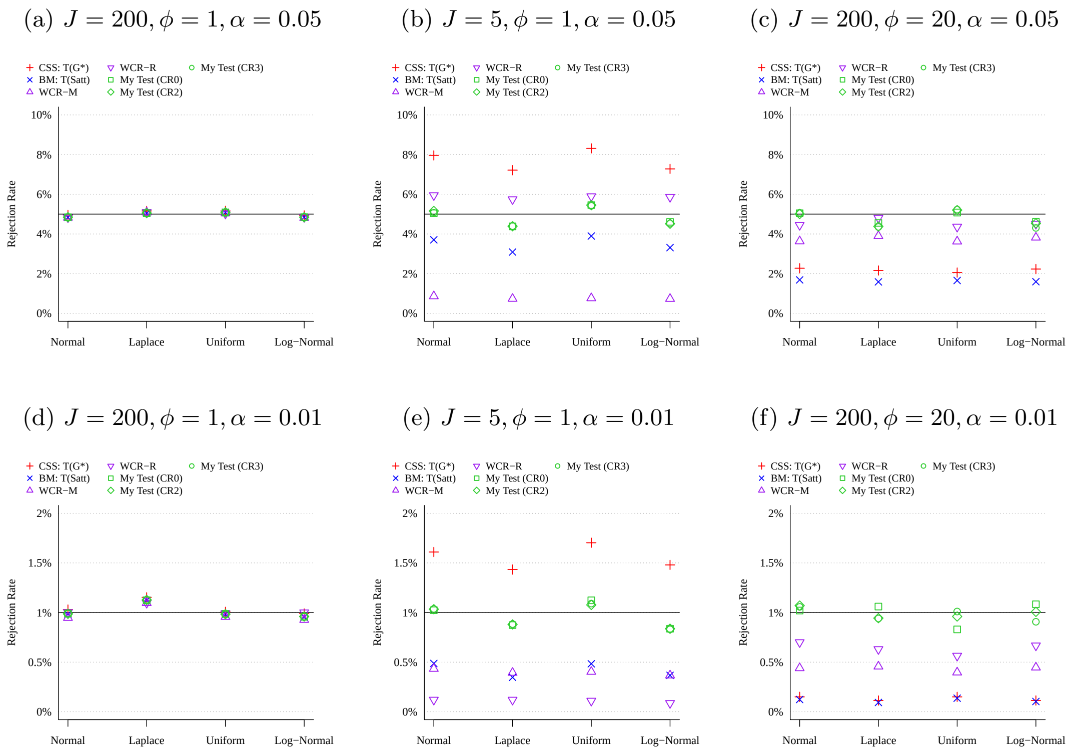
Notes: In these simulations, $G = 200$, $J = 200$, and $N_1 = 5$. The treatment variable x_{1ig} is distributed as $x_{1ig} = \mathbb{1}(g \leq J) \times \phi^{\mathbb{1}(g=1)} \times \frac{x^*_{1ig}-8}{4}$, where $x^*_{1ig} \sim \chi^2_8$. The errors are i.i.d. standard normal. Each specification has 30,000 simulations.

Following that paper as well as several other simulation studies of cluster-robust inference ((Cameron, Gelbach, et al., 2008; Djogbenou et al., 2019)), I also test robustness to the error variance differing across clusters.

In Figure 5, I plot rejection rates when the errors have a normal distribution and when they have non-normal distributions. I showed above that the degree of a given test's over- or under-rejection often depends on the test size α, so I continue to give results for both 1% and 5% tests here. I selected distributions with substantially different third and fourth moments than the normal distribution. The Laplace distribution is leptokurtic, the uniform distribution is platykurtic, and the log-normal distribution is skewed right. When $J = 200$ and $\phi = 1$ (panels 5a and 5d), so that there are many effective clusters, the different error distributions do not matter and every test rejects at the correct rate because all of the tests (including my test) are asymptotically consistent. When $J = 5$ (panels 5b and 5e) and when $\phi = 20$ (panels 5c and 5f), my test is not quite exact for non-normal error distributions, but it performs better than any of the other tests.

Next, I check test performance when the errors are serially correlated. Figure 6 shows the results of simulations where the errors are distributed as a stationary AR(1) process: $\epsilon_{i,g} = \psi\epsilon_{i-1,g} + \sqrt{1-\psi^2}\epsilon^*_{i,g}$, where $\epsilon^*_{i,g} \sim N(0,1)$. It bears repeating that when $J = 200$ and $\phi = 1$ (panels 6a and 6d), every test rejects at the correct rate because all of the tests (including my test) are asymptotically consistent. When $J = 5$ (panels 6b and 6e) and when $\phi = 20$ (panels 6c and 6f), my test is still nearly exact.

Figure 5: Comparison of Hypothesis Tests When the Error Distribution Varies



(a) $J = 200, \phi = 1, \alpha = 0.05$

(b) $J = 5, \phi = 1, \alpha = 0.05$

(c) $J = 200, \phi = 20, \alpha = 0.05$

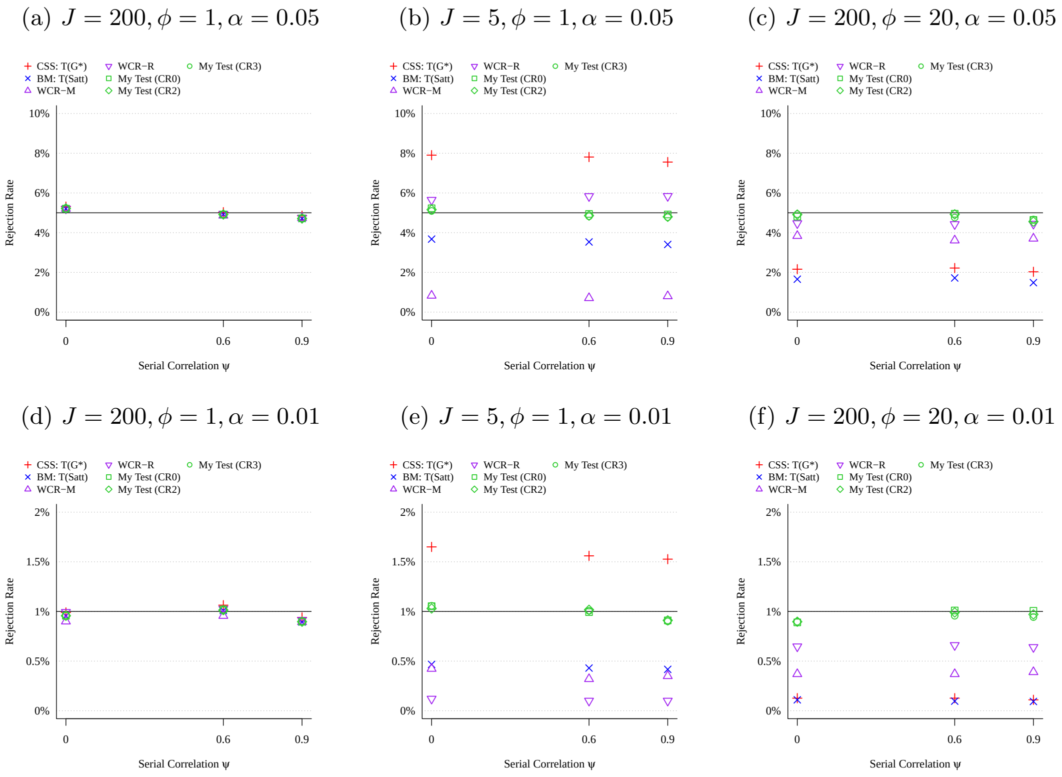
(d) $J = 200, \phi = 1, \alpha = 0.01$

(e) $J = 5, \phi = 1, \alpha = 0.01$

(f) $J = 200, \phi = 20, \alpha = 0.01$

Notes: In these simulations, $G = 200$ and $N_1 = 5$. The errors are distributed as recentered and rescaled normal, Laplace, uniform, and log-normal distributions such that $\mathbb{E}(\epsilon_{ig}) = 0$ and $V(\epsilon_{ig}) = 1$. Each specification has 30,000 simulations.

Figure 6: Comparison of Hypothesis Tests When the Errors are AR(1)


(a) $J = 200, \phi = 1, \alpha = 0.05$

(b) $J = 5, \phi = 1, \alpha = 0.05$

(c) $J = 200, \phi = 20, \alpha = 0.05$

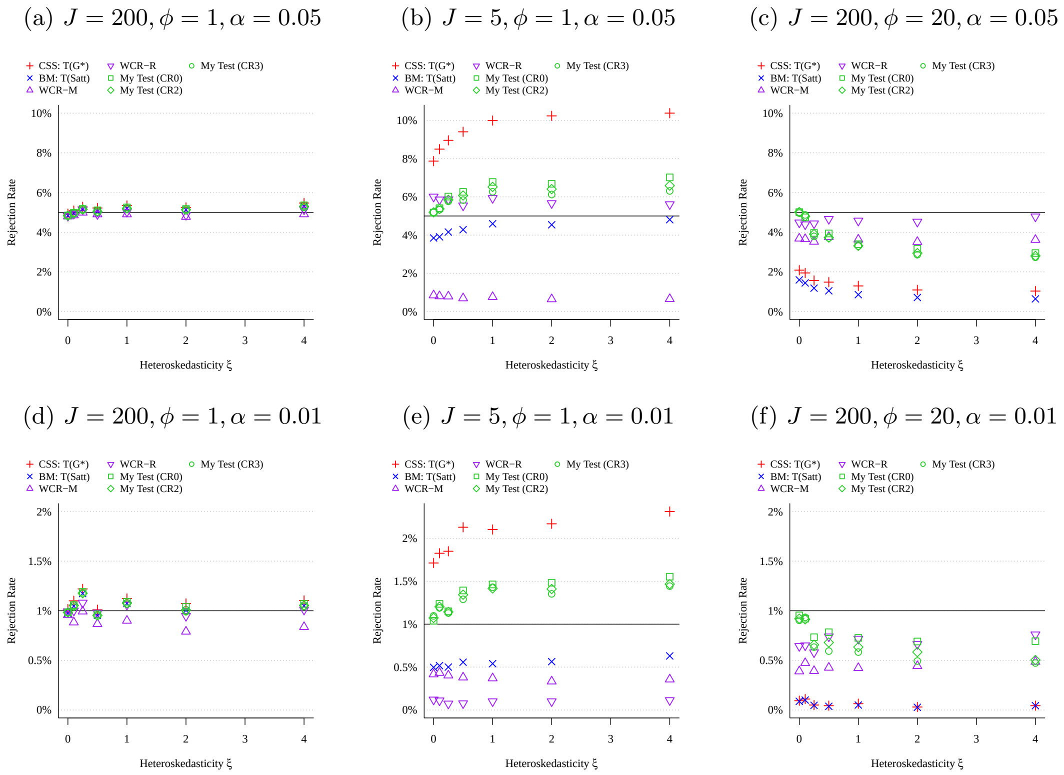
(d) $J = 200, \phi = 1, \alpha = 0.01$

(e) $J = 5, \phi = 1, \alpha = 0.01$

(f) $J = 200, \phi = 20, \alpha = 0.01$

Notes: In these simulations, $G = 200$ and $N_1 = 5$. The errors are distributed as $\epsilon_{i,g} = \psi\epsilon_{i-1,g} + \sqrt{1-\psi^2}\epsilon^*_{i,g}$, where $\epsilon^*_{i,g} \sim N(0,1)$. Each specification has 30,000 simulations.

Figure 7: Comparison of Hypothesis Tests When the Errors are Heteroskedastic



(a) $J = 200, \phi = 1, \alpha = 0.05$

(b) $J = 5, \phi = 1, \alpha = 0.05$

(c) $J = 200, \phi = 20, \alpha = 0.05$

(d) $J = 200, \phi = 1, \alpha = 0.01$

(e) $J = 5, \phi = 1, \alpha = 0.01$

(f) $J = 200, \phi = 20, \alpha = 0.01$

Notes: In these simulations, $G = 200$ and $N_1 = 5$. The errors are distributed as $\epsilon_{ig} = (1 + \xi(x^*_{1ig})^2))^{\frac{1}{2}}\epsilon^*_{ig}$, where $\epsilon^*_{ig} \sim N(0, 1)$. Each specification has 30,000 simulations.

Finally, in Figure 7, I show the rejection rates for heteroskedastic errors. Specifically, I use a variant of the error distribution used to explore heteroskedasticity in Djogbenou et al. (2019); I let $\epsilon_{ig} = (1 + \xi(x^*_{1ig})^2))^{\frac{1}{2}}\epsilon^*_{ig}$, where $\epsilon^*_{ig} \sim N(0,1)$. The error variance is higher for observations with greater magnitudes of x_{1ig}. When $J = 5$ (panels 7b and 7e), my test overrejects when ξ is high. When $\phi = 20$ (panels 7c and 7f), my test underrejects when ξ is high. Even in these cases, my test performs about as well as any other test.

My test is only plausibly vulnerable to violations of Assumption 1 when the sample is not asymptotically large. I have shown in this section that several straightforward violations of Assumption 1 don't seem to affect the performance of my test very much. In particular, my test is less affected by these violations than the other tests are affected by few effective clusters.

6 Empirical Application

Many applied economics papers utilize hypothesis tests with critical values based on asymptotic inference alone in settings with small effective sample sizes. Below, I describe a formal measure of effective sample size, and then I use that measure in a discussion of state panels and other natural experiments exploiting cross-state treatment variation. I show that these research designs, fundamental for measuring the impact of public policy in the United States, are both popular and sensitive to inference specification. Finally, I apply my test to the empirical setting from one such paper.

First, recall that previous hypothesis tests from the literature reject at incorrect rates when the effective sample size is small, which occurs when there are few clusters, few clusters with treatment variation, cluster size outliers, or treatment intensity outliers. Figure 8 shows rejection rates for a 5% test using my test and the previous tests in scenarios with these features. These were generated using simulations described in Appendix D. In Section 5, I present detailed simulation evidence comparing the performance of my test and previous tests. Figure 8 summarizes that work using a simpler data generating process. There is a pre-determined design matrix X, so that only the error terms vary across simulation instances within a scenario.

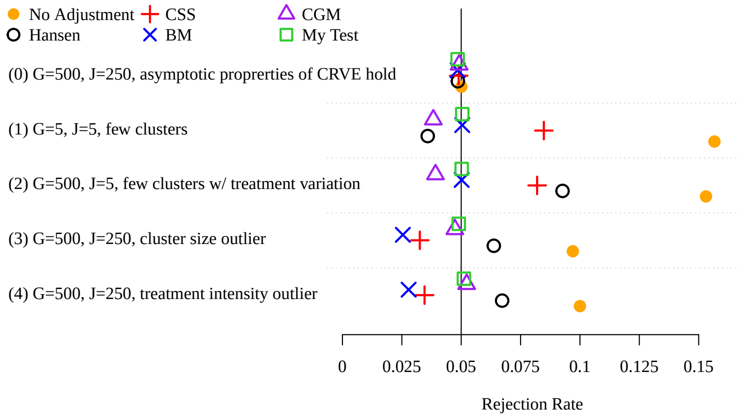
I engineered these different scenarios to highlight how different features of the design matrix cause different previous hypothesis tests fail. In order to ensure that the features would distort previous tests to a similar degree, I use the effective number of clusters measure G^* from Carter et al. (2017) to compare the intensity of different kinds of features:[7]

$$G^* = \frac{(\sum_g \gamma_g)^2}{\sum_g(\gamma_g^2)} \tag{5}$$

$$\text{where } \gamma_g = c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}_g^T\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0$$

[7] The definition in Carter et al. (2017) is slightly more complex, but it simplifies to equation 5 after fixed effects absorption.

Figure 8: Rejection Rates With Various Features Causing Small G^*



Notes: The data generating process for the design matrix X can be found in Appendix D. $G^* = 5$ by construction in scenarios (1)-(4). The errors are distributed $\epsilon_{ig} \sim N(0,1)$. Each specification has 30,000 simulations.

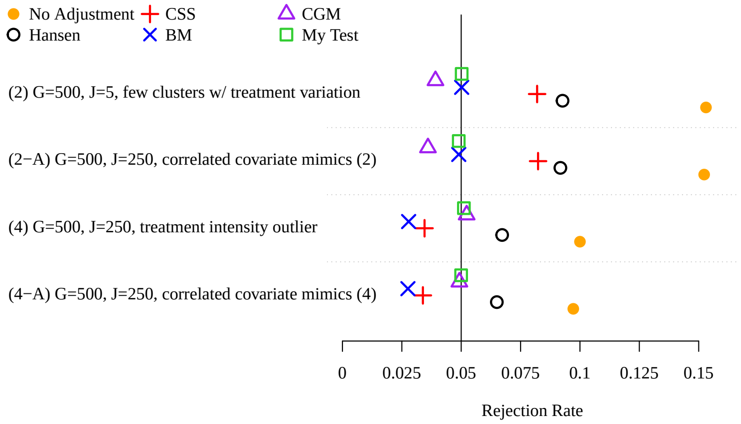
Two key facts about G^* are that it is always less than or equal to the actual number of clusters and that it is related to the asymptotic convergence of the distribution of the test statistic in a similar way to how the sample size is related to asymptotic convergence in i.i.d. samples (Carter et al., 2017).

I set the parameters of these simulations so that $G^* = 5$ in each scenario other than scenario (0). Scenario (0) is a baseline simulation that has a large effective size and where all tests reject at the correct rate. In the other scenarios, my test continues to reject at the correct rate, while previous tests have behavior that differs across scenarios.

These environments may seem easy to recognize. However, a design matrix can have a mix of the features that challenge previous cluster-robust inference, and those features may be obscured by the presence of covariates. Figure 9 demonstrates this with two additional scenarios. Scenario (2-A) has many clusters with treatment variation, but for nearly all clusters, the treatment variable is perfectly correlated with a covariate. The rejection rates are almost identical to Scenario (2). Scenario (4-A) has many clusters with the same treatment intensity, but the treatment variable is correlated with a covariate in a more complex pattern (see Appendix D for details). The rejection rates are almost identical to Scenario (4).

Applied economic studies often have research designs that are vulnerable to these potentially hidden features. One such design that is especially common involves treatment varying across U.S. states with standard errors clustered at the level of the state. These papers often (but not always) involve a policy implemented in a patchwork of states over time. I conducted a literature review to

Figure 9: Hidden Features



Notes: The data generating process for the design matrix X can be found in Appendix D. $G^* = 5$ by construction in all scenarios. The errors are distributed $\epsilon_{ig} \sim N(0,1)$. Each specification has 30,000 simulations.

surface a sample of these state studies, identifying papers that met the following criteria:

1. The paper was published between 2018 and 2022, inclusive.
2. The paper was published in the American Economic Journal: Economic Policy, the American Economic Journal: Applied Economics, or the Review of Economics and Statistics.[8]
3. One of the main specifications in the paper evaluated an empirical setting in the United States.
4. That specification used state-level fixed effects[9]
5. That specification clustered standard errors at the state level.

Among 66 total issues containing 859 articles meeting criteria (1) and (2), I identified 47 papers that additionally met criteria (3)-(5). I list them in Appendix Table E.1.

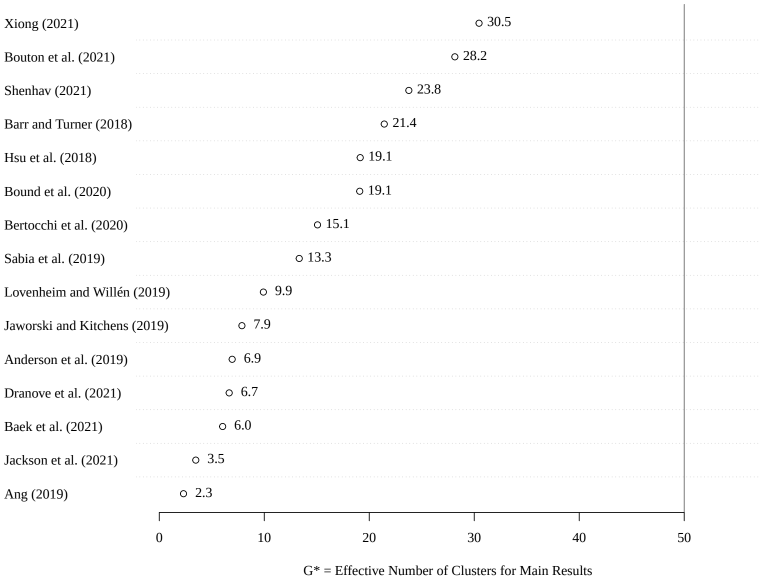
Many of these papers are vulnerable to incorrect inference because their effective sample size is small. Where feasible, I replicate the main result of each paper and calculate the effective number of clusters G^* for that specification.[10]

[8] I chose these journals because they are reputable journals that focus heavily or primarily on applied work and that have robust replication policies and archives.

[9] Specifications where other fixed effects are nested within state fixed effects are included.

[10] Many estimates could not be replicated because the data were not publicly accessible, because the programs were computationally expensive, or because the code could not be interpreted.

Figure 10: Effective Number of Clusters G^* in 50-State Studies



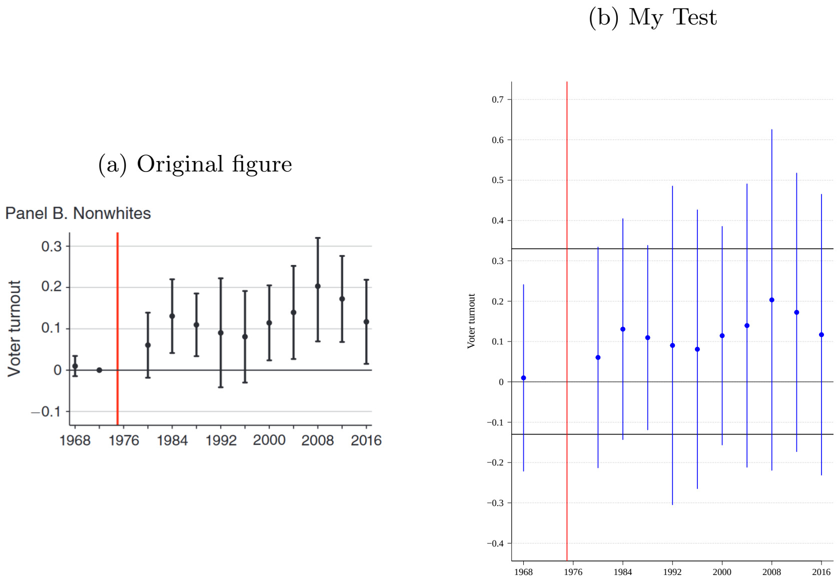
Notes: Includes all papers meeting criteria (1)-(5) where calculation of G^* was feasible. I refer to the papers loosely as 50-state studies to emphasize the contrast between the research design based on US states and the effective number of clusters G^*. The number of clusters in these studies G ranges from 12 to 51, with the largest including the District of Columbia.

Appendix Table E.1 notes 15 papers that were included in this G^* analysis. Since the design matrix varies by specification, Appendix Table E.1 also describes where to find the estimate corresponding to my G^* calculation. Where possible, I chose estimates that were mentioned in abstracts or introductions or estimates in tables with titles like "Main Results." Where there were different specifications to choose from, I chose specifications with fewer covariates, and when there were multiple treatment variables, I examine the one whose linear hypothesis leads to a smaller G^*.

To give a sense of the types of papers included, Appendix Table E.2 shows the subject areas (JEL codes) of the 32 AEJ: Applied and AEJ: Policy papers in the literature sample. These papers were primarily but not exclusively concentrated in the (overlapping) categories of "Health, Education, and Welfare," "Labor and Demographic Economics," and "Public Economics."

I calculate G^* for 15 papers. Figure 10 shows the results, in descending order of G^*. Because these papers all cluster by state, $G^* \leq 50$. Driven by a mix of the features described above, treatment variation is concentrated in a smaller number of clusters, and so $G^* < 35$ in every instance and $G^* < 20$ in 11 of the 15 papers. There is no discontinuous threshold below which inference using previous tests suddenly falls apart, but Figure 8 showed that by $G^* = 5$,

Figure 11: Confidence Intervals in Ang (2019)



(a) Original figure

(b) My Test

Notes: Panel 11a shows Panel B of Figure 4 from Ang (2019). Panel 11b shows an updated version using my test to calculate 95% confidence intervals. The y-axis boundaries of the original version are marked with black horizontal lines in my version.

the rejection rates for all previous tests are distorted by at least 10% in at least one scenario.

In two papers, Jackson et al. (2021) and Ang (2019), $G^* < 5$, so their hypothesis tests (and corresponding p-values and confidence intervals) are likely to be even more distorted than those in Figure 8. Why is G^* so low in these two papers? Jackson et al. (2021) has a small number of clusters with treatment variation. In that paper, the authors estimate the impact of spending cuts in schools during the Great Recession using an instrumental variables strategy that divides states into treatment groups based on the fraction of K-12 revenues that came from state appropriations in the 2007-2008 school year. One of the three treatment groups only contained 3 states (Illinois, Nebraska, and the District of Columbia), while another only contained 4 states (Arkansas, Hawaii, New Mexico, and Vermont). Jackson et al. (2021) do not report confidence intervals or p-values for the estimates I focused on.

Ang (2019) estimates the impact of the 1975 preclearance oversight amendment to the Voting Rights Act on voter turnout over the following decades. Column 3 of Table 3 estimates the impact for nonwhites. The treatment group consists of only 2 states: Arizona and Texas. Ang (2019) displays 95% confidence intervals graphically in Panel B of Figure 4 of that paper. Figure 11 shows the original figure alongside my replication that uses my test to generate confidence intervals. According to my test, none of the coefficients are significantly different from zero; on the contrary, all but two of the 95% confidence intervals contain the boundaries of the original graph.

I have shown that the important features of this empirical setting are not unique. Staggered changes in law and policy across states or counties provide useful natural experiments for learning about the impact of public policy. A good hypothesis test is necessary for understanding how much can be learned from a given experiment.

7 Conclusion

In this paper, I have proposed a hypothesis test for inference in clustered samples with cluster-level fixed effects. My test is robust to samples with few clusters, few clusters with treatment variation, cluster size outliers, or treatment intensity outliers. In addition to being consistent in large samples, my test is also exact under normal, homoskedastic errors, and in simulations it performs well compared to other methods from the literature under other assumptions about the errors.

Many samples are large enough that the conventional cluster-robust test will work fine, drawing critical values from the standard normal distribution or from a t-distribution. However, some samples are not as large as they seem, in the sense of the distribution of the test statistic.

Samples with many observations but few clusters require adjustment to the conventional test. Samples with many clusters but only a few with treatment variation require adjustment. And samples where the residual treatment variation (conditional on covariates) is concentrated in a small number of clusters require adjustment. Covariates can hide the features that change the behavior of the test statistic. It is worthwhile to use a test that is simply robust to these features.

References

Adhvaryu, Achyuta et al. (May 2020). "When It Rains It Pours: The Long-Run Economic Impacts of Salt Iodization in the United States". en. In: *The Review of Economics and Statistics* 102.2, pp. 395–407. DOI: 10.1162/rest_a_00822.

Allcott, Hunt and Charlie Rafkin (Nov. 2022). "Optimal Regulation of E-cigarettes: Theory and Evidence". en. In: *American Economic Journal: Economic Policy* 14.4, pp. 1–50. DOI: 10.1257/pol.20200805.

Alpert, Abby, David Powell, and Rosalie Liccardo Pacula (Nov. 2018). "Supply-Side Drug Policy in the Presence of Substitutes: Evidence from the Introduction of Abuse-Deterrent Opioids". en. In: *American Economic Journal: Economic Policy* 10.4, pp. 1–35. DOI: 10.1257/pol.20170082.

Anderson, D. Mark et al. (Apr. 2019). "Was the First Public Health Campaign Successful?" en. In: *American Economic Journal: Applied Economics* 11.2, pp. 143–175. DOI: 10.1257/app.20170411.

Ang, Desmond (July 2019). "Do 40-Year-Old Facts Still Matter? Long-Run Effects of Federal Oversight under the Voting Rights Act". en. In: *American Economic Journal: Applied Economics* 11.3, pp. 1–53. DOI: 10.1257/app.20170572.

Arellano, M. (Nov. 1987). "Computing Robust Standard Errors for Within-groups Estimators". en. In: *Oxford Bulletin of Economics and Statistics* 49.4, pp. 431–434. DOI: 10.1111/j.1468-0084.1987.mp49004006.x.

Baek, ChaeWon et al. (Dec. 2021). "Unemployment Effects of Stay-at-Home Orders: Evidence from High-Frequency Claims Data". en. In: *The Review of Economics and Statistics* 103.5, pp. 979–993. DOI: 10.1162/rest_a_00996.

Barr, Andrew and Sarah Turner (Aug. 2018). "A Letter and Encouragement: Does Information Increase Postsecondary Enrollment of UI Recipients?" en. In: *American Economic Journal: Economic Policy* 10.3, pp. 42–68. DOI: 10.1257/pol.20160570.

Bastian, Jacob (Aug. 2020). "The Rise of Working Mothers and the 1975 Earned Income Tax Credit". en. In: *American Economic Journal: Economic Policy* 12.3, pp. 44–75. DOI: 10.1257/pol.20180039.

Bell, Robert M. and Daniel F. McCaffrey (2002). "Bias reduction in standard errors for linear regression with multi-stage samples". In: *Survey Methodology* 28.2. ISBN: 0714-0045, pp. 169–182.

Bernecker, Andreas, Pierre C. Boyer, and Christina Gathmann (May 2021). "The Role of Electoral Incentives for Policy Innovation: Evidence from the US Welfare Reform". en. In: *American Economic Journal: Economic Policy* 13.2, pp. 26–57. DOI: 10.1257/pol.20190690.

Bertocchi, Graziella et al. (Aug. 2020). "Youth Enfranchisement, Political Responsiveness, and Education Expenditure: Evidence from the US". en. In: *American Economic Journal: Economic Policy* 12.3, pp. 76–106. DOI: 10.1257/pol.20180203.

Bertrand, M., E. Duflo, and S. Mullainathan (Feb. 2004). "How Much Should We Trust Differences-In-Differences Estimates?" en. In: *The Quarterly Journal of Economics* 119.1, pp. 249–275. DOI: 10.1162/003355304772839588.

Binder, Carola and Christos Makridis (Mar. 2022). "Stuck in the Seventies: Gas Prices and Consumer Sentiment". en. In: *The Review of Economics and Statistics* 104.2, pp. 293–305. DOI: 10.1162/rest_a_00944.

Borgschulte, Mark and Paco Martorell (July 2018). "Paying to Avoid Recession: Using Reenlistment to Estimate the Cost of Unemployment". en. In: *American Economic Journal: Applied Economics* 10.3, pp. 101–127. DOI: 10.1257/app.20160257.

Bound, John et al. (Feb. 2020). "A Passage to America: University Funding and International Students". en. In: *American Economic Journal: Economic Policy* 12.1, pp. 97–126. DOI: 10.1257/pol.20170620.

Bouton, Laurent et al. (Mar. 2021). “The Tyranny of the Single-Minded: Guns, Environment, and Abortion”. en. In: *The Review of Economics and Statistics* 103.1, pp. 48–59. DOI: 10.1162/rest_a_00897.

Buchmueller, Thomas C. and Colleen Carey (Feb. 2018). “The Effect of Prescription Drug Monitoring Programs on Opioid Utilization in Medicare”. en. In: *American Economic Journal: Economic Policy* 10.1, pp. 77–112. DOI: 10.1257/pol.20160094.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (Aug. 2008). “Bootstrap-Based Improvements for Inference with Clustered Errors”. en. In: *Review of Economics and Statistics* 90.3, pp. 414–427. DOI: 10.1162/rest.90.3.414.

Cameron, A. Colin and Douglas L. Miller (2015). “A Practitioner’s Guide to Cluster-Robust Inference”. en. In: *Journal of Human Resources* 50.2, pp. 317–372. DOI: 10.3368/jhr.50.2.317.

Carey, Colleen M., Sarah Miller, and Laura R. Wherry (Oct. 2020). “The Impact of Insurance Expansions on the Already Insured: The Affordable Care Act and Medicare”. en. In: *American Economic Journal: Applied Economics* 12.4, pp. 288–318. DOI: 10.1257/app.20190176.

Carpenter, Christopher S. and Emily C. Lawler (Feb. 2019). “Direct and Spillover Effects of Middle School Vaccination Requirements”. en. In: *American Economic Journal: Economic Policy* 11.1, pp. 95–125. DOI: 10.1257/pol.20170067.

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (Oct. 2017). “Asymptotic Behavior of a t -Test Robust to Cluster Heterogeneity”. en. In: *The Review of Economics and Statistics* 99.4, pp. 698–709. DOI: 10.1162/REST_a_00639.

Dickert-Conlin, Stacy, Todd Elder, and Keith Teltser (Oct. 2019). “Allocating Scarce Organs: How a Change in Supply Affects Transplant Waiting Lists and Transplant Recipients”. en. In: *American Economic Journal: Applied Economics* 11.4, pp. 210–239. DOI: 10.1257/app.20170476.

Djogbenou, Antoine A., James G. MacKinnon, and Morten Ørregaard Nielsen (Oct. 2019). “Asymptotic theory and wild bootstrap inference with clustered errors”. en. In: *Journal of Econometrics* 212.2, pp. 393–412. DOI: 10.1016/j.jeconom.2019.04.035.

Dranove, David, Christopher Ody, and Amanda Starc (Jan. 2021). “A Dose of Managed Care: Controlling Drug Spending in Medicaid”. en. In: *American Economic Journal: Applied Economics* 13.1, pp. 170–197. DOI: 10.1257/app.20190165.

Dube, Arindrajit (Oct. 2019). “Minimum Wages and the Distribution of Family Incomes”. en. In: *American Economic Journal: Applied Economics* 11.4, pp. 268–304. DOI: 10.1257/app.20170085.

Evans, William N., Ethan M. J. Lieber, and Patrick Power (Mar. 2019). “How the Reformulation of OxyContin Ignited the Heroin Epidemic”. en. In: *The Review of Economics and Statistics* 101.1, pp. 1–15. DOI: 10.1162/rest_a_00755.

Ganapati, Sharat, Joseph S. Shapiro, and Reed Walker (Apr. 2020). "Energy Cost Pass-Through in US Manufacturing: Estimates and Implications for Carbon Taxes". en. In: *American Economic Journal: Applied Economics* 12.2, pp. 303–342. DOI: 10.1257/app.20180474.

Ganong, Peter and Jeffrey B. Liebman (Nov. 2018). "The Decline, Rebound, and Further Rise in SNAP Enrollment: Disentangling Business Cycle Fluctuations and Policy Changes". en. In: *American Economic Journal: Economic Policy* 10.4, pp. 153–176. DOI: 10.1257/pol.20140016.

Garthwaite, Craig, Tal Gross, and Matthew J. Notowidigdo (Jan. 2018). "Hospitals as Insurers of Last Resort". en. In: *American Economic Journal: Applied Economics* 10.1, pp. 1–39. DOI: 10.1257/app.20150581.

Goldin, Jacob, Tatiana Homonoff, and Katherine Meckel (Feb. 2022). "Issuance and Incidence: SNAP Benefit Cycles and Grocery Prices". en. In: *American Economic Journal: Economic Policy* 14.1, pp. 152–178. DOI: 10.1257/pol.20190777.

Hansen, Bruce E. (2021). *Econometrics*. Princeton, New Jersey: Princeton University Press.

Hansen, Christian B. (Dec. 2007). "Asymptotic properties of a robust variance matrix estimator for panel data when is large". en. In: *Journal of Econometrics* 141.2, pp. 597–620. DOI: 10.1016/j.jeconom.2006.10.009.

Hausman, Naomi and Kurt Lavetti (Apr. 2021). "Physician Practice Organization and Negotiated Prices: Evidence from State Law Changes". en. In: *American Economic Journal: Applied Economics* 13.2, pp. 258–296. DOI: 10.1257/app.20180078.

Hsu, Po-Hsuan et al. (Oct. 2018). "Natural Disasters, Technology Diversity, and Operating Performance". en. In: *The Review of Economics and Statistics* 100.4, pp. 619–630. DOI: 10.1162/rest_a_00738.

Imbens, Guido W. and Michal Kolesár (Oct. 2016). "Robust Standard Errors in Small Samples: Some Practical Advice". en. In: *Review of Economics and Statistics* 98.4, pp. 701–712. DOI: 10.1162/REST_a_00552.

Imhof, J. P. (Dec. 1961). "Computing the Distribution of Quadratic Forms in Normal Variables". In: *Biometrika* 48.3/4, p. 419. DOI: 10.2307/2332763.

Jackson, C. Kirabo, Cora Wigger, and Heyu Xiong (May 2021). "Do School Spending Cuts Matter? Evidence from the Great Recession". en. In: *American Economic Journal: Economic Policy* 13.2, pp. 304–335. DOI: 10.1257/pol.20180674.

Jaworski, Taylor and Carl T. Kitchens (Dec. 2019). "National Policy for Regional Development: Historical Evidence from Appalachian Highways". en. In: *The Review of Economics and Statistics* 101.5, pp. 777–790. DOI: 10.1162/rest_a_00808.

Johnson, Janna E. and Morris M. Kleiner (Aug. 2020). "Is Occupational Licensing a Barrier to Interstate Migration?" en. In: *American Economic Journal: Economic Policy* 12.3, pp. 347–373. DOI: 10.1257/pol.20170704.

Johnson, Rucker C. and C. Kirabo Jackson (Nov. 2019). "Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public

School Spending". en. In: *American Economic Journal: Economic Policy* 11.4, pp. 310–349. DOI: 10.1257/pol.20180510.

Karaivanov, Alexander et al. (July 2021). "Face masks, public policies and slowing the spread of COVID-19: Evidence from Canada". en. In: *Journal of Health Economics* 78, p. 102475. DOI: 10.1016/j.jhealeco.2021.102475.

Kose, Esra, Elira Kuka, and Na'ama Shenhav (Aug. 2021). "Women's Suffrage and Children's Education". en. In: *American Economic Journal: Economic Policy* 13.3, pp. 374–405. DOI: 10.1257/pol.20180677.

Kroft, Kory et al. (Feb. 2020). "Optimal Income Taxation with Unemployment and Wage Responses: A Sufficient Statistics Approach". en. In: *American Economic Journal: Economic Policy* 12.1, pp. 254–292. DOI: 10.1257/pol.20180033.

Kuka, Elira (June 2020). "Quantifying the Benefits of Social Insurance: Unemployment Insurance and Health". en. In: *The Review of Economics and Statistics* 102.3, pp. 490–505. DOI: 10.1162/rest_a_00865.

Kuka, Elira, Na'ama Shenhav, and Kevin Shih (Feb. 2020). "Do Human Capital Decisions Respond to the Returns to Education? Evidence from DACA". en. In: *American Economic Journal: Economic Policy* 12.1, pp. 293–324. DOI: 10.1257/pol.20180352.

Kuziemko, Ilyana, Katherine Meckel, and Maya Rossin-Slater (Aug. 2018). "Does Managed Care Widen Infant Health Disparities? Evidence from Texas Medicaid". en. In: *American Economic Journal: Economic Policy* 10.3, pp. 255–283. DOI: 10.1257/pol.20150262.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach (Apr. 2018). "School Finance Reform and the Distribution of Student Achievement". en. In: *American Economic Journal: Applied Economics* 10.2, pp. 1–26. DOI: 10.1257/app.20160567.

Leung, Justin H. (Aug. 2021). "Minimum Wage and Real Wage Inequality: Evidence from Pass-Through to Retail Prices". en. In: *The Review of Economics and Statistics*, pp. 1–16. DOI: 10.1162/rest_a_00915.

Liang, Kung-Yee and Scott L. Zeger (1986). "Longitudinal data analysis using generalized linear models". en. In: *Biometrika* 73.1, pp. 13–22. DOI: 10.1093/biomet/73.1.13.

Lovenheim, Michael F. and Alexander Willén (Aug. 2019). "The Long-Run Effects of Teacher Collective Bargaining". en. In: *American Economic Journal: Economic Policy* 11.3, pp. 292–324. DOI: 10.1257/pol.20170570.

MacKinnon, James G. and Matthew D. Webb (June 2018). "The wild bootstrap for few (treated) clusters". en. In: *The Econometrics Journal* 21.2, pp. 114–135. DOI: 10.1111/ectj.12107.

Mayda, Anna Maria, Giovanni Peri, and Walter Steingress (Jan. 2022). "The Political Impact of Immigration: Evidence from the United States". en. In: *American Economic Journal: Applied Economics* 14.1, pp. 358–389. DOI: 10.1257/app.20190081.

Modestino, Alicia Sasser, Daniel Shoag, and Joshua Ballance (Oct. 2020). "Upskilling: Do Employers Demand Greater Skill When Workers Are Plentiful?"

en. In: *The Review of Economics and Statistics* 102.4, pp. 793–805. DOI: 10.1162/rest_a_00835.

Myers, Caitlin Knowles (Dec. 2017). “The Power of Abortion Policy: Reexamining the Effects of Young Women’s Access to Reproductive Control”. en. In: *Journal of Political Economy* 125.6, pp. 2178–2224. DOI: 10.1086/694293.

Niccodemi, Gianmaria et al. (2020). *Refining clustered standard errors with few clusters*. English. WorkingPaper. University of Groningen, SOM research school.

Renkin, Tobias, Claire Montialoux, and Michael Siegenthaler (Sept. 2022). “The Pass-Through of Minimum Wages into U.S. Retail Prices: Evidence from Supermarket Scanner Data”. en. In: *The Review of Economics and Statistics* 104.5, pp. 890–908. DOI: 10.1162/rest_a_00981.

Sabia, Joseph J., M. Melinda Pitts, and Laura M. Argys (Mar. 2019). “Are Minimum Wages a Silent Killer? New Evidence on Drunk Driving Fatalities”. en. In: *The Review of Economics and Statistics* 101.1, pp. 192–199. DOI: 10.1162/rest_a_00761.

Satterthwaite, F. E. (Dec. 1946). “An Approximate Distribution of Estimates of Variance Components”. In: *Biometrics Bulletin* 2.6, p. 110. DOI: 10.2307/3002019.

Shenhav, Na’ama (May 2021). “Lowering Standards to Wed? Spouse Quality, Marriage, and Labor Market Responses to the Gender Wage Gap”. en. In: *The Review of Economics and Statistics* 103.2, pp. 265–279. DOI: 10.1162/rest_a_00919.

Siemer, Michael (Mar. 2019). “Employment Effects of Financial Constraints during the Great Recession”. en. In: *The Review of Economics and Statistics* 101.1, pp. 16–29. DOI: 10.1162/rest_a_00733.

Stuart, Bryan A. (Jan. 2022). “The Long-Run Effects of Recessions on Education and Income”. en. In: *American Economic Journal: Applied Economics* 14.1, pp. 42–74. DOI: 10.1257/app.20180055.

White, Halbert (1984). *Asymptotic Theory for Econometricians*. Economic Theory, Econometrics, and Mathematical Economics. Orlando, Florida: Academic Press.

Wilson, Riley (Sept. 2022). “The Impact of Social Networks on EITC Claiming Behavior”. en. In: *The Review of Economics and Statistics* 104.5, pp. 929–945. DOI: 10.1162/rest_a_00995.

Xiong, Heyu (Oct. 2021). “The Political Premium of Television Celebrity”. en. In: *American Economic Journal: Applied Economics* 13.4, pp. 1–33. DOI: 10.1257/app.20190147.

A Proof of Theorem 1

The null hypothesis H_0 is true if $c_0^T\beta = a_0$. Assumption 1 holds when $\epsilon_g \sim N(0, \sigma^2 I_g)$ (σ may be unknown).

Theorem 1 says that when H_0 is true and Assumption 1 holds, the CDF of the squared test statistic is a known function of the design matrix X and the hypothesis H_0:

$$P(t_0^2 < q \mid X) = L(q; X, H_0)$$

In this section, I will prove Theorem 1 by deriving $L(.;.,.)$. I give some notation to help with the analysis, I prove that $L(.;.,.)$ is known in principle, and I show how $L(.;.,.)$ can be calculated quickly in practice.

A.1 Notation

Let N_g be the number of observations in cluster g, and let $N = \sum_g N_g$. Then let I_g be an identity matrix of size N_g, and let ι_g be a column vector of length N_g whose elements are all 1. So then let $M_g = I_g - \frac{1}{N_g}\iota_g\iota_g^T$, and note that $M_g\iota_g = 0$.

Recall that:

$$y_{ig} = x_{ig}\beta + \gamma_g + \epsilon_{ig}$$

For identification, we have that $\mathbb{E}(\epsilon_{ig} \mid x_{jg}) = 0$, and for inference, we have that $\mathbb{E}(\epsilon_{ig}\epsilon_{jg'}) = 0$. Y_g is $(N_g \times 1)$, stacking up the dependent variable within cluster g, and X_g is $(N_g \times K)$, stacking up the covariates within cluster g. Then $\epsilon_g = Y_g - X_g\beta - \gamma_g\iota_g$.

For fixed effects absorption – that is, to absorb γ_g – the cluster-level means of y_{ig} and x_{ig} are subtracted from the individual-level y_{ig} and x_{ig}, respectively. Stated another way:

$$\begin{aligned}
\ddot{Y}_g &= M_g Y_g \\
\ddot{X}_g &= M_g X_g \\
\ddot{\epsilon}_g &= M_g\epsilon_g = M_g(Y_g - X_g - \gamma_g\iota_g) = \ddot{Y}_g - \ddot{X}_g\beta
\end{aligned}$$

Also recall that $\hat{\beta}$ is the fixed effects estimator of β, and $\hat{V}(\hat{\beta})$ is the CRVE:

$$\begin{aligned}
\hat{\beta} &= (\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\ddot{Y} \\
\hat{\epsilon}_g &= \ddot{Y}_g - \ddot{X}_g\hat{\beta} \\
\hat{V}(\hat{\beta}) &= (\ddot{X}^T\ddot{X})^{-1}\left(\sum_g \ddot{X}_g^T\hat{\epsilon}_g\hat{\epsilon}_g^T\ddot{X}_g\right)(\ddot{X}^T\ddot{X})^{-1}
\end{aligned}$$

Finally, recall the hypothesis H_0 and the test statistic t_0:

$$\begin{aligned}
H_0 &: c_0^T\beta = a_0 \\
t_0 &= \frac{c_0^T\hat{\beta} - a_0}{\sqrt{c_0^T\hat{V}(\hat{\beta})c_0}}
\end{aligned}$$

A.2 Main Proof

First, I show that the CDF of t_0^2 at a particular quantile q can be written as the CDF at 0 of a linear combination of χ^2 random variables. Second, I show that it is possible to determine the coefficients of that linear combination. Third, I give a formula for $L(q; X, H_0)$.

Since the hypothesis H_0 holds:

$$\begin{aligned}
t_0^2 &= \frac{(c_0^T\hat{\beta} - a_0)^2}{c_0^T\hat{V}(\hat{\beta})c_0} \\
&= \frac{(c_0^T(\hat{\beta} - \beta))^2}{c_0^T\hat{V}(\hat{\beta})c_0} \\
&= \frac{c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\ddot{\epsilon}\ddot{\epsilon}^T\ddot{X}(\ddot{X}^T\ddot{X})^{-1}c_0}{c_0^T(\ddot{X}^T\ddot{X})^{-1}\left(\sum_g \ddot{X}_g^T\hat{\epsilon}_g\hat{\epsilon}_g^T\ddot{X}_g\right)(\ddot{X}^T\ddot{X})^{-1}c_0}
\end{aligned}$$

Let $H = \ddot{X}(\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T$, let I be an identity matrix of size N, let $(I - H)_g$ be the rows of $(I - H)$ corresponding to cluster g. Then I continue:

$$\begin{aligned}
t_0^2 &= \frac{c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\ddot{\epsilon}\ddot{\epsilon}^T\ddot{X}(\ddot{X}^T\ddot{X})^{-1}c_0}{c_0^T(\ddot{X}^T\ddot{X})^{-1}\left(\sum_g \ddot{X}_g^T(I-H)_g\ddot{\epsilon}\ddot{\epsilon}^T(I-H)_g^T\ddot{X}_g\right)(\ddot{X}^T\ddot{X})^{-1}c_0} \\
&= \frac{\ddot{\epsilon}^T\ddot{X}(\ddot{X}^T\ddot{X})^{-1}c_0c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\ddot{\epsilon}}{\ddot{\epsilon}^T\left(\sum_g (I-H)_g^T\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}_g^T(I-H)_g\right)\ddot{\epsilon}}
\end{aligned}$$

Now let $d_0 = \ddot{X}(\ddot{X}^T\ddot{X})^{-1}c_0$, and for $g \geq 1$, let $d_g = (I-H)_g^T\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0$. Also, let M be an $(N \times N)$ block-diagonal matrix where the g-th block is M_g. Note here that, since M_g is idempotent:

$$\begin{aligned}
\ddot{X}_g^T M_g &= X_g^T M_g M_g = \ddot{X}_g^T \\
\ddot{X}_g^T(I-H)_g M &= X_g^T M_g M_g - \ddot{X}_g\ddot{X}_g^T(\ddot{X}^T\ddot{X})^{-1}X^T M M \\
&= \ddot{X}_g^T(I-H)_g \\
d_0^T M &= d_0^T \\
d_g^T M &= d_g^T
\end{aligned}$$

Then it follows that:

$$\begin{aligned}
t_0^2 &= \frac{\ddot{\epsilon}^T d_0 d_0^T \ddot{\epsilon}}{\ddot{\epsilon}^T \left(\sum_g d_g d_g^T\right) \ddot{\epsilon}} \\
P(t_0^2 < q \mid X) &= P\left(\frac{\ddot{\epsilon}^T d_0 d_0^T \ddot{\epsilon}}{\ddot{\epsilon}^T \left(\sum_g d_g d_g^T\right) \ddot{\epsilon}} < q \mid X\right) \\
&= P\left(\frac{1}{q}\ddot{\epsilon}^T \left(d_0 d_0^T\right) \ddot{\epsilon} - \ddot{\epsilon}^T \left(\sum_g d_g d_g^T\right) \ddot{\epsilon} < 0 \mid X\right) \\
&= P\left(\ddot{\epsilon}^T \left(\frac{1}{q} d_0 d_0^T - \sum_g d_g d_g^T\right) \ddot{\epsilon} < 0 \mid X\right)
\end{aligned}$$

then

$$P(t_0^2 < q \mid X) = P\left(\ddot{\epsilon}^T \left(\frac{1}{q} d_0 d_0^T - \sum_g d_g d_g^T\right) \ddot{\epsilon} < 0 \mid X\right) \tag{6}$$

Now using Assumption 1, the errors before fixed effects absorption are distributed $\epsilon_g \sim N(0, \sigma^2 I_g)$. Then the errors after absorption, $\ddot{\epsilon}_g = M_g \epsilon_g$, are distributed $\ddot{\epsilon}_g \sim N(0, \sigma^2 M_g M_g)$. Let $D_+ = [d_0 \quad d_1 \ldots d_g \ldots d_G]$ and $D_- = [\frac{1}{q} d_0 \quad -d_1 \cdots - d_g \cdots - d_G]$. Then let η be joint normal with $\eta \sim N(0, I)$, such that $\ddot{\epsilon} = \sigma M \eta$. Substituting this into (6):

$$\begin{aligned}
L(q; X, H_0) &= P(t_0^2 < q \mid X) \\
&= P\left(\eta^T M \sigma \left(\frac{1}{q} d_0 d_0^T - \sum_g d_g d_g^T\right) \sigma M \eta < 0 \mid X\right) \\
&= P\left(\eta^T M \left(\frac{1}{q} d_0 d_0^T - \sum_g d_g d_g^T\right) M \eta < 0 \mid X\right) \\
&= P\left(\eta^T D_+ D_-^T \eta < 0 \mid X\right)
\end{aligned}$$

At this point, the right-hand side is a function only of known quantities and random variables with known distributions. The unknown σ does not appear; the behavior of the test statistic t_0 does not depend on it.

A.3 Calculating $L(q; X, H_0)$ in Practice

I have shown that $L(q; X, H_0) = P\left(\eta^T D_+ D_-^T \eta < 0 \mid X\right)$. In this section, I will explain a method for calculating $L(q; X, H_0)$ quickly in practice.

Let $Q_{(N \times N)} = D_+ D_-^T$. Then let S be the orthogonal[11] matrix of eigenvectors of Q, let λ^* be an $(N \times 1)$ column vector whose elements are the eigenalues of

[11] An orthogonal S can always be found because $Q = \left(\frac{1}{q} d_0 d_0^T - \sum_g d_g d_g^T\right)$ is symmetric.

Q, and let Λ be an $(N \times N)$ diagonal matrix whose diagonal elements are also the eigenvalues of Q. Note that since S is orthogonal, $S\eta \sim \eta$. Then:

$$\begin{aligned} P(t_0^2 < q \mid X) &= P\left(\eta^T S \Lambda S^T \eta < 0 \mid X\right) \\ &= P\left(\eta^T \Lambda \eta < 0 \mid X\right) \end{aligned}$$

Let w be an N-vector of independent random variables such that $\forall i, w_i \sim \chi_1^2$. Then:

$$P(t_0^2 < q \mid X) = P\left(w^T \lambda^* < 0 \mid X\right) \tag{7}$$

Thus, I have shown that the CDF of t_0^2 at q can be written as the CDF at 0 of a linear combination of independent χ_1^2 random variables. Next, I find the non-zero elements of λ^*; it will be the case that λ^* has no more than $G+1$ non-zero elements.

In principle, the vector of eigenvalues λ^* can be found by eigendecomposing Q. However, since Q is $(N \times N)$, that might be inconvenient in practice. Instead, it is sufficient to find the non-zero eigenvalues of $D_-^T D_+$, which are the same as the non-zero eigenvalues of $Q = D_+ D_-^T$.

To see why, suppose that λ_j is a non-zero eigenvalue of $D_+ D_-^T$ corresponding to the eigenvector s_j. Therefore:

$$\begin{aligned} D_+ D_-^T s_j &= \lambda_j s_j \\ D_-^T D_+ D_-^T s_j &= D_-^T \lambda_j s_j \\ D_-^T D_+ (D_-^T s_j) &= \lambda_j (D_-^T s_j) \end{aligned}$$

Thus, λ_j is an eigenvalue of $D_-^T D_+$ corresponding to the eigenvector $D_-^T s_j$. And since $D_-^T D_+$ is a $(G+1 \times G+1)$ matrix, it has no more than $G+1$ non-zero eigenvalues. Letting λ be a $(G+1 \times 1)$ vector whose elements are the eigenvalues of $D_-^T D_+$, and in an abuse of notation letting w now be $(G+1 \times 1)$, we have that:

$$P(t_0^2 < q \mid X) = P\left(w^T \lambda < 0 \mid X\right) \tag{8}$$

Introducing yet more notation:

$$\begin{aligned} \chi_g &= \ddot{X}_g^T \ddot{X}_g \\ \delta_g &= (\ddot{X}^T \ddot{X})^{-\frac{1}{2}} \chi_g (\ddot{X}^T \ddot{X})^{-1} c_0 \\ \omega_g &= c_0^T (\ddot{X}^T \ddot{X})^{-1} \chi_g (\ddot{X}^T \ddot{X})^{-1} c_0 \\ \Delta &= \begin{bmatrix} \delta_1 & \delta_2 & ... & \delta_G \end{bmatrix} \\ \Omega &= \begin{bmatrix} \omega_1 & 0 & ... & 0 \\ 0 & \omega_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \omega_G \end{bmatrix} \end{aligned}$$

Then $D_-^T D_+$ can be further simplified because:

$$\begin{aligned}
d_0^T d_0 &= c_0^T (\ddot{X}^T \ddot{X})^{-1} \ddot{X}^T \ddot{X} (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&= c_0^T (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&\forall g > 0,\ d_0^T d_g = \\
&= c_0^T (\ddot{X}^T \ddot{X})^{-1} \ddot{X}^T (I - H)_g^T \ddot{X}_g S_g (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&= c_0^T (\ddot{X}^T \ddot{X})^{-1} \ddot{X}^T \left(\begin{bmatrix} 0 \\ I_g \\ 0 \end{bmatrix} - \ddot{X} (\ddot{X}^T \ddot{X})^{-1} \ddot{X}_g^T \right) \ddot{X}_g S_g (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&= c_0^T (\ddot{X}^T \ddot{X})^{-1} \left(\ddot{X}_g^T - \ddot{X}^T \ddot{X} (\ddot{X}^T \ddot{X})^{-1} \ddot{X}_g^T \right)_g^T \ddot{X}_g (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&= c_0^T (\ddot{X}^T \ddot{X})^{-1} (0) \ddot{X}_g (\ddot{X}^T \ddot{X})^{-1} c_0 = 0 \\
d_g^T d_g &= c_0^T (\ddot{X}^T \ddot{X})^{-1} (\ddot{X}_g^T \ddot{X}_g) (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&\quad - c_0^T (\ddot{X}^T \ddot{X})^{-1} (\ddot{X}_g^T \ddot{X}_g) (\ddot{X}^T \ddot{X})^{-1} (\ddot{X}_g^T \ddot{X}_g) (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&= \omega_g - \delta_g^T \delta_g
\end{aligned}$$

And for $g \neq g^{'}$,

$$\begin{aligned}
d_{g'}^T d_g &= -c_0^T (\ddot{X}^T \ddot{X})^{-1} (\ddot{X}_{g'}^T \ddot{X}_{g'}) (\ddot{X}^T \ddot{X})^{-1} (\ddot{X}_g^T \ddot{X}_g) (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&= \delta_{g'}^T \delta_g
\end{aligned}$$

Therefore:

$$D_-^T D_+ = \begin{bmatrix} \frac{1}{q} c_0^T (\ddot{X}^T \ddot{X})^{-1} c_0 & 0 \\ 0 & \Delta^T \Delta - \Omega \end{bmatrix}$$

The CDF of a linear combination of independent χ_1^2 random variables $w^T \lambda$ is given by Imhof (1961)[12]:

$$P\left(w^T \lambda < 0 \mid X\right) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\sin\left(\frac{1}{2} \sum_{j=1}^{G+1} \tan^{-1}\left(\lambda_j u\right)\right)}{u \prod_{j=1}^{G+1} \left(1 + \lambda_j^2 u^2\right)^{\frac{1}{4}}} du$$

So the CDF of t_0^2 at q can be written as the CDF at 0 of a linear combination of $G+1$ independent χ_1^2 random variables, and it is possible to calculate the coefficients λ as a function of the design matrix X, the hypothesis H_0, and the

[12] This can be calculated quickly by numerical integration, with a high degree of precision, using the *imhof()* function from the R package *CompQuadForm*.

quantile q. In summary:

$$\begin{aligned}
&\chi_g = \ddot{X}_g^T \ddot{X}_g \\
&\delta_g = (\ddot{X}^T \ddot{X})^{-\frac{1}{2}} \chi_g (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&\omega_g = c_0^T (\ddot{X}^T \ddot{X})^{-1} \chi_g (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&\Delta = \begin{bmatrix} \delta_1 & \delta_2 & ... & \delta_G \end{bmatrix} \\
&\Omega = \begin{bmatrix} \omega_1 & 0 & ... & 0 \\ 0 & \omega_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \omega_G \end{bmatrix} \\
&\lambda \text{ are } \frac{1}{q} c_0^T (X^T X)^{-1} c_0 \text{ and the eigenvalues of } \Delta^T \Delta - \Omega, \text{ and}
\end{aligned}$$

$$\begin{aligned}
L(q; X, H_0) &= P(t_0^2 < q \mid X) \\
&= \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\sin\left(\frac{1}{2} \sum_{j=1}^{G+1} \tan^{-1}(\lambda_j u)\right)}{u \prod_{j=1}^{G+1} \left(1 + \lambda_j^2 u^2\right)^{\frac{1}{4}}} du
\end{aligned}$$

And this is what I set out to find.

My test involves selecting a critical value:

$$q^*(\alpha; X, H_0) = \sqrt{L^{-1}(1 - \alpha; X, H_0)}$$

And I reject H_0 if $|t_0|> q^*$. Thus, under Assumption 1, my test is exact, so that the rate at which a true hypothesis is rejected is equal to the nominal size of the test:

$$\begin{aligned}
P\left(|t_0|> q^*\right) &= P\left(t_0^2 > (q^*)^2\right) = 1 - L((q^*)^2; X, H_0) \\
&= \alpha
\end{aligned}$$

B Asymptotic Consistency

In this section, I show that my test is asymptotically consistent (without Assumption 1). This result draws on the first theorem from Carter et al. (2017).

The null hypothesis H_0 is true if $c_0^T \beta = a_0$. Assumption 2 requires that the errors have fourth moments, and Assumption 3 requires that the observations aren't too concentrated in a small number of clusters.

Let $\alpha^* = P\left(|t_0| > q^*(\alpha; X, H_0)\right)$ be the rejection rate of my test – the rate at which my test rejects a true hypothesis H_0.

Theorem 2 says that when H_0 is true and Assumptions 2 and 3 hold, the rejection rate of my test converges to the nominal size of the test:

$$\alpha^* \xrightarrow{p} \alpha$$

Recall that $\mathbb{E}(\epsilon_{ig}\epsilon_{jg'}) = 0$, so that the errors are uncorrelated across clusters. According to the first theorem in Carter et al. (2017), it follows that $t_0 \xrightarrow{d} N(0,1)$.

Now, consider a counterfactual data generating process:

$$\tilde{y}_{ig} = x_{ig}\beta + \gamma_g + \tilde{\epsilon}_{ig}$$

where $\tilde{\epsilon} \sim N(0, I)$, so that the counterfactual errors are normal, i.i.d., and homoskedastic. Let $\tilde{t_0}$ be the test statistic that would be generated for H_0 using $\hat{V}_{CR0}$ in the counterfactual data generating process where the errors are $\tilde{\epsilon}$ rather than ϵ:

$$\begin{aligned}
\tilde{\beta} &= \beta + (\ddot{X}^T \ddot{X})^{-1} \ddot{X}^T M \tilde{\epsilon} \\
\hat{\tilde{\epsilon}} &= \ddot{Y} - \ddot{X}\tilde{\beta} \\
\tilde{V} &= (\ddot{X}^T \ddot{X})^{-1} \left(\sum_g \ddot{X}_g^T \hat{\tilde{\epsilon}}_g \hat{\tilde{\epsilon}}_g^T \ddot{X}_g \right) (\ddot{X}^T \ddot{X})^{-1} \\
\tilde{t_0} &= \frac{c_0^T(\tilde{\beta} - \beta)}{\sqrt{c_0^T \tilde{V} c_0}}
\end{aligned}$$

Since $\tilde{\epsilon}$ meets the conditions of Assumptions 2 and 3, the first theorem in Carter et al. (2017) applies, so $\tilde{t_0} \xrightarrow{p} N(0,1)$. And whereas $\tilde{\epsilon}$ also meets the conditions of Assumption 1, it is also the case that Theorem 1 applies, so that $P\left(|\tilde{t_0}| \geq q^*(\alpha; X, H_0) \mid X\right) = \alpha$.

Note that for the case of stochastic X with PDF $f(.)$, it follows from Theorem 1 and the law of total probability that:

$$\begin{aligned}
P(|\tilde{t_0}| \geq q^*(\alpha; X, H_0)) &= \int P(|\tilde{t_0}| \geq q^* \mid X) f(X) dX \\
&= \int \alpha f(X) dX \\
&= \alpha
\end{aligned}$$

Therefore:

$$\begin{aligned}
\alpha^* - \alpha &= P\left(|t_0| \geq q^*\right) - P\left(|\tilde{t}_0| \geq q^*\right) \\
&= \left(1 - P\left(t_0 < q^*\right) + P\left(t_0 < -q^*\right)\right) \\
&\quad - \left(1 - P\left(\tilde{t}_0 < q^*\right) + P\left(\tilde{t}_0 < -q^*\right)\right) \\
&\xrightarrow{p} \left(\Phi(-q^*) - \Phi(q^*)\right) - \left(\Phi(-q^*) - \Phi(q^*)\right) \\
\alpha^* - \alpha &\xrightarrow{p} 0 \\
\alpha^* &\xrightarrow{p} \alpha
\end{aligned}$$

where $\Phi(.)$ is the CDF of the standard normal distribution. In summary:

1. The true test statistic t_0 and the counterfactual test statistic $\tilde{t}_0$ converge to the same distribution.

2. My test is exact for the counterfactual DGP.

3. My test converges to an exact test for any DGP meeting Assumptions 2 and 3.

C Approximating t_0 as $T(\upsilon)$

In this section, I will discuss how previous tests have selected critical values for the test statistic t_0 by approximating its distribution as $T(\upsilon)$, a t-distribution with υ degrees of freedom. Bell and McCaffrey (2002) use $T(m)$, where m is the degrees of freedom from an approximation to a t-distribution based on Satterthwaite (1946). Carter et al. (2017) use $T(G^*)$, where G^* is the effective number of clusters. The purpose of this section is to make sense of the assumptions and simplifications that are necessary to rationalize those tests in a finite sample.

Consider the definition of a t-distributed random variable with υ degrees of freedom as τ:

$$\begin{aligned} & \tau \sim T(\upsilon) \\ \tau = \frac{z}{\sqrt{\Upsilon}}, \quad & \text{where } z \sim N(0,1), \\ & \upsilon\Upsilon \sim \chi^2_\upsilon, \text{ and} \\ & \Upsilon \perp z \end{aligned} \tag{9}$$

The intuition behind the approximations below is that, under Assumption 1, the test statistic resembles this structure superficially; t_0 is a ratio of a normal random variable divided by the square root of a sum of squared normals. For simplicity, assume a non-stochastic covariate design matrix X. Let $z_0 = \frac{c_0^T(\hat{\beta}-\beta)}{\sqrt{c_0^T V(\hat{\beta})c_0}}$ and let $\Upsilon_0 = \frac{c_0^T \hat{V}(\hat{\beta})c_0}{c_0^T V(\hat{\beta})c_0}$. Now consider the test statistic for a true hypothesis $H_0 : c_0^T t_0 = a_0$:

$$\begin{aligned} t_0 &= \frac{c_0^T\hat{\beta} - a_0}{\sqrt{c_0^T \hat{V}(\hat{\beta})c_0}} \\ &= \frac{c_0^T(\hat{\beta}-\beta)}{\sqrt{c_0^T \hat{V}(\hat{\beta})c_0}} \\ &= \frac{c_0^T(\hat{\beta}-\beta)}{\sqrt{c_0^T V(\hat{\beta})c_0}} \times \sqrt{\frac{c_0^T V(\hat{\beta})c_0}{c_0^T \hat{V}(\hat{\beta})c_0}} \\ &= \frac{z_0}{\sqrt{\Upsilon_0}} \end{aligned}$$

If it were the case that $z_0 \sim N(0,1)$, that $(\upsilon_0\Upsilon_0) \sim \chi^2_{\upsilon_0}$ for some υ_0, and that $z_0 \perp \Upsilon_0$, then t_0 would in fact have a t-distribution with υ_0 degrees of freedom.

C.1 Approximation in Bell and McCaffrey (2002)

In this section, I will show how the approximation to a t-distribution is constructed in Bell and McCaffrey (2002). They calculate a test statistic using the variance estimator $\hat{V}_{CR2}$:[13]

$$
\begin{aligned}
A_{jg} &= (X^T X)^{-\frac{1}{2}} \left(I_k - (X^T X)^{-\frac{1}{2}} X_g^T X_g (X^T X)^{-\frac{1}{2}} \right)^{-\frac{1}{2}} (X^T X)^{\frac{1}{2}} \\
\hat{V}_{CR2}(\hat{\beta}) &= (\ddot{X}^T \ddot{X})^{-1} \left(\sum_g A_g^T \ddot{X}_g^T \hat{\epsilon}_g \hat{\epsilon}_g^T \ddot{X}_g A_g \right) (\ddot{X}^T \ddot{X})^{-1}
\end{aligned}
$$

As in Appendix A, I define N_g as the number of observations in cluster g, I_g as an identity matrix of size N_g, and ι_g as a column vector of length N_g whose elements are all 1. Additionally, $M_g = I_g - \frac{1}{N_g}\iota_g \iota_g^T$, and M is a block-diagonal matrix whose g-th block is M_g. Also, recall that $\ddot{\epsilon}_g \sim N(0, \sigma^2 M_g)$. Furthermore, define:

$$
\begin{aligned}
& d_g = (I - H)_g^T \ddot{X}_g A_g (\ddot{X}^T \ddot{X})^{-1} c_0, \\
& D_{(N \times G)} = [d_1 \ldots d_g \ldots d_G] \\
& \lambda^{BM}_{(G \times 1)} \text{ are the eigenvalues of } D^T D \\
& m = \frac{\left(\sum_g \lambda_g^{BM} \right)^2}{\sum_g (\lambda_g^{BM})^2}
\end{aligned}
$$

Bell and McCaffrey (2002) select critical values from $T(m)$. They assume that $\epsilon \sim N(0, \sigma^2 I_N)$; since Assumption 1 is sufficient (and weaker), I will refer to that assumption. Assumption 1 implies that:

1. $\hat{V}_{CR2}$ is unbiased: $\mathbb{E}\left(\hat{V}_{CR2}\right) = V(\hat{\beta})$
2. $\hat{\epsilon} \perp \hat{\beta}$

It follows immediately that $z_0 \perp \Upsilon_0$. Furthermore, observe that under Assumption 1, $z_0 \sim N(0,1)$. This follows from the normality of the errors; $c_0^T \hat{\beta}$ is normal with mean $c_0^T \beta$ and variance $c_0^T V(\hat{\beta}) c_0$. Now following a similar logic as

[13] Bell and McCaffrey (2002) use $X_g^T \tilde{A}_{jg}$ in place of $A_{jg}^T X_g^T$, where $\tilde{A}_g = (I_g - H_{gg})^{\frac{1}{2}}$, but Niccodemi et al. (2020) show that these are equivalent and that A_{jg} is smaller and easier to compute than $\tilde{A}_{jg}$.

in Appendix A:

$$\begin{aligned}
c_0^T \hat{V}(\hat{\beta}) c_0 &= c_0^T (\ddot{X}^T \ddot{X})^{-1} \left(\sum_g A_g^T \ddot{X}_g^T \hat{\epsilon}_g \hat{\epsilon}_g^T \ddot{X}_g A_g \right) (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&= c_0^T (\ddot{X}^T \ddot{X})^{-1} \left(\sum_g A_g^T \ddot{X}_g^T (I-H)_g \ddot{\epsilon} \ddot{\epsilon}^T (I-H)_g^T \ddot{X}_g A_g \right) (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&= \sum_g \ddot{\epsilon}^T (I-H)_g^T \ddot{X}_g A_g (\ddot{X}^T \ddot{X})^{-1} c_0 c_0^T (\ddot{X}^T \ddot{X})^{-1} A_g^T \ddot{X}_g^T (I-H)_g \ddot{\epsilon} \\
&= \ddot{\epsilon}^T \left(\sum_g d_g d_g^T \right) \ddot{\epsilon}
\end{aligned}$$

Again recalling from Appendix A that $\ddot{\epsilon} \sim N(0, \sigma^2 M)$, let $\eta = N(0, I)$ such that $\ddot{\epsilon} = \sigma M \eta$ and let w be a G-vector of independent χ_1^2 random variables. Also, the non-zero eigenvalues of MDD^TM are the same as the non-zero eigenvalues of D^TD. Then:

$$\begin{aligned}
c_0^T \hat{V}(\hat{\beta}) c_0 &= \sigma^2 \eta^T M D D^T M \eta \\
&\sim \sigma^2 w^T \lambda^{BM}
\end{aligned}$$

Now since the variance estimator is unbiased:

$$\begin{aligned}
c_0^T V(\hat{\beta}) c_0 &= \mathbb{E}\left(c_0^T \hat{V}(\hat{\beta}) c_0 \right) \\
&= \mathbb{E}\left(\sigma^2 w^T \lambda^{BM} \right) \\
&= \sigma^2 \sum_g \mathbb{E}(w_g) \lambda_g^{BM} \\
&= \sigma^2 \sum_g \lambda_g^{BM}
\end{aligned}$$

It follows that:

$$\begin{aligned}
\Upsilon_0 &= \frac{c_0^T \hat{V}(\hat{\beta}) c_0}{c_0^T V(\hat{\beta}) c_0} = \frac{\sum_g w_g \lambda_g^{BM}}{\sum_g \lambda_g^{BM}} \\
\mathbb{E}(\Upsilon_0) &= 1 \\
V(\Upsilon_0) &= \sum_g V(w_g) \left(\frac{\lambda_g^{BM}}{\sum_{g'} \lambda_{g'}^{BM}} \right)^2 = \frac{2}{m}
\end{aligned}$$

Now, Bell and McCaffrey (2002) can apply the Satterthwaite approximation of $\Upsilon_0 \approx \Upsilon^{BM}$, where $m\Upsilon^{BM} \sim \chi_m^2$. The first two moments match, so that $\mathbb{E}(\Upsilon^{BM}) = 1$ and $V(\Upsilon^{BM}) = \frac{2}{m}$. Let $\tau^{BM} = \frac{z_0}{\sqrt{\Upsilon^{BM}}}$ be the approximated test statistic with Υ^{BM} substituted for Υ_0. Then it is in fact the case that $\tau^{BM} \sim T(m)$.

In summary, Bell and McCaffrey (2002) make (a stronger version of) Assumption 1. Then they apply a Satterthwaite approximation to the denominator of the test statistic. As a result, they find that the approximated test statistic t^{BM} has a t distribution with m degrees of freedom, where $m = \frac{\left(\sum_g \lambda_g^{BM}\right)^2}{\sum_g (\lambda_g^{BM})^2}$.

C.2 Approximation in Carter et al. (2017)

I will now show how Carter et al. (2017) approximate the test statistic t_0 as a t-distribution. Their method also involves a Satterthwaite approximation. They assume that $\forall g \epsilon_g \sim N(0, \sigma^2 \iota_g \iota_g^T)$; since Assumption 1 is once again sufficient (and weaker), I will refer to that assumption.

In Carter et al. (2017), the test statistic is calculated using the variance estimator $\hat{V}_{CR0}$ (so $A_g = I_k$). Then, critical values are selected from $T(G^*)$, where G^* is what they refer to as the "effective number of clusters":

$$\begin{aligned}
\lambda_g^{CSS} &= c_0^T (\ddot{X}^T \ddot{X})^{-1} \ddot{X}_g^T \mathbb{E}(\ddot{\epsilon}_g \ddot{\epsilon}_g^T) \ddot{X}_g (\ddot{X}^T \ddot{X})^{-1} c_0 \\
&= \sigma^2 c_0^T (\ddot{X}^T \ddot{X})^{-1} \ddot{X}_g^T M_g \ddot{X}_g (\ddot{X}^T \ddot{X})^{-1} c_0 \\
G^* &= \frac{(\sum_g \lambda_g^{CSS})^2}{\sum_g (\lambda_g^{CSS})^2}
\end{aligned}$$

Recall that $\mathbb{E}(\ddot{\epsilon}_g \ddot{\epsilon}_g^T) = \sigma^2 M_g$. Returning to the test statistic, $t_0 = \frac{z_0}{\sqrt{\Upsilon_0}}$, we can decompose Υ_0 into two parts:

$$\begin{aligned}
\Upsilon_0 &= \frac{c_0^T \hat{V}(\hat{\beta}) c_0}{c_0^T V(\hat{\beta}) c_0} \\
&= \frac{c_0^T (\ddot{X}^T \ddot{X})^{-1} \left(\sum_g \ddot{X}_g^T A_g \hat{\epsilon}_g \hat{\epsilon}_g^T A_g \ddot{X}_g\right) (\ddot{X}^T \ddot{X})^{-1} c_0}{c_0^T (\ddot{X}^T \ddot{X})^{-1} \ddot{X}^T \mathbb{E}(\ddot{\epsilon}\ddot{\epsilon}^T) \ddot{X} (\ddot{X}^T \ddot{X})^{-1} c_0} \\
&= \frac{c_0^T (\ddot{X}^T \ddot{X})^{-1} \left(\sum_g \ddot{X}_g^T (I-H)_g \ddot{\epsilon}\ddot{\epsilon}^T (I-H)_g^T \ddot{X}_g\right) (\ddot{X}^T \ddot{X})^{-1} c_0}{c_0^T (\ddot{X}^T \ddot{X})^{-1} \left(\sum_g \ddot{X}_g^T \mathbb{E}(\ddot{\epsilon}_g \ddot{\epsilon}_g^T) \ddot{X}_g\right) (\ddot{X}^T \ddot{X})^{-1} c_0} \\
&= \frac{c_0^T (\ddot{X}^T \ddot{X})^{-1} \left(\sum_g \ddot{X}_g^T \ddot{\epsilon}_g \ddot{\epsilon}_g^T \ddot{X}_g\right) (\ddot{X}^T \ddot{X})^{-1} c_0}{\sum_g \lambda_g^{CSS}} \\
&+ \frac{c_0^T (\ddot{X}^T \ddot{X})^{-1} \left(\sum_g \ddot{X}_g^T \left(H_g \ddot{\epsilon}\ddot{\epsilon}^T H_g^T - H_g \ddot{\epsilon}\ddot{\epsilon}_g^T - \ddot{\epsilon}_g \ddot{\epsilon}^T H_g^T\right) \ddot{X}_g\right) (\ddot{X}^T \ddot{X})^{-1} c_0}{\sum_g \lambda_g^{CSS}} \\
&= \Upsilon_1 + \Upsilon_2, \quad \text{ where } \Upsilon_1 = \frac{c_0^T (\ddot{X}^T \ddot{X})^{-1} \left(\sum_g \ddot{X}_g^T \ddot{\epsilon}_g \ddot{\epsilon}_g^T \ddot{X}_g\right) (\ddot{X}^T \ddot{X})^{-1} c_0}{\sum_g \lambda_g^{CSS}}, \\
\Upsilon_2 &= \left(\sum_g \lambda_g^{CSS}\right)^{-1} c_0^T (\ddot{X}^T \ddot{X})^{-1} (\sum_g \ddot{X}_g^T (H_g \ddot{\epsilon}\ddot{\epsilon}^T H_g^T \\
&\quad - H_g \ddot{\epsilon}\ddot{\epsilon}_g^T - \ddot{\epsilon}_g \ddot{\epsilon}^T H_g^T) \ddot{X}_g) (\ddot{X}^T \ddot{X})^{-1} c_0
\end{aligned}$$

If the residuals $\hat{\epsilon}$ are close to the errors $\ddot{\epsilon}$, then Υ_2 will be small. One of the approximations necessary for Carter et al. (2017) is to approximate Υ_0 as Υ_1:

$$\Upsilon_0 \approx \Upsilon_1 = \frac{c_0^T(\ddot{X}^T\ddot{X})^{-1}\left(\sum_g \ddot{X}_g^T\ddot{\epsilon}_g\ddot{\epsilon}_g^T\ddot{X}_g\right)(\ddot{X}^T\ddot{X})^{-1}c_0}{\sum_g \lambda_g^{CSS}}$$

$$\Upsilon_1 = \frac{\sum_g \ddot{\epsilon}_g^T\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}_g^T\ddot{\epsilon}_g}{\sum_g \lambda_g^{CSS}}$$

Let $\eta_g \sim N(0, I_g)$ such that $\ddot{\epsilon}_g = \sigma M_g \eta$:

$$\Upsilon_1 = \sigma^2\frac{\sum_g \eta_g\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}_g^T\eta_g}{\sum_g \lambda_g^{CSS}}$$

Notice that $\eta_g\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}_g^T\eta_g$ is a linear combination of χ_1^2 random variables with coefficients equal to the single non-zero eigenvalues of:

$$\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}_g^T$$

There is only one such non-zero eigenvalue, and that is:

$$c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}_g^T\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0$$

Let $w_g \sim \chi_1^2$. It follows that:

$$\begin{aligned}\Upsilon_1 &\sim \sigma^2\frac{\sum_g w_g c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}_g^T M_g\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0}{\sum_g \lambda_g^{CSS}} \\ &\sim \frac{\sum_g w_g\lambda_g^{CSS}}{\sum_g \lambda_g^{CSS}}\end{aligned}$$

The second approximation applied (implicitly) by Carter et al. (2017) is the Satterthwaite approximation. Υ_1 is a linear combination of χ_1^2 random variables, and it is approximated by Υ^{CSS}, where $G^*\Upsilon^{CSS} \sim \chi^2_{G^*}$. Just like with Υ^{BM} above, the first two moments of Υ_1 and Υ^{CSS} match each other. So then let $\tau^{CSS} = \frac{z_0}{\sqrt{\Upsilon^{CSS}}}$ be the approximated test statistic with Υ^{CSS} substituted for Υ_0. We have that $\tau^{CSS} \sim T(G^*)$.

Much of what I've ascribed to Carter et al. (2017) is implicit in the description of their test rather than explicit in their analysis. My purpose here was to explain what simplifications and approximations separate the reference distribution $T(G^*)$ from the exact distribution of the test statistic t_0.

D Data Generating Process For Summary Simulation

This section describes the data generating process for the simulations whose results are shown in Figures 8 and 9. The notation for this DGP matches the notation in section 5:

- $G =$ number of clusters
- $J =$ number of clusters with treatment variation
- $N_1 =$ size of first cluster
- $\phi =$ treatment intensity of first cluster

Additionally, I use h to index observations within a cluster, so that $h_i = a$ for $a \leq N_g$ is true for some observation i in cluster g. In 8, I use the following data-generating process:

$$\begin{aligned}
x_{1ig} &= \mathbb{1}(g \leq J) \times \phi^{-\mathbb{1}(g>1)} \times \mathbb{1}(h_i \leq \frac{N_g}{2}) \\
x_{2ig} &= \mathbb{1}(g \leq J) \times \mathbb{1}(h_i = N_g) \\
\epsilon_{ig} &\sim N(0,1) \\
\beta_0 &= 3, \beta_1 = 2, \beta_2 = 1 \\
y_{ig} &= \beta_0 + \beta_1 x_{1ig} + \beta_2 x_{2ig} + \gamma_g + \epsilon_{ig}
\end{aligned}$$

For $g > 1$, $N_g = 5$. Within 8, scenarios are parameterized as:

(0) $G = 500$, $J = 250$, $N_1 = 5$, $\phi = 1$

(1) Same as (0), but $G = 5$, $J = 5$

(2) Same as (0), but $J = 5$

(3) Same as (0), but $N_g = 799$ (selected so that $G^* = 5$)

(4) Same as (0), but $\phi = 13.092198$ (selected so that $G^* = 5$)

Then in 9, scenarios are parameterized the same as (0), except that:

(2-A) $x_{2ig} = \mathbb{1}(5 < g \leq J) \times \mathbb{1}(h_i \leq \frac{N_g}{2})$

(4-A) $\phi = 7.556576$ (selected so that $G^* = 5$), and

$$x_{2ig} = \begin{cases} 0 & \text{if } g = 1 \ , \\ (1 - \phi^{-1}) \times \mathbb{1}(h_i \leq \dfrac{N_g}{2}) & \text{if } 1 < g \leq \frac{J}{2} \ , \\ \mathbb{1}(h_i \leq \dfrac{N_g}{2}) & \text{if } \frac{J}{2} < g \leq J \ . \end{cases}$$

E Additional Tables

Table E.1: Empirical Literature Meeting Inclusion Criteria

Paper (1)	Journal (2)	G^* Analysis (3)	Focus Estimate (4)
Adhvaryu et al. (2020)	RESTAT	No	
Allcott and Rafkin (2022)	AEJ: Policy	No	
Alpert et al. (2018)	AEJ: Policy	No	
Anderson et al. (2019)	AEJ: Applied	Yes	Table 2, column 1
Ang (2019)	AEJ: Applied	Yes	Table 3, column 3
Baek et al. (2021)	RESTAT	Yes	Table 3, column 1
Barr and Turner (2018)	AEJ: Policy	Yes	Table 2, column 1
Bastian (2020)	AEJ: Policy	No	
Bernecker et al. (2021)	AEJ: Policy	No	
Bertocchi et al. (2020)	AEJ: Policy	Yes	Table 2, column 5
Binder and Makridis (2022)	RESTAT	No	
Borgschulte and Martorell (2018)	AEJ: Applied	No	
Bound et al. (2020)	AEJ: Policy	Yes	Table 2, column 1
Bouton et al. (2021)	RESTAT	Yes	Table 1, column 1
Buchmueller and C. Carey (2018)	AEJ: Policy	No	
C. M. Carey et al. (2020)	AEJ: Applied	No	
Carpenter and Lawler (2019)	AEJ: Policy	No	
Dickert-Conlin et al. (2019)	AEJ: Applied	No	
Dranove et al. (2021)	AEJ: Applied	Yes	Figure 2, $x = 0$
Dube (2019)	AEJ: Applied	No	
Evans et al. (2019)	RESTAT	No	
Ganapati et al. (2020)	AEJ: Applied	No	
Ganong and Liebman (2018)	AEJ: Policy	No	
Garthwaite et al. (2018)	AEJ: Applied	No	
Goldin et al. (2022)	AEJ: Policy	No	
Hausman and Lavetti (2021)	AEJ: Applied	No	
Hsu et al. (2018)	RESTAT	Yes	Table 3, column 1
Jackson et al. (2021)	AEJ: Policy	Yes	Table 2, column 7
Jaworski and Kitchens (2019)	RESTAT	Yes	Table 1, column 2
J. E. Johnson and Kleiner (2020)	AEJ: Policy	No	
R. C. Johnson and Jackson (2019)	AEJ: Policy	No	
Kose et al. (2021)	AEJ: Policy	No	
Kroft et al. (2020)	AEJ: Policy	No	
Kuka et al. (2020)	AEJ: Policy	No	
Kuka (2020)	RESTAT	No	
Lafortune et al. (2018)	AEJ: Applied	No	
Leung (2021)	RESTAT	No	
Lovenheim and Willén (2019)	AEJ: Policy	Yes	Table 2, column 3

Continued on next page

Table E.1: Empirical Literature Meeting Inclusion Criteria (Continued)

Paper (1)	Journal (2)	G^* Analysis (3)	Focus Estimate (4)
Mayda et al. (2022)	AEJ: Applied	No	
Modestino et al. (2020)	RESTAT	No	
Renkin et al. (2022)	RESTAT	No	
Sabia et al. (2019)	RESTAT	Yes	Table 2, column 3
Shenhav (2021)	RESTAT	Yes	Table 2, column 1
Siemer (2019)	RESTAT	No	
Stuart (2022)	AEJ: Applied	No	
Wilson (2022)	RESTAT	No	
Xiong (2021)	AEJ: Applied	Yes	Table 4, columns 1

Notes: Includes all papers meeting criteria: (1) published between 2018 and 2022, inclusive; (2) published in AEJ: Policy, AEJ: Applied, or RESTAT; (3) one of the main specifications in the paper evaluated an empirical setting in the United States; (4) that specification used state-level fixed effects, or fixed effects for groups nested within states; and (5) that specification clustered standard errors at the state level.

Table E.2: Subject Areas of Empirical Literature

JEL Code	Description	All Papers	G^* Analysis Only
D	Microeconomics	10	3
E	Macroeconomics and Monetary Economics	7	3
G	Financial Economics	4	1
H	Public Economics	19	5
I	Health, Education, and Welfare	24	7
J	Labor and Demographic Economics	16	5
K	Law and Economics	4	2
L	Industrial Organization	6	2
N	Economic History	2	1
Q	Agriculture, Natural Resources, Environment	1	0
R	Urban, Rural, Regional, Real Estate, and Transportation	2	0
Z	Other Special Topics	2	1

Notes: Includes all papers meeting criteria: (1) published between 2018 and 2022, inclusive; (2) published in AEJ: Policy or AEJ: Applied (RESTAT excluded here); (3) one of the main specifications in the paper evaluated an empirical setting in the United States; (4) that specification used state-level fixed effects, or fixed effects for groups nested within states; and (5) that specification clustered standard errors at the state level.