Inference with Cluster Imbalance: The Case of State Corporate Laws *

Allen Hu Holger Spamann

December 28, 2020

Abstract

A workhorse research design identifies the effects of corporate governance by changes in state laws, clustering standard errors by state of incorporation. Asymptotic inference using these standard errors, however, dramatically understates false positives: in a typical specification, randomly generated placebo laws have 1/5/10%-level significant estimated treatment effects 9/21/30% of the time. This poor finite sample performance is due to unequal cluster sizes, especially Delaware's concentration of half of all incorporations. Bootstrap or permutation tests mostly fix the problem, common robustness checks less so. The placebo law approach can also be used to calculate power, which will be acceptably high only for substantial effect sizes.

Keywords— Anti-Takeover Laws, Bootstrap, Corporate Governance, Cluster-Robust Inference, Monte Carlo, Placebo Laws, Permutation Test

JEL Classification— C12, G34, G38, K22

*Hu is with Yale School of Management; and Spamann is with Harvard Law School. Emails: allen.hu@yale.edu; hspamann@law.harvard.edu. For very helpful comments and suggestions, we thank Bobby Bartlett, Emiliano Catan, Jonah Gelbach, Paul Goldsmith-Pinkham, James G. MacKinnon, Justin McCrary, Matt Webb, Michael Wittry, Alwyn Young, and seminar participants at Harvard Law School and Yale SOM. Spamann thanks the Wissenschaftskolleg zu Berlin for its hospitality. A predecessor paper in 2019 was circulated under the title "On Inference When Using State Corporate Laws for Identification."

1 Introduction

In the United States, many important corporate governance features are laid down in state laws. A large empirical literature in corporate finance studies the effects of these laws on corporate actions and performance with firm-level data, usually focusing on legislative changes in a difference-in-difference framework. Interest is in the laws for their own sake or, more commonly, as exogenous variation in general economic determinants such as managerial slack. In 2018, Karpoff and Wittry counted 78 published articles and working papers using changes in anti-takeover laws alone; this list has kept growing rapidly.[1] The standard design is a linear firm-level panel regression with firm fixed effects and a variety of possibly confounding time-varying factors. Following Bertrand, Duflo, and Mullainathan 2004, it has become standard econometric practice in these papers to cluster the standard errors by state of incorporation before performing asymptotic inference using the normal distribution or the t-distribution with degrees of freedom equal to the number of clusters minus 1.[2]

This paper shows, however, that the conventional approach to inference dramatically overrejects in this setting. False positives are far more likely than conventional p-values and significance levels suggest. In a typical difference-in-difference design with real data, tests of randomly generated "placebo laws" reject the true null hypothesis of no effect at the 1/5/10% level *at least* 3/9/16% of the time, depending on the number of "treated" states and using the popular Tobin's q as dependent variable. The mean (median) false positive rate across numbers of "treated" states and whether Delaware is "treated" is 9/21/30% (8/19/28%) for tests that are supposed to hold that rate to 1/5/10%. Other popular dependent variables like leverage or ROA yield similarly excessive false positive rates. Theory and simulations show that the source of the problem is cluster size imbalance, which is extreme in this setting: Delaware contains half of all U.S. firms. Most fixes recommended in the literature cannot be used with state corporate law difference-in-difference designs (due to staggered adoption), or fail to correct the overrejection. Acceptable results obtain with the cluster wild bootstrap (Cameron, Gelbach, and Miller 2008) or, better, with a permutation test on the treatment indicators (similar to DiCiccio and Romano 2017; MacKinnon and Webb 2020). Nevertheless, obtaining 80% power requires effect sizes on the order of 0.1 standard deviations of the raw dependent variable.

This paper is structured as follows. Section 2 reviews conventional, asymptotic cluster-robust inference in the context of the popular firm-level linear regression design with state corporate law as the key independent variable, focusing on the design's difference-in-difference panel variant. Section 3 uses "placebo laws" to demonstrate empirically that the conventional approach grossly understates false positive rates. The results were summarized above. Section 4 explains this failure with a review of the asymptotic theory and more Monte Carlo evidence. Essentially, cluster size imbalance drastically slows the rate of convergence of the conventional cluster-robust variance estimator. Section 5 considers possible fixes, showing that only the cluster wild bootstrap and a permutation test work in this context. Section 6 simulates power for these tests. Section 7 concludes.

[1] See, e.g., Bharath and Hertzel 2019; He and Hirshleifer 2019; Demiroglu, Iskenderoglu, and Ozbas 2019; Gutiérrez Urtiaga and Vazquez 2019; Cremers, Guernsey, and Sepe 2019. Another example of a statute popular with finance researchers is universal demand laws.

[2] On the necessity to cluster by state of incorporation, see section 2.2 below.

1.1 Related Literature

The present paper is similar in spirit to Bertrand, Duflo, and Mullainathan 2004, Petersen 2009, and others who use Monte Carlo simulations to demonstrate the practical importance of properly accounting for serial and cross-sectional correlation in the error term. When the number of clusters is above 42 or 50, as in regressions using U.S. state laws, the standard advice (e.g., Bertrand, Duflo, and Mullainathan 2004; Petersen 2009; Angrist and Pischke 2008) is to use the clustered "sandwich" variance estimator (White 1984; Liang and Zeger 1986). MacKinnon and Webb 2017 show that this approach fails when cluster sizes are unequal, the more so the further away the fraction of treated clusters is from $\frac{1}{2}$, and instead point to the cluster wild bootstrap proposed by Cameron, Gelbach, and Miller 2008 as the solution. With one exception, however, none of the prior literature consider cluster size imbalance as extreme as that in the corporate governance context: More than half of all publicly traded U.S. corporations are incorporated in Delaware, whereas, e.g., MacKinnon and Webb 2020 considered even 19% to be "quite extreme" for the largest cluster. When one cluster contains half the observations, both the sandwich estimator and the cluster wild bootstrap spectacularly fail to control size, as shown by Monte Carlo evidence with a continuous regressor in Djogbenou, MacKinnon, and Nielsen 2019 and for the cluster treatment assignment model in this paper (even if the fraction of treated states is exactly $\frac{1}{2}$). Theory explaining this failure, reviewed in section 4, is provided in Carter, Schnepel, and Steigerwald 2017, MacKinnon and Webb 2017, and Djogbenou, MacKinnon, and Nielsen 2019. The latter papers are part of an active theoretical econometric literature on inference with clustered data (see surveys of Cameron and Miller 2015; MacKinnon and Webb 2019; revised May 2020), briefly reviewed in section 5. More specifically, this paper's proposal of a permutation test is related to a recent surge in interest in permutation tests for regression coefficients (DiCiccio and Romano 2017), including in clustered regression (Canay, Romano, and Shaikh 2017; Hagemann 2019a; Hagemann 2019b; MacKinnon and Webb 2020).

Fixing inference in tests with state corporate laws is also the concern of Heath et al. 2020's investigation of the multiplicity problem with "reusing natural experiments." They use simulations and actual data to show that critical values have to be adjusted upwards sharply in the study of one law's effect on *multiple* outcomes if it is desired to control the familywise error rate (FWER). The FWER is the probability of falsely rejecting *at least one* of the individual null hypotheses with respect to individual outcomes, or equivalently, the probability of falsely rejecting the *composite* null hypothesis that the law has no effect on any outcome. Heath et al. 2020 control the FWER asymptotically using the stepdown method of Romano and Wolf 2005. By contrast, the present paper is concerned with tests of simple hypotheses in finite samples.

Other papers consider not inference but specification issues and bias. Karpoff and Wittry 2018 point out that various anti-takeover statutes share similar purposes and are often adopted in close temporal proximity, i.e., they are not independent and thus must be controlled to avoid omitted variable bias (also see Coates 2000; Catan and Kahan 2016; Cain, McKeon, and Solomon 2017). In our "placebo law" tests, we control for all the laws identified by Karpoff and Wittry 2018.

A wave of recent papers in the general econometrics literature (Chaisemartin and D'Haultfœuille 2020; Goodman-Bacon 2020; Imai and Kim 2020; Sun and Abraham 2020; Borusyak and Jaravel 2018) points out specification problems with linear twoway fixed effects models (i.e., models with unit and time fixed effects), which are the workhorse of the corporate finance literature. The recent papers emphasize that the workhorse model with a simple treatment indicator is misspecified if treatment effects are dynamic, i.e., accumulate or

otherwise change over time rather than occur instantaneously, and that even more general treatment dummies cannot be interpreted as average treatment effects when treatment effects are heterogeneous. While our Monte Carlo evidence uses the workhorse model, we sidestep these specification issues: our treatment effects are always instantaneous and homogenous (of size 0 in case of the "placebo laws"). In this sense, our Monte Carlo evidence illustrates the theoretical point that the inference problem is separate from the specification problem. Nevertheless, fixing the inference problem should have a salutory effect for the specification problem because correct inference makes it less likely that nonsensical estimates produced by misspecified models will be mistaken for true effects.

2 The Typical Study Design

This section reviews the typical study design. Our focus in this paper is on inference (2.2). In principle, the inference problems generated by the extreme imbalance in incorporation cluster sizes are not specific to any particular estimated equation or estimator. Nevertheless, to fix ideas and to be specific about the context for our Monte Carlo evidence, we begin by explicitly setting out the typical twoway fixed effect equation and estimator (2.1). Readers familiar with the typical twoway fixed effect model and the conventional "sandwich" cluster-robust variance estimator of White 1984; Liang and Zeger 1986 may skip ahead to section 3.

2.1 Estimated Equation and Estimator

A typical study estimates a linear panel model of the following type:[3]

$$y_{ij\ldots st} = \alpha_i + \beta D_{st} + \delta' \mathbf{x_{it}} + \gamma' \mathbf{z_{j\ldots st}} + \epsilon_{ij\ldots st} \tag{1}$$

where β is the coefficient of interest, and

- the subscripts are $i \in \{1, \ldots, I\}$ for firms, $j \in \{1, \ldots, J\}$ for industries, $s \in \{1, \ldots, S\}$ for incorporation states, $t \in \{1, \ldots, T\}$ for time (years), and "$\ldots$" in a subscript stand for further possible groupings (e.g., location of firm headquarters);
- D_{st} is a dummy variable for the "treatment," i.e., whether state s, where firm i is incorporated, has the provision in question in year t (usually, the dummy switches on in some year t^* and stays on for the remainder of the sample period $t \geq t^*$ in that state);
- $y_{ij\ldots st}$ is the outcome variable for firm i in year t;
- α_i is a firm-specific intercept, i.e., a fixed effect for firm i;
- $\mathbf{x_{it}}$ is a vector of time-varying firm-level controls;
- $\mathbf{z_{j\ldots st}}$ is a vector of time-varying industry etc. level controls; and
- $\epsilon_{ij\ldots st}$ is a firm-year specific error term (which is allowed to be correlated within firm and within incorporation state, as discussed below).

[3] Cross-sectional models, where $T = 1$ and hence $t = 1$ constant for all observations, are far less popular because they must have $\alpha_i = \alpha \forall i$ and hence do not allow controlling for unobserved time-invariant firm-level heterogeneity. In any event, we have verified that similar overrejection occurs in simple cross-sectional data.

Following Gormley and Matsa 2013, most studies include (industry-)time fixed effects in $\mathbf{z_{j\ldots st}}$, such that the model is a twoway fixed effect model.[4] The model is usually estimated with the fixed effect (FE) estimator. That is, one estimates by OLS the firm-demeaned equation:

$$\ddot{y}_{ij\ldots st} = \beta \ddot{D}_{st} + \delta' \ddot{\mathbf{x}}_{\mathbf{it}} + \gamma' \ddot{\mathbf{z}}_{\mathbf{j\ldots st}} + \ddot{\epsilon}_{ij\ldots st} \tag{2}$$

where double dots denote time-specific deviations from the firm-specific mean for firm i.

There are several important specification issues, many mentioned in 1.1, that we can sidestep in our analysis of inference. In particular, our Monte Carlo evidence using placebo laws or entirely simulated data avoids omitted variable bias and problems from heterogeneous and/or dynamic treatment effects by construction.

2.2 Inference

As already mentioned, it is now standard practice to cluster the standard errors by state after estimating (2), i.e., to account for the likely non-zero covariance between residuals for the same firm over time and for multiple firms within the same state of incorporation. Clustering at the state level is indicated because the treatment assignment is clustered at the state level (Abadie et al. 2017): treatment is perfectly correlated within state-years, and very highly serially correlated within state due to legislative inertia.[5]

The usual way to cluster is the "sandwich" estimator of the coefficient variance matrix introduced by White 1984; Liang and Zeger 1986 (CRVE):

$$\hat{\mathbf{V}} \equiv (\mathbf{W'W})^{-1} \sum_{s=1}^{S} (\mathbf{W}_\mathbf{s}' \hat{\epsilon}_\mathbf{s} \hat{\epsilon}_\mathbf{s}' \mathbf{W}_\mathbf{s}) (\mathbf{W'W})^{-1} \tag{3}$$

where $\hat{\epsilon}_\mathbf{s}$ is a column vector of regression residuals for the observations in state s, $\mathbf{W_s}$ is a matrix of covariates for the observations in state s with typical row $\left(\ddot{D}_{st}, \ddot{\mathbf{x}}_{\mathbf{it}}', \ddot{\mathbf{z}}_{\mathbf{j\ldots st}}'\right)$, and $\mathbf{W} \equiv \left(\mathbf{W_1}' \ldots \mathbf{W_S}'\right)'$. Let $\hat{V}(\hat{\beta})$ denote the estimated variance of $\hat{\beta}$, which is of course the appropriate diagonal element of $\hat{\mathbf{V}}$. For hypothesis tests, focus in the literature and in this paper is on the resulting t-statistic

$$\hat{t} \equiv \frac{\hat{\beta}}{\sqrt{\hat{V}(\hat{\beta})}}. \tag{4}$$

Carter, Schnepel, and Steigerwald 2017, B. E. Hansen and S. Lee 2019, and Djogbenou, MacKinnon, and Nielsen 2019 generalize to the case of cluster heterogeneity earlier results that $\hat{\mathbf{V}}$ consistently estimates the true variance and that $\hat{t}$ is asymptotically normal as $S \to \infty$. Nevertheless, it is standard to use critical

[4] The computational challenge of estimating the many fixed effects was solved by the `Stata` package `reghdfe` of Correia 2016.

[5] Ignoring the clustered treatment assignment would be harmless if errors were uncorrelated within clusters. Errors are virtually guaranteed to be correlated within clusters, however, since a myriad of state court decisions and statutory amendments affect all or many firms within the state simultaneously. Indeed, some papers use individual decisions or amendments for identification (e.g., Cohen and Wang 2013; Cain, McKeon, and Solomon 2017). But one cannot hope to identify and to control for all possibly relevant state-level shocks. Empirically, unreported placebo law tests with firm-level clustering show even worse overrejection than that reported with state-level clustering in Section 3, controls for the five second-generation anti-takeover statutes from Karpoff and Wittry 2018 notwithstanding.

values from a t-distribution and to apply an adjustment factor to $\hat{\mathbf{V}}$ to correct for finite sample bias (see Donald and Lang 2007 for intuition and an exact result under more restrictive assumptions). In particular, `Stata` multiplies $\hat{\mathbf{V}}$ by $\frac{N-1}{N-k} \times \frac{S}{S-1}$, where k is the number of regressors excluding nested fixed effects, and uses critical values from the t-distribution with $S-1$ degrees of freedom. This paper does so too.

3 Placebo Laws

In this section, we use placebo laws to demonstrate failure of the conventional cluster robust inference in finite samples with state corporate laws, asymptotic consistency notwithstanding. The next section will explain why this happens (cluster size imbalance).

We study the "effect" of *random* dummy variables ("placebo laws") in a typical data set. Even though the Placebo laws are random and hence have no real effect by construction, the conventional tests reject the null of no effect at rates far higher than their nominal level. That is, the false positive rate is far higher than the chosen test size or p-value would make one believe.

The base data are all firm-year observations of US-listed firms from the CRSP/Compustat merged database for the years 1983-2018 excluding financials and utilities and ADRs. The years are chosen to avoid interference from first-generation anti-takeover laws (Karpoff and Wittry 2018, section II.A). As dependent variable y_{ijst}, we consider Tobin's q (e.g., Gompers, Ishii, and Metrick 2003) and the seven dependent variables in Karpoff and Wittry 2018: ROA, Capex, PPE growth, asset growth, cash, SGA expenses, and leverage.[6] All dependent variables are 95%-winsorized. Following Giroud and Mueller 2010; Karpoff and Wittry 2018, the firm-level control variables $\mathbf{x_{it}}$ include (only) firm age, age squared, size, and size squared. The group control variables $\mathbf{z_{jst}}$ always include industry-year jt fixed effects (using the Fama-French 49 industry coarsening of the SIC classification).[7] To emphasize that the inference problem is separate from the specification issues considered by other papers, we show results with and without additional dummy controls for each the five second-generation anti-takeover statutes from Karpoff and Wittry 2018.[8] Standard errors are clustered by 51 states s of incorporation (the 50 U.S. states and the District of Columbia) using historic incorporation data from (Spamann and Wilkinson 2019) (backfilled for years prior to 1995). As the citations indicate, this basic setup corresponds to widespread practice in the literature, but in any event, nothing hinges on the specifics, as we will obtain very similar results with purely simulated data in 4.

Starting from the base data just described, we then generate random placebo laws (i.e., a dummy variable equalling one for "treated" state-years) in random states in random years. Once enacted, placebo laws stay "in effect" throughout the sample period, as is usually the case in the real world. The simulations are run separately 2,000 times for each combination of (a) dependent variable, (b) every number of treated states between 1 and 51, (c) whether or not Delaware (DE) is among the treated states (if it is not generally among the treated, it nevertheless will be once the number of treated states is 51), and (d) whether anti-takeover

[6] Tobin's q is constructed as (total assets + market equity - book equity) divided by total assets, and market and book equity are as defined on Ken French's website https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/variable_definitions.html.

[7] The 49 industry definitions are from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/Siccodes49.zip.

[8] Karpoff and Wittry 2018's five second-generation anti-takeover statutes are control share acquisition laws, business combination laws, fair price laws, directors' duties laws, and poison pill laws.

statute dummies from Karpoff and Wittry 2018 (KW) are included as controls. For each run, the requisite number of "treated" states are picked at random (subject to ensuring that Delaware is or is not treated, as the case may be), and then a first treatment year is drawn independently for each of the treated states from a uniform discrete distribution over all sample years. (Unreported results forcing the first treatment year to be year 2 or beyond look virtually identical.) Equation (2.1) is then estimated using OLS, the t-statistic calculated as in (4) using (3), and then compared to critical values from a t-distribution with 50 degrees of freedom. Rejection rates are then calculated for each combination of treated number, Delaware status, and controls.

[Insert Figure 1 Here.]

Figure 1 graphs the results for Tobin's q. To facilitate comparison of the tests' nominal size (i.e., their ostensible false positive rate) to their empirical size (i.e., their actual false positive rate), the vertical axis has been rescaled by taking square roots of the rejection rates, and horizontal lines drawn at 1%, 5%, and 10%. Evidently, the conventional test using (4) grossly overrejects.

A natural question to ask if there is something special about Tobin's q. There is not. With variation in the details, unreported graphs for the other seven dependent variables mentioned above look similarly bad. The full results are summarized in the Table 1.

[Insert Table 1 Here.]

Overrejection tends to be a little less bad when many clusters are treated, when Delaware is among the treated clusters, and when KW controls are included. The main takewaway, however, is that conventional cluster-robust inference fails spectacularly in this setting. Tests with nominal size (i.e., false positive rates) of 10/5/1% have empirical size (i.e., actual false positive rates) on the order of 20/15/5%. In the next section, we provide intuition for why this happens.

4 Why Conventional Inference Fails

The reason why conventional inference fails in this context is that its justification is asymptotic, whereas the convergence of the "sandwich" cluster-robust variance is very slow when clusters are of unequal size (Carter, Schnepel, and Steigerwald 2017; MacKinnon and Webb 2017).

[Insert Figure 2 Here.]

Incorporation clusters are of very unequal size, as shown in figure 2. Delaware contains 62% of firm-years in CRSP/Compustat data 1983-2018. Even without Delaware, however, the distribution is very unequal. The second-largest cluster, California, contains 3.4% of the observations, which is 9% of the non-Delaware data and 383 times as many as the smallest cluster, New Hampshire, which contains only 0.01% of the observations. By contrast, classical papers on the performance of the CRVE had operated with equal cluster sizes (Bertrand, Duflo, and Mullainathan 2004). Recent work showing failure of the CRVE with 50 clusters of "wildly different cluster sizes" (MacKinnon and Webb 2017) used a cluster size distribution mimicking the population distribution of U.S. states, which resembles the incorporation cluster size distribution *without* Delaware.

To understand the intuition behind the importance of cluster size imbalance, consider the formula for the "sandwich" cluster-robust variance estimator (3) (CRVE). The summation over the outer product of the error terms in each cluster $\hat{\epsilon}_s\hat{\epsilon}_s'$ estimates the within-cluster covariance matrix $\mathbf{\Omega_s}$, in each case sandwiched between the cluster-specific covariates $\mathbf{W_s}$, (i.e., the sum is over the outer products of the cluster-specific score vector $\mathbf{W_s}'\hat{\epsilon}_s$). The attraction of the CRVE is that it makes no assumptions on the structure of $\mathbf{\Omega_s}$, up to the point of perfect correlation within cluster. The price to pay for this is that the number of independent observations to average over in estimating the CRVE is the number of clusters (i.e., states), not the number of individual observations (i.e., firm-years). Asymptotic results hold for $S \to \infty$, not $N \to \infty$ (Carter, Schnepel, and Steigerwald 2017; Djogbenou, MacKinnon, and Nielsen 2019).[9]

The formula (3) also shows that the weight of individual clusters in calculating the CRVE tends to increase as the square of cluster size because a typical element of $\mathbf{W_s'}\hat{\epsilon}_\mathbf{s}\hat{\epsilon}_\mathbf{s}'\mathbf{W_s}$ is a weighted sum over all N_s^2 elements of $\hat{\epsilon}_\mathbf{s}\hat{\epsilon}_\mathbf{s}'$. This implies that there tends to be less "averaging" and hence slower convergence when cluster sizes are unequal. Nevertheless, the true rate of convergence also depends on the weights (i.e, the independent variables) and on the degree of intra-cluster correlation, which interact in complicated ways (cf. Djogbenou, MacKinnon, and Nielsen 2019, section 2.1, B. E. Hansen and S. Lee 2019, section 4). In particular, Carter, Schnepel, and Steigerwald 2017, result 1.a show that the variance for a particular coefficient is estimated only by the between-cluster variation in the coefficient estimate. For a treatment model with individual or cluster fixed effects, this means that the variance of the treatment coefficient is estimated only from clusters in which the treatment coefficient is identified, i.e., clusters that experience a change in treatment status. This explains why section 3 tended to find worse overrejection with fewer treated clusters. At the same time, it is important to recall that adding more untreated clusters, or observations in untreated clusters, will change the weights in the formula and is thus not irrelevant for the variance estimation.

Carter, Schnepel, and Steigerwald 2017 introduce the feasible effective number of clusters as a guide to the complicated behavior of the CRVE and resulting t-statistics. In the setting of section 3 with Tobin's q as dependent variable, the feasible effective number of clusters for testing the effects of one of the five modern anti-takeover statutes of Karpoff and Wittry 2018 is between 1.3 and 3.2.[10] This is not only much lower than the nominal number of clusters, 51, but also much too small for conventional inference. To be sure, the feasible effective number of clusters is based on a worst-case assumption of perfect within-cluster correlation, and we show below that it is conservative. Nevertheless, it underscores that one should thus expect the variance estimate 3 to be very poor and the resulting t-statistic 4 to be very erratic.

To isolate the importance of cluster size imbalance, we repeat our placebo law tests on simulated data with different degrees of cluster imbalance: 2 (perfect balance), 22, 42, 62 (actual data), or 82% of the firms are in the largest cluster/state (and the rest divided equally among the remaining 50 states). The simulated data match the first and second moments of the real data, but abstract from other features of the real data that might complicate inference, such as non-zero higher moments. As in the real data, we simulate 8,413 firms over 36 years but with perfect balance, i.e., every firm is in the data for all 36 years. We divide the firms among 51 states, and treat 31 of them at independently and randomly chosen years (31 being the average of the number of states having adopted one of the five KW anti-takeover statutes). We again

[9] Equivalently, asymptotic results can be stated in terms of $N \to \infty$ but under a restriction on relative cluster sizes that implicitly guarantees $S \to \infty$, as in B. E. Hansen and S. Lee 2019.

[10] We calculate these numbers using a package by C. H. Lee and Steigerwald 2018.

simulate separately data in which Delaware is "treated", and data in which it is not. We model firm and time fixed effects, and add an error term that is the sum of two AR(1) processes, one at the firm and another at the state level; we do not include industry or other additional effects. Specifically, our data-generating process (DGP) is

$$
\begin{aligned}
y_{ist} &= \alpha_i + \alpha_t + \eta_{ist} \\
\eta_{ist} &= \mu_{it} + \mu_{st} \\
\mu_{it} &= \rho_1 \mu_{it-1} + \epsilon_{it} \\
\mu_{st} &= \rho_2 \mu_{st-1} + \epsilon_{st} \\
\alpha_i &\overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2_{\alpha_i}\right), \quad \alpha_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2_{\alpha_t}\right) \\
\epsilon_{it} &\overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2_1\right), \quad \epsilon_{st} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2_2\right)
\end{aligned}
\tag{5}
$$

Note that the DGP does not contain a treatment effect.

We calibrate the DGP using 1983-2018 CRSP/Compustat data, from which we estimate:

Parameter	Value
σ_{α_i}	1.080
σ_{α_t}	0.236
ρ_1	0.488
ρ_2	0.491
σ_1	0.782
σ_2	0.113

We then use the FE estimator to estimate $y_{ist} = \alpha_i + \alpha_t + \beta D_{st} + \eta_{ist}$, calculate the t-statistic (4) from the CRVE (3), and compared it to critical values from the t-distribution with 50 degrees of freedom, as before. This procedure is run 2,000 times per cluster imbalance and Delaware treatment status.

[Insert Figure 3 Here.]

Figure 3 shows the results. When clusters are balanced (Delaware = 2% of the sample), tests are not too far from their nominal size. As cluster imbalance increases, however, overrejection becomes greater and eventually severe as the sample becomes as unbalanced as in reality, where Delaware contains 62% of all firm-years.

5 Solutions

This section examines possible solutions, i.e., methods for valid inference with state corporate laws. Focus is on robustness checks commonly used in the empirical corporate finance literature, the wild cluster boostrap (Cameron, Gelbach, and Miller 2008), and permutation tests on the treatment indicators. The common robustness checks are dropping Delaware and/or examining lead/lag indicators of treatment; both perform

poorly. The wild cluster bootstrap does fine by itself but completely fixes the problem only when Delaware is dropped from the sample, as might be expected given results in MacKinnon and Webb 2017; Djogbenou, MacKinnon, and Nielsen 2019). The permutation test presents the advantage of being exact against a sharp null hypothesis (as in the Fisher exact test).

Before continuing our examination of these three solutions, we pause to mention a few other methods that one might consider but that do not work in the present context. A simple fix could be to adjust the degrees of freedom in the t-distribution used to obtain critical values. In particular, one could set the degrees-of-freedom equal to the feasible effective number of clusters of Carter, Schnepel, and Steigerwald 2017, which, as reported above, is between 1 and 3 in the present context. As Carter, Schnepel, and Steigerwald 2017 note and Cameron and Miller 2015; MacKinnon and Webb 2017 find in other contexts, however, this method is likely to be excessively conservative because it is based on a worst-case assumption, and this is exactly what we found in unreported results: with this degree-of-freedom correction, the conventional test almost never rejects. We have found the opposite problem with the degrees-of-freedom and variance bias corrections of Young 2016: unreported results show equal or even more overrejection than in the conventional approach in our setting.

Several other recently developed methods for cluster-robust inference are conceptually unsuitable for the present setting, for example because they require an equal or large number of observations in each cluster (e.g., Donald and Lang 2007; Bester, Conley, and C. B. Hansen 2011; Ibragimov and Müller 2016), because the parameter of interest must be identified within each cluster (e.g., Canay, Romano, and Shaikh 2017[11]), or because one must at least be able to collect a clearly defined post-treatment indicator from each cluster (Hagemann 2019b; Hagemann 2019a (which is not the case in the difference-in-difference setting when treatment years vary[12]).[13]

5.1 Typical Robustness Checks in Empirical Corporate Finance

We begin by considering the performance of typical robustness checks in the empirical corporate finance literature: dropping Delaware from the sample, and/or examining a dynamic "event study" specification with lead/lag treatment dummies. For each robustness check, We repeat the placebo law exercise from section 3 to gauge performance.

[11] When treatment is assigned at the cluster level and hence not identified within cluster, Canay, Romano, and Shaikh 2017 implement their method by considering pairs of treated and untreated clusters. In the current setting, however, their postulate of pairs "suggested" by the treatment assignment is not met, such that this implementation seems unappealing.

[12] As in Ibragimov and Müller 2016, fn. 10, one might overcome this problem by considering only years before the first and after the last state adopted the statute in question. Given that at least some adoptions occur many years after others, however, this would either entail very considerable data loss or require dropping very early or late adopters or years. Of course, it might be preferable to restrict the estimates to a narrower window. This leads into broader questions of research design that are beyond the scope of the present paper.

[13] A related technical difference between the first two approaches and those considered below is in the type of asymptotics: the former consider asymptotics $n_s \to \infty$ for fixed S, whereas the latter, including the conventional cluster-robust inference exposited above, consider $S \to \infty$. To the extent $S \to \infty$ asymptotics have been assumed based on the nature of the setting rather than analytical convenience, an appropriate "fix" should retain that assumption.

5.1.1 Dynamic "Event Study" Test

For the dynamic specification, we estimate two specifications and combine the results. We first estimate the baseline estimating equation (2.1) as before. If the conventional CRVE test rejects the null hypothesis at the specified level, we then additionall estimate a modified specification where we replace the single treatment dummy with a dummy for the treatment year, dummies for each of the two years before treatment, a dummy for the first year after treatment, and a dummy for all subsequent years.[14] Denoting that last period "2+", we thus estimate the following equation:

$$y_{ij\ldots st} = \alpha_i + \sum_{\tau=-2}^{2+} \beta_\tau D_{st+\tau} + \delta' \mathbf{x_{it}} + \gamma' \mathbf{z_{j\ldots st}} + \epsilon_{ij\ldots st}$$

We reject the null hypothesis if and only if:

1. The treatment dummy β is statistically significant at the specified level (i.e., 1, 5, or 10%) in the baseline specification; *and*
2. The lead year dummies β_{-1}, β_{-2} are jointly statistically insignificant at the specified level; *and*
3. The two post dummies β_1, β_{2+} are jointly statistically significant at the specified level.

We also consider a variant where we test not simple joint significance but significance of the sum of the dummies (in steps 2. and 3., respectively), which will tend to eliminate cases where the two dummies are of opposite sign.

[Insert Figure 4 Here.]

Figure 4 shows the results, which are mixed. Even the joint test overrejects for smaller numbers of treated clusters. As the number of treated clusters increases to about 25 or more, however, the joint test achieves more or less nominal size. The test using the sums of the pre- and post-dummies, respectively, does a little better than the simple joint significance test, especially when Delware is not treated (left column). For 10% nominal test size (top row), the joint test even overrejects at larger numbers of treated clusters, especially when Delaware is treated (right column). but it asymp, however, this is far from being the case.

There is an issue with the joint dynamic tests, however, of which the underrejection at size 10% is a sign. If the joint tests were independent, then the combination of three tests (baseline, pre, post) should have $\alpha^3 \ll \alpha$ size. For example, the 1% test should falsely reject the null only with probability 10^-6. In actuality, the pre and post tests are not independent from one another because errors are allowed to be serially correlated, and neither is independent of the baseline test because the latter tests a linear combination of the former. Nevertheless, one should expect the joint size to be less than the nominal size of the ingredient tests, not equal to it. The coincidence of nominal and empirical size for larger numbers of treated clusters is thus partly a coincidence. Moreover, the stringency of the triple-test is likely to reduce power, which we indeed find below.

[14] This is the same as the specification used by Bertrand and Mullainathan 2003; Giroud and Mueller 2010 except that we added a second lead dummy, $D_{s,-2}$.

5.1.2 Dropping Delaware

A simpler alternative is to drop Delaware from the sample. This eliminates the worst part of the cluster size imbalance. However, as mentioned above and shown in figure 2, even without Delaware, incorporation cluster sizes are "wildly different" of a degree that has been found to generate severe overrejection in other settings (MacKinnon and Webb 2017). Unsurprisingly, overrejection remains large here after dropping Delaware. This is shown in the bottom left panel of figure 5.

[Insert Figure 5 Here.]

5.2 Cluster Wild Bootstrap

The top row of figure 5 shows results for the cluster wild bootstrap introduced by Cameron, Gelbach, and Miller 2008. The wild bootstrap uses restricted estimates and Rademacher weights in the bootstrap data generation process, since these generally perform better than alternatives in simulations of MacKinnon and Webb 2017; Djogbenou, MacKinnon, and Nielsen 2019; the number of bootstrap replications is set to 999.

The wild bootstrap's performance is much better than that of conventional CRVE inference analyzed in section 3. Nevertheless, there is still overrejection for all but small numbers of treated clusters when Delware is not treated (top left plot) or small numbers of treated clusters when Delware is treated (top right plot). There is also slight overrejection when moderate or large numbers of clusters are treated, more so when Delware is not treated.

These results are not surprising in as much as Djogbenou, MacKinnon, and Nielsen 2019, fig. 3 also find overrejection of the cluster wild bootstrap when one cluster contains half of all observations (using a continuous regressor). Recall that Delaware contains 62% of the observations. By contrast, MacKinnon and Webb 2017 show that the cluster wild bootstrap works well with degrees of cluster imbalance resembling incorporation cluster size imbalance *without* Delaware.

This suggests that the combination of the cluster wild bootstrap and dropping Delaware should lead to valid inference. The bottom right plot of figure 5 shows that this is indeed the case. The exception is very low numbers of treated clusters, a well-known special case presenting a separate problem (Imbens and Kolesár 2016; MacKinnon and Webb 2020). The main drawback would appear to be that dropping Delaware sacrifices 2/3 of the available data that could otherwise have been used to estimate the control variables with more precision, with a resulting loss in power (to which we return below).

5.3 Permutation Test on the Treatment Dummy

Finally, we consider a permutation test on the treatment dummy (DiCiccio and Romano 2017; MacKinnon and Webb 2020; Young 2020. The intuition of the test is a simple inversion of the placebo law logic of the preceding sections. That logic was that a valid test should not find a "significant" result with random placebo laws–a false positive–more frequently than the test's nominal size. While the preceding sections used this logic to criticize conventional tests, it can also be used constructively to formulate a valid test: the null hypothesis should be rejected if and only if the actual test statistic (4) is in the relevant tail of the empirical distribution of equivalent test statistics generated by placebo laws.[15]

[15]By contrast, Hagemann 2019b; Hagemann 2019a permutes cluster-year-specific intercept estimates, not placebo laws (i.e., independent variables) themselves. As mentioned above, the difficulty with this approach is that when

The main attraction of the permutation test is that it is exact in finite samples against the sharp null hypothesis that the treatment dummy has no effect on the distribution of the errors (e.g., Lehmann and Romano 2005, Theorem 15.2.2).[16] The sharp null is stronger than the null of no average treatment effect. In particular, it excludes the possibility that treated units have higher variance. Nevertheless, the sharp null seems attractive for a minimum hurdle test in the sense that if even a sharp null cannot be rejected, then any more specific claims inconsistent with the sharp null (e.g., of a particular treatment effect) should be viewed with deep suspicion. In simpler, univariate contexts, researchers frequently adopt this position implicitly by using the Fisher exact test, which is also valid (only) against a sharp null.

In addition, asymptotically, the permutation test on the treatment dummy can be conjectured to be asymptotically valid against the less stringent null hypothesis of zero average treatment effect, provided the test statistic is studentized. DiCiccio and Romano 2017, Theorem 3.3 prove this for permutation of a subset of regressors in a simple cross-sectional regression, and Young 2020 proves it for clustered regression under certain restrictive conditions. For this reason, and to eliminate mechanical uninteresting heteroskedasticity from cluster sizes etc; we use as the test-statistic the t-statistic 4 derived from the CRVE 3. MacKinnon and Webb 2020 investigate the performance of such a permutation test on a continuous regressor using the t-statistic (which they call the randomization-inference t-test, or RI-t), and find that it performs well in simulations even when the largest cluster contains half of all observations.

We permute the treatment dummy while holding Delaware's treatment status and the number of treated states constant. The reason to do so is that these factors have a mechanical effect on the distribution of the t-statistic, as shown by all the simulations thus far. Similarly, we hold and the years of treatment fixed in our permutation tests, because those too will have mechanical impacts on the test statistics by shifting weights (cf; eġ; Chaisemartin and D'Haultfœuille 2020). Mechanically, the permutation test will always have correct size against the sharp null no matter what is being permuted. However, there is no reason to discard the information that some treatment assignments will tend to generate larger test statistics *even under the null*, i.e., irrespective of any treatment effects. Holding such ancillary statistics fixed will also increase power. See Cox 1958; MacKinnon and Webb 2020 for related discussion.[17]

To investigate the finite sample performance of the permutation test, we can resort to more Monte Carlo simulations. To gain insight about the behavior of the permutation test when the randomization hypothesis does not hold, we generate cross-sectional data [**Note to readers: this will eventually be extended to panel data, but we have not been able to perform the extensive computations yet.**] with zero average treatment effect but treatment heteroskedasticity. A standard normal variable ξ constant within state is added to the data generating process only for treated observations, such that the error variance is about 10% higher and the intra-state correlation about double in treated relative to untreated states. 5,100 firms are divided between Delaware (2,550 firms) and 50 other states of 51 firms each. The data generating

treatment occurs in different years, it is not clear which cluster-year dummy to retain from each cluster.

[16] [**Note to readers of this draft:** DiCiccio and Romano 2017 formulate the permutation test for regression coefficients under the assumption that the treatment is independent of all the covariates, not only the error term. This seems unnecessarily strong but is something we are still looking into.]

[17] MacKinnon and Webb 2020, section 3.2 also argue for conditioning the permutation on cluster sizes. The difficulty with doing so is that this conditioning will always be imperfect. Empirically, MacKinnon and Webb 2020 find that this conditioning makes very little difference.

process is now

$$y_{ijs1} = \mu_{s1} + \eta_{ijs1} + D\xi_{s1}, \quad \mu, \xi \overset{iid}{\sim} \mathcal{N}(0,1), \quad \eta \overset{iid}{\sim} \mathcal{N}(0,9), \tag{6}$$

with μ, η, ξ, D, z mutually independent. We then estimate with OLS

$$y_{ijs1} = \alpha + \beta D \tag{7}$$

and calculate the t-statistic (4) using (3). Unlike before, however, the critical values for each test are now taken not from the t-distribution but from the empirical distribution of t-statistics generated by re-randomizing D across the 51 states (we use 1,000 re-randomizations per run of the simulation). Delaware is never treated, but this is without loss of generality because with cross-sectional data, the cases of S_1 treated clusters including Delaware are equivalent to $51 - S_1$ treated clusters excluding Delaware, and vice versa. We run the simulation 10,000 times per number of treated clusters.

[Insert Figure 6 Here.]

The top-right panel of figure 6 shows the results. For comparison, we also show results using the conventional CRVE approach (bottom left), the wild cluster bootstrap (bottom right), and a permutation test that does not preserve Delaware's treatment status (top left). While our preferred permutation test is not perfect (nor should it be, since the sharp null does not hold!), it does much better than the other tests. [**Note to readers: We have yet to reconcile the poor performance of the cluster wild bootstrap in these simulated cross-sectional data with its good performance with the real panel data.**]

6 Power

Finally, we consider power. To do so, we start with the CRSP/Compustat data described in section 3, focusing on (winsorized) Tobin's q. However, we augment the data by adding to the outcome variable, Tobin's q, for state-years randomly chosen to be treated. That is, we again generate random laws, but now they are not placebos but actually do have an effect. We add 3, 5, or 10% of a standard deviation of Tobin's q, which is 1.47 (the mean of Tobin's q in our data is 2.04). We do so 1,600 times for each number of treated states and whether Delaware is treated, not treated, or dropped, and for each of the three effect sizes. We then apply the respective test (permutation, wild bootstrap, or conventional CRVE), focusing on nominal test size 5%.

[Insert Figure 7 Here.]

Figure 7 graphs the power. The rows correspond to effect sizes (from top to bottom: 3, 5, 10%), whereas the columns correspond to Delaware's treatment status (from left to right: not treated, treated, dropped). The three lines within every plot correspond to the three different tests, as described in the legend.

Three results stand out. First, power is unacceptably low for 0.03 s.d. effect sizes and arguably also for 0.05 s.d. effect sizes, regardless of the test, i.e., even with the conventional CRVE test. When the true effect size is 0.03 s.d., power barely reaches 50% when all states are treated, including Delaware, and is mostly far below that. The rule of thumb of 80% power is never attained with 0.03 s.d. effect size, and attained for 0.05 s.d. effect size only for 40 treated clusters or more using the conventional CRVE (which, of course,

dramatically overrejects), or using the cluster wild bootstrap when all but Delaware are treated (where the bootstrap overrejects, see 5).

Second, even for effect sizes of 0.10 s.d., reliable inference–i.e., the permutation test, or the bootstrap when Delaware is dropped–attains 80% power only when Delaware is treated, or if at least 20 other states are treated.

Third, when they are both reliable, there is not much separating the power of the wild cluster bootstrap and the permutation test. In particular, there power is virtually identical when Delaware is dropped, which is the only situation when the bootstrap is true to size.

Finally, we compare the power of the permutation test to the power of the (joint) "dynamic" robustness check of lead/lag treatment dummies described in section 5.1.1. This is shown in figure 8, again for tests of nominal size 5%. As hinted in section 5.1.1, the joint dynamic test is conservative. Its power is uniformly less than that of the permutation test, except for low numbers of treated clusters where the joint dynamic test overrejects.

[Insert Figure 8 Here.]

7 Conclusion

Using Monte Carlo simulations, this paper demonstrates severe problems with conventional inference when using state corporate laws for identification of corporate governance effects in firm-level data, in particular the popular difference-in-difference panel approach. The paper shows that even the cluster wild bootstrap struggles to deal with the extreme imbalance in incorporation state cluster sizes. The paper proposes a permutation test to address this problem along the lines of DiCiccio and Romano 2017; MacKinnon and Webb 2020. The permutation test is exact under the randomization hypothesis, and shows promising performance superior to alternative tests in Monte Carlo simulations even when the randomization hypothesis does not hold.

Whether or not the permutation test proposed here will ultimately be adopted, researchers need to do something to address the severe inferential challenge posed by unequal cluster sizes. Importantly, while this paper has focused on demonstrating the worst problem originating from Delaware's dominant size, simply omitting Delaware firms from the sample will not solve all issues. Even without Delaware, incorporation cluster sizes are very unequal, and conventional inference still strongly overrejects; only when paired with the wild cluster bootstrap does dropping Delaware solve the problem.

Beyond the specifics of the tests, this paper can also be read as another cautionary tale about trying to find relatively small effects in noisy data using complex methods such as high-dimensional fixed effect models (cf. Young 2019 on instrumental variables). The high number of firm-year observations may mislead one into thinking that even small effects should be detectable. Once it is realized that the number of clusters is the relevant degrees of freedom for inference, however, it becomes clear that power will often be an issue. This is especially so because, as Carter, Schnepel, and Steigerwald 2017 have shown, the rate of convergence of the variance estimator is governed by the *effective* number of clusters, which with incorporation clusters is generally in the low single digits. The specifics will depend on the hypothesized effect size, the noisiness of the dependent variable, the distribution of the treatment assignment, and the ability to control for known predictors. Fortunately, modern computing power offers the ability to perform

custom-made power calculations even for complex problems with relative ease. Researchers can and should also check the performance of their statistical tests specifically under the conditions that they are studying using methods such as those discussed in this paper.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge (2017). "When Should You Adjust Standard Errors for Clustering?"

Angrist, Joshua D. and Jörn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). "How Much Should We Trust Differences-In-Differences Estimates?" In: *The Quarterly Journal of Economics* 119.1, pp. 249–275.

Bertrand, Marianne and Sendhil Mullainathan (2003). "Enjoying the Quiet Life? Corporate Governance and Managerial Preferences". In: *Journal of Political Economy* 111.5, pp. 1043–1075.

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen (2011). "Inference with dependent data using cluster covariance estimators". In: *Journal of Econometrics* 165.2, pp. 137–151.

Bharath, Sreedhar T and Michael Hertzel (2019). "External Governance and Debt Structure". In: *The Review of Financial Studies* 32.9, pp. 3335–3365.

Borusyak, Kirill and Xavier Jaravel (2018). "Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume".

Cain, Matthew D., Stephen B. McKeon, and Steven Davidoff Solomon (2017). "Do takeover laws matter? Evidence from five decades of hostile takeovers". In: *Journal of Financial Economics* 124.3, pp. 464–485.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2008). "Bootstrap-Based Improvements for Inference with Clustered Errors". In: *The Review of Economics and Statistics* 90.3, pp. 414–427.

Cameron, A. Colin and Douglas L. Miller (2015). "A Practitioner's Guide to Cluster-Robust Inference". In: *The Journal of Human Resources* 50.2, pp. 317–372.

Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh (2017). "Randomization Tests Under an Approximate Symmetry Assumption". In: *Econometrica* 85.3, pp. 1013–1030.

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

Catan, Emiliano M. and Marcel Kahan (2016). "The Law and Finance of Antitakeover Statutes". In: *Stanford Law Review* 68, pp. 629–680.

Chaisemartin, Clément de and Xavier D'Haultfœuille (2020). "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects". In: *American Economic Review* 110.9, pp. 2964–96.

Coates IV, John C. (2000). "Takeover Defenses in the Shadow of the Pill: A Critique of the Scientific Evidence". In: *Texas Law Review* 79.2, pp. 271–382.

Cohen, Alma and Charles C.Y. Wang (2013). "How do staggered boards affect shareholder value? Evidence from a natural experiment". In: *Journal of Financial Economics* 110.3, pp. 627–641.

Correia, Sergio (2016). *Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator*. Tech. rep. Working Paper.

Cox, D. R. (1958). "Some Problems Connected with Statistical Inference". In: *Ann. Math. Statist.* 29.2, pp. 357–372.

Cremers, K.J. Martijn, Scott B. Guernsey, and Simone M. Sepe (2019). "Stakeholder Orientation and Firm Value".

Demiroglu, Cem, Cansu Iskenderoglu, and Oguzhan Ozbas (2019). "Managerial Discretion and Efficiency of Internal Capital Markets".

DiCiccio, Cyrus J. and Joseph P. Romano (2017). "Robust Permutation Tests For Correlation And Regression Coefficients". In: *Journal of the American Statistical Association* 112.519, pp. 1211–1220.

Djogbenou, Antoine A., James G. MacKinnon, and Morten Ørregaard Nielsen (2019). "Asymptotic theory and wild bootstrap inference with clustered errors". In: *Journal of Econometrics* 212 (2), pp. 393–412.

Donald, Stephen G. and Kevin Lang (2007). "Inference with Difference-in-Difference and Other Panel Data". In: *The Review of Economics and Statistics* 89, pp. 221–233.

Giroud, Xavier and Holger M. Mueller (2010). "Does corporate governance matter in competitive industries?" In: *Journal of Financial Economics* 95.3, pp. 312–331.

Gompers, Paul, Joy Ishii, and Andrew Metrick (2003). "Corporate Governance and Equity Prices". In: *The Quarterly Journal of Economics* 118.1, pp. 107–156.

Goodman-Bacon, Andrew (2020). "Difference-in-Differences with Variation in Treatment Timing".

Gormley, Todd A. and David A. Matsa (2013). "Common Errors: How to (and Not to) Control for Unobserved Heterogeneity". In: *The Review of Financial Studies* 27.2, pp. 617–661.

Gutiérrez Urtiaga, María and Antonio B. Vazquez (2019). "Boards of Directors' Legal Incentives and Firm Outcomes".

Hagemann, Andreas (2019a). "Permutation Inference with a Finite Number of Heterogeneous Clusters".

— (2019b). "Placebo inference on treatment effects when the number of clusters is small". In: *Journal of Econometrics* 213 (1), pp. 190–209.

Hansen, Bruce E. and Seojeong Lee (2019). "Asymptotic theory for clustered samples". In: *Journal of Econometrics* 210.2, pp. 268–290.

He, Zhaozhao and David A. Hirshleifer (2019). "The Exploratory Mindset and Corporate Innovation".

Heath, Davidson, Matthew C. Ringgenberg, Mehrdad Samadi, and Ingrid M. Werner (2020). "Reusing Natural Experiments".

Ibragimov, Rustam and Ulrich K. Müller (2016). "Inference with Few Heterogeneous Clusters". In: *The Review of Economics and Statistics* 98.1, pp. 83–96.

Imai, Kosuke and In Song Kim (2020). "On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data".

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

Karpoff, Jonathan M. and Michael D. Wittry (2018). "Institutional and Legal Context in Natural Experiments: The Case of State Antitakeover Laws". In: *The Journal of Finance* 73.2, pp. 657–714.

Lee, Chang Hyung and Douglas G. Steigerwald (2018). "Inference for Clustered Data". In: *The Stata Journal* 18.2, pp. 447–460.

Lehmann, Erich L. and Joseph P. Romano (2005). *Testing Statistical Hypotheses*. Springer-Verlag New York.

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

MacKinnon, James G. and Matthew D. Webb (2019; revised May 2020). *When and How to Deal with Clustered Errors in Regression Models*. Working Paper 1421. Kingston, Ontario: Queen's Economics Department.

— (2017). "Wild Bootstrap Inference for Wildly Different Cluster Sizes". In: *Journal of Applied Econometrics* 32.2, pp. 233–254.

— (2020). "Randomization Inference For Difference-in-differences With Few Treated Clusters". In: *Journal of Econometrics* 218.

Petersen, Mitchell A. (2009). "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches". In: *Review of Financial Studies* 22.1, pp. 435–480.

Romano, Joseph P. and Michael Wolf (2005). "Stepwise Multiple Testing as Formalized Data Snooping". In: *Econometrica* 73.4, pp. 1237–1282.

Spamann, Holger and colby Wilkinson (2019). *Historic State-of-Incorporation Data 1994-2019 (data and EDGAR scraping R script)*. Version 1.0.

Sun, Liyang and Sarah Abraham (2020). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". In: *Journal of Econometrics*.

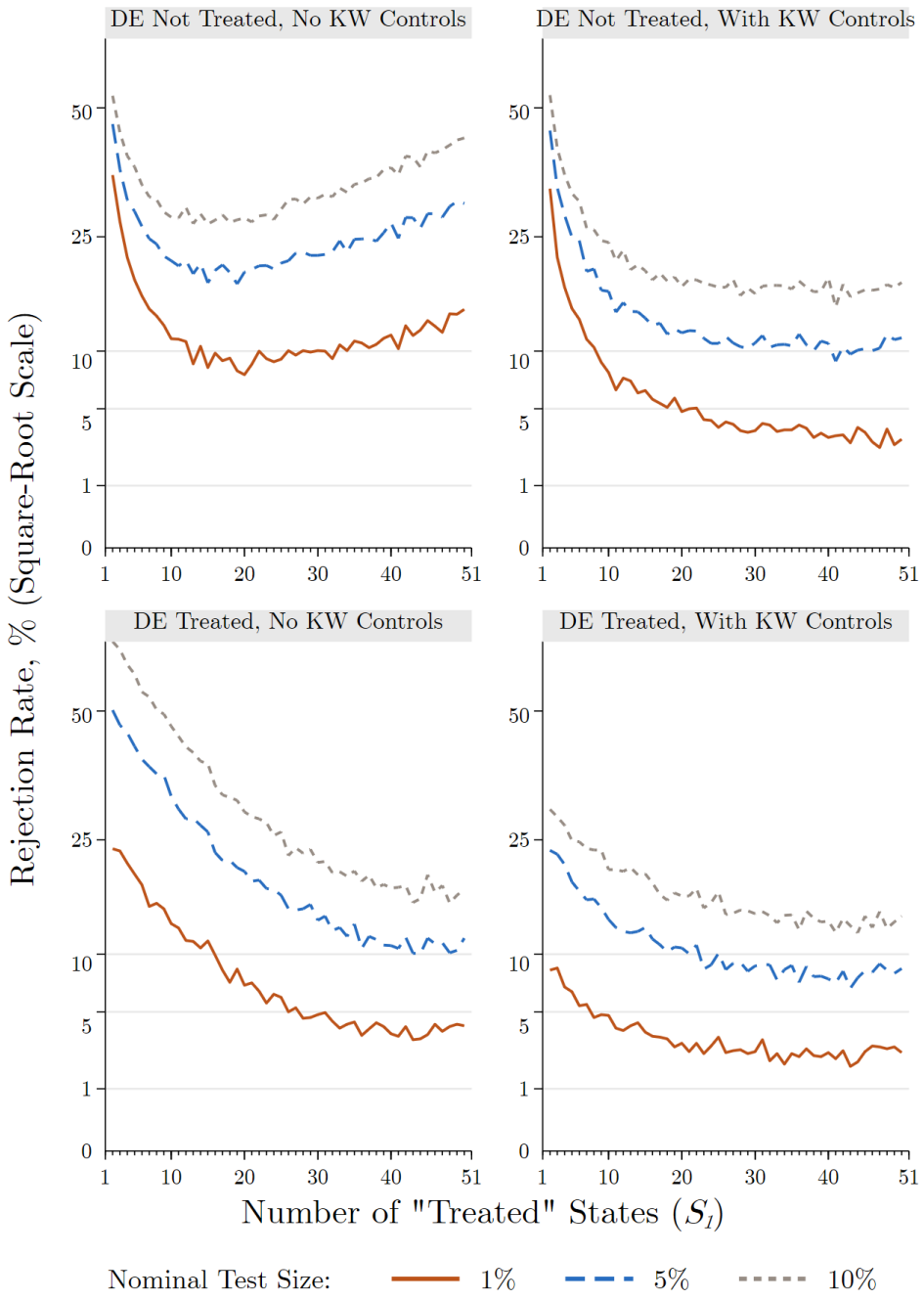
White, Halbert (1984). *Asymptotic Theory for Econometricians*. Academic Press.

Young, Alwyn (2016). "Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices using Effective Degrees of Freedom Corrections".

— (2019). "Consistency Without Inference: Instrumental Variables in Practical Application".

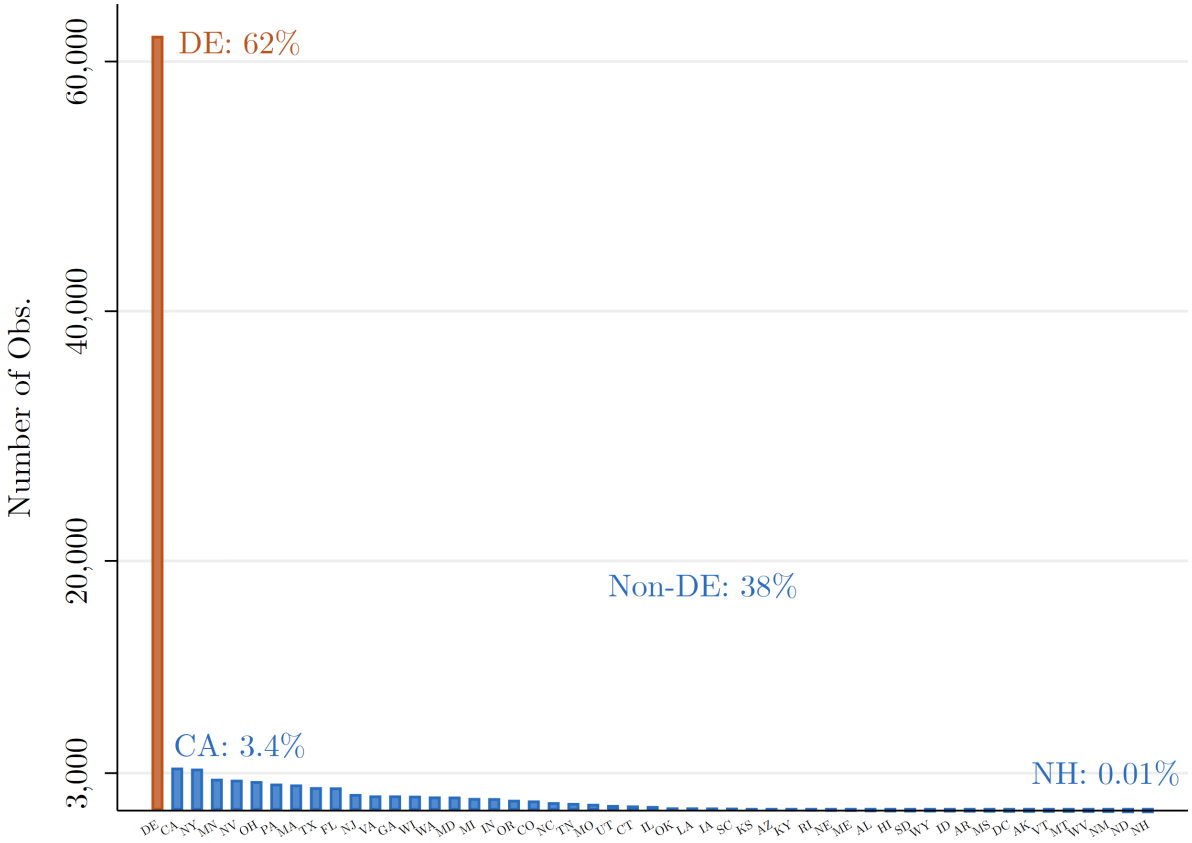
— (2020). "Asymptotically Robust Randomization Confidence Intervals for Parametric OLS Regression".

Figure 1. Rejection Rates for Placebo Laws (CRVE, Real Data)



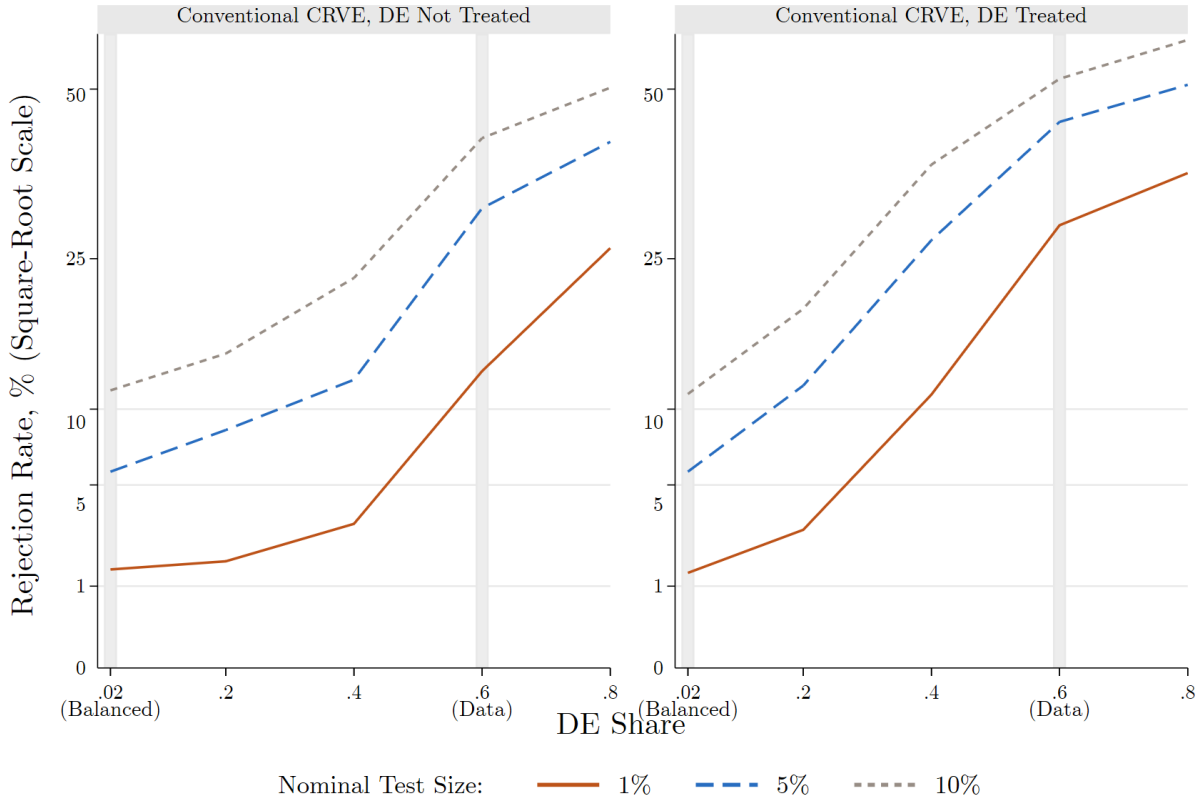
Notes. This figure presents the empirical rejection rates of conventional tests in real CRSP/Compustat data 1983-2018 using critical values from the t-distribution with 50 degrees of freedom and t-statistics calculated from the "sandwich" variance estimator clustered by state of incorporation (CRVE). The dependent variable is Tobin's q. The outcome variable is winsorized at the 2.5% and 97.5% levels. Rejection rates are calculated from 2,000 randomly generated placebo laws for each number of "treated" states, each DE "treatment" status, and whether KW controls are included or not in the specification.

Figure 2. Cluster Sizes for States of Incorporation



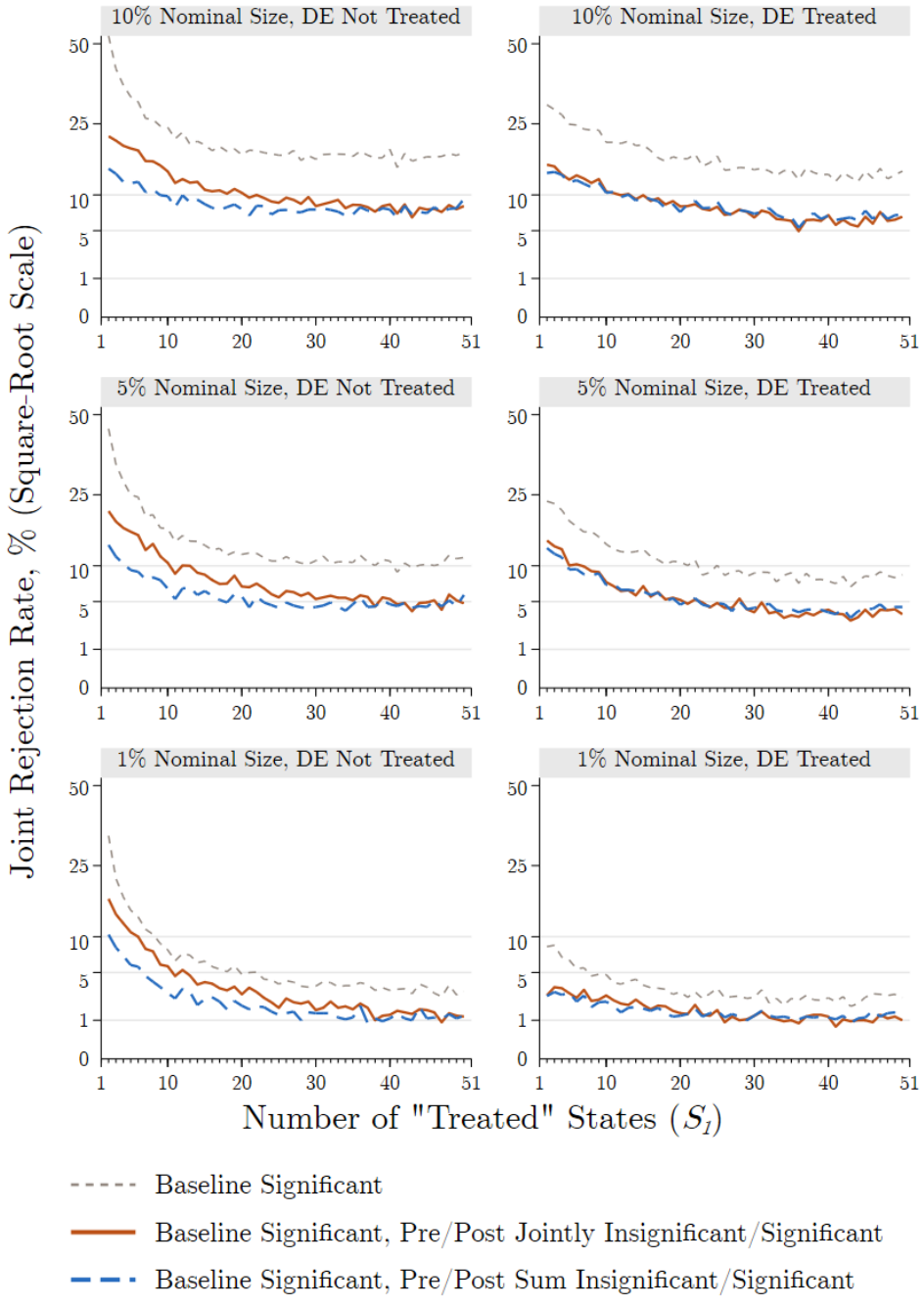
Notes. This figure presents the distribution of incorporation cluster sizes, i.e., how many firm-years in CRSP/Compustat data 1983-2018 belong to each of the 50 states and the District of Columbia. States are ordered by cluster size.

Figure 3. Cluster Imbalance and Overrejection (CRVE, Simulated Data)



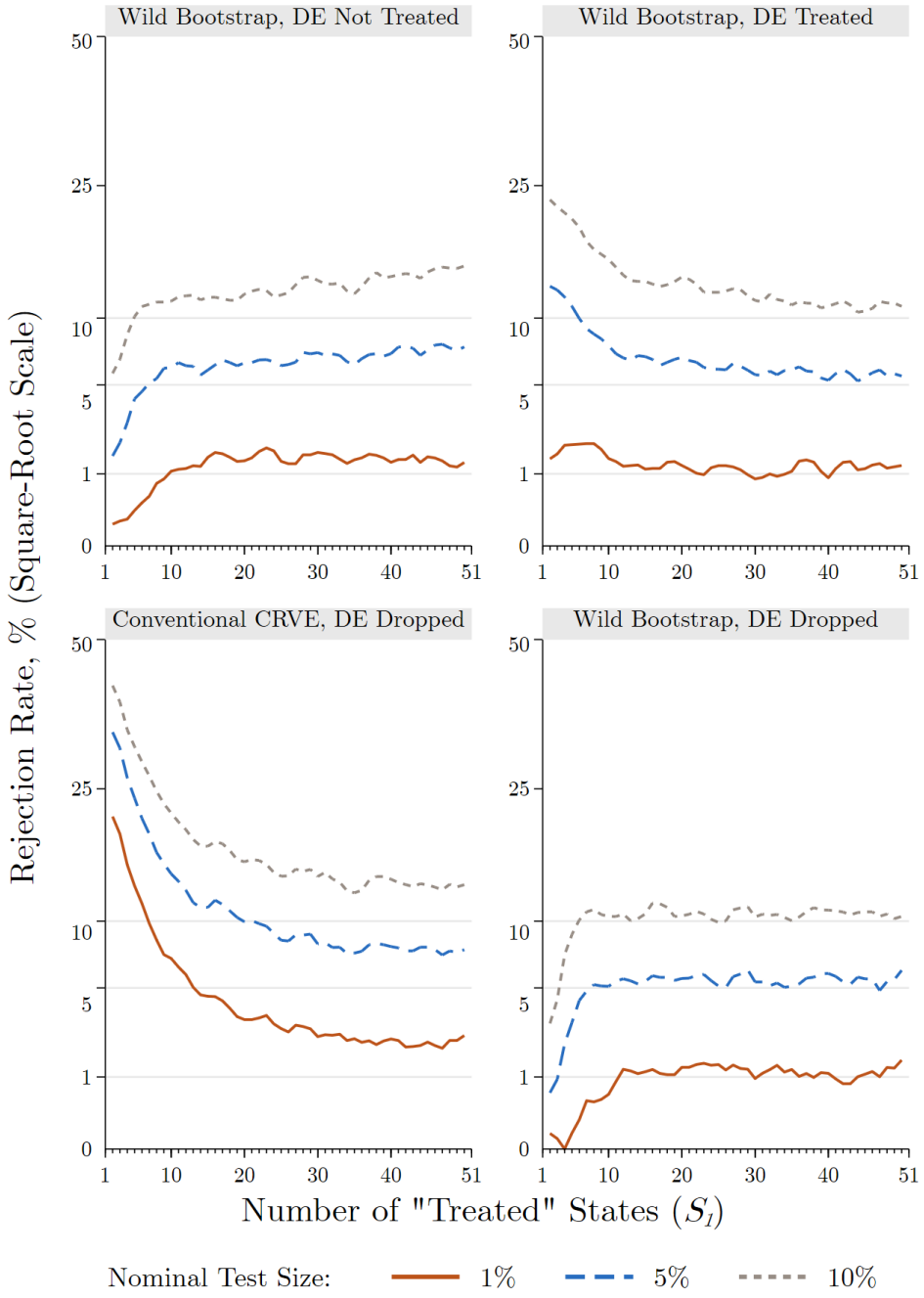
Notes. This figure presents the relation between the degree of cluster imbalance and the degree of overrejection of conventional inference with the CRVE. The empirical rejection rates are calculated from 2,000 simulated panels by conventional tests using critical values from the t-distribution with 50 degrees of freedom and t-statistics calculated from the "sandwich" variance estimator clustered by state of incorporation (CRVE). Degree of cluster imbalance is governed by the share of firms incorporated in DE.

Figure 4. Rejection Rates for Placebo Laws
(Joint Tests With Dynamic Specification)



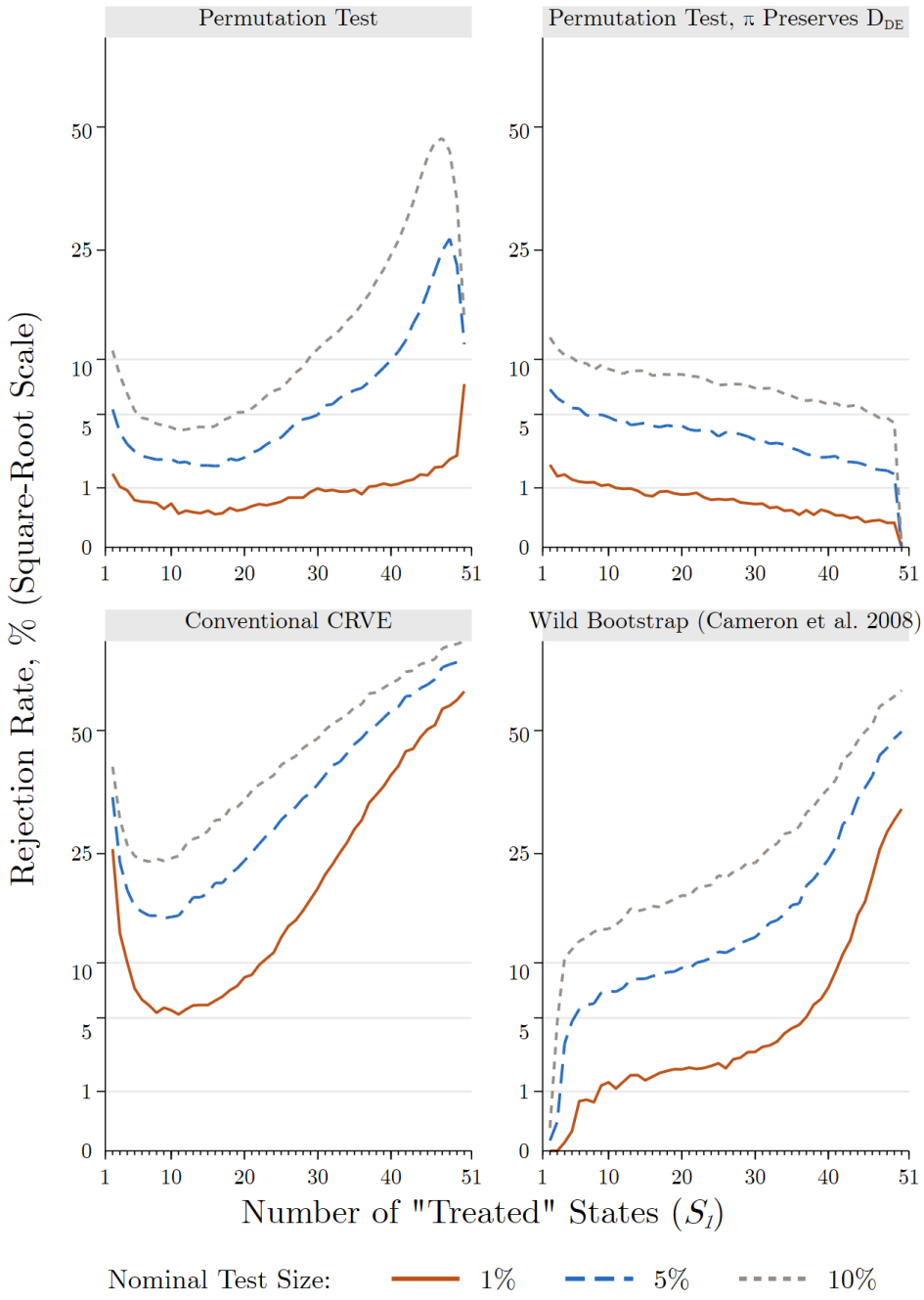
Notes. This figure presents the empirical rejection rates for the joint test of baseline specification and dynamic specification. The dependent variable is Tobin's q. The outcome variable is winsorized at the 2.5% and 97.5% levels. KW controls are included in the specification. Rejection rates are calculated from 2,000 randomly generated placebo laws for each number of "treated" states and each DE "treatment" status.

Figure 5. Rejection Rates for Placebo Laws (Wild Bootstrap and Drop DE)



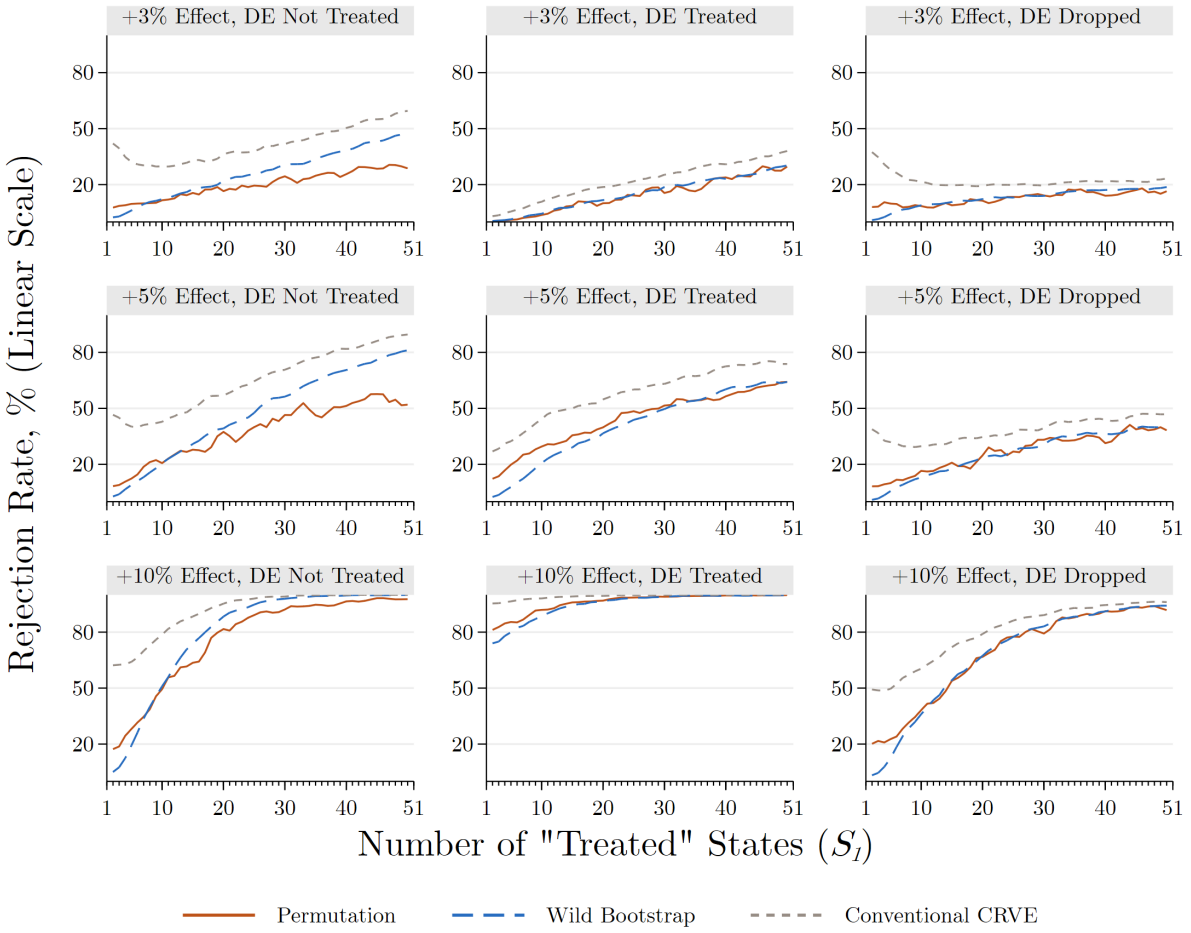
Notes. This figure presents the empirical rejection rates of cluster Wild Bootstrap (by state of incorporation) and empirical rejection rates of CRVE on the DE-dropped sample. The dependent variable is Tobin's q. The outcome variable is winsorized at the 2.5% and 97.5% levels. KW controls are included in the specification. Rejection rates are calculated from 2,000 randomly generated placebo laws for each number of "treated" states and each DE "treatment" status.

Figure 6. Rejection Rates for Heteroskedasticity (Permutation Tests and Alternative Tests)



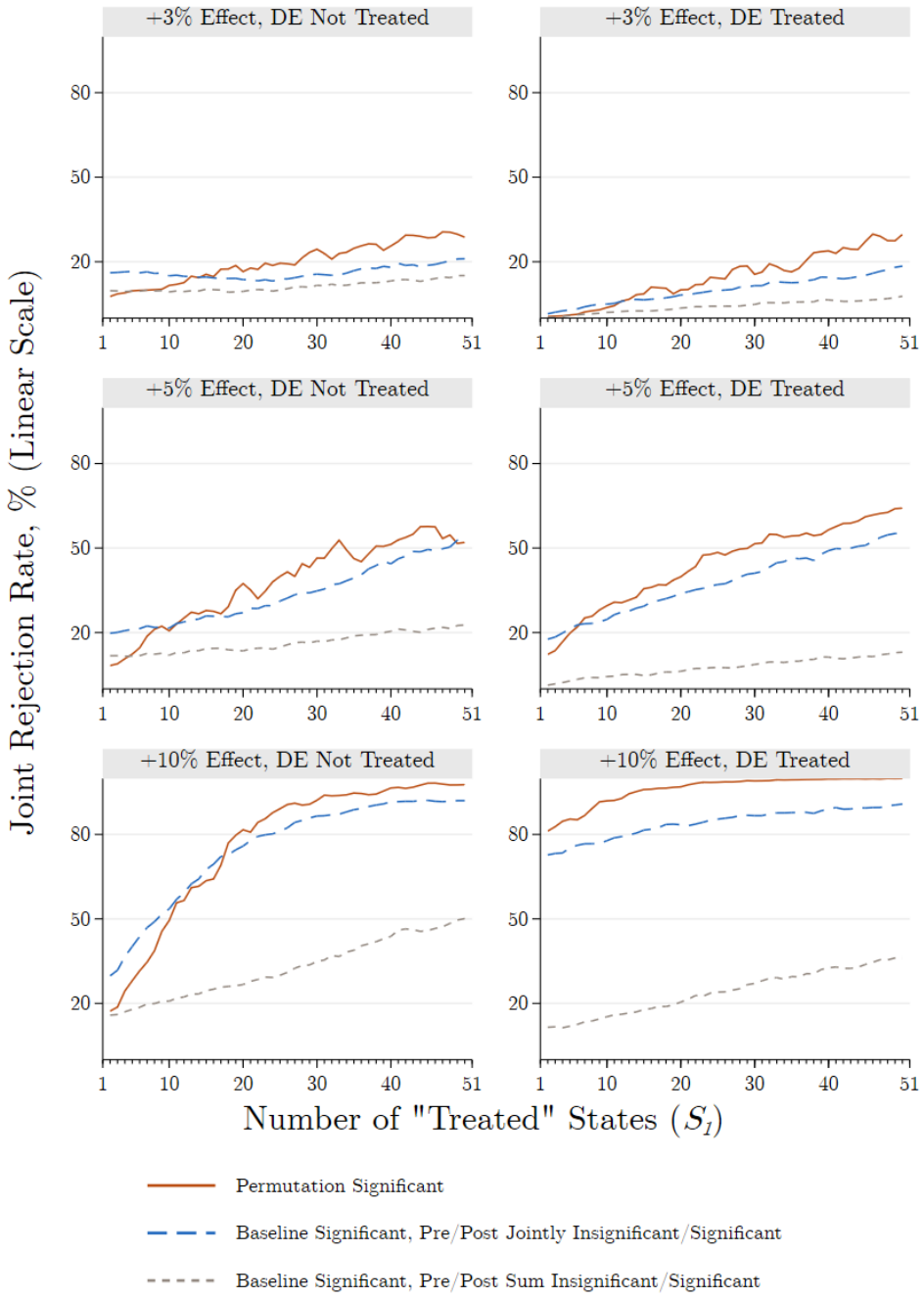
Notes. This figure presents the empirical rejection rates of permutation test, CRVE, and cluster Wild Bootstrap (by state of incorporation) on simulated cross-sectional data. Rejection rates are calculated from 10,000 simulated data each method and each number of treated states.

Figure 7. Power: Permutation Test, Wild Bootstrap, and CRVE



Notes. This figure presents the empirical power for the permutation test, Wild Bootstrap, and CRVE. The dependent variable is Tobin's q plus an additional 3/5/10% effect induced by placebo laws. The outcome variable is winsorized at the 2.5% and 97.5% levels. KW controls are included in the specification. The nominal test size is 5%. Powers are calculated from 1,600 randomly generated laws for each number of "treated" states, each DE "treatment" status, and each effect size.

Figure 8. Power: Joint Tests With Dynamic Specification



Notes. This figure presents the empirical power for the the joint test of baseline specification and dynamic specification. The dependent variable is Tobin's q plus an additional 3/5/10% effect induced by placebo laws. The outcome variable is winsorized at the 2.5% and 97.5% levels. KW controls are included in the specification. The nominal test size is 5%. Powers are calculated from 1,600 randomly generated laws for each number of "treated" states, each DE "treatment" status, and each effect size.

Table 1. Average Rejection Rates for Placebo Laws (Other Outcomes)

	10% Nominal Size		5% Nominal Size		1% Nominal Size	
Dep. Var	DE Treated	DE Not Treated	DE Treated	DE Not Treated	DE Treated	DE Not Treated
ROA	0.47	0.38	0.38	0.27	0.23	0.13
	(0.11)	(0.10)	(0.12)	(0.09)	(0.13)	(0.06)
Leverage	0.23	0.22	0.14	0.15	0.04	0.07
	(0.07)	(0.09)	(0.05)	(0.09)	(0.01)	(0.08)
Capex	0.25	0.20	0.16	0.13	0.06	0.05
	(0.07)	(0.09)	(0.06)	(0.09)	(0.04)	(0.08)
PPE Growth	0.24	0.16	0.16	0.10	0.07	0.04
	(0.02)	(0.09)	(0.02)	(0.09)	(0.01)	(0.07)
Asset Growth	0.17	0.19	0.10	0.13	0.03	0.06
	(0.01)	(0.11)	(0.01)	(0.10)	(0.01)	(0.09)
Cash	0.21	0.21	0.13	0.15	0.05	0.07
	(0.02)	(0.09)	(0.01)	(0.09)	(0.01)	(0.07)
SGA Expense	0.06	0.15	0.03	0.10	0.004	0.04
	(0.02)	(0.10)	(0.01)	(0.10)	(0.002)	(0.08)

Notes. This table shows the average empirical rejection rates of conventional tests (CRVE) across different numbers of "treated" states (from 1 to 51), using critical values from the t-distribution with 50 degrees of freedom and t-statistics calculated from the "sandwich" variance estimator clustered by state of incorporation. Standard deviations of rejection rates are displayed in parentheses. The outcome variables include ROA, leverage, Capex, PPE growth, asset growth, cash, and SGA expense. ROA equals EBITDA divided by total assets; Capex, PPE, cash, and SGA expenses are scaled by total assets; PPE and asset growth are the percentage change in PPE and total assets, respectively; and leverage equals debt divided by total assets. All outcome variables are winsorized at the 2.5% and 97.5% levels. KW controls are included in the specification. Rejection rates are calculated from 2,000 randomly generated placebo laws for each number of "treated" states and each DE "treatment" status.