Practical and Effective Approaches to Dealing with Clustered Data*

Justin Esarey† and Andrew Menger‡

October 2, 2017

Abstract

Cluster-robust standard errors (as implemented by the eponymous `cluster` option in Stata) can produce misleading inferences when the number of clusters G is small, even if the model is consistent and there are many observations in each cluster. Nevertheless, political scientists commonly employ this method in data sets with few clusters. The contributions of this paper are: (a) developing new and easy-to-use Stata and R packages that implement alternative uncertainty measures robust to small G, and (b) explaining and providing evidence for the advantages of these alternatives, especially cluster-adjusted t-statistics based on Ibragimov and Müller (2010). To illustrate these advantages, we reanalyze recent work by Grosser, Reuben and Tymula (2013), Lacina (2014), and Hainmueller, Hiscox and Sequeira (2015) whose results are based on cluster-robust standard errors.

Introduction

The cluster-robust standard error (CRSE) first proposed by Liang and Zeger (1986) has become ubiquitous in applied quantitative work in political science since it was implemented in Stata by Rogers (1993). The reasons for its ubiquity are straightforward: the problems that clustered data present for statistical analysis are well-known to political scientists (Moulton, 1986, 1990), and CRSEs are extremely simple to estimate and useful when added to a research design involving fixed effects estimation (Bertrand, Duflo and Mullainathan, 2004).

*The authors thank Ulrich Müller, Carlisle Rainey, Jonathan Kropko, Matthew Webb, Neal Beck, Jens Hainmueller, Shuai Jin, Jens Grosser, Ernesto Reuben, our anonymous reviewers, and participants at the 2015 Annual Meeting of the Midwest Political Science Association, the 2015 Annual Meeting of the Society for Political Methodology, and the 2016 Annual Meeting of the Southern Political Science Association for helpful comments and suggestions on earlier drafts of this paper.

†Assistant Professor of Political Science, Rice University. Corresponding author: justin@justinesarey.com.

‡Department of Political Science, Rice University.

But it is not as widely understood that CRSEs have significant limitations, and that using CRSEs without regard to these limitations can produce very misleading inferences. Research has shown that CRSE confidence intervals are too narrow and false positive rates[1] are substantially in excess of the nominal size of a statistical hypothesis test when the number of clusters in a data set is small (Green and Vavreck 2008; Cameron, Gelbach and Miller 2008; Harden 2011; Angrist and Pischke 2009, Chapter 8). There is no universal cutoff for how many clusters is "small;" prior research tends to put the threshold of elevated concern somewhere around 40 clusters, although this threshold can be much higher in some circumstances (MacKinnon and Webb, 2017). Nevertheless, we find that political scientists are still very likely to employ CRSEs in this situation. We speculate that CRSEs are still used in data sets with few clusters because (a) no alternatives have been made as easy to implement in Stata and R, and (b) the limitations of CRSEs have not yet been sufficiently publicized in political science.

Our paper makes two contributions to the political science literature. Our primary contribution is to make it easy for substantive researchers to use alternatives to the CRSE as a normal part of their workflow, particularly cluster-adjusted t-statistics (or CATs) based on Ibragimov and Müller (2010). Toward this end, we create and make available pre-packaged routines for estimating CATs, pairs cluster bootstrapped t-statistics (PCBSTs) (Bertrand, Duflo and Mullainathan, 2004; Cameron, Gelbach and Miller, 2008; Harden, 2011), and wild cluster bootstrapped t-statistics (WCBSTs) (Cameron, Gelbach and Miller, 2008) for common models. These routines are in the `clusterSEs` package for R and the `clustse` and `clusterbs` ado files for Stata.[2]

We also explain why these alternatives perform better than CRSEs in data sets with a small number of clusters and provide simulation evidence in favor of this argument. In our

[1] A false positive occurs when a true null hypothesis is rejected.

[2] Note that the cluster wild bootstrap is already introduced in a Stata .do file by Doug Miller (http://www.econ.ucdavis.edu/faculty/dlmiller/statafiles/), and was implemented by Judson Caskey (https://sites.google.com/site/judsoncaskey/data) in a straightforward package. We therefore use Caskey's code for WCBSTs in Stata, and provide our own code for R.

simulations, we find that CATs are usually more effective (and always at least as effective) at limiting false positives compared to CRSEs, PCBSTs, and WCBSTs. CATs are also often more powerful at detecting *true* positives (rejections of the null hypothesis when the null is false) compared to the alternatives. Although CATs cannot be estimated in any cluster with unidentified coefficients (e.g., without variation on the dependent variable), we provide both an analytic argument and simulation evidence that simply dropping these clusters allows CATs to be effectively used in most cases. Finally, although random effects models have better size and power characteristics in our simulation compared to any cluster adjustment alternative for very small numbers of clusters when the model is correctly specified, cluster adjustments do substantially better when the assumptions of the random effects model are not satisifed.

To substantively illustrate how CRSEs can be misleading in data sets with few clusters, we re-examine three recently published analyses by Grosser, Reuben and Tymula (2013), Lacina (2014), and Hainmueller, Hiscox and Sequeira (2015) using alternative approaches to clustering. Grosser, Reuben and Tymula (2013) observed that candidates for office in an experiment lower their proposals for redistribution in response to increased donations from a rich voter; our reanalysis with CATs finds that this relationship is more uncertain than shown by the original CRSE-based analysis (and in some cases is statistically insignificant). In Lacina (2014), we find little evidence in the replication dataset for a link between political representation and civil unrest in India when using PCBSTs; this is contrary to Lacina's original conclusion derived using CRSEs. Our reanalysis of the evidence in Hainmueller, Hiscox and Sequeira (2015) generally supports the authors' conclusion that consumers are willing to pay more for products with a "fair trade" label; however, we find that there is substantially more uncertainty in the magnitude of this relationship using CATs or PCBSTs than the original analysis using CRSEs would indicate.

The analysis of clustered data: problems and solutions

Data in political science is frequently grouped by some sort of structure; as one example, survey observations of individual respondents are often clustered by geographical units (counties, states, countries, etc.). We expect that respondents inside of a cluster are related to one another in complex ways that may be difficult to understand or model. For statistical analysis, this poses the difficulty that the random component of observed outcomes cannot be treated as independently and identically distributed within a cluster. Accounting for this statistical dependence between observations is a well-known problem in the statistical literature. In datasets where the data exhibits some dependency between observations in the same cluster, ignoring this dependence can greatly underestimate the true standard errors for parameters of interest when this clustering structure is associated with the independent variable (Moulton, 1986, 1990). This can lead to researchers falsely rejecting the null hypothesis of a statistical significance test too frequently, resulting in an excess of published papers with conclusions that are not supported by the data.

There are many approaches to dealing with clustered structure in a data set of interest. When a fixed effects model is desired (e.g., as a part of a difference-in-difference research design), it is common to correct the standard errors for residual cluster dependency (Bertrand, Duflo and Mullainathan, 2004). In this scenario, researchers commonly employ the `cluster` option in Stata (developed by Rogers, 1993). This procedure is a modification of White's (1980) robust standard errors, altering the White "sandwich estimator" to allow for dependence between observations inside a cluster. CRSEs were described in the context of generalized estimating equations by Liang and Zeger (1986), and were implemented in Stata by Rogers (1993).[3] The formula for the cluster-robust variance-covariance matrix in OLS

[3]See also Arellano (1987).

regression is (Cameron and Miller, 2015, pp. 8-9):

$$\text{var}\hat{\beta} = (X'X)^{-1}\left[\sum_{g=1}^{G}\left[X_g'\hat{u}_g\hat{u}_g'X_g\right]\right](X'X)^{-1} \tag{1}$$

where $\hat{\beta}$ are OLS estimates, G is the number of clusters, X is the $N \times m$ matrix of independent variables (for N observations and m variables), and $\hat{u}_g$ is the vector of residuals $y_g - X_g\hat{\beta}$ in cluster g. Theorem 2 in Liang and Zeger (1986, p. 16) demonstrates that, for models with intra-cluster error dependence but inter-cluster independence, the maximum likelihood estimator of $\hat{\beta}$ is consistent and multivariate Gaussian as $G \to \infty$; the covariance of this distribution is consistently estimated by equation (1) as $G \to \infty$. The result also holds (with suitable adjustment of equation (1)) for other GLM models.

Since the introduction of cluster robust standard errors to social science research, a series of papers have emphasized their ability to obtain accurate measures of uncertainty (and appropriately sized statistical significance tests) in a wide variety of scenarios (Liang and Zeger, 1993; Donner, 1998; Williams, 2000; Klar and Donner, 2001; Kezdi, 2004; Bertrand, Duflo and Mullainathan, 2004). Of course, CRSEs are not a cure-all. Differences between the typical maximum likelihood standard errors and CRSEs can indicate harmful misspecification problems that are not addressed by adjusting the variance-covariance matrix (Hardin and Hilbe, 2003, pp. 33-34). In these cases, CRSEs can be used as a check on the appropriateness of the model's specification (King and Roberts, 2014). CRSEs can also be useful in cases where explicitly modeling some aspect of the data generating process is problematic but a simplified model is still consistent (see, e.g., Cameron and Trivedi, 2005, pp. 147-150). For example, Bertrand, Duflo and Mullainathan (2004) report that parametrically modeling residual serial correlation results in excess false positive rates for placebo treatments in a Monte Carlo analysis of Current Population Survey data and simulated data; CRSEs are more effective at limiting false positives in their study.[4]

[4]Another alternative to dealing with serial dependence, explicitly including a lagged dependent variable, can be problematic if it creates "Nickell bias" due to the presence of fixed effects in the model (Nickell, 1981;

Unfortunately, evidence reveals a major problem with using CRSEs in datasets that have a small number of clusters: using CRSEs when the number of clusters is small can cause models to find statistically significant relationships where no relationships actually exist. That is, when only a small number of clusters are used in an analysis with clustered standard errors, the CRSEs are biased downward (Mancl and DeRouen, 2001; Cameron, Gelbach and Miller, 2008; Donald and Lang, 2007; Angrist and Pischke, 2009; Ibragimov and Müller, 2010; Imbens and Kolesar, 2012). The CRSE procedure depends on an asymptotic justification that the number of clusters G (and *not* the number of observations per cluster) approaches infinity (Hansen, 2007; Cameron, Gelbach and Miller, 2008). This can be seen in equation (1), where the center summation happens over the number of clusters G and not over the number of observations N; as a result, the accuracy and stability of the estimate relies on having access to many clusters, *not* many observations, because consistency of the center summation depends on $G \to \infty$ (Cameron and Miller, 2015, pp. 7-9). Intuitively, if one has a data set with N many observations in G many clusters but is unwilling to assume that observations in different clusters are identically distributed, one is (in a sense) estimating standard errors based on G many items of information.[5] Consequently, any asymptotically-derived results for the distribution of $\hat{\beta}$ will not apply unless G is very large, and significance tests or confidence intervals based on these asymptotically-derived distributions will be inaccurate when G is small.

There is no hard-and-fast rule for how few clusters is too few for using CRSEs. The work of Cameron, Gelbach and Miller (2008), Angrist and Pischke (2009), and Harden (2011) suggests that data sets with fewer than about 40 clusters are at substantively elevated risk of having downward-biased CRSEs. However, this threshold can vary under different circumstances; for example, simulations by MacKinnon and Webb (2017) indicate that CRSEs

Gaibulloev, Sandler and Sul, 2014; Beck, Katz and Mignozzetti, 2014).

[5]Indeed, one very simple proposal to correct downward bias in CRSEs is to use $G - 1$ degrees of freedom (rather than $N - k$, with k the number of estimated parameters) when conducting t-tests; unfortunately, this procedure still results in excess false positive results (Cameron and Miller, 2015, p. 29), as we confirm in our own Monte Carlo simulations. See footnote 14.

can be problematic with as many as 100 clusters if the clusters have very different numbers of observations.

How often are CRSEs used with few clusters in Political Science?

How often do political scientists use CRSEs in data sets with few clusters in published quantitative work? To answer this question, we gathered data from four highly visible general interest political science journals: *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, and *International Studies Quarterly* (hereafter known by their initials). We examined the use of cluster robust standard errors in every published article starting with the most recent issue (as of July 2014) and going back four years; we only go back four years to account for the fact that many of the articles highlighting the undesirable small-sample properties of CRSEs were published in the mid- to late-2000s. For each article, we recorded the subfield, whether CRSEs were used, the number of clusters, and the number of observations. In the case of multiple models per article, we recorded the article that used the fewest clusters.

Table 1 displays the summary statistics for each journal. Between 20-27% of articles in each journal used CRSEs, with AJPS showing the highest prevalence. This journal also displayed the highest rate of models with fewer than 40 clusters (7.54% of all articles and 27.94% of articles which used clustering). AJPS also had a substantial proportion of papers with fewer than 20 clusters, representing 13.24% of all articles that used CRSEs. APSR and JOP fare somewhat better, with between 3-5% of all articles using clustering in data sets with fewer than 40 clusters (between 14-19% of all articles which used CRSEs). ISQ showed the lowest rate of articles with fewer than 40 clusters at 1.57% (6.90% of CRSE articles), but this journal showed the highest rate of models with an unknown number of clusters at 10.63% (46.55% of CRSE articles).

Table 1: Use of Cluster Robust Standard Errors in General Interest Journals in Political Science (2010-2014)

	AJPS (July 2010-April 2014)			APSR (May 2010-February 2014)		
	# of articles	% of CRSEs	% of total	# of articles	% of CRSEs	% of total
total articles	252	.	.	185	.	.
CRSEs	68	.	26.98%	42	.	22.70%
$<$ 40 clusters	19	27.94%	7.54%	6	14.29%	3.24%
$<$ 20 clusters	9	13.24%	3.57%	2	4.76%	1.08%
unknown clusters	19	27.94%	7.54%	7	16.67%	3.78%

	JOP (October 2010 - July 2014)			ISQ (June 2010 - March 2014)		
	# of articles	% of CRSEs	% of total	# of articles	% of CRSEs	% of total
total articles	326	.	.	254	.	.
CRSEs	81	.	24.85%	58	.	22.83%
$<$ 40 clusters	15	18.52%	4.60%	4	6.90%	1.57%
$<$ 20 clusters	8	9.88%	2.45%	2	3.45%	0.79%
unknown clusters	17	20.99%	5.21%	27	46.55%	10.63%

What are the alternatives to CRSEs and why might they work better for a small number of clusters?

According to the prior literature on the subject, using CRSEs in a data set with a small number of clusters is ill-advised. We focus on three alternatives for cluster-adjustment that are relatively versatile and simple. These options are:

1. pairs cluster bootstrapped t-statistics (PCBSTs), variants of which are studied by Bertrand, Duflo and Mullainathan (2004), Cameron, Gelbach and Miller (2008), and Harden (2011)

2. wild cluster boostrapped t-statistics (WCBSTs), as proposed by Cameron, Gelbach and Miller (2008)

3. cluster-adjusted t-statistics (CATs), based on the work of Ibragimov and Müller (2010) and also studied by Canay, Romano and Shaikh (2014).

Our primary contribution in this paper is the creation of a streamlined statistical package for implementing these alternative uncertainty estimators.[6] The replication materials for this paper include code for all these procedures for GLM and multinomial logit models in Stata and R; the code is available at the CRAN repository for R (under library name `clusterSEs`) and the SSC repository for Stata (under the names `clustse` and `clusterbs`).[7] We describe the principles behind each procedure in the main body of the text, providing technically detailed step-by-step procedures for each one in an online appendix.

After describing these potential options and how their performance in data with a small number of clusters may differ, we conduct a Monte Carlo analysis to evaluate their performance in relation to CRSEs and the usual maximum likelihood standard errors (which we

[6] We thank an anonymous reviewer for suggesting this language.

[7] Note that our Stata software does not have an option for estimating CATs for multinomial logit models; this is, however, available in our R software.

refer to hereafter as "vanilla" standard errors). We also compare these techniques to a standard random effects model (Wooldridge, 2002, pp. 257-265); this allows us to compare the performance of cluster-adjustment techniques to that of a potentially more efficient model that is also potentially more sensitive to violations of its assumptions (Clark and Linzer, 2015).

Pairs cluster bootstrapped t-statistics (PCBSTs)

The pairs cluster bootstrapped t-statistic is a variation on the typical bootstrap procedure that accounts for the clustered structure of data. A typical bootstrapping process draws a bootstrap data set of size N with replacement from the original data set (also of size N), estimates the model of interest on the bootstrap data set, saves a quantity of interest (such as $\hat{\beta}$) from this model, repeats the process K times for a large value of K, and then examines the empirical distribution of K-many values of $\hat{\beta}_k$, $k = 1 \ldots K$. The bootstrap distribution of $\hat{\beta}_k$ will approximate the distribution of $\hat{\beta}$ (Efron, 1979; van der Vaart, 1998, Chapters 19 and 23). If the bootstrapped quantity is $\hat{\beta}$, its standard error can be estimated by the standard error of the replicates; alternatively, a 95% confidence interval can be estimated using the 2.5th and 97.5th quantiles of $\hat{\beta}_k$.

The PCBST modifies this procedure (1) to sample *clusters* with replacement, rather than individual *observations* with replacement, and (2) to sample the test statistic $t = \hat{\beta}/\hat{\sigma}$ instead of $\hat{\beta}$. The first modification to the usual bootstrap procedure is necessary to recognize that observations within a cluster are not independently distributed and thus we cannot resample at the level of the individual observation and still preserve the distribution of $\hat{\beta}$. The second modification is used because t is *pivotal* (its large-sample distribution does not depend on the unknown true values of β and $\sigma_{\hat{\beta}}$) and therefore its performance in small samples can in some cases be better than that for bootstrapping of non-pivotal statistics such as $\hat{\beta}$ (Liu and Singh 1987; Horowitz 1997). For calculating 95% confidence intervals, the 95th percentile value of t_z may be used as a part of the normal formula for confidence intervals, $\hat{\beta} \pm (t_{1-\alpha})\,\hat{\sigma}$.

Thus, PCBSTs treat clusters rather than observations as the fundamental unit of analysis and implement the bootstrap at this level. Based on a Monte Carlo analysis, Harden (2011) recommends that "state politics researchers use [the pairs cluster bootstrap] to conduct statistical inference with clustered data" (p. 224).[8] Bertrand, Duflo and Mullainathan (2004) study and apply pairs cluster bootstrapped t-statistics (which they call the "block bootstrap") and find that they are effective in controlling the size of hypothesis tests with serially correlated panel data for a moderate number of clusters in a difference-in-difference design, but produce excess false positives for a small number of clusters. However, they use vanilla standard errors rather than CRSEs for the bootstrap replicates; Cameron, Gelbach and Miller (2008) shows that CRSE replicates perform better with a small number of clusters.

Wild cluster bootstrapped t-statistics (WCBSTs)

Wild cluster bootstrapped t-statistics are similar to PCBSTs, but are based on the idea of "wild bootstrapping" the residuals of a regression rather than bootstrapping observations directly (Wu, 1986). The wild cluster bootstrap procedure relies on construction of error terms $\hat{\varepsilon}$ from the original linear model $y = X\hat{\beta} + \hat{\varepsilon}$, then creating new bootstrap data sets by sampling on clusters and assigning new values of the error term equal to the original estimated error term multiplied by weights randomly selected from a set, such as the two-point Rademacher weights $\{1, -1\}$ (Liu, 1988).[9] As with any bootstrapping technique, the procedure relies on approximating the distribution of $\hat{\beta}$ using repeated resampling of the data and re-estimation of the model on the resampled data. But for WCBSTs, the model's estimated error terms are treated as the source of variation in the observations. Cameron,

[8] It appears that Harden uses the standard error of $\hat{\beta}$ estimates for hypothesis testing, rather than of the pivotal t statistic, in his implementation of the procedure (Harden, 2011, pp. 227-229).

[9] The construction of appropriate confidence intervals using WCBSTs is somewhat more involved than with the other options discussed here due to the possibility of "imposing the null hypothesis" of $\beta = 0$ when the bootstrap replicates are drawn; see the discussion of WCBSTs in the online appendix for more information.

Gelbach and Miller (2008, p. 425) conclude that "this [wild cluster] bootstrap works well in our own simulation exercise and when applied to the data of Bertrand, Duflo and Mullainathan (2004)." Note that this procedure depends on estimates of the residuals $\hat{\varepsilon}$, and is therefore unsuited for GLM models with non-standard residuals (e.g., the probit).[10]

Cluster-adjusted t-statistics (CATs)

Cluster-adjusted t-statistics were first suggested as an approach to modeling clustered data by Ibragimov and Müller (2010). Intuitively, the contribution of Ibragimov and Müller (2010) (and the proofs in Bakirov and Szekely (2006) that underlie their work) is in determining the small-sample properties for an estimator that accounts for intra-cluster dependence among observations. This allows us to improve on the performance of CRSEs in a small number of clusters, wherein the performance of the CRSE estimator is analytically unknown.

The key theoretical insight for CATs (as described in Ibragimov and Müller (2010) on pp. 455-456, which we repeat here) comes in realizing that, when the number of observations N_g in every cluster g is large and cluster-level estimates $\hat{\beta}_g$ are (asymptotically) independent, then for many common statistical estimators each cluster estimate $\hat{\beta}_g$ should take an asymptotic distribution (as $N_g \to \infty$) of:

$$\sqrt{N_g}\left(\hat{\beta}_g - \beta\right) \overset{asym}{\sim} \Phi\left(0, \sigma_g^2\right) \tag{2}$$

This property flows from the well-known asymptotic properties of many estimators, including and especially the OLS regression estimator and maximum likelihood estimator for GLM-family models. Consequently, the vector of group-specific estimates $\hat{\beta}_{\mathbf{G}}$ takes the asymptotic distribution (as $N_g \to \infty$ for all $g \in \{1, ..., G\}$):

$$\sqrt{N}\left(\hat{\beta}_{\mathbf{G}} - \beta\right) \overset{asym}{\sim} \Phi(\mathbf{0}, \mathbf{\Sigma_G}) \tag{3}$$

[10] See Hu and Kalbfleisch (2000) for an application of the wild bootstrap to non-linear models via bootstrapping of individual observations' contribution to the score of a non-linear model.

with $N = \sum_{g=1}^{G} N_g$, $\hat{\beta}_{\mathbf{G}}^{T} = \left[\begin{array}{ccc} \hat{\beta}_1 & ... & \hat{\beta}_G \end{array} \right]$ and $\mathbf{\Sigma_G} = \text{diag}\,(\sigma_1, ..., \sigma_G)$; the block-diagonal nature of $\mathbf{\Sigma_G}$ flows from the assumption that observations in clusters i and j are independent whenever $i \neq j$.[11] Ibragimov and Müller (2010) use a formal proof from Bakirov and Szekely (2006) to show that, under these conditions, a two-tailed t-test of the grand mean of cluster estimates $\bar{\beta}_{\mathbf{G}} = (1/G) \sum_{g=1}^{G} \hat{\beta}_g$ against the null of $\beta = 0$ with $G - 1$ degrees of freedom is valid for $\alpha = 0.05$ for a two-tailed test with $G \geq 2$ (pp. 455-458). Note that this property does not necessarily hold for larger α, including and especially two-tailed $\alpha = 0.10$; for this reason, we recommend against using CATs for hypothesis tests with $\alpha > 0.05$ or confidence intervals below the 95% level. A small panel data simulation study by Ibragimov and Müller (2010, p. 460) shows that their approach rejects a true null at close to the nominal level for an $\alpha = 0.05$ test, with better power characteristics than CRSEs or fixed effects with standard errors based on Arellano (1987); however, this study examines only ten panels with fifty observations each ($G = 10$, $N_g = 50$) and does not examine the performance of pairs cluster or wild cluster bootstrapped standard errors. Another simulation study of time series and difference-in-difference data conducted by Canay, Romano and Shaikh (2014) shows that CATs can sometimes be too conservative (rejecting a true null at a rate less than the α level of the test), although CATs' power to reject false null hypotheses is still at least as good as the alternatives the authors examined (which did not include either PCBSTs or WCBSTs).

In short, CATs involve simply running the target model separately in every cluster, saving the $\hat{\beta}_g$ estimates in each cluster, then calculating confidence intervals and test statistics using the mean and variance of the collection of cluster-specific $\hat{\beta}_g$ values. A t-statistic can be calculated as $\hat{t}_{\mathbf{G}} = \bar{\beta}_{\mathbf{G}} / \hat{s}_{\mathbf{G}}$, where $\bar{\beta}_{\mathbf{G}}$ is the mean of the cluster level coefficients and $\hat{s}_{\mathbf{G}}$ is their estimated standard error. Note that the variance-covariance matrix of $\hat{\beta}$ is recovered in this procedure as the variance-covariance matrix of $\hat{\beta}_g$. This allows us to calculate 95% confidence intervals as $\bar{\beta}_{\mathbf{G}} \pm (t_{\alpha, G-1})\,(\hat{s}_{\mathbf{G}})$; it also allows us to calculate standard errors on interaction terms as prescribed by Brambor, Clark and Golder (2006). Note that $\bar{\beta}_{\mathbf{G}}$ and $\hat{\beta}$

[11] See equation (4) in Ibragimov and Müller (2010), p. 456.

will often not be equivalent;[12] therefore 95% CIs formed with this procedure will often not be centered on $\hat{\beta}$.

The method relies on $\hat{\beta}_g$ existing for every value of $g = 1 \ldots G$. If there is no variation in a limited dependent variable for one or more clusters such that a GLM model cannot be estimated in that cluster, then CATs cannot be calculated. In an online appendix, we analytically examine the conditions under which dropping these clusters preserves the distribution of equation (3) so that the results of Ibragimov and Müller (2010) still apply; as a rule of thumb, dropping unidentified clusters is acceptable when the probability of an unidentified cluster is low or there is small heterogeneity in the probability of dropping a cluster across plausible values of $\hat{\beta}_g$. We also perform Monte Carlo simulations estimating CATs after dropping unidentified cluster coefficients to confirm this result.

More challengingly, CATs *also* cannot be estimated for any independent variables that do not vary within the cluster. This limitation is not a problem for using CATs in a fixed effects model where the fixed effect is at the same level as the cluster: the fixed effect simply gets absorbed into the constant term without loss of generality. It *is* a problem when there is a substantively relevant independent variable that does not vary within clusters and for which we are interested in uncertainty in its relationship with the dependent variable. This limitation of CATs is similar to the well-known proviso that fixed effects models cannot be used with variables that do not vary within units.

Small-cluster properties of each procedure

Why would we expect these alternatives to outperform CRSEs when the data contains a small number of clusters G? CRSEs depend on results for the distribution of $\hat{\beta}$ for asymptotically large G. As an illustration, consider the simple example of estimating a mean using clustered data with an equal number of observations in every cluster; this example corresponds to a

[12] In an e-mail correspondence, Ulrich Müller indicated that $\bar{\beta}_{\mathbf{G}}$ and $\hat{\beta}$ "have a complicated dependence structure" that involves several factors, and that "nothing general can be said about their relationship."

regression using a constant only:

$$y_{gi} = \beta + \varepsilon_{gi}$$

where g indexes clusters $g = 1 \dots G$ and i indexes individual observations within a cluster $i = 1 \dots N_g$ and a total number of observations $\sum_{g=1}^{G} N_g = N$. For this simple example:

$$\hat{\beta}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} y_{gi}$$

(or the average of observations in the cluster) and the overall estimate of $\hat{\beta}$ is:

$$\hat{\beta} = \frac{1}{N} \sum_{g=1}^{G} N_g \hat{\beta}_g$$

For an example like this, the Liang and Zeger CRSE from equation (1) is:

$$\text{var}\hat{\beta} = N^{-2} \sum_{g=1}^{G} \left[\left(\sum_{i=1}^{N_g} \hat{u}_{gi} \right)^2 \right]$$

where $\hat{u}_{gi} = \hat{\beta} - y_{gi}$. If we replace the squared sum of cluster-level deviations with $s_g^2 = \left(\sum_{i=1}^{n_g} \hat{u}_{gi} \right)^2$ to represent the squared sum:

$$\text{var}\hat{\beta} = N^{-2} \sum_{g=1}^{G} s_g^2$$

it becomes apparent that any asymptotics for the summed term depend on $G \to \infty$.

Although the cluster bootstrap estimator of variance is constructed differently, it too relies on asymptotically large G. Cluster bootstrap samples are created by randomly drawing G-many clusters from the data with replacement, then recomputing $\hat{\beta}$ as above to create a bootstrap replicate estimate $\hat{\beta}_k$ for the kth bootstrap replicate. This bootstrap resampling and estimation procedure is repeated K many times; we can set K to be arbitrarily large.

The cluster bootstrap estimate of $\hat{\beta}$ is:

$$\begin{aligned}\hat{\beta}^{\star} &= \frac{1}{KGn_g}\sum_{k=1}^{K}\sum_{g_k=1}^{G}\sum_{i=1}^{n_g} y_{g_k i} \\ &= \frac{1}{KG}\sum_{k=1}^{K}\sum_{g_k=1}^{G}\bar{y}_{g_k\bullet} \\ &= \frac{1}{K}\sum_{k=1}^{K}\bar{y}_{\bullet\bullet}\end{aligned}$$

where g_k indexes the bootstrap resampled clusters in $1...G_k$ for replicate k and $\bullet$ indicates that the mean (indicated by the bar notation) is being taken over the bulleted index.[13] Under these conditions (and for clusters with equal numbers of observations), Field and Welsh (2007, pp. 383-385) demonstrate that "the cluster bootstrap mean and variance of $\hat{\beta}^{\star}$ are $\hat{\beta}$ and $n_g G^2 S_{B2}$ respectively" where:

$$S_{B2} = n_g \sum_{g=1}^{G}\left(\bar{y}_{g\bullet} - \bar{y}_{\bullet\bullet}\right)^2$$

Furthermore, "the cluster bootstrap variances of $\hat{\beta}^{\star}$... and the covariance between the sums of squares are asymptotically correct as $G \to \infty$ with N_g fixed." Similar conclusions should hold for cluster bootstrapped t statistics as well; in fact bootstrapping the t statistic can yield faster convergence to an appropriate asymptotic distribution (Liu and Singh, 1987). Similar principles should also apply for wild cluster bootstrapping, though wild cluster bootstrapping may converge faster to asymptotics with a careful choice of resampling distribution (Liu, 1988).

CATs are distinct from both CRSEs and cluster bootstrap procedures when the number of clusters is small because the small-sample characteristics of the underlying test statistic are *known*; this knowledge is thanks to analytical work from Bakirov and Szekely (2006) that is utilized by Ibragimov and Müller (2010). As shown in equations (2) and (3), asymptotic

[13] Note that we adapt the Field and Welsh (2007) notation here that they describe on p. 372, with some modifications to match this paper's notation.

arguments for CATs depend on $N_g \to \infty$, *not* on $G \to \infty$. Thus, when clusters are chosen so that observations are independent across clusters, the number of observations per cluster is large, the estimator is asymptotically normal, and the key independent variables vary within the cluster (so that models can be estimated), we might reasonably expect CATs to perform better than CRSEs and cluster bootstrapped standard errors when the number of clusters is small.

Assessing techniques for the analysis of clustered data

We assess the performance of statistical significance tests using ordinary (vanilla) standard errors, CRSEs,[14] CATs, and PCBSTs in continous and binary dependent variable scenarios; for continuous dependent variables, we also assess the performance of WCBSTs and linear random effects (RE) models.[15] For PCBSTs, we further compare using vanilla SE replicates to CRSE replicates for calculating t_k. We denote the dependent variable as y. We are interested in the proportion of the time that correctly specified models reject the null hypothesis of no relationship between y and x (an independent variable whose values are correlated with the cluster structure), or y and z (an independent variable uncorrelated with the cluster structure). For each type of dependent variable, we examine two subcases: one where x and z have no relationship with y, and one where they do have a relationship. For GLMs, the data generating process is:

$$y_{gi} = f\left(\beta_x x_{gi} + \beta_z z_{gi} + \beta_w w_{gi} + \gamma_g + \varepsilon_{gi}\right)$$

[14] CRSEs can differ in whether they use a multiplicative small sample correction, what kind of correction they use, and in the number of degrees of freedom used for the t-density (Cameron and Miller, 2015). Our programs use a multiplicative correction of $G/(G-1)$ and a t-density with $G-1$ degrees of freedom; these are the defaults for Stata's `cluster` option when using maximum likelihood models.

[15] For vanilla SEs, CRSE, and RE models, ordinary t-tests are used to test the statistical significance of the relevant coefficient. CATs, PCBSTs, and WCBSTs are used as described above.

where $f(\bullet)$ is the identity or probit link depending on the structure of the dependent variable; i indexes observations and g indexes the group identity that forms the clustering structure of the data. When x and z are unrelated to the data, $\beta_z = \beta_x = 0$ and the proportion of the time that the null hypothesis of $\beta = 0$ is rejected is our measure of the false positive rate. For dependent variables where x and z are related to the data, $\beta_x = \beta_z = 0.25$ and the proportion of the time that the null hypothesis of $\beta = 0$ is rejected is our measure of the true positive rate. $\beta_w = 1$ in all cases. The unit effects are $\gamma_g \sim \Phi(\mu = 0, \sigma = 1)$, where Φ represents the normal distribution. $\varepsilon \sim \Phi(\mu = 0, \sigma = 1)$ for continuous dependent variables and $= 0$ for binary dependent variables. Correlation with the cluster structure is created by drawing values for $x \sim \Phi(\mu = \mu_g, \sigma = 1)$ where μ_g is shared by all members of cluster g; by comparison, $z \sim \Phi(\mu = 0, \sigma = 1)$. $\mu_g \sim U[1, 5]$ in these simulations; this implies that the intra-cluster correlation coefficient for x is $\rho = \left(\frac{1}{12}(5-1)^2\right) / \left(1 + \frac{1}{12}(5-1)^2\right) \approx 0.57$. Note, however, that there is no relationship between the common group error component γ_g and the cluster-average value for x, μ_g; if this were true, then estimates of $\hat{\beta}$ would be biased and simply correcting the standard errors would be inappropriate.

This data structure is a very close match to the structure of a typical random effects model (Wooldridge, 2002, pp. 257-265). This gives us an opportunity to compare the performance of linear link GLM with cluster-corrected standard errors to random effects models for continuous dependent variables under ideal conditions for the RE model. In so doing, we are able to examine the degree to which achieving accurate model specification (as opposed to a robust approximation) improves size and power (King and Roberts, 2014). Models without random effects are estimated using the `glm` function in R, while random effects models are estimated using the `lme4` package (Bates et al., 2014).

We also wish to examine cases where the RE model is *not* an ideal match, to compare the performance of a somewhat misspecified RE model to fixed effects (FE) models with cluster-adjusted SEs (Clark and Linzer, 2015). Accordingly, for continuous dependent variables, we also generate data sets for which there is both (a) correlation between the value of the

regressor x and the group-level effect γ_g, specifically, $\mu_g = 1.5\gamma_g$, and (b) serial correlation inside of a cluster. To accomplish this, the observations inside each simulated cluster were arranged in a temporal order, $t = 1...T$; this replaces the individual indexing i. For each group g and time t, we then set $\varepsilon_{gt} = 0.9\varepsilon_{g(t-1)} + \psi_{gt}$ (with $\psi_{gt} \sim \Phi(\mu = 0, \sigma = 0.1)$) and $x_{gt} = 0.9x_{g(t-1)} + \omega_{gt}$ (with $\omega_{gt} \sim \Phi(\mu = \mu_g, \sigma = 0.1)$. Note that each group's x comes from a distribution with a different mean.[16] This data structure imitates the residual serial correlation that is commonly encountered in difference-in-difference research designs (Bertrand, Duflo and Mullainathan, 2004). Under these circumstances, we use the `plm` package in R (Croissant and Millo, 2008) to add a fixed effect for g to models before correcting the standard errors for clustering. There are adjustments to the random effects model that could enable it to adapt to these scenarios (e.g., Beck and Katz, 1995; Bafumi and Gelman, 2006); we could also directly specify a model for the serial correlation. But the point of this comparison is not to demonstrate that cluster-adjustment is always the best model—it often isn't—but to demonstrate that the cluster-adjustment procedure can be less efficient than a near-perfectly specified model but more robust to violations of accurate specification.

We also look at multinomial logit models:

$$\begin{aligned}
\Pr(y_{gi} = 1) &= \frac{1}{1 + \sum_{J\setminus 1} \exp(\beta_{xj}x_{gi} + \beta_{zj}z_{gi} + \beta_{wj}w_{gi} + \gamma_{gj})} \\
\Pr(y_{gi} = k) &= \frac{\exp(\beta_{xk}x_{gi} + \beta_{zk}z_{gi} + \beta_{wk}w_{gi} + \gamma_{gk})}{1 + \sum_{J\setminus 1} \exp(\beta_{xj}x_{gi} + \beta_{zj}z_{gi} + \beta_{wj}w_{gi} + \gamma_{gj})} \text{ for } k > 1
\end{aligned}$$

8where $J \in \{1, 2, 3\}$ and the base category is 1. For other categories $j \in \{2, 3\}$, $\gamma_{gj} \sim \Phi(\mu = 0, \sigma = 1)$. When x and z are related to the dependent variable, $\beta_{x2} = \beta_{z2} = 1$ and $\beta_{w3} = 1$, and $\beta_{x3} = \beta_{z3} = \beta_{w3} = 0$. $\mu_g \sim U[-2, 2]$ in these simulations. When x and z are unrelated to the dependent variable, we set $\beta_{x2} = \beta_{z2} = \beta_{x3} = \beta_{z3} = 0$, $\beta_{w2} = 1$, $\beta_{w3} = -1.5$, and $\mu_g \sim U[1, 5]$.[17] We estimate these models using the `mlogit` package in R (Croissant, 2015).

[16]The smaller standard deviations for ω and ψ are calculated to give the (highly autocorrelated) distributions of ε and x standard deviations of 1.

[17]In the multinomial simulation, the β values were chosen so as to provide some qualitative variation in y across the substantively relevant variables.

The data are structured so that there are varying numbers of clusters G; we examine cases with $G = \{3, 6, 15, 21, 30, 60, 75, 90, 120\}$. In each case, all clusters had an equal number of observations (40 observations per cluster). To save space, the main text discusses results for continuous dependent variables in detail and only summarizes the binary and multinomial results; these other results are detailed in an online appendix. We also conducted supplemental simulations for the linear dependent variable where we divided the clusters so that there are an equal number with 20, 40, and 60 observations; other researchers have found that CRSEs can be more sensitive to small numbers of clusters when the cluster sizes are unequal (MacKinnon and Webb, 2017). The qualitative findings of this analysis were similar to those for analysis with equally sized clusters; we relegate the discussion of these findings to an online appendix.

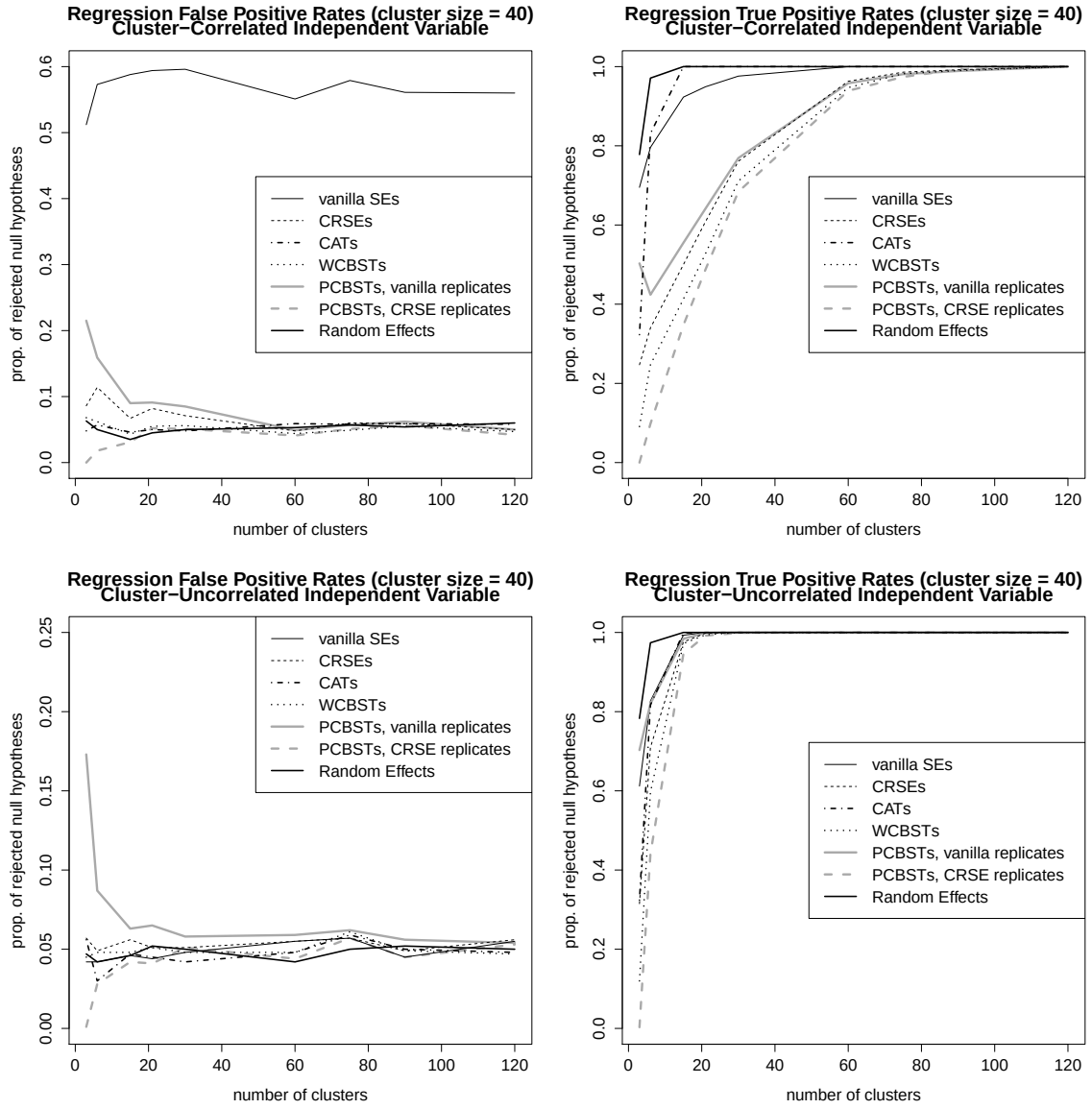
Simulation results for linear models without μ_g/γ_g correlation or serial dependence

We begin with an assessment of size and power for a continuous dependent variable with no correlation between μ_g and γ_g; these are shown in Figure 1. As the plots show, the ordinary SEs produced by `glm` (referred to as vanilla SEs in the figure) produce excess false positives for the cluster-correlated independent variable x when $\beta_x = 0$. On the other hand, the same SEs have a false positive rate for the cluster-uncorrelated variable z that is a good match for the $\alpha = 0.05$ value of the test. CRSEs produce excess false positives for $G \leq 30$ for both x and z but not for $G \geq 60$. These results are consistent with the prior findings of Angrist and Pischke (2009), Cameron, Gelbach and Miller (2008), and Harden (2011).[18]

PCBSTs offer better false positive performance than vanilla SEs or CRSEs for ≤ 21 clusters, but only when CRSE replicates are used. The proportion of rejected null hypotheses is close to the nominal 5% value of the test when using the CRSE replicates, except for very

[18]We also examine the possibility of estimating both vanilla SEs and CRSEs and using the maximum of the two for inference (Green and Vavreck, 2008); the results, reported in an online appendix, show that the procedure results in excess false positives in simulations with a small number of clusters.

Figure 1: Size and power assessment for linear dependent variables



The graphs on the left show the proportion of rejected null hypotheses ($\beta = 0$) out of 1000 simulations for parameters whose true values are $\beta_x = \beta_z = 0$ in the linear model $y_i = \beta_x x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i$ with cluster dependency; this is a measure of the false positive rate. Each model (except random effects) is a correctly specified linear link GLM (estimated using `glm`) with a different method of calculating statistical significance, as indicated in the legend; random effects models are correctly specified linear RE models estimated using `lme4`. The hypothesis tests are conducted at $\alpha = 0.05$, so the false positive rate should ideally equal 0.05. The top graph shows the false positive rate for a variable (x) that is correlated with the cluster structure, while the bottom graph shows the false positive rate for a variable (z) that is uncorrelated with the cluster structure by design. The graphs on the right show the proportion of rejected null hypotheses out of 1000 simulations for parameters whose true values are $\beta_x = \beta_z = 0.25$ in the same linear model; this is a measure of the true positive rate. One simulation is dropped for random-effects models with 60 clusters due to estimation failure and the rejection rate is calculated out of 999 simulations for that case. For a method to have adequate power to conduct significance tests, the true positive rate should ideally equal 1. The top graph shows the true positive rate for a variable (x) that is correlated with the cluster structure, while the bottom graph shows the true positive rate for a variable (z) that is uncorrelated with the cluster structure by design.

small numbers of clusters (where the rejection rate is lower than 5%). But this improvement in false positive performance is offset by worse performance in the detection of true positives, which is the worst for PCBSTs with CRSE replicates compared to any of the other options that we considered.

WCBSTs have false positive rates close to the target $\alpha = 0.05$ for even the smallest number of clusters. On the other hand, the true positive detection performance of WCBSTs is worse than all other techniques except PCBSTs with CRSE replicates. For the cluster-correlated independent variable, just over 70% of estimated $\hat{\beta}_x$ values are statistically significant when the true $\beta_x = 0.25$ when there are 30 clusters (for a total sample size of 1200 observations).

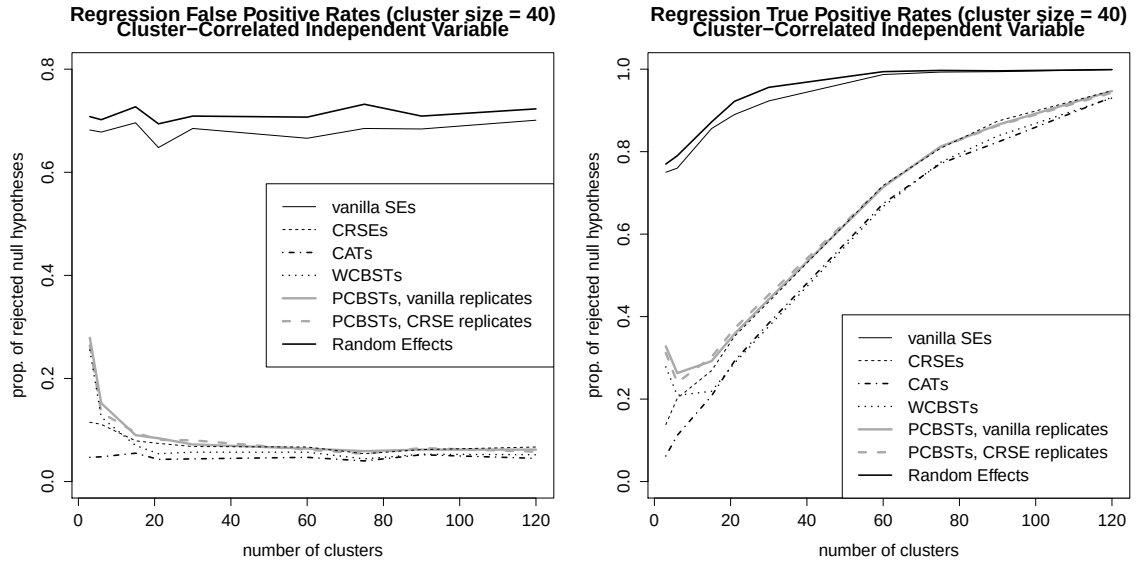
The CATs based on Ibragimov and Müller (2010) have false positive rates that are close to the 5% α value of the test with appropriate size over the entire range of G. The CATs are also substantially better than all the other cluster-correction options we examine in terms of the power of hypothesis tests to detect true positives. While power is diminished for all the techniques for small numbers of clusters, the CATs achieve near-100% power rates more quickly than any of the clustering alternatives.

The best performance is achieved by RE models, which maintain a 5% false positive rate across the entire range of G and also achieve extremely high true positive detection rates at even the very smallest number of clusters. While there is little difference between CAT and RE true positive or false positive detection rates when the number of clusters is ≥ 15, for 3 or 6 clusters the RE model's true positive performance is substantially better, $\approx 78\%$ even with only three clusters.

Simulation results for linear models with μ_g/γ_g correlation and serial dependence

The performance of the random effects estimator (shown in Figure 2) is substantially worsened when the data generating process includes (a) correlation between the average value of

Figure 2: Size and power assessment for linear dependent variables with fixed effects and serial dependence



The graph on the left shows the proportion of rejected null hypotheses ($\beta = 0$) out of 1000 simulations for parameters whose true values is $\beta_x = 0$ in the linear model $y_i = \beta_x x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i$ with (a) correlation between the group-specific mean of x ($= \mu_g$) and the group-level effect γ_g and (b) within-group serial dependence in ε and x; this is a measure of the false positive rate. Each model (except random effects) is a correctly specified linear fixed effects model estimated using `plm` with a different method of calculating statistical significance, as indicated in the legend; random effects models are linear RE models with correct variable specification (but no fixed effects) estimated using `lme4`. The hypothesis tests are conducted at $\alpha = 0.05$, so the false positive rate should ideally equal 0.05. The graph on the right shows the proportion of rejected null hypotheses out of 1000 simulations for parameters whose true values are $\beta_x = 0.25$ in the same linear model; this is a measure of the true positive rate. For a method to have adequate power to conduct significance tests, the true positive rate should ideally equal 1. See the online appendix for results for the cluster-uncorrelated independent variable (z).

x and the group effect γ_g and (b) serial dependence in x and y. In this environment, the random effects model produces a false positive for the cluster-correlated independent variable $\approx 70\%$ of the time for all values of G, even worse than a fixed effects model with vanilla SEs in terms of excess false positive results. CRSEs, WCBSTs, and both forms of PCBSTs all perform better than these two options, but tend to over-reject the null hypothesis for $G \leq 15$. CATs have a false positive rate near 5% across all values of G.

The power characteristics of all models with appropriate size characteristics (particularly fixed effects models with CATs) are substantially reduced in this simulation for the serially dependent and cluster-correlated variable x. However, power characteristics improve sub-

stantially when the signal is made stronger (relative to noise[19]) so that $\beta_x = \beta_z = 0.5$. In this alternative environment, CATs achieve 78% detection of true positives with 21 clusters.

Summary of binary and multinomial simulation results

The results of our simulations for binary and multinomial dependent variables are broadly comparable to those for linear models: CATs and PCBSTs with CRSE replicates have appropriate false positive rates, even for small numbers of clusters, while the other options have elevated false positive rates in datasets with less than or equal to 30 clusters. CATs have better power performance than PCBSTs with CRSE replicates. Details on these simulation results are provided in an online appendix.

There is one important caveat: CATs cannot be calculated whenever any cluster has no variation on the dependent variable, which can often occur with a categorical dependent variable. In addition, these models often produce a small number of outlying cluster coefficient estimates; this can happen, e.g., when an independent variable perfectly predicts the dependent variable. We handle these problems by dropping clusters with missing and outlying estimates and calculating the CATs using the number of clusters that are not dropped. Theoretically, this corresponds to the idea that the dropped clusters contain no information and the remaining clusters still have the same distribution.[20] We provide an analytic argument in the online appendix that dropping these clusters will not distort inferences in many

[19]Simply rescaling a variable X will *not* improve the power characteristics of an estimator; we change the *true* value of β **relative to the scale of** X in our simulations.)

[20]If any cluster fails to estimate for every coefficient in a probit model (due, e.g., to non-variation in the DV), both the R and Stata software packages return an error by default. The alternative behavior for both software packages drops these clusters from consideration, but uses the results from models that were successfully estimated. In some cases, individual variables' coefficients cannot be estimated in a particular cluster due to non-variation of the variable in that cluster, but the model can still be estimated if that variable is dropped. By default, both packages report the final result as missing for any variable that is dropped from any cluster-specific model (but still reports the results for variables whose coefficients could be estimated in every cluster). The packages' alternative behavior will exclude all results from any cluster for which *any* variable's results cannot be estimated. For the Monte Carlo analysis shown in Figure 9, all clusters where any variable's coefficient could not be estimated were dropped (the alternative behavior). The software for both R and Stata also has an option to drop any cluster with any beta estimate whose distance to the mean is more than 6 times the inter-quartile range; the results in Figure 9 enable this option. The R software for multinomial logit models always excludes clusters for which any variable's results cannot be estimated, but allows dropping of clusters with outlying estimates as an option.

cases, particularly if the number of dropped clusters is relatively small.[21] By dropping the clusters with missing and/or outlying estimates, we are able to estimate CATs on every data set with more than 6 clusters in our simulation[22] and our false positive rate remains very close to the nominal value of 5% with good power characteristics. We describe the practical consequences of dropping (and not dropping) clusters from the CAT procedure in greater detail in an online appendix.

Applied Examples

Our simulation results indicate that how we handle clustered data with a small number of clusters can greatly affect our inferences, and close examination of recently published examples of the analysis of clustered data underscores this finding. To this end, we re-analyzed the results of three recent publications that used CRSEs in data with few clusters: Grosser, Reuben and Tymula (2013), Lacina (2014), and Hainmueller, Hiscox and Sequeira (2015). In each case, our reanalysis using alternative cluster corrections yields meaningfully different results. To save space, we describe the Grosser, Reuben and Tymula (2013) replication in detail and only summarize the findings from Lacina (2014) and Hainmueller, Hiscox and Sequeira (2015); the details of these replications are present in an online appendix.

[21] Specifically, the difference between the true distribution of parameters across clusters and the distribution of parameters conditional on the clusters not being dropped will be small if:

1. the overall probability of dropping a cluster is small, or
2. the degree of heterogeneity in the probability of a dropped cluster for different possible values of the parameters is small (i.e., all plausible values of the parameters have a roughly equal probability of producing a data set that does not identify the parameters).

See the online appendix for more details.

[22] In the probit simulations, one model could not be estimated for the false positive simulations with three clusters, and nine models could not be estimated for the true positive simulations with three clusters. One model could not be estimated for the probit true positive simulations with six clusters.

Political quid pro quo agreements (Grosser, Reuben and Tymula, 2013)

In a recent *American Journal of Political Science* article, Grosser, Reuben and Tymula (2013) conduct a laboratory experiment designed to study the relationship between contributions to political candidates and the policies that are enacted by those candidates. In the experiment, a subject's wealth endowment is randomly assigned by the experimenters; one "rich" voter is grouped with three "poor" voters, with the rich voter having 13 times more money than each of the poor voters; two non-voting "candidate" subjects are also assigned to the group. In the experimental condition that we study in this replication, groups are fixed[23] for the duration of the experiment; 17 such groups are present in the data set. The rich voter is allowed to give some of his or her own endowment to each of the two candidate subjects; we denote the rich voter's monetary transfer to candidate i as m_i. After receiving these transfers, each candidate i publicly proposes a redistributive tax rate τ_i between 0 and 100 percent. The rich voter and the three poor voters then choose a candidate via majority vote; the selected candidate receives a fixed bonus payoff for winning. The winning candidate's proposal ($\tau^\star$) is adopted: each voter pays $\tau^\star$ percent of their income into a common fund, which is divided evenly among the voters. Thus, a voter with initial endowment e stands to lose $\tau^\star(e - \bar{e})$ if rich, and gain $\tau^\star(\bar{e} - e)$ if poor; $\bar{e}$ is the average endowment among all four voters in the group. This process is repeated in each of 15 periods. Explicit agreements can neither be made nor enforced in the experiment; for example, the rich voter cannot condition his or her donation on candidate behavior. Thus, any relationship between the transfers of the rich voter and the candidates's tax proposals must be the result of a *tacit* agreement.

Grosser, Reuben and Tymula (2013, Figure 2) show a strong relationship between lower proposed tax rates and higher transfers to candidates from the rich voter despite electoral competition and the majority's strong interest in complete redistribution. They are also

[23] Analysis by Grosser, Reuben and Tymula (2013) convincingly demonstrates that the tacit agreements between rich voters and candidates that are of interest to the authors only materialize in treatments with repeated interaction between fixed groups; see, e.g., Figure 2 on p. 590 of their paper for details.

interested in how these tacit agreements are formed, speculating that rich voters might first offer high transfers that are then reciprocated by lowered tax proposals. As they explain on p. 591:

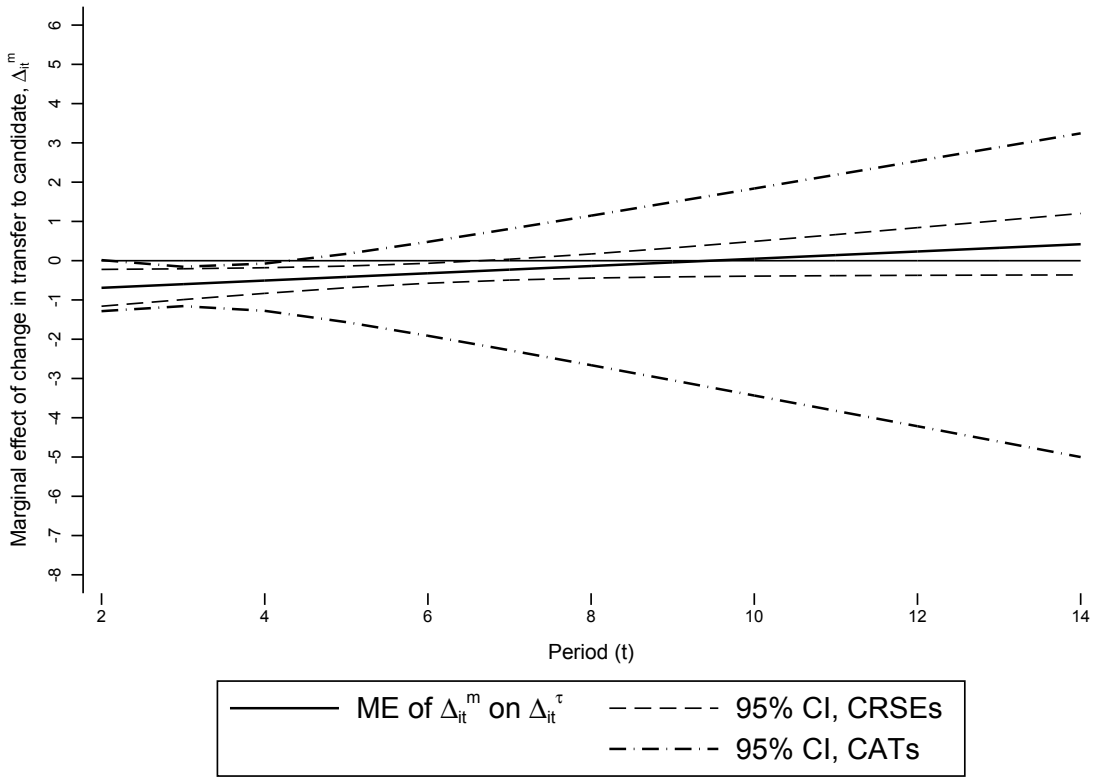
> Are tacit agreements the result of mutual reciprocation between the rich voter and the two candidates? To answer this question, we use regression analysis to test whether *changes* in transfers can predict subsequent changes in tax policies and vice versa.

That is, they run a regression modeling the change in a candidate i's proposed tax rate between period t and the previous period $t-1$, $\Delta^{\tau}_{it} = \tau_{it} - \tau_{i(t-1)}$, that is associated with a change in the transfer received by that candidate from the rich voter between t and $t-1$, $\Delta^{m}_{it} = m_{it} - m_{i(t-1)}$. The authors also include a period counter t and interact Δ^{m}_{it} with t in order to determine whether tacit agreements unravel over time. The difference between each candidate's previous tax choice and the opponent's choice, $\tau_{i(t-1)} - \tau_{j(t-1)}$, is included as two variables, one for a positive difference ($D^{+}_{ij} = \max\left(\tau_{i(t-1)} - \tau_{j(t-1)}, 0\right)$) and one for a negative difference ($D^{-}_{ij} = \max\left(\tau_{j(t-1)} - \tau_{i(t-1)}, 0\right)$); these variables are designed to study how candidates react to one anothers' proposals.

The regression analysis of Grosser, Reuben and Tymula (2013) uses cluster-robust standard errors calculated on the 17 groups of subjects; there are two candidate observations per group in each of 14 periods (excluding the first period, as changes are undefined for that period) for a total of 476 observations in the data set (28 total observations per group). The authors report strong evidence that "candidates do reciprocate the actions of the rich by decreasing (increasing) their tax policies in proportion to a previous increase (decrease) in received transfers (the coefficient of $[\Delta^{m}_{it}]$ is always statistically significant)" (p. 592). However, the authors also report considerable group-to-group heterogeneity; tacit agreements between rich voters and candidates to exchange transfers for low tax rates evidently did not emerge in every experimental group (pp. 590-591).[24]

[24]See the online appendix for more details.

Figure 3: Marginal Effects Plot, Effect of Changed in Recieved Transfer (Δ_{it}^{m}) on Changes in Proposed Tax Rate (Δ_{it}^{τ}) by Period



Dependent variable: change in candidate's proposed tax rate (Δ_{it}^{τ}). Marginal effects are calculated using the results in online appendix Table 2.

It is possible that uncertainty in the relationship between changes in transfers and changes in proposals might be understated by cluster-robust standard errors on data with only 17 groups. Therefore, we reproduce the regression analysis of Grosser, Reuben and Tymula (2013) using their original CRSEs as well as pairs cluster bootstrapped t-statistics and cluster-adjusted t statistics. A marginal effect plot[25] (Figure 3) for the marginal effect of changes in transfers on changes in tax policy in each period using these results[26] shows that the relationship is statistically significant in periods 2-6 using CRSEs, but only in periods 3 and 4 using CATs.

Grosser, Reuben and Tymula (2013, p. 290) also separately examine the behavior of

[25]See Brambor, Clark and Golder (2006).

[26]A coefficient table is shown in online appendix Table 2.

"high tax" groups, which have high winning tax policies (between 90% and 100%), and "low tax" groups, which have much lower winning tax policies (between 33.3% and 83.7%). We reproduce their analysis for high and low tax groups using CRSEs as well as PCBSTs and CATs; marginal effects plots for this analysis are shown in Figure 4.[27] For high tax groups, CRSEs indicate a statistically significant relationship between changes in transfers and changes in tax policy in periods 2-8. By comparison, CATs find no statistically significant relationship in any period.[28] For low tax groups, transfers are associated with lower tax proposals between periods 2 and 5 with CRSEs but are statistically indistinguishable from zero in every period except 2 with CATs.

In summary, although the evidence presented by Grosser, Reuben and Tymula (2013) shows a possible link between Δ_{it}^{m} and Δ_{it}^{τ}, the statistical significance of this link is affected when we use CATs instead of CRSEs. The strong relationship the authors find between lower proposed tax rates and larger transfers from the rich voter might be explained by candidates reciprocating increased donations with decreased tax rates in low tax societies, but the experimental evidence for this explanation is more uncertain than CRSEs indicate.

Other examples

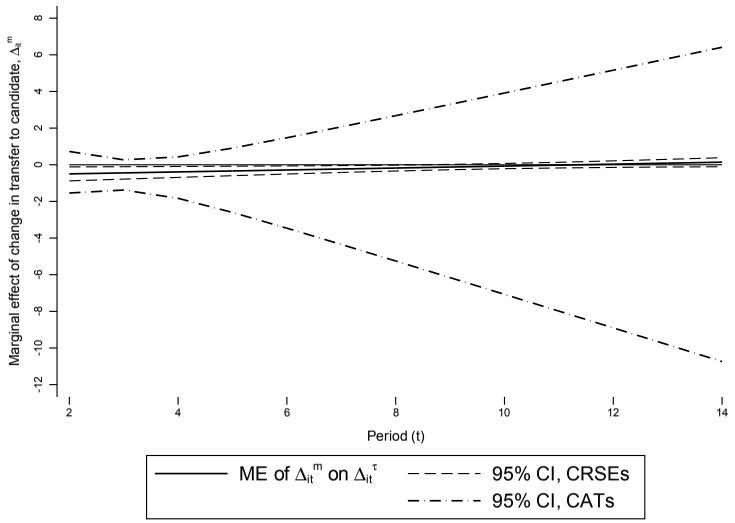
Our replications of Lacina (2014) and Hainmueller, Hiscox and Sequeira (2015), described more fully in an online appendix, generally serve to reiterate the issues with using CRSEs with few clusters that are illustrated in the Grosser, Reuben and Tymula (2013) replication. Lacina (2014) uses a multinomial logit model and CRSEs to provide evidence for a link between political representation and civil unrest in India; specifically, she argues that linguistic groups that could have become Indian states were peacefully accommodated, ignored, or resorted to civil violence in relation to their representation in the Indian National Congress party. But the relationship between outcomes and representation that she describes

[27] Coefficient tables are shown in online appendix Tables 3 and 4.

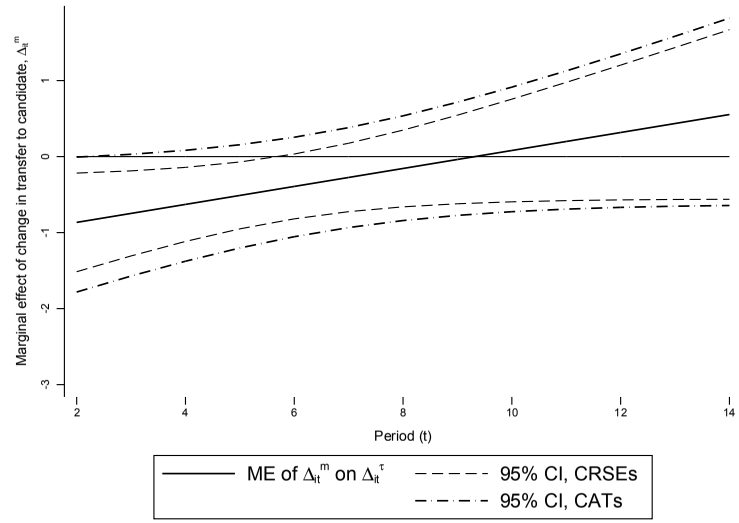
[28] The high tax group results for CATs are arguably more supportive of the authors' theory than their original CRSE findings because candidate-voter reciprocity is the mechanism for sustaining *low* tax rates.

Figure 4: Marginal Effects Plot, Effect of Changed in Recieved Transfer (Δ_{it}^{m}) on Changes in Proposed Tax Rate (Δ_{it}^{τ}) by Period in High and Low Tax Groups

(a) "High Tax" groups



(b) "Low Tax" groups



Dependent variable: change in candidate's proposed tax rate (Δ_{it}^{τ}). Marginal effects are calculated using the results in online appendix Table 3 for panel 4a and online appendix Table 4 for panel 4b.

is not robust to alternative methods of cluster adjustment for the standard errors. More optimistically, we find that the substantive conclusion of Hainmueller, Hiscox and Sequeira (2015) is robust to alternative clustering adjustments: based on their experiment, consumers are willing to pay more for products with a "fair trade" label. However, we also find that CRSEs understate the uncertainty of the substantive magnitude of this effect: estimated 95% confidence intervals are 20% wider when using PCBSTs and 46% wider when using CATs compared to the original results using CRSEs.

Conclusion

Political scientists often use cluster-robust standard errors to analyze clustered data where the structure of relationships inside of the cluster is uncertain or unknown. Unfortunately, past research indicates that CRSEs produce downward-biased standard errors when the number of clusters is small; this can creates a hazard of excessive false positive results. Our research indicates that political scientists are still in the habit of using CRSEs in this scenario, possibly because alternatives are difficult to implement in common statistical packages.

Our findings and the prior literature suggest that substantive researchers should consider alternatives to the CRSE; we make it easy for them to do so with our `clusterSEs` package for R and the `clustse` and `clusterbs` ado files for Stata. Our simulation analysis finds that cluster-adjusted t-statistics (CATs) (based on the work of Ibragimov and Müller, 2010) are the best choice among the options we examine for correcting standard errors for clustering in data sets with a small number of clusters. CATs require that a model must be separately estimable in every cluster, which is not always possible; however, dropping a small number of clusters where the model cannot be estimated and using the rest to estimate CATs can still produce valid inferences under many circumstances. If CATs cannot be estimated (e.g., because a key independent variable does not vary within clusters), pairs cluster bootstrapped t-statistics (PCBSTs) with CRSE replicates and wild cluster bootstrapped

t-statistics (WCBST) generally provide better performance than vanilla standard errors or CRSEs. Finally, in our simulations an accurate random effects model of intra-cluster heterogeneity provides better performance than any cluster adjustment technique, but the cluster adjustment techniques perform better in the event of misspecification.

Our final recommendation is that researchers think carefully about the amount of information contained in a data set with a small number of clusters, even if the number of observations in each cluster is large. If a researcher does not want to rely on assumptions about the structure of relationships inside of a cluster, then in some sense s/he is deciding to ignore the intra-cluster variation in the data set when estimating standard errors. Political scientists would be wary of deriving strong conclusions from a data set with $N = 15$ or 20 observations and 5 independent variables, even on the basis of a strongly statistically significant result. Based on our research and the findings of prior simulation studies, we believe that the same caution is warranted in the scenario with large N but only 15 or 20 clusters.

References

Anderson, T. W. 2003. *An Introduction to Multivariate Statistical Analysis.* 3rd ed. Wiley.

Angrist, Joshua D. and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press.

Arellano, Manuel. 1987. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics* 49(4):431–434.

Bafumi, Joseph and Andrew Gelman. 2006. "Fitting Multilevel Models When Predictors and Group Effects Correlate." Online. URL: http://goo.gl/usvQsn.

Bakirov, N. K. and G. J. Szekely. 2006. "Student's t-test for Gaussian scale mixtures." *Journal of Mathematical Sciences* 139(3):6497–6505.

Bates, Douglas, Martin Maechler, Ben Bolker and Steven Walker. 2014. *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.1-7.
URL: *http://CRAN.R-project.org/package=lme4*

Beck, Nathaniel and Jonathan N. Katz. 1995. "What To Do (And Not To Do) With Time-Series Cross-Section Data." *American Political Science Review* 89(3):634–647.

Beck, Nathaniel L., Jonathan N. Katz and Umberto G. Mignozzetti. 2014. "Of Nickell Bias and its Cures: Comment on Gaibulloev, Sandler, and Sul." *Political Analysis* 22(2):274–278.
URL: *http://pan.oxfordjournals.org/content/22/2/274*

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics* 119(1):249–275.

Brambor, Thomas, William Roberts Clark and Matthew Golder. 2006. "Understanding interaction models: Improving empirical analyses." *Political Analysis* pp. 1–20.

Cameron, A. Colin and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* Forthcoming.

Cameron, A. Colin, Jonah B. Gelbach and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90(3):414–427.

Cameron, A.C. and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press.

Canay, Ivan A., Joseph P. Romano and Azeem M. Shaikh. 2014. "Randomization Tests under an Approximate Symmetry Assumption." Working Paper (version: December 19, 2014). URL: `https://goo.gl/TUEQee` accessed 1/29/2017.

Clark, Tom S. and Drew A. Linzer. 2015. "Should I Use Fixed or Random Effects?" *Political Science Research and Methods* 3(2):399–408.

Croissant, Yves. 2015. "Package 'mlogit'." CRAN.
URL: *http://cran.r-project.org/web/packages/mlogit/mlogit.pdf*

Croissant, Yves and Giovanni Millo. 2008. "Panel Data Econometrics in R: The plm Package." *Journal of Statistical Software* 27(2).

Donald, Stephen G. and Kevin Lang. 2007. "Inference with Difference-in-Differences and Other Panel Data." *The Review of Economics and Statistics* 89(2):221–233.

Donner, Allan. 1998. "Some aspects of the design and analysis of cluster randomization trials." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47(1):95–113.

Efron, Bradley. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7(1):1–26.

Field, C. A. and A. H. Welsh. 2007. "Bootstrapping Clustered Data." *Journal of the Royal Statistical Society: Series B* 69(3):369–390.

Gaibulloev, Khusrav, Todd Sandler and Donggyu Sul. 2014. "Dynamic Panel Analysis under Cross-Sectional Dependence." *Political Analysis* 22:258–273.

Green, Donald P. and Lynn Vavreck. 2008. "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches." *Political Analysis* 16(2):138–152.

Grosser, Jens, Ernesto Reuben and Agnieszka Tymula. 2013. "Political Quid Pro Quo Agreements: An Experimental Study." *American Journal of Political Science* 57:582–597.

Hainmueller, Jens, Michael Hiscox and Sandra Sequeira. 2015. "Consumer Demand for the Fair Trade Label: Evidence from a Field Experiment." *Review of Economics and Statistics* forthcoming.

Hansen, Christian B. 2007. "Asymptotic properties of a robust variance matrix estimator for panel data when T is large." *Journal of Econometrics* 141(2):597–620.

Harden, Jeffrey J. 2011. "A Bootstrap Method for Conducting Statistical Inference with Clustered Data." *State Politics & Policy Quarterly* 11(2):223–246.

Hardin, James W. and Joseph M. Hilbe. 2003. *Generalized Estimating Equations*. Boca Raton: Chapman & Hall/CRC.

Horowitz, Joel L. 1997. Bootstrap methods in econometrics: theory and numerical performance. In *Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress*, ed. David M. Kreps and Kenneth F. Wallis. Cambridge University Press pp. 189–222.

Hu, Feifang and John D. Kalbfleisch. 2000. "The estimating function bootstrap." *Canadian Journal of Statistics* 28(3):449–481.

Ibragimov, Rustam and Ulrich K. Müller. 2010. "t-Statistic Based Correlation and Heterogeneity Robust Inference." *Journal of Business & Economic Statistics* 28(4):453–468.

Imbens, Guido W. and Michal Kolesar. 2012. "Robust Standard Errors in Small Samples: Some Practical Advice." Working Paper.

Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lutkepohl and Tsoung-Chao Lee. 1988. *Introduction to the Theory and Practice of Econometrics*. Wiley.

Kezdi, Gabor. 2004. "Robust standard error estimation in fixed-effects panel models." *Hungarian Statistical Review* 9:95–116.

King, Gary and Margaret E. Roberts. 2014. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It." *Political Analysis* 23:159–179.

Klar, Neil and Allan Donner. 2001. "Current and future challenges in the design and analysis of cluster randomization trials." *Statistics in Medicine* 20(24):3729–3740.

Lacina, Bethany. 2014. "How Governments Shape the Risk of Civil Violence: India's Federal Reorganization, 1950-56." *American Journal of Political Science* 58(3):720–738.

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

Liang, Kung-Yee and Scott L. Zeger. 1993. "Regression Analysis for Correlated Data." *Annual Review of Public Health* 14(1):43–68.

Liu, Regina Y. 1988. "Bootstrap Procedures under some Non-I.I.D. Models." *The Annals of Statistics* 16(4):1696–1708.

Liu, Regina Y. and Kesar Singh. 1987. "On a Partial Correction by the Bootstrap." *The Annals of Statistics* 15(4):1713–1718.

MacKinnon, James G. 2015. "Wild cluster bootstrap confidence intervals." *L'Actualité économique* 91(1-2):11–33.

MacKinnon, James G. and Matthew D. Webb. 2017. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Journal of Applied Econometrics* 32(2):233–254.

Mancl, L. A. and T. A. DeRouen. 2001. "A covariance estimator for GEE with improved small-sample properties." *Biometrics* 57(1):126–134.

Moulton, Brent R. 1986. "Random group effects and the precision of regression estimates." *Journal of Econometrics* 32(3):385–397.

Moulton, Brent R. 1990. "An illustration of a pitfall in estimating the effects of aggregate variables on micro units." *The Review of Economics and Statistics* pp. 334–338.

Nickell, Stephen. 1981. "Biases in dynamic models with fixed effects." *Econometrica* 49:1417–1426.

Rogers, William. 1993. "Regression standard errors in clustered samples." *Stata Technical Bulletin* 13:19–23.

van der Vaart, A.W. 1998. *Asymptotic Statistics*. Cambridge University Press.

White, Halbert. 1980. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica* 48(4):817–838.

Williams, Rick L. 2000. "A note on robust variance estimation for cluster-correlated data." *Biometrics* 56(2):645–646.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Wu, C. F. J. 1986. "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis." *The Annals of Statistics* 14(4):1261–1295.

Appendix A: Detailed descriptions of alternative cluster-robust uncertainty calculation procedures

In this appendix, we provide detailed step-by-step procedures for calculating pairs cluster bootstrapped t-statistics (PCBSTs), wild cluster bootstrapped t-statistics (WCBSTs), and cluster-adjusted t-statistics (CATs).

Pairs cluster bootstrapped t-statistics (PCBSTs)

We present this procedure for a data set of size N with G clusters as it is described in Cameron, Gelbach and Miller (2008, p. 427), with some adjustment of presentation and notation where necessary.

1. From the original sample, calculate $t = \hat{\beta}/\hat{\sigma}_{\hat{\beta}}$ using a statistical model, where $\hat{\beta}$ is an estimated model parameter of interest and $\hat{\sigma}_{\hat{\beta}}$ (the standard error of the estimated $\hat{\beta}$) is computed using either the usual non-clustered formula or CRSEs.

2. For $k = 1 \ldots K$:

   (a) draw a bootstrap data set of G clusters by resampling with replacement G times from the original sample.

   (b) estimate $\hat{\beta}_k$ using the cluster bootstrapped data set and the model from step 1.

   (c) calculate $t_k = \left[\left(\hat{\beta}_k - \hat{\beta}\right)/\hat{\sigma}_{\hat{\beta}_k}\right]$ where $\hat{\sigma}_{\hat{\beta}_k}$ (the standard error of the estimate of $\hat{\beta}_k$) is computed using the same formula as in step 1. $\hat{\beta}$ is subtracted from $\hat{\beta}_k$ in order to determine the distribution of t in repeated sampling under the null.

3. Reject the null hypothesis $\beta = 0$ at level α if and only if $|t| > t_{1-\alpha}$ where t_z is the z^{th} quantile of the absolute value of the K-many bootstrap draws, $|t_k|$.[29]

[29] Cameron, Gelbach and Miller (2008) describes using the $\frac{\alpha}{2}$ and $(1 - \frac{\alpha}{2})$ quantiles of the signed values of t_k; we opt to "fold" the draws of t_k over $t = 0$ as described in Cameron and Miller (2015, p. 26) to make more efficient use of a smaller number of draws.

Wild cluster bootstrapped t-statistics (WCBSTs)

Wild cluster bootstrapping (which we present here, with some adaptation and adjustment of notation) is described in Cameron, Gelbach and Miller (2008, p. 427) as follows:

1. From the original sample, calculate $t = \hat{\beta}/\hat{\sigma}_{\hat{\beta}}$ from the linear model $y = X\hat{\beta} + Z\hat{\alpha} + \hat{\varepsilon}$ where $\hat{\sigma}_{\hat{\beta}}$ (the standard error of the estimated $\hat{\beta}$) is computed using the usual non-clustered formula (or CRSEs). X is a $1 \times N$ vector; if there is more than one variable of interest, the procedure is repeated for each variable separately (putting all other variables into the Z term).

2. Estimate the model from step 1 including all necessary variables *except* the variable of interest, $y = X_k 0 + Z\hat{\alpha} + \hat{\varepsilon}$; this imposes the null hypothesis that $\beta = 0$ so that the bootstrap simulates repeated sampling under the null. Save the residuals $\hat{\varepsilon}$ from the model as a part of the data set.

3. For $k = 1 \ldots K$:

   (a) draw G many cluster-level weights w_{gk} from the set $\{-1, 1\}$, with probability $1/2$ for each possible weight.[30]

   (b) for each observation $i = 1 \ldots N$, set $\hat{\varepsilon}^{\star}_{ik} = \hat{\varepsilon}_{ik} w_{g(i)k}$ using the weight for the cluster to which observation i corresponds $g(i)$. Then calculate $\hat{y}^{\star}_{k} = Z\hat{\alpha} + \hat{\varepsilon}^{\star}_{k}$. This creates a wild cluster bootstrapped data set of N dependent variable observations $\hat{y}^{\star}_{k}$ and independent variable observations X_k and Z_k.

   (c) estimate $\hat{\beta}_k$ using the wild bootstrap data set and the model $\hat{y}^{\star}_{k} = X_k\hat{\beta}_k + Z_k\hat{\alpha}_k + \hat{\gamma}_k$, where $\hat{\gamma}_k$ is an error term.

   (d) calculate $t_k = \hat{\beta}_k/\hat{\sigma}_{\hat{\beta}_k}$ where $\hat{\sigma}_{\hat{\beta}}$ (the standard error of the estimate of $\hat{\beta}_k$) is computed using the same formula as in step 1.

[30] These are Rademacher weights; other weights are possible, as described in Cameron, Gelbach and Miller (2008, p. 427).

4. Reject the null hypothesis $\beta = 0$ at level α if and only if $t > |t_{1-\alpha}|$ where t is the z^{th} quantile of the K-many bootstrap draws of t_k.[31]

Note that the procedure described above imposes the null hypothesis that $\beta = 0$ for the cofficient of interest. Bootstrapping in this way produces accurate p-values for statistical hypothesis testing, but using the bootstrapped critical t-statistic from this procedure will produce confidence intervals with inaccurate coverage; consequently, our R software package does not report confidence intervals for WCBSTs when the null is imposed. To create accurate confidence intervals, one must either bootstrap without imposing the null hypothesis (an option available with our software) or follow the procedure described in MacKinnon (2015, pp. 15-18) to impose appropriate null hypotheses for the boundaries of the confidence interval.

Cluster-adjusted t-statistics (CATs)

The results of Ibragimov and Müller (2010) suggest the following procedure for hypothesis testing in the presence of clustered data:

1. Estimate a model, saving an estimated parameter of interest $\hat{\beta}$.

2. For each cluster $g = 1, ..., G$, estimate the model from step 1 on the observations in the cluster only, saving a model parameter of interest $\hat{\beta}_g$.

3. Calculate the average $\hat{\beta}_g$ over the G-many cluster estimates, $\bar{\beta}_{\mathbf{G}}$. Calculate $\tilde{\beta}_g = \hat{\beta}_g - \bar{\beta}_{\mathbf{G}}$ for all g. Subtracting the grand mean $\bar{\beta}_{\mathbf{G}}$ enables us to consider each cluster as a sample from the distribution of possible clusters centered on the null hypothesis $\beta = 0$.

4. Calculate the standard error of $\bar{\beta}_{\mathbf{G}}$, $\hat{s}_{\mathbf{G}} = \left[\left(\frac{1}{G}\right)\left(\frac{1}{G-1}\right)\sum_{g=1}^{G}\left(\tilde{\beta}_g^2\right)\right]^{1/2}$.

5. Calculate $\hat{t}_{\mathbf{G}} = \bar{\beta}_{\mathbf{G}}/\hat{s}_{\mathbf{G}}$.

[31] Cameron, Gelbach and Miller (2008) describes using the $\frac{\alpha}{2}$ and $(1-\frac{\alpha}{2})$ quantiles of the signed values of t_k; we opt to "fold" the draws of t_k over $t = 0$ as described in Cameron and Miller (2015, p. 27) to make more efficient use of a smaller number of draws.

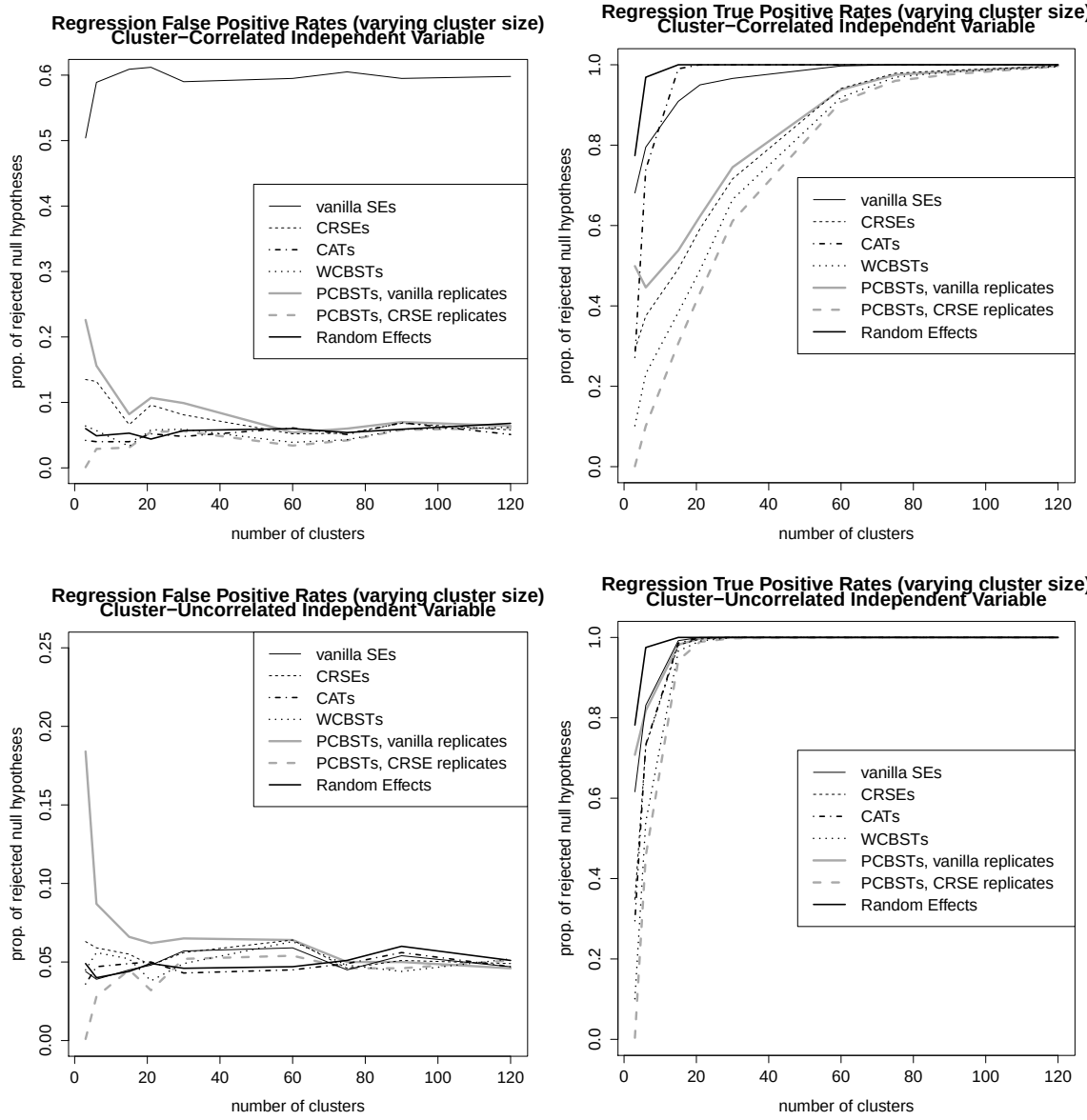
6. Reject the null hypothesis $\beta = 0$ at level α if and only if $\left|\hat{t}_{\mathbf{G}}\right| > t_{\alpha, G-1}$ where $t_{\alpha, G-1}$ is the critical-t statistic for a two-tailed hypothesis test at level α with $G - 1$ degrees of freedom.

Note that the variance-covariance matrix of $\hat{\beta}$ is recovered in this procedure as $\hat{s}_{\mathbf{G}}$. This allows us to calculate 95% confidence intervals as $\bar{\beta}_{\mathbf{G}} \pm (t_{\alpha, G-1})(\hat{s}_{\mathbf{G}})$; it also allows us to calculate standard errors on interaction terms as prescribed by Brambor, Clark and Golder (2006). Note that $\bar{\beta}_{\mathbf{G}}$ and $\hat{\beta}$ will often not be equivalent; therefore 95% CIs formed with this procedure will often not be centered on $\hat{\beta}$.

Appendix B: Replication of simulations in Figure 1 with varying cluster size

Figure 5 presents the results of simulations identical to those from Figure 1 with one key difference: instead of setting all clusters to have 40 observations, we divided the clusters so that there are an equal number with 20, 40, and 60 observations. The qualitative findings of this simulation are identical to those of Figure 1.

Figure 5: Size and power assessment for linear dependent variables, varying cluster size

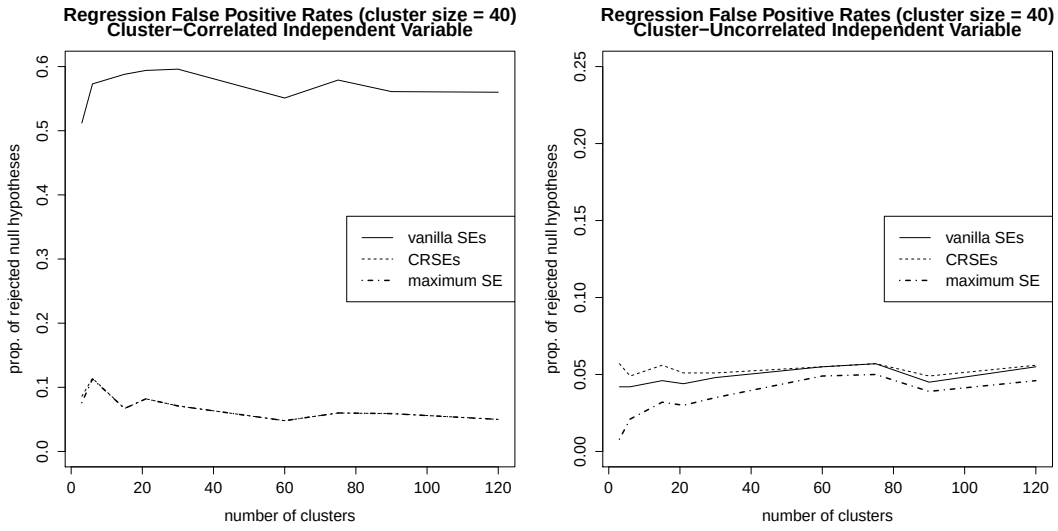


The graphs on the left show the proportion of rejected null hypotheses out of 1000 simulations for parameters whose true values are $\beta_x = \beta_z = 0$ in the linear model $y_i = \beta x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i$ with cluster dependency; this is a measure of the false positive rate. Each model (except random effects) is a correctly specified linear link GLM (estimated using `glm`) with a different method of calculating statistical significance, as indicated in the legend; random effects models are correctly specified linear RE models estimated using `lme4`. One simulation is dropped for random-effects models with 21 clusters due to estimation failure and the rejection rate is calculated out of 999 simulations for that case. The hypothesis tests are conducted at $\alpha = 0.05$, so the false positive rate should ideally equal 0.05. The top graph shows the false positive rate for a variable (x) that is correlated with the cluster structure, while the bottom graph shows the false positive rate for a variable (z) that is uncorrelated with the cluster structure by design. The graphs on the right show the proportion of rejected null hypotheses out of 1000 simulations for parameters whose true values are $\beta_x = \beta_z = 0.25$ in the same linear model; this is a measure of the true positive rate. One simulation is dropped for random-effects models with 90 clusters due to estimation failure and the rejection rate is calculated out of 999 simulations for that case. For a method to have adequate power to conduct significance tests, the true positive rate should ideally equal 1. The top graph shows the true positive rate for a variable (x) that is correlated with the cluster structure, while the bottom graph shows the true positive rate for a variable (z) that is uncorrelated with the cluster structure by design.

Appendix C: Selecting the minimum of vanilla and cluster robust standard errors

If there is correlation between the independent variable of interest and the cluster structure, CRSEs are a flawed tool but are better at limiting false positives when compared to vanilla standard errors. If the cluster structure is unassociated with the independent variable, then vanilla standard errors are much better at limiting false positives. A safe course may be to estimate both, then use whatever standard error is largest to draw any inferences (Green and Vavreck, 2008); we show the outcome of applying this process to our continuous dependent variable simulations from Figure 1 in Figure 6. Ths procedure still produces substantial excess false positives for cluster-correlated independent variables with ≤ 30 clusters.

Figure 6: Result of using the maximum of vanilla and cluster-robust SEs for inference



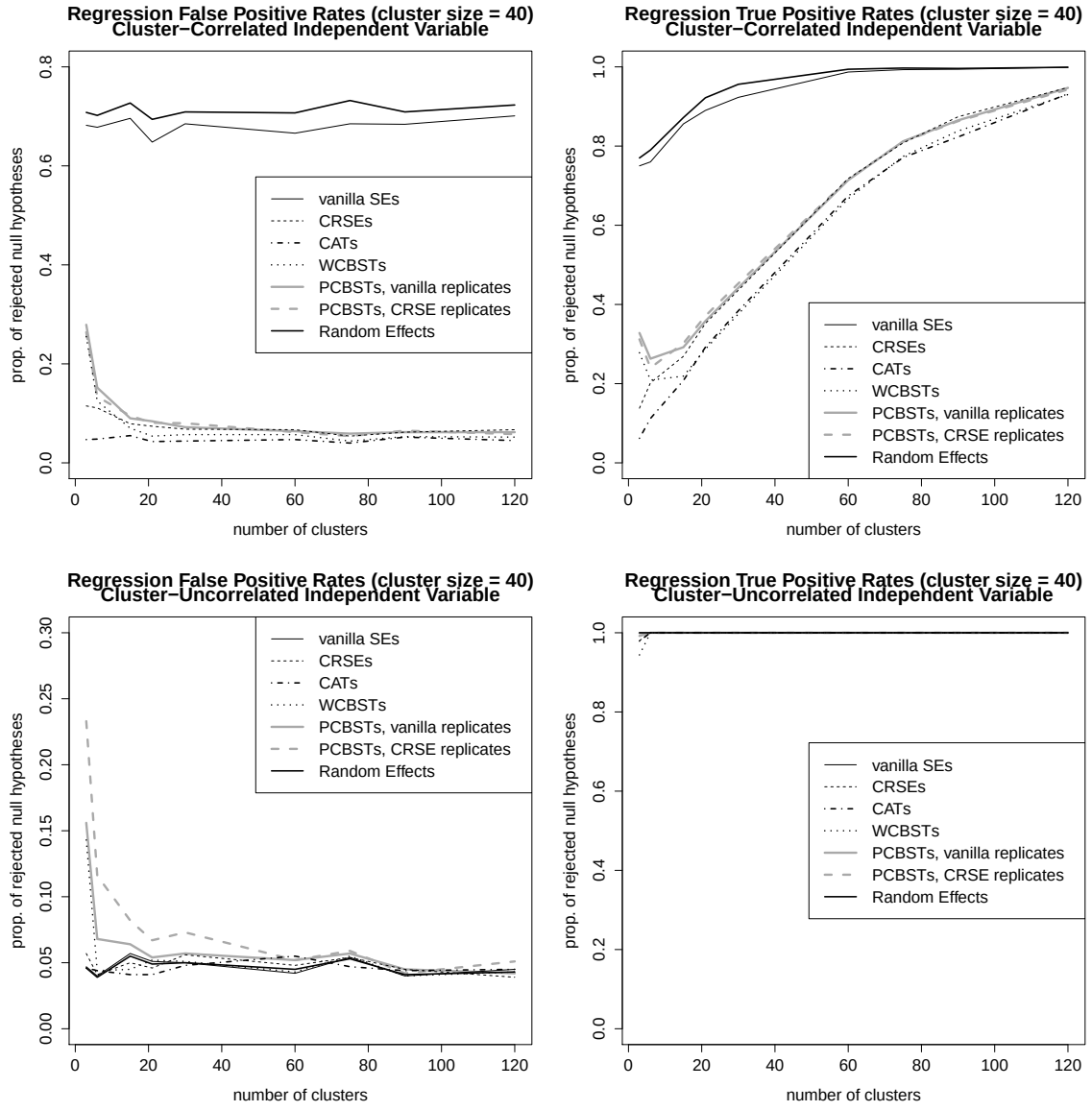
These graphs show the proportion of rejected null hypotheses out of 1000 simulations for parameters whose true values are $\beta_x = \beta_z = 0$ in the linear model $y_i = \beta_x x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i$ with cluster dependency; this is a measure of the false positive rate. Each model is a correctly specified linear link GLM with a different method of calculating statistical significance, as indicated in the legend (maximum SE indicates using the maximum of vanilla and CRSE values for each simulated data set). The hypothesis tests are conducted at the value $\alpha = 0.05$, so the false positive rate should ideally equal 0.05. The graph on the left shows the false positive rates for a variable (x) that is correlated with the cluster structure, while the graph on the right shows the false positive rate for a variable (z) that is not correlated with the cluster structure by design.

Appendix D: Additional simulation results for linear dependent variables with μ_g/γ_g correlation and serial dependence

Figure 7 shows the results of our simulations that include (a) correlation between the average value of x and the group effect γ_g and (b) serial dependence in x and y. The random effects models are the worst performers across all values of G in terms of falsely rejecting the null hypothesis for the cluster-correlated independent variable x, while CATs achieve appropriate null rejection rates for all values of G. The power of all the cluster-adjustment techniques to correctly reject a null hypothesis for the cluster-correlated independent variable is substantially smaller in this simulation compared to the simulation without μ_g/γ_g correlation, particularly for small G.

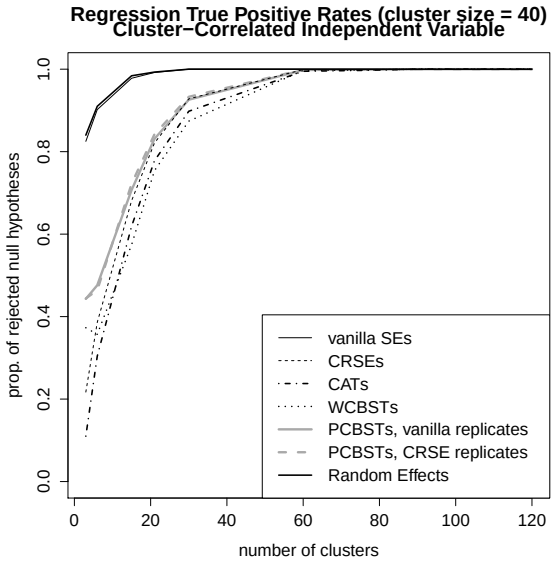
Figure 8 presents the results of simulations identical to those from Figure 7 with β_x and β_z increased to 0.50 from their original setting of 0.25. The detection of true positives for fixed effects models with CATs, WCBSTs, PCBSTs, and CRSEs are all improved relative to the original simulation.

Figure 7: Size and power assessment for linear dependent variables with fixed effects and serial dependence



The graphs on the left show the proportion of rejected null hypotheses ($\beta = 0$) out of 1000 simulations for parameters whose true values are $\beta_x = \beta_z = 0$ in the linear model $y_i = \beta_x x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i$ with (a) correlation between the group-specific mean of x ($= \mu_g$) and the group-level effect γ_g and (b) within-group serial dependence in ε and x; this is a measure of the false positive rate. Each model (except random effects) is a correctly specified linear fixed effects model estimated using `plm` with a different method of calculating statistical significance, as indicated in the legend; random effects models are linear RE models with correct variable specification (but no fixed effects) estimated using `lme4`. The hypothesis tests are conducted at $\alpha = 0.05$, so the false positive rate should ideally equal 0.05. The top graph shows the false positive rate for a variable (x) whose mean is correlated with the group-level effect, while the bottom graph shows the false positive rate for a variable (z) that is uncorrelated with the cluster structure by design. The graphs on the right show the proportion of rejected null hypotheses out of 1000 simulations for parameters whose true values are $\beta_x = \beta_z = 0.25$ in the same linear model; this is a measure of the true positive rate. For a method to have adequate power to conduct significance tests, the true positive rate should ideally equal 1. The top graph shows the true positive rate for a variable (x) whose mean is correlated with the group-level effect, while the bottom graph shows the true positive rate for a variable (z) that is uncorrelated with the cluster structure by design.

Figure 8: Size and power assessment for linear dependent variables with fixed effects and serial dependence, stronger signal for x

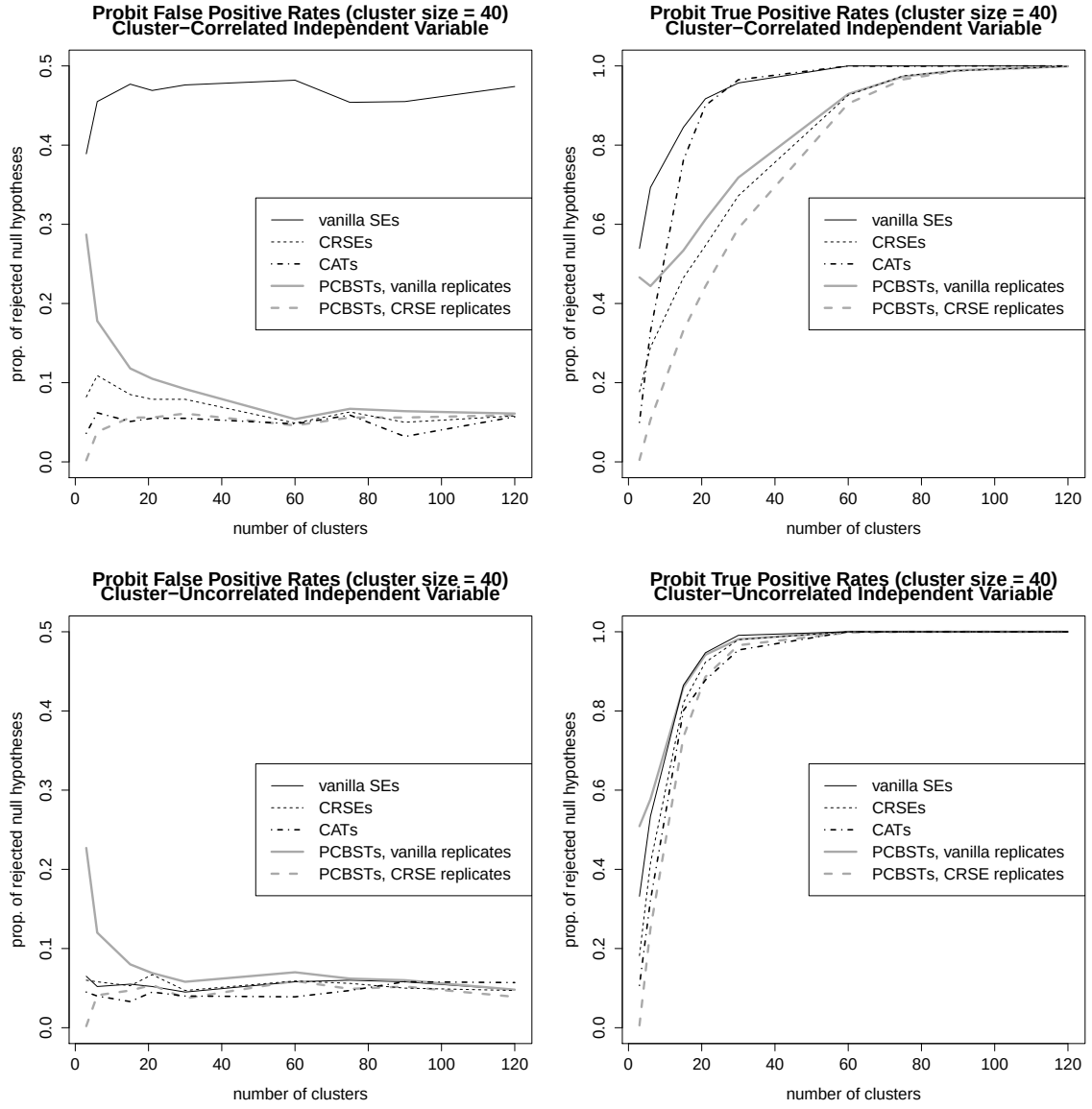


The graph shows the proportion of rejected null hypotheses ($\beta = 0$) out of 1000 simulations (with true values $\beta_x = \beta_z = 0.5$) for the x parameter whose mean is correlated with the group-level effect in the linear model $y_i = \beta_x x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i$ with (a) correlation between the group-specific mean of x (μ_g) and the group-level effect γ_g and (b) within-group serial dependence in ε and x; this is a measure of the true positive rate. Each model (except random effects) is a correctly specified linear fixed effects model estimated using `plm` with a different method of calculating statistical significance, as indicated in the legend; random effects models are linear RE models with correct variable specification (but no fixed effects) estimated using `lme4`. The hypothesis tests are conducted at $\alpha = 0.05$; the true positive rate should ideally equal 1.

Appendix E: Detailed results for binary dependent variables

Figure 9 shows the result of a size/power assessment identical to that for Figure 1 (without fixed effects or serial dependence), but using a binary dependent variable and a probit model in place of the continuous dependent variable with linear model. Just as in the continuous case, CATs have false positive rates that are consistently near the nominal 5% α value of the test across the entire range of cluster sizes with good true positive detection performance (albeit somewhat worse than alternatives for the cluster-uncorrelated independent variable z). By contrast, CRSEs and PCBSTs with vanilla replicates have false positive rates that are substantially higher than α for $\leq$ 30 clusters. PCBSTs with CRSE replicates have false positive rates of less than or equal to 5%, but also have poor true positive detection performance for the cluster-correlated independent variable x.

Figure 9: Size and power assessment for binary dependent variables



The graphs on the left show the proportion of rejected null hypotheses ($\beta = 0$) out of 1000 simulations for parameters whose true values are $\beta_x = \beta_z = 0$ in the probit model $\Pr(y_i = 1) = \Phi(\beta_x x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i)$ with cluster dependency; this is a measure of the false positive rate. Each model is a correctly specified probit link GLM model (estimated using `glm`) with a different method of calculating statistical significance, as indicated in the legend. The hypothesis tests are conducted at the value $\alpha = 0.05$, so the false positive rate should ideally equal 0.05. The top graph shows the false positive rate for a variable (x) that is correlated with the cluster structure, while the bottom graph shows the false positive rate for a variable (z) that is uncorrelated with the cluster structure by design. The graphs on the right show the proportion of rejected null hypotheses out of 1000 simulations for parameters whose true values are $\beta_x = \beta_z = 0.25$ in the same probit model; this is a measure of the true positive rate. For a method to have adequate power to conduct significance tests, the true positive rate should ideally equal 1. The top graph shows the true positive rate for a variable (x) that is correlated with the cluster structure, while the bottom graph shows the true positive rate for a variable (z) that is uncorrelated with the cluster structure by design. Note that, for CAT estimates with 3 clusters, 1 data set is discarded for the false positive simulations and 9 are discarded for the true positive simulations; 1 data set is discarded for 6 clusters in the power simulations. The denominator for the false and true positive rates is adjusted to exclude these results.

Anomaly 1: Failed cluster estimates

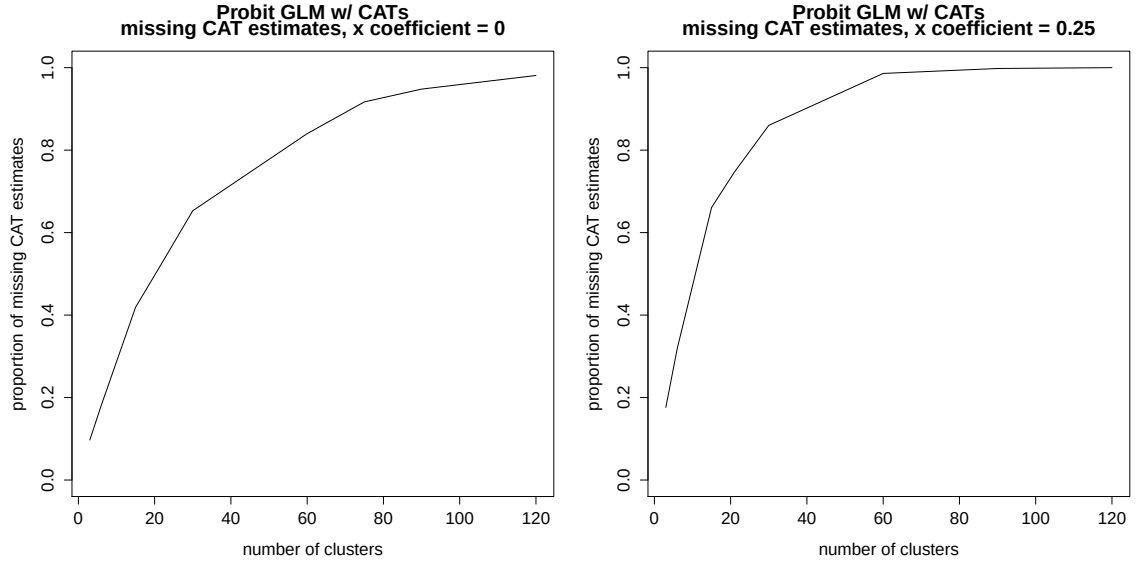
We note an anomaly: CATs often cannot be estimated in the probit model with a large number of clusters. For example, CATs cannot be calculated whenever any cluster has no variation on the dependent variable because the probit model is unidentified in this cluster and will thus fail to converge to an appropriate solution. This problem is far more likely to occur with a binary dependent variable (where noise in a latent continuous propensity does not always cause variation in the observed dichotomous outcome) than with a continuous dependent variable. This severity of the problem varies depending on the value of β, the distribution of γ_g and ε, and the number of clusters; recall that a failure of the probit model to estimate in only *one cluster* results in a failure to estimate CATs. We handle the problem of failed clusters by simply dropping these clusters from the analysis in Figure 9 and calculating the CATs using the number of clusters in which the model was successfully estimated; theoretically, this corresponds to the idea that the dropped clusters contain no information and the remaining clusters still have the same distribution. We provide an argument to support this idea in the subsequent appendix.[32]

Figure 10 shows the proportion of the time that CATs fail in our simulation study (in Figure 9) when individual failed clusters are *not* dropped; instead, a missing result is returned for any variable for which an estimate cannot be obtained in at least one cluster. The percentage of failed (missing) CAT results grows in the number of clusters. In our simulation

[32]If any cluster fails to estimate for every coefficient in a probit model (due, e.g., to non-variation in the DV), both the R and Stata software packages return an error by default. The alternative behavior for both software packages drops these clusters from consideration, but uses the results from models that were successfully estimated. In some cases, individual variables' coefficients cannot be estimated in a particular cluster due to non-variation of the variable in that cluster, but the model can still be estimated if that variable is dropped. By default, both packages report the final result as missing for any variable that is dropped from any cluster-specific model (but still reports the results for variables whose coefficients could be estimated in every cluster). The packages' alternative behavior will exclude all results from any cluster for which *any* variable's results cannot be estimated. For the Monte Carlo analysis shown in appendix Figure 9, all clusters where any variable's coefficient could not be estimated were dropped (the alternative behavior). The software also has an option to drop any cluster with any beta estimate whose distance to the mean is more than 6 times the inter-quartile range; the results in appendix Figure 9 enable this option. The R software for multinomial logit models always excludes clusters for which any variable's results cannot be estimated, but allows dropping of clusters with outlying estimates as an option.

with true positives, CATs fail well over 90% of the time for 120 clusters. The reason is simple: as the number of clusters rises, the probability that at least one cluster will have no variation in the dependent variable *also* rises; even *one* cluster with unidentified $\hat{\beta}$ estimates will cause CATs to fail.

Figure 10: CAT failure rates



The graphs show the proportion of probit models for which CAT estimates could not be computed for β_x out of 1000 simulations. The graph on the left shows the proportion of missing CAT estimates for simulations where $\beta_x = 0$, while the graph on the right shows the proportion of missing CAT estimates where $\beta_x = 0.25$.

Anomaly 2: Outlying β estimates

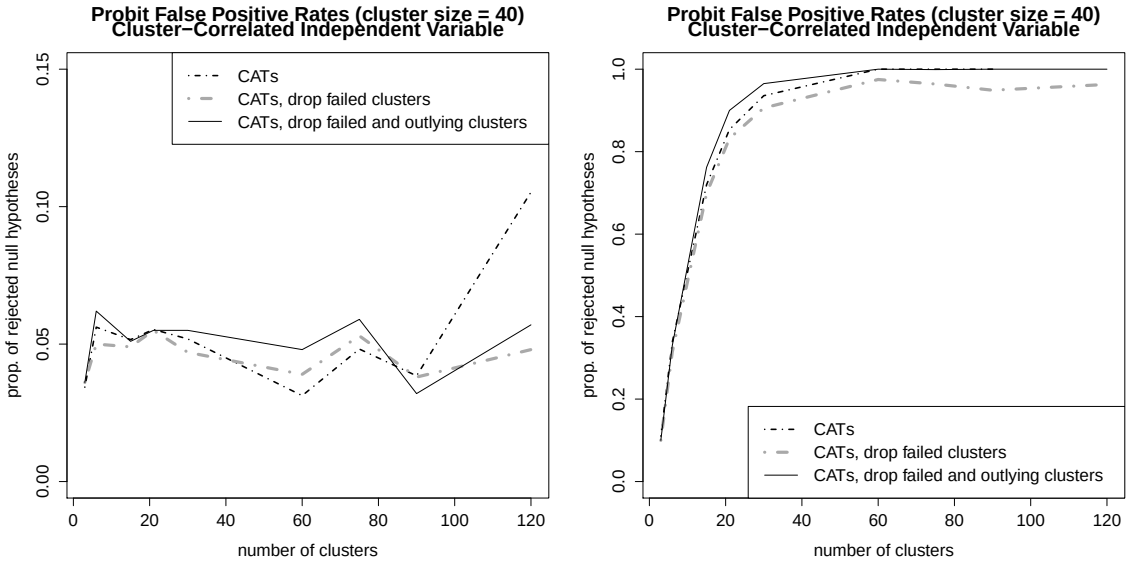
We note another anomaly: even if the model is technically identified in every cluster, extreme outlier β estimates can be produced in cases where perfect or near-perfect separation of the outcomes is predicted by one or more independent variables. The result is that the cluster-level distribution of β has a distribution that is too wide, and consequently too many results are rejected. We address this problem by dropping clusters whose beta estimates are extreme outliers from the distribution of cluster-specific betas. Specifically, we drop any cluster with any beta estimate whose distance to the mean is more than 6 times the inter-quartile range; the subsequent analytic appendix addresses this possibility. These results in Figure 9 include

the use of this procedure.

Analysis with and without anomaly corrections

Figure 11 compares the results of CATs under three alternative procedures: (a) dropping non-converged clusters, (b) dropping non-converged clusters and clusters with outlying β estimates, and (c) excluding any results with non-converged clusters and without dropping outlying β estimates. By dropping the clusters without successfully estimated models and outlying β estimates, we are able to estimate CATs on every data set with more than 6 clusters in our simulation with excellent power and a false positive rate remains very close to the nominal value of 5%. Note that power rates suffer when CATs drop failed clusters but not outlying estimates; dropping the outliers improves the power characteristics.

Figure 11: Size and power assesment for binary dependent variables with and without dropped clusters



The graph on the left show the proportion of rejected null hypotheses ($\beta = 0$) out of 1000 simulations for a parameter whose true value is $\beta_x = 0$ in the probit model $\Pr(y_i = 1) = \Phi(\beta_x x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i)$ with cluster dependency; this is a measure of the false positive rate. Each model is a correctly specified probit link GLM model (estimated using `glm`) with a different method of calculating statistical significance, as indicated in the legend. CATs are either discarded (not calculated) if all coefficients could not be estimated in a cluster ("CATs"), or are calculated by dropping any clusters in which a model failed to estimate and using the remainder ("CATs with dropped clusters"). The graph on the left shows the false positive rate for a variable (x) that is correlated with the cluster structure. The hypothesis tests are conducted at the value $\alpha = 0.05$, so the false positive rate should ideally equal 0.05. The graph on the right shows the proportion of rejected null hypotheses out of the total number of successful simulations for a parameter whose true value is $\beta_x = 0.25$ in the same probit model; this is a measure of the true positive rate. For a method to have adequate power to conduct significance tests, the true positive rate should ideally equal 1. The graph on the right shows the true positive rate for a variable (x) that is correlated with the cluster structure. Note that a small number of models could not be estimated for CATs with dropped failed and outlying clusters: one model could not be estimated for the false positive simulations with three clusters, and nine models could not be estimated for the true positive simulations with three clusters (with one additional model failure for six clusters when outlying estimates are dropped). Failed estimation rates for CATs without dropped clusters are shown in in Figure 11. In all cases, the denominator for the false and true positive rates is adjusted to exclude these results.

Appendix F: Analytic Examination of CATs with Missing Clusters

The Monte Carlo results from the previous appendix indicate that, with limited dependent variable models, researchers may occasionally encounter clusters where the $\hat{\beta}$ coefficients are unidentified because there is no variation in the dependent variable y, or because an independent variable perfectly predicts y ("separation"). This presents a potential problem for CATs (Ibragimov and Müller, 2010), because cluster-level estimates must be calculated in order to use CATs to conduct hypothesis tests and construct confidence intervals. In this appendix, we demonstrate conditions under which simply dropping the clusters with separation does not interfere with inference using CATs.

We begin by quoting the key theorem from Ibragimov and Müller (2010, p. 455), based on a theorem first proved in Bakirov and Szekely (2006):

Theorem 1. *Let x_j, $j = 1...G$ with $G \geq 2$, be independent Gaussian random variables with common mean $E[x_j] = \mu$ and variances $V[x_j] = \sigma_j^2$. Let $t = \sqrt{G}\frac{\bar{x}}{s_x}$ with $\bar{x} = G^{-1}\sum_{j=1}^{G} x_j$ and $s_x^2 = (G-1)^{-1}\sum_{j=1}^{G}(x_j - \bar{x})$. Let $cv_G(\alpha)$ be the critical value of the usual two-sided t-test based on (1) of level α, that is, $P(|T_{G-1}| > cv_G(\alpha)) = \alpha$, and let Φ denote the cumulative density function of a standard normal random variable. If $\alpha \leq 2\Phi\left(-\sqrt{3}\right) \approx 0.083$, then for all $G \geq 2$:*

$$\begin{aligned} \sup_{\{\sigma_1,\ldots,\sigma_G\}} P(|t| > cv_G(\alpha)|H_0) &= P(|T_{G-1}| > cv_G(\alpha)) \\ &= \alpha \end{aligned}$$

Φ denotes the normal distribution. The results of Ibragimov and Müller (2010) depend on cluster-level estimation results $\hat{\beta}_g$ taking a particular distribution as the number of observations in each cluster $N_g \to \infty$:

$$\sqrt{N_g}\left(\hat{\beta}_g - \beta\right) \overset{asym}{\sim} \Phi\left(0, \sigma_g^2\right) \tag{4}$$

which results in the overall vector of results from all clusters $\hat{\beta}_{\mathbf{G}}$ having the distribution:

$$\sqrt{N}\left(\hat{\beta}_{\mathbf{G}} - \beta\right) \overset{asym}{\sim} \Phi(\mathbf{0}, \mathbf{\Sigma_G}) \tag{5}$$

with $N = \sum_{g=1}^{G} N_g$, $\hat{\beta}_{\mathbf{G}}^T = \begin{bmatrix} \hat{\beta}_1 & ... & \hat{\beta}_G \end{bmatrix}$, and $\mathbf{\Sigma_G} = \text{diag}\left(\sigma_1, ..., \sigma_G\right)$. This is consistent with the asymptotic distribution of many estimators under the assumption that observations are uncorrelated between clusters and have constant correlation σ_g within each cluster g. Thus, for Theorem 1 to apply after dropping clusters with unidentified $\hat{\beta}$, the remaining clusters must still be independently and identically distributed normal as in equation (5).

If the clusters are dropped in a way that is uncorrelated with cluster-level value of $\hat{\beta}_g$, then the Ibragimov and Muller procedure remains valid. To demonstrate this, let the set of G_u dropped clusters be designated $\mathbf{G_u}$, and $\mathbf{G_i} = \mathbf{G} \setminus \mathbf{G_u}$ be the set of G_i many clusters that are not dropped.

Proposition 1. *Let $D = [diag(d_1, d_2, ..., d_G)]$ be a $G \times G$ matrix of indicator variables for the identified clusters where $d_g = 1$ when the quantity $\left(\hat{\beta}_g - \beta\right)$ is identified and $= 0$ otherwise. Assuming the conditions of Theorem 1 in Ibragimov and Müller (2010, p. 455) and the asymptotic distribution of cluster-specific coefficients in equation (5), if D is statistically independent from $\hat{\beta}_{\mathbf{G}}$ then $\sqrt{G_i}\left(\hat{\beta}_{\mathbf{G_i}} - \beta\right) \overset{asym}{\sim} \Phi(\mathbf{0}, \mathbf{\Sigma_{G_i}})$.*

Proof. If $\hat{\beta}_{\mathbf{G}}$ and D are statistically independent, then $f(\hat{\beta}_{\mathbf{G}}|D) = f(\hat{\beta}_{\mathbf{G}})$. Define $D_i = [\text{diag}(d_1, d_2, ..., d_{G_i})|0]$, a $G_i \times G$ matrix containing the elements of D for which $d_g = 1$. By Theorem 2.4.4 in Anderson (2003, p. 30),[33] if X is distributed according to $\Phi(\mu, \Sigma)$, then $Z = AX$ is distributed $\Phi\left(A\mu, A\Sigma A'\right)$ where A is an $k \times m$ matrix of rank $k \leq m$. Consequently, $\sqrt{G_i}\left(D_i\left(\hat{\beta}_{\boldsymbol{G}} - \beta\right)|D_i\right) \overset{asym}{\sim} \Phi(\mathbf{0}, \mathbf{\Sigma_{G_i}})$ But $f(\hat{\beta}_{\mathbf{G}}|D) = f(\hat{\beta}_{\mathbf{G}})$; therefore $\sqrt{G_i}\left(D_i\left(\hat{\beta}_{\mathbf{G}} - \beta\right)\right) = \sqrt{G_i}\left(\hat{\beta}_{\mathbf{G_i}} - \beta\right) \overset{asym}{\sim} \Phi(\mathbf{0}, \mathbf{\Sigma_{G_i}})$. □

[33] See also Judge et al. (1988, p. 50).

However, the value of the cluster-level $\hat{\beta}_g$ *will* typically be associated with the probability that the cluster is dropped. Consider the marginal distribution of non-missing coefficients for a single cluster, $f(\hat{\beta}_g)$; the joint distribution just stacks these marginals, as each cluster is independent from the others by assumption. The difference between the true distribution $f(\hat{\beta}_g)$ and what we observe after dropping clusters with unidentified values of $\hat{\beta}$ at any particular value of $\hat{\beta}_g$ is:

$$f(\hat{\beta}_g) - \frac{\pi(\hat{\beta}_g, X) f(\hat{\beta}_g)}{\int \pi(\hat{\beta}_g, X) f(\hat{\beta}_g) d\hat{\beta}_g} = f(\hat{\beta}_g) \left(1 - \frac{\pi(\hat{\beta}_g, X)}{\int \pi(\hat{\beta}_g, X) f(\hat{\beta}_g) d\hat{\beta}_g} \right)$$

where $\pi\left(\hat{\beta}_g, X\right)$ is the probability that a cluster with coefficient values $\hat{\beta}_g$ and data set X will produce set of dependent variable values that identify the coefficient estimates in that cluster (or one minus the probability of missingness). The denominator is the overall probability of non-missingness over all values of $\hat{\beta}_g$ in the cluster. Distortion is minimized when:

$$\pi(\hat{\beta}_g, X) \approx \int \pi(\hat{\beta}_g, X) f(\hat{\beta}_g) d\hat{\beta}_g$$
$$0 \approx \pi(\hat{\beta}_g, X) - \int \pi(\hat{\beta}_g, X) f(\hat{\beta}_g) d\hat{\beta}_g$$

Consequently, it appears that the difference between the true distribution of $\hat{\beta}_{\mathbf{G}}$ and the distribution after dropping non-identified clusters will be minimal if:

1. the overall probability of dropping is small for every cluster g, $\int \pi(\hat{\beta}_g, X) f(\hat{\beta}_g) d\hat{\beta}_g \approx 1$ (implying that $\pi(\hat{\beta}_g, X) \approx 1$ for all values of $\hat{\beta}_g$) , or:

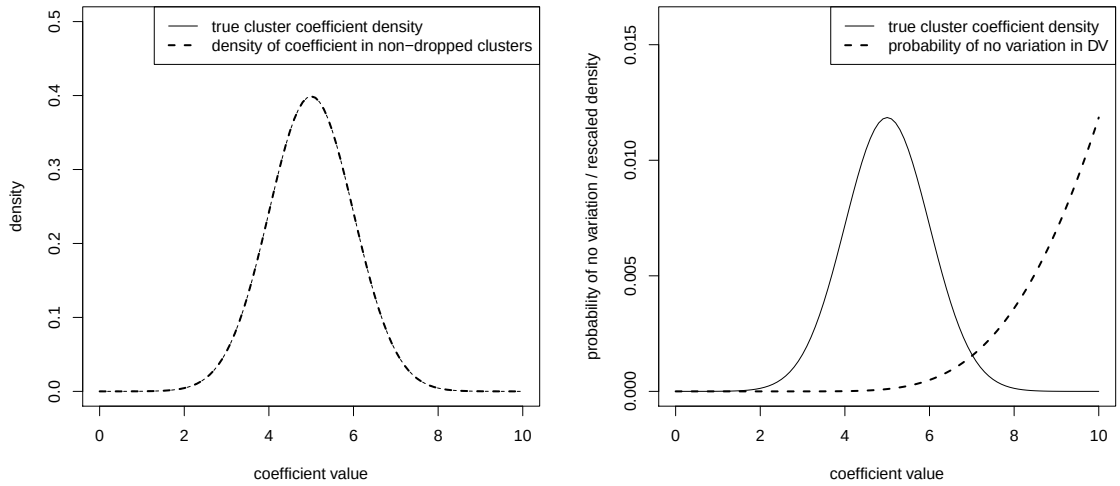
2. the degree of heterogeneity in the probability of dropping the cluster for different values of $\hat{\beta}_g$ is small, or $\pi(\hat{\beta}_g^a, X) - \pi(\hat{\beta}_g^b, X) \approx 0$ for all values of a and b so that $\pi(\hat{\beta}_g, X) \approx \int \pi(\hat{\beta}_g, X) f(\hat{\beta}_g) d\hat{\beta}_g$.

As a rough rule of thumb, by rule 1 above the distortion of results is likely to be small if the number of missing clusters is also relatively small. It is possible to formally assess the probability of missingness for each cluster under some assumptions about $\hat{\beta}_g$ and to attempt to estimate the resulting distortion, if a greater degree of rigor is desired, but we leave this task to future research.

To illustrate when dropped clusters may or may not be a problem, we have prepared two examples, depicted in Figures 12 and 13. The figures assume a distribution for a single cluster coefficient $\hat{\beta}_g$, then determine the probability that there is no variation in the dependent variable under the probit model $\Phi(X\hat{\beta})$; a cluster-level estimate will not be identified under this condition. This probability is particular to the data set, so we create a simple data set where X is a sequence of values $\{0.01, 0.02, ..., 0.99, 1\}$ to use in all calculations. The figures depict the source distribution of $\hat{\beta}_g$ as well as the density of non-missing values of $\hat{\beta}_g$ in the left panel while the probability that a cluster observation is dropped (as a function of $\hat{\beta}_g$) is shown in the right panel. Based on the discussion above, we expect minimal distortion when (a) the overall probability of missingness is low, or (b) the probability of missingness is consistent across different values of $\hat{\beta}_g$.

As Figure 12 shows, when $f(\hat{\beta}_g) \sim \phi(\mu = 5, \sigma = 1)$, there is almost no difference between the distribution of $\hat{\beta}_g$ with and without dropped clusters. This is because (as shown in the right panel) the probability of dropping a cluster is near zero across most of the high-density values of $f(\hat{\beta}_g)$ and is close to zero throughout. The picture is much different in Figure 13, where the density of $\hat{\beta}_g$ in non-dropped clusters is substantially different than the source distribution $f(\hat{\beta}_g) \sim \phi\,(\mu = 25, \sigma = 12)$. The distortion is caused because high values of $\hat{\beta}_g$ are likely to produce no variation in the dependent variable and therefore be dropped, while lower values of $\hat{\beta}_g$ are unlikely to do so.

Figure 12: Cluster dropping due to no DV variation with small distortion

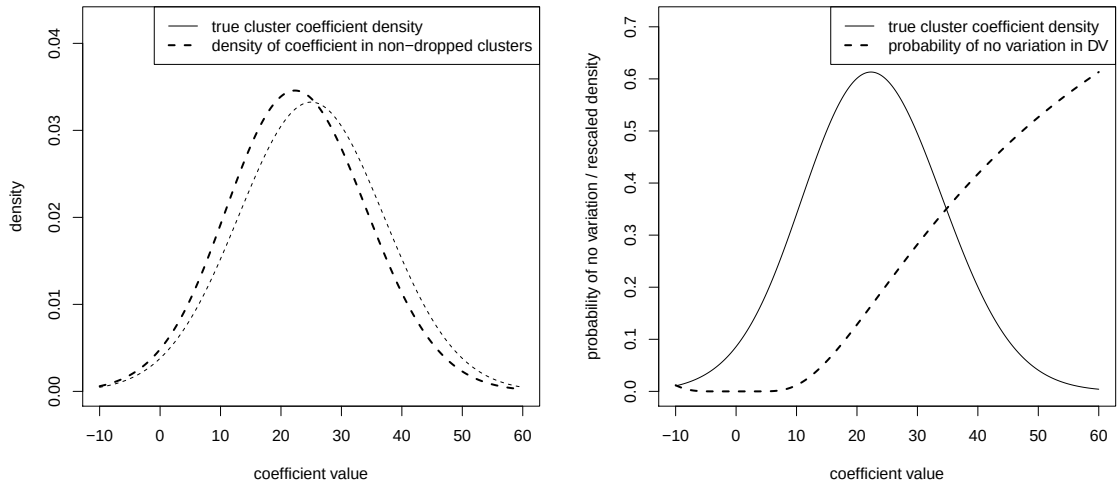


These graphs depict the result of calculating the probability of no variation in the dependent variable y under a model $y = \Phi\left(X\hat{\beta}\right)$ with cluster-level estimates $\hat{\beta}_g$ distributed according to $f(\hat{\beta}_g) = \phi(\mu = 5, \sigma = 1)$. The dataset X is a sequence of values $\{0.01, 0.02, ..., 0.99, 1\}$ with each value representing a distinct observation on a single variable in the cluster. The probability of missingness in the cluster is calculated as $\pi(\hat{\beta}_g) = \prod_i \left[\Phi\left(X_i\hat{\beta}_g\right)\right] + \prod_i \left[1 - \Phi\left(X_i\hat{\beta}_g\right)\right]$. The density of non-missing coefficients is calculated as $g(\hat{\beta}_g|\text{non-missing}) \propto \left[1 - \pi(\hat{\beta}_g)\right] f(\hat{\beta}_g)$.

Appendix G: Detailed results for multinomial dependent variables

Our results for the multinomial case are listed in Figure 14. The performance of each type of standard error is qualitatively similar to our results in linear and probit models. We conclude that applying CATs (or PCBSTs with CRSE replicates) is a valid way of limiting the false positive rate when estimating uncertainty and conducting hypothesis tests for multinomial models with a small number of clusters. As with the probit models, we drop any clusters for which any coefficient cannot be estimated or with any beta estimate whose distance to the inter-cluster mean is more than 6 times the inter-quartile range. The results of this procedure are depicted in Figure 14.

Figure 13: Cluster dropping due to no DV variation with large distortion



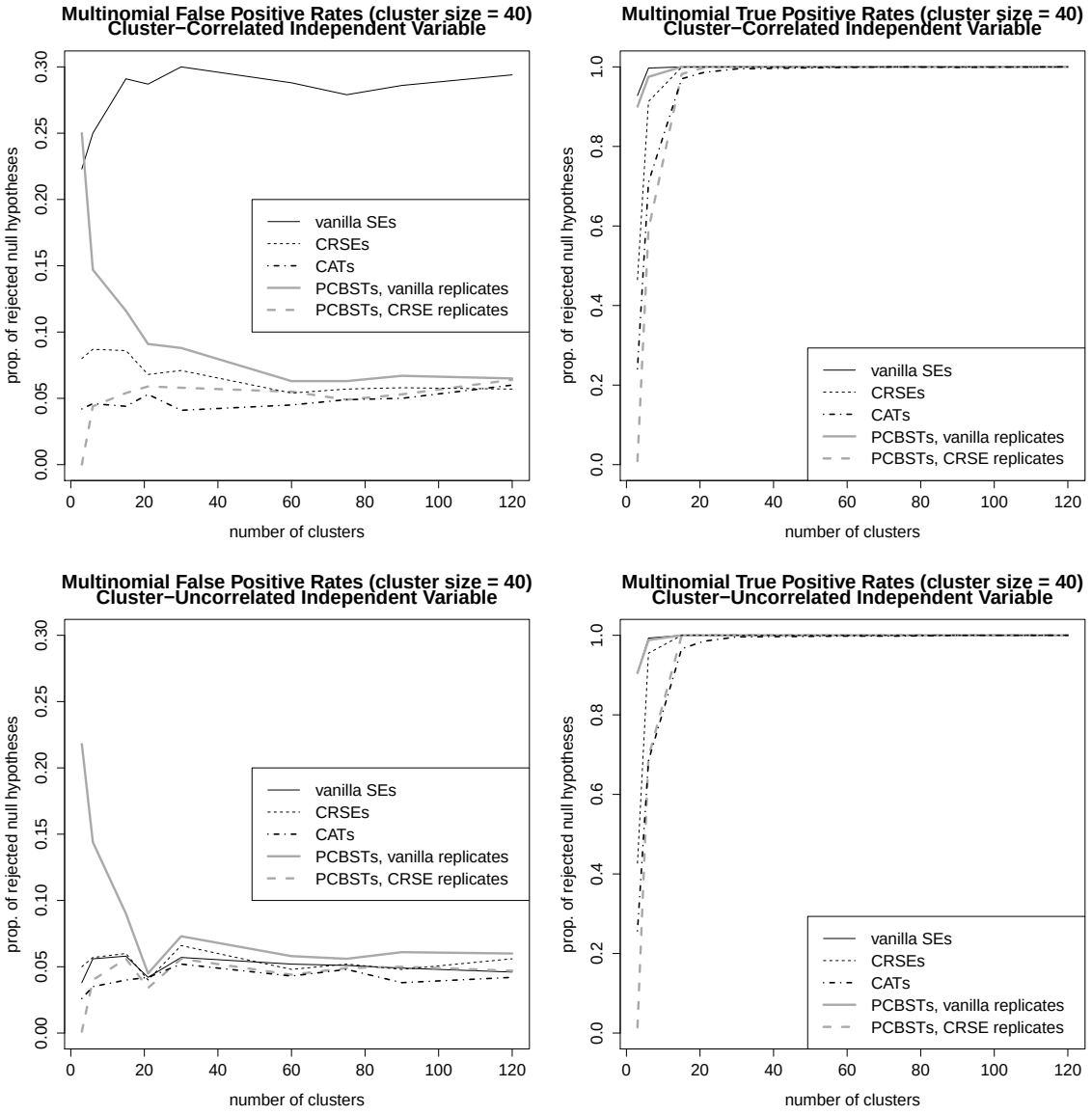
These graphs depict the result of calculating the probability of no variation in the dependent variable y under a model $y = \Phi\left(X\hat{\beta}\right)$ with cluster-level estimates $\hat{\beta}_g$ distributed according to $f(\hat{\beta}_g) = \Phi(\mu = 25, \sigma = 12)$. The dataset X is a sequence of values $\{0.01, 0.02, ..., 0.99, 1\}$ with each value representing a distinct observation on a single variable in the cluster. The probability of missingness in the cluster is calculated as $\pi(\hat{\beta}_g) = \prod_i \left[\Phi\left(X_i\hat{\beta}_g\right)\right] + \prod_i \left[1 - \Phi\left(X_i\hat{\beta}_g\right)\right]$. The density of non-missing coefficients is calculated as $g(\hat{\beta}_g|\text{non-missing}) \propto \left[1 - \pi(\hat{\beta}_g)\right] f(\hat{\beta}_g)$.

Appendix H: Additional results for Grosser, Reuben and Tymula (2013)

A bivariate analysis of the relationship between changes in transfers and changes in candidate tax proposals indicates that group level heterogeneity exists in how candidates react to the rich voter's behavior; this is shown in Figure 15. In some of the groups (e.g., groups 7 and 11) there is a reasonably clear negative relationship between the change in how much money a candidate received from the rich voter and the contemporaneous change in that candidate's tax proposal. But, as shown in Figure 15b; many other groups seem to have relationships clustered around zero, with some slightly less than zero and some slightly greater than zero.

Table 2 reproduces the regression analysis of Grosser, Reuben and Tymula (2013) using their original CRSEs as well as pairs cluster bootstrapped t-statistics and cluster-adjusted t statistics. The CRSE uncertainty measures support the authors' original interpretation

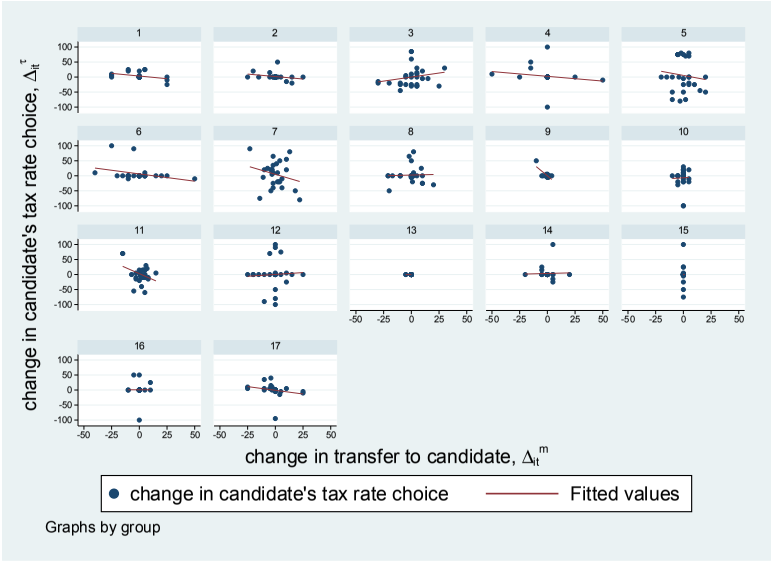
Figure 14: Size and power assessment for multinomial dependent variables



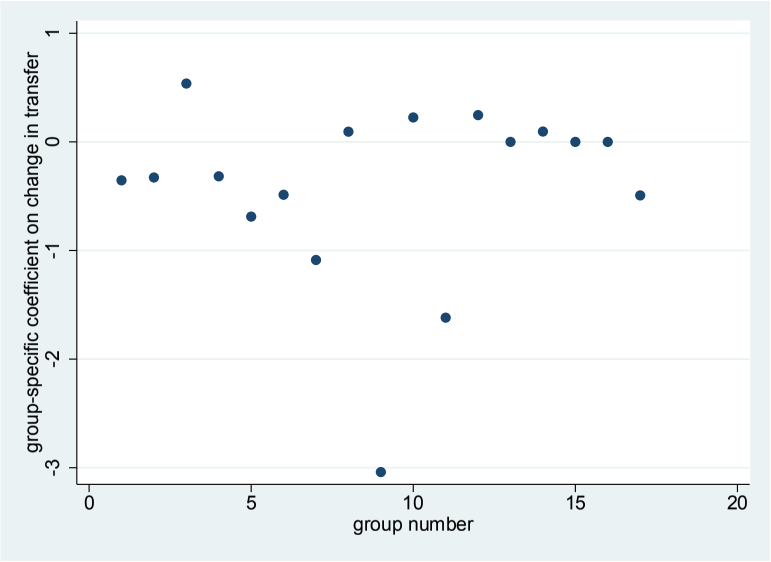
The graphs on the left show the proportion of rejected null hypotheses ($\beta = 0$) out of 1000 simulations for parameters whose true values are $\beta_{x2} = \beta_{z2} = 0$ in the multinomial logit model with cluster dependency; this is a measure of the false positive rate. Each model is a correctly specified multinomial logit model estimated with `mlogit` with a different method of calculating statistical significance, as indicated in the legend. The hypothesis tests are conducted at the value $\alpha = 0.05$, so the false positive rate should ideally equal 0.05. The top graph shows the false positive rate for a variable (x) that is correlated with the cluster structure, while the bottom graph shows the false positive rate for a variable (z) that is uncorrelated with the cluster structure by design. The graphs on the right show the proportion of rejected null hypotheses out of 1000 simulations for parameters whose true values are $\beta_{x2} = \beta_{z2} = 1$ in the same multinomial model; this is a measure of the true positive rate. For a method to have adequate power to conduct significance tests, the true positive rate should ideally equal 1. The top graph shows the true positive rate for a variable (x) that is correlated with the cluster structure, while the bottom graph shows the true positive rate for a variable (z) that is uncorrelated with the cluster structure by design.

Figure 15: Bivariate analysis of relationship between transfers and proposed tax policy

(a) Scatterplot of changes in transfers to candidates (Δ_{it}^{m}) and changes in candidate proposed tax rate (Δ_{it}^{τ}), by group



(b) Two-variable regression coefficient for relationship between changes in transfers to candidates (Δ_{it}^{m}) and changes in candidate proposed tax rate (Δ_{it}^{τ}), by group



The left graph shows each candidate's change in tax rate proposal between the present and past period (Δ_{it}^{τ}) on the y-axis and the change in transfers received by that candidate over the same time (Δ_{it}^{m}) on the x-axis, with fitted regression lines over the scatterplots of observations; each panel indicates a different group. The right graph displays the bivariate regression coefficient on change in transfers in a regression on change in tax policy on the observations in each group.

that a candidate who receives increased transfers from the rich voter tends to subsequently propose a reduced tax rate. However, both PCBSTs (with CRSE replicates) and CATs fail to reject the null of no relationship for the coefficients on Δ_{it}^{m} and the interaction term $(\Delta_{it}^{m} * t)$ using an $\alpha = 0.05$ test, two-tailed.

Tables 3 and 4 respectively contain the analysis of "high tax" and "low tax" groups in the experiment of Grosser, Reuben and Tymula (2013). These regressions are identical to the regression in Table 2 presented in the main text of the manuscript, except on subsamples of the subjects defined to be in "high tax" or "low tax" groups according to the criteria specified in the text. The results of these tables are used to produce the marginal effects plots in Figures 4a and 4b that are shown in the main text.

Table 2: Determinants of Tax Policy Changes (Table 2, Column 1 from Grosser, Reuben and Tymula (2013))

		uncertainty estimates (95% CIs and two-tailed p-values)		
	coefficient	CRSE	PCBST	CAT
change in received transfer (Δ^m_{it})	-0.876	[-1.50, -0.247] $p = 0.009$	[-1.81, 0.0599] $p = 0.059$	[-1.86, 0.675] $p = 0.332$
change in received transfer X period ($\Delta^m_{it} * t$)	0.0925	[-0.00101, 0.186] $p = 0.052$	[-0.0508, 0.236] $p = 0.149$	[-0.396, 0.355] $p = 0.909$
positive diff. in previous tax policy (D^+_{ij})	-0.201	[-0.474, 0.0732] $p = 0.140$	[-0.804, 0.403] $p = 0.185$	[-0.269, 0.327] $p = 0.837$
negative diff. in previous tax policy (D^-_{ij})	0.777	[0.562, 0.991] $p < 0.001$	[0.477, 1.08] $p = 0.004$	[0.595, 1.34] $p < 0.001$
period (t)	-0.0340	[-0.684, 0.616] $p = 0.913$	[-0.678, 0.610] $p = 0.909$	[-0.585, 1.37] $p = 0.404$

Dependent variable: change in candidate's proposed tax rate (Δ^τ_{it}). This table reports the results of a fixed effects linear regression model. The constant and subject-level fixed effects were included in the analysis but omitted in this table. 2 groups were automatically dropped from the CAT analysis because at least one parameter could not be estimated in the group.

Table 3: Determinants of Tax Policy Changes in "High Tax" Groups (Table 2, Column 2 from Grosser, Reuben and Tymula (2013))

		uncertainty estimates (95% CIs and two-tailed p-values)		
	coefficient	CRSE	PCBST	CAT
change in received transfer (Δ_{it}^{m})	-0.604	[-1.05, -0.156] $p = 0.014$	[-1.47, 0.261] $p = 0.128$	[-2.59, 2.36] $p = 0.915$
change in received transfer X period ($\Delta_{it}^{m} * t$)	0.0532	[0.00907, 0.0973] $p = 0.023$	[-0.0228, 0.129] $p = 0.110$	[-0.923 0.630] $p = 0.669$
positive diff. in previous tax policy (D_{ij}^{+})	0.0216	[-0.0173, 0.0604] $p = 0.241$	[-0.0220, 0.0651] $p = 0.248$	[-0.132, 0.489] $p = 0.216$
negative diff. in previous tax policy (D_{ij}^{-})	0.807	[0.473, 1.14] $p < 0.001$	[-0.0743, 1.69] $p = 0.065$	[0.583, 1.38] $p = 0.001$
period (t)	-0.377	[-1.25, 0.496] $p = 0.354$	[-1.53, 0.777] $p = 0.346$	[-1.71, 0.839] $p = 0.445$

Dependent variable: change in candidate's proposed tax rate (Δ_{it}^{τ}). This table reports the results of a fixed effects linear regression model. The constant and subject-level fixed effects were included in the analysis but omitted in this table. 2 groups were automatically dropped from the CAT analysis because at least one parameter could not be estimated in the group.

Table 4: Determinants of Tax Policy Changes in "Low Tax" groups (Table 2, Column 3 from Grosser, Reuben and Tymula (2013))

		uncertainty estimates (95% CIs and two-tailed p-values)		
	coefficient	CRSE	PCBST	CAT
change in received transfer (Δ_{it}^{m})	-1.102	[-1.96, -0.246] $p = 0.020$	[-2.00, -0.205] $p = 0.034$	[-2.23, -0.0484] $p = 0.043$
change in received transfer X period ($\Delta_{it}^{m} * t$)	0.118	[-0.00840, 0.245] $p = 0.062$	[-0.0942, 0.331] $p = 0.134$	[-0.0137, 0.261] $p = 0.070$
positive diff. in previous tax policy (D_{ij}^{+})	-0.448	[-0.933, 0.0373] $p = 0.065$	[-1.06, 0.162] $p = 0.103$	[-0.762, 0.479] $p = 0.597$
negative diff. in previous tax policy (D_{ij}^{-})	0.709	[0.346, 1.07] $p - 0.003$	[-0.136, 1.55] $p = 0.067$	[0.131, 1.78] $p = 0.030$
period (t)	0.574	[-0.563, 1.71] $p = 0.263$	[-0.368, 1.51] $p = 0.246$	[-0.149, 2.83] $p = 0.070$

Dependent variable: change in candidate's proposed tax rate (Δ_{it}^{τ}). This table reports the results of a fixed effects linear regression model. The constant and subject-level fixed effects were included in the analysis but omitted in this table.

Appendix I: How governments shape the risk of civil violence (Lacina, 2014)

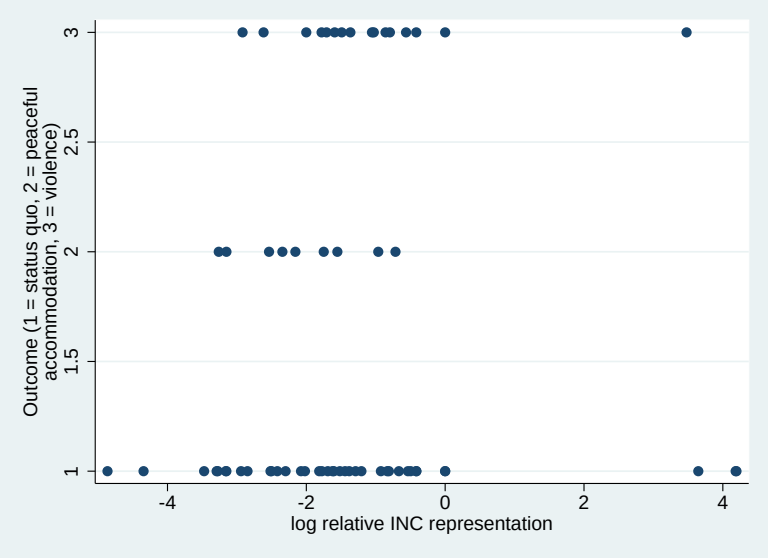
In the 2014 volume of the *American Journal of Political Science*, Lacina (2014) uses data from India to argue that "representation in the ruling party conditioned the likelihood of a violent statehood movement" (p. 720). Her primary empirical evidence comes from a data set consisting of 63 "language enclaves" that could have become states inside of the Indian federal system examined between 1950-1956. The idea is to determine whether there is a relationship between civil unrest in these enclaves and the balance with which conflicting viewpoints about statehood were represented in the governing Indian National Congress Party (INC) inside of these enclaves. Relative representation in the INC "is the ratio of the Congress representation of the opponents of statehood to the Congress representation of proponents" (p. 728). State outcomes are coded as status quo (= 1), peaceful accommodation (= 2), or violence (= 3). According to Lacina's coding rules, accommodation occurs when "an enclave becom[es] a state (or part of a state) where the enclave's largest language is also the state's majority language" (p. 729), while violence occurs when a statehood-related incident involving injuries or deaths is reported in the Bombay edition of the *Times of India* during the time period under study.

The conclusion that the author draws from this data set is stated clearly in the abstract:

> I show that representation in the ruling party conditioned the likelihood of a violent statehood movement. Prostatehood groups that were politically advantaged over the interests opposed to them were peacefully accommodated. Statehood movements similar in political importance to their opponents used violence. Very politically disadvantaged groups refrained from mobilization, anticipating repression. (Lacina, 2014, 720)

This conclusion is supported by the results of a multinomial logit model using clustered standard errors, which we replicate in the first column of Table 5. As the table indicates, the

Figure 16: Bivariate plot of relative INC representation and outcomes in Indian language enclaves, 1950-1956, based on data from Lacina (2014)



relationship between the log of relative INC representation and its square have a statistically significant relationship with the violence and peaceful accommodation outcomes. However, we also have reason to suspect that these results will be sensitive to the structure of the standard errors. The original CRSEs are clustered on the 25 pre-existing states in which the language enclaves are located, and our simulation results indicate that this puts the result at an elevated risk of being a false positive. In our view, an examination of the bivariate relationship between INC representation and outcomes in the language enclaves (shown in Figure 16) suggests that this may be a false positive result driven by the use of CRSEs with a small number of clusters. The plot indicates little apparent relationship between outcomes and the log of relative INC representation. We therefore proceed with a re-analysis of the multinomial logit model using alternative cluster-robust measures of uncertainty.

Because some of the 25 clusters have only one observation each and there are only 63 observations total, we cannot feasibly estimate CATs on this data set; there are not enough degrees of freedom in each cluster to actually estimate the multinomial logistic model in most clusters. Consequently, we rely on PCBSTs with CRSE replicates for inference as a fallback measure; 37 bootstrap replicates (out of 1000 estimated) would not estimate, but

Table 5: Effect of INC representation on violent transition to statehood (Table 5, Model 1 from Lacina (2014))

		uncertainty estimates (95% CIs and two-tailed p-values)		
	coefficient	CRSE	PCBST	Vanilla SEs
Outcome: Peaceful accommodation				
ln relative INC representation	-4.92	[-9.38, -0.448] $p = 0.031$	[-86.4, 76.5] $p = 0.337$	[-11.3, 1.50] $p = 0.133$
ln relative INC representation sq.	-1.17	[-2.12, -0.223] $p = 0.015$	[-18.7, 16.4] $p = 0.298$	[-2.72, 0.370] $p = 0.136$
Outcome: Violence				
ln relative INC representation	0.609	[0.115, 1.10] $p = 0.016$	[-8.75, 9.97] $p = 0.219$	[-0.133, 1.35] $p = 0.108$
ln relative INC representation sq.	-0.341	[-0.545, -0.138] $p = 0.001$	[-4.19, 3.51] $p = 0.188$	[-0.630, -0.0523] $p = 0.021$

The base category of this multinomial logit model is a "status quo" outcome. Other variables included in the model but not listed here are: demographic polarization, ln enclave plurality group's INC representation, ln enclave plurality group's population, agricultural labor share in enclave, landless rate in enclave, Hindu share in enclave, ln km to New Delhi, and a constant.

we use the rest for our analysis.

Table 5 shows our results. As you can see, pairs cluster bootstrapped t-statistics decisively fail to reject the null hypothesis for all the independent variables of interest. Moreover, the 95% CIs around these effects are quite large; this reflects the fact that the tails of the bootstrap distribution are very wide because we have such a small number of clusters that contain a relatively small amount of information. We also note that the vanilla standard errors indicate considerably more uncertainty in the results than the CRSEs; only one of the coefficients is significant at the $\alpha = 0.05$ level, two-tailed.

Our conclusion is that Lacina's (2014) data set is probably too small to support an analysis that accounts for the clustered structure of the data. If we must draw a conclusion, a multinomial model with pairs cluster bootstrap standard errors fails to reject the null of no relationship between INC representation and the presence of a violent statehood movement. Moreover, an analysis with no cluster correction (using vanilla SEs) yields a similar result.

Appendix J: Consumer demand for the fair trade label (Hainmueller, Hiscox and Sequeira, 2015)

Even when the choice of clustering method does not change inferences, it can influence the degree of uncertainty in the substantive size of a finding. For example, Hainmueller, Hiscox and Sequeira (2015) conducted a field experiment testing the response of consumers to coffee bearing a "fair trade" label compared to a standard (non-fair trade) label. The experiment is designed to see whether purchasing behavior is genuinely influenced by appeals to the ethical preferences of consumers, including whether these appeals are drowned out when the ethical product is higher-priced. We focus on the portion of their experiment designed to detect whether fair trade labels increased coffee sales. In this experiment, the researchers attached a fair trade label to certain bulk coffee bins in some randomly selected stores, but not in others. They then compared sales of this coffee from stores with the label to sales from stores where the label was not applied. The research design is predicated on the assumption that, on average and at any given time, nothing differs between the two sets of stores or the coffee in those stores except the application of the fair trade label.

The dependent variable in Hainmueller et al.'s analysis is:

$$\delta_{jt} = \log(s_{jt}) - \log(s_{0t})$$

where s_{jt} is coffee brand j's market share in week t for a particular store and s_{0t} is the proportion of the latent market share not captured by any brand (viz., the portion of the potential coffee market occupied by other non-coffee goods). Each observation in the data set is a brand-store-week. The authors calculate market share "by converting volume sales to pounds and dividing by the total potential number of pounds of coffee in a given market. The potential coffee market is assumed to be equal to one cup of coffee per customer per day in a given store-week" (Hainmueller, Hiscox and Sequeira, 2015, p. 19) There are two bulk

coffees where the fair trade label is manipulated, and five other bulk coffees never labeled as fair trade. 26 stores are observed over eight weeks in the data set, but a few brand-store-week observations are discarded "because of occasional stock outs and/or bulk bin rotations" (p. 19). The resulting model is:

$$\delta_{jt} = \beta_0 + \beta_1 L + \xi_{sj} + \xi_t$$

where L is an indicator variable for bulk coffees where the label is applied, ξ_{sj} is a fixed effect for product j in store s, and ξ_t is a time dummy for week. As Hainmueller, Hiscox and Sequeira (2015) show, this statistical model can be deduced from a theoretical random utility model where individual-level utility is a function of L and random noise.

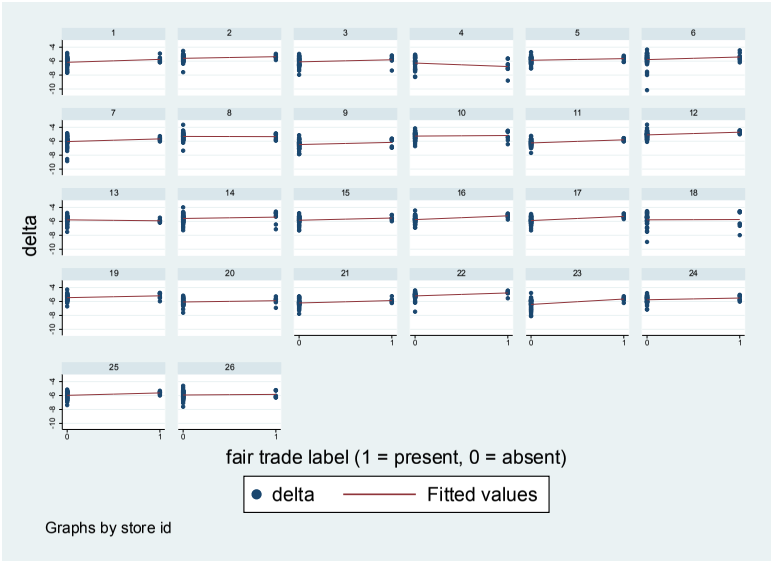
Visual assessment of the bivariate relationship between market share and fair trade labeling (in Figure 17a) seems to suggest that there is a small, positive relationship between fair trade labeling and market share in 23 out of 26 stores. This is confirmed in a plot of store-specific regression coefficients of fair trade against market share in Figure 17b. However, these coefficients vary substantially in magnitude, and there are three coefficients less than zero (one of which is *substantially* less than zero).

Hainmueller, Hiscox and Sequeira (2015) originally used CRSEs in their model, clustering on the 26 stores that participated in their experiment; we replicated these results exactly and report them in Table 6. As they report, "sales increased by about 10% with the Fair Trade label ($p < 0.01$)." However, we also calculated 95% confidence intervals and p-values using PCBSTs (with CRSE replicates) and CATs, again clustering on store. Table 6 makes it apparent that the results are more variable when using PCBSTs and CATs compared to CRSEs; the 95% confidence intervals are 20% wider for PCBSTs and 46% wider for CATs compared to CRSEs. However, none of these confidence intervals cross zero, allowing us to reject the null hypothesis of no effect in every case.

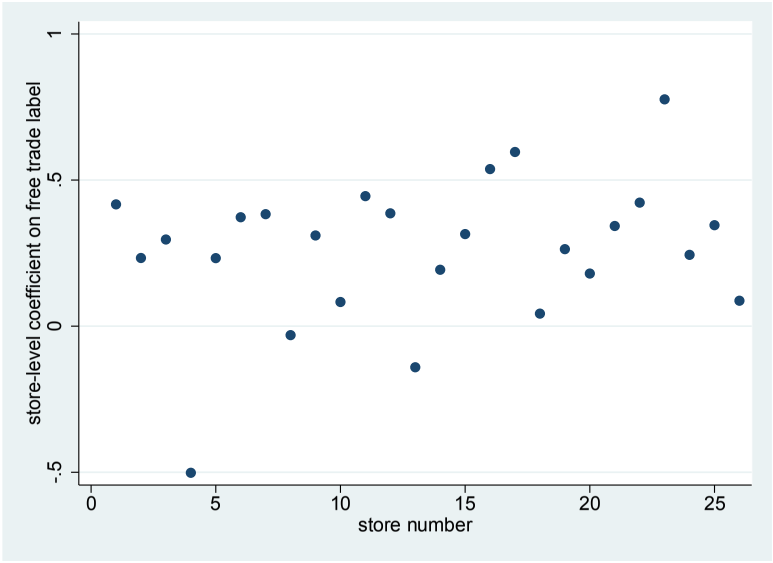
On the basis of this evidence, we conclude that the data collected in Hainmueller, Hiscox and Sequeira (2015) are generally supportive of their claim that fair trade labeling increases market share. There is somewhat greater uncertainty associated with the substantive

Figure 17: Bivariate Analysis of Effect of Fair Tade Labeling on Market Share (δ_{jt})

(a) Scatterplot of Fair Trade Label against Market Share, by Store



(b) Two-variable regression coefficient for effect of Fair Trade Label on Market Share, by Store



The left graph shows the market share δ_{jt} on the y-axis and the presence of the fair trade label on the x-axis, with fitted regression lines over the scatterplots of observations; each panel indicates a different store. The right graph displays the bivariate regression coefficient between the fair trade label and the market share δ_{jt} for observations within each store.

Table 6: Effect of Fair Trade Label of Sales of Test Coffees (Table 5, Column 1 from Hainmueller, Hiscox and Sequeira (2015))

		uncertainty estimates (95% CIs and two-tailed p-values)		
	coefficient	CRSE	PCBST	CAT
fair trade label	0.103	[0.0425, 0.163] $p = 0.007$	[0.0303, 0.175] $p = 0.007$	[0.0486, 0.225] $p = 0.004$

This table reports the results of a fixed effects linear regression model. The constant, week fixed effects, and product-store fixed effects were included in the analysis but omitted in this table.

magnitude of the relationship than would be implied by CRSEs. Fortunately, this greater uncertainty does not change the results of a t-test, even at the $\alpha = 0.01$ level.