Political Science Research and Methods Vol 7, No. 3, 541–559 July 2019
© The European Political Science Association, 2018
doi:10.1017/psrm.2017.42

Practical and Effective Approaches to Dealing With Clustered Data*

JUSTIN ESAREY AND ANDREW MENGER

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. To illustrate these advantages, we reanalyze recent work where results are based on cluster-robust standard errors.

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). But it is not as widely understood that using CRSEs without regard to their 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 (Cameron, Gelbach and Miller 2008; Green and Vavreck 2008; Angrist and Pischke 2009, Chapter 8; Harden 2011). 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 are as easy to implement in Stata and R, and (b) the limitations of CRSEs are not 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. Toward this end, we create and make available pre-packaged routines for estimating cluster-adjusted *t*-statistics (CATs) (Ibragimov and Müller 2010), pairs

* Justin Esarey is an Assistant Professor of Political Science, Rice University, 6100 Main St, MS-24, Houston, TX 77005 (justin@justinesarey.com). Andrew Menger, Ph.D. Candidate, Department of Political Science, Rice University, 6100 Main St, MS-24, Houston, TX 77005 (Andrew.M.Menger@rice.edu). 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. To view supplementary material for this article, please visit https://doi.org/10.1017/psrm.2017.42

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

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

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 simulations, we find that CATs are usually more effective (and always at least as effective) at limiting false positives compared with 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 with 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. Finally, although random effects (RE) models have better size and power characteristics in our simulation compared with any cluster adjustment alternative when the model is correctly specified, cluster adjustments do substantially better when the assumptions of the RE model are not satisfied.

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. 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 data set 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 are frequently grouped; 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 data sets where the data exhibits some dependency between observations in the same cluster, ignoring this dependence can greatly underestimate the true standard errors (SEs) for parameters of interest when this clustering structure is associated with the independent variable (Moulton 1986; Moulton 1990). This can lead to researchers falsely rejecting the null hypothesis of a statistical significance test too

[2] Note that the cluster wild bootstrap was already introduced in a Stata. Ado file by Doug Miller (http://www.econ.ucdavis.edu/faculty/dlmiller/statafiles/), and 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.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

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 SEs 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 SEs, 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) (see also Arellano 1987). The formula for the cluster-robust variance-covariance matrix in ordinary least squares (OLS) regression is (Cameron and Miller 2015, 8–9):

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

where $\hat{\beta}$ are OLS estimates, G the number of clusters, X the $N \times m$ matrix of independent variables (for N observations and m variables), and $\hat{u}_g$ the vector of residuals $y_g - X_g\hat{\beta}$ in cluster g. Theorem 2 in Liang and Zeger (1986, 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 generalized linear model (GLM) models.

Following the introduction of CRSE 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 SEs and CRSEs can indicate harmful misspecification problems that are not addressed by adjusting the variance-covariance matrix (Hardin and Hilbe 2003, 33–34). In these cases, CRSEs might be useful 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, 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.[3]

Unfortunately, evidence reveals a major problem with using CRSEs in data sets 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 SEs, the CRSEs are biased downward (Mancl and DeRouen 2001; Donald and Lang 2007; Cameron, Gelbach and Miller 2008; 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).

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

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

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 $G \rightarrow \infty$ (Cameron and Miller 2015, 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 SEs based on G many items of information. 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 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* (APSR), *American Journal of Political Science* (AJPS), *Journal of Politics* (JOP), and *International Studies Quarterly* (ISQ) (hereafter known by their initials). We examined the use of CRSEs 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 and 27 percent 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 percent of all articles and 27.94 percent of articles that used clustering). AJPS also had a substantial proportion of papers with fewer than 20 clusters, representing 13.24 percent of all articles that used CRSEs. APSR and JOP fare somewhat better, with between 3 and 5 percent of all articles using clustering in data sets with fewer than 40 clusters (between 14 and 19 percent of all articles that used CRSEs). ISQ showed the lowest rate of articles with fewer than 40 clusters at 1.57 percent (6.90 percent of CRSE articles), but this journal showed the highest rate of models with an unknown number of clusters at 10.63 percent (46.55 percent of articles using CRSEs).

WHAT ALTERNATIVES TO CRSES ARE AVAILABLE?

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 as follows:

1. PCBSTs, variants of which are studied by Bertrand, Duflo and Mullainathan (2004), Cameron, Gelbach and Miller (2008), and Harden (2011).
2. WCBSTs, as proposed by Cameron, Gelbach and Miller (2008).

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

TABLE 1 *Use of Cluster-Robust Standard Errors (CRSE) in General Interest Journals in Political Science (2010–2014)*

	No. of articles	% of CRSEs	% of total	No. of articles	% of CRSEs	% of total
	AJPS (July 2010 to April 2014)			APSR (May 2010 to February 2014)		
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 to July 2014)			ISQ (June 2010 to March 2014)		
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

AJPS = *American Journal of Political Science*; APSR = *American Political Science Review*; JOP = *Journal of Politics*; ISQ = *International Studies Quarterly.*

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

Our primary contribution is the creation of a streamlined statistical package for implementing these alternative uncertainty estimators.[4] 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 Comprehensive R Archive Network repository for R (under library name clusterSEs) and the Statistical Software Components archive repository for Stata (under the names clustse and clusterbs).[5] 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 sets with a small number of clusters may differ,[6] we conduct a Monte Carlo analysis to evaluate their performance in relation to CRSEs and the usual maximum likelihood SEs (which we refer to hereafter as "vanilla" SEs). We also compare these techniques to a standard RE model (Wooldridge 2002, 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).

PCBSTs

The PCBST 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

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

[5] 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.

[6] See the Online Appendix for a more thorough analytic explanation of each approach's performance in data sets with few clusters.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

$\hat{\beta}_k$, $k = 1 \ldots K$. 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 SEs can be estimated by the SEs of the replicates; alternatively, a 95 percent 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 percent 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 recommends that "state politics researchers use [the pairs cluster bootstrap] to conduct statistical inference with clustered data" (2011, 224).[7] Bertrand, Duflo and Mullainathan (2004) study and apply PCBSTs (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 SEs rather than CRSEs for the bootstrap replicates; Cameron, Gelbach and Miller (2008) shows that CRSE replicates perform better with a small number of clusters.

WCBSTs

WCBSTs 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 $\hat{\epsilon}$ from the original linear model $y = X\hat{\beta} + \hat{\epsilon}$, 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).[8] 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, Gelbach and Miller 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)" (2008, 425). Note that this procedure depends on estimates of the residuals $\hat{\epsilon}$, and is therefore unsuited for GLM models with non-standard residuals (e.g., the probit).[9]

CATs

CATs 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

[7] 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, 227–229).

[8] See the discussion of WCBSTs in the Online Appendix for more information.

[9] See Hu and Kalbeisch (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.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

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, 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 $g \in \{1, \ldots, G\}$):

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

with $N = \sum_{g=1}^{G} N_g$, $\hat{\beta}_{\mathbf{G}}^T = \left[\hat{\beta}_1 \ \ldots \ \hat{\beta}_G\right]$ and $\sum_{\mathbf{G}} = \text{diag}(\sigma_1, ..., \sigma_G)$; the block-diagonal nature of $\sum_{\mathbf{G}}$ flows from the assumption that observations in clusters i and j are independent whenever $i \neq j$.[10] 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 $\overline{\beta}_{\mathbf{G}} = (1/G)\sum_{g=1}^{G} \hat{\beta}_g$ against the null of $\beta = 0$ with $G - 1$ df 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 percent level. A small panel data simulation study by Ibragimov and Müller (2010, 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 SEs based on Arellano (1987); however, this study examines only ten panels with 50 observations each ($G = 10$, $N_g = 50$) and does not examine the performance of pairs cluster or wild cluster bootstrapped SEs. 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 $\hat{t}_{\mathbf{G}} = \overline{\beta}_{\mathbf{G}} / \hat{s}_{\mathbf{G}}$, where $\overline{\beta}_{\mathbf{G}}$ is the mean of the cluster level coefficients and $\hat{s}_{\mathbf{G}}$ their estimated SE. 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 percent confidence intervals as $\overline{\beta}_{\mathbf{G}} \pm \left(t_{\alpha, G-1}\right)\left(\hat{s}_{\mathbf{G}}\right)$; it also allows us to calculate SEs on interaction terms as prescribed by Brambor, Clark and Golder (2006). Note that $\overline{\beta}_{\mathbf{G}}$ and $\hat{\beta}$ will often not be equivalent;[11] therefore 95 percent confidence intervals formed with this procedure will often not be centered on $\hat{\beta}$.

[10] See Equation 4 in Ibragimov and Müller (2010, 456).

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

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

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 little 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.

ASSESSING TECHNIQUES FOR THE ANALYSIS OF CLUSTERED DATA

We assess the performance of statistical significance tests using ordinary (vanilla) SEs, CRSEs,[12] CATs, and PCBSTs in continuous and binary dependent variable scenarios; for continuous dependent variables, we also assess the performance of WCBSTs and linear RE models.[13] 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:

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

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, $z \sim \Phi(\mu = 0,\ \sigma = 1)$. $\mu_g \sim U[1,5]$

[12] 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.

[13] 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.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

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 SEs would be inappropriate.

This data structure is a very close match to the structure of a typical RE model (Wooldridge 2002, 257–265). This gives us an opportunity to compare the performance of linear link GLM with cluster-corrected SEs to RE 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 RE are estimated using the glm function in R, while RE 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 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, $t = 1 \ldots T$; 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.[14] 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 SEs for clustering. There are adjustments to the RE 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 is not— 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:

$$\Pr(y_{gi} = 1) = \frac{1}{1 + \sum_{j \in \{2,3\}} \exp(\beta_{xj} x_{gi} + \beta_{zj} z_{gi} + \beta_{wj} w_{gi} + \gamma_{gj})}$$

$$\Pr(y_{gi} = k) = \frac{\exp(\beta_{xk} x_{gi} + \beta_{zk} z_{gi} + \beta_{wk} w_{gi} + \gamma_{gk})}{1 + \sum_{j \in \{2,3\}} \exp(\beta_{xj} x_{gi} + \beta_{zj} z_{gi} + \beta_{wj} w_{gi} + \gamma_{gj})} \text{ for } k > 1,$$

where $J \in \{1,2,3\}$ and the base category is 1. For other categories $j \in \{2,3\}$, $\gamma_{g,j} \sim \Phi(\mu = 0, \sigma = 1)$. When x and z are related to the dependent variable, $\beta_{x2} = \beta_{z2} = 1$ and $\beta_{w3} = 1$, and $\beta_{x3} = \beta_{z3} = \beta_{w3} = 0$. $\mu_g \sim \mathrm{U}[-2, 2]$ 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 $\mu_g \sim U[1,5]$.[15] We estimate these models using the mlogit package in R (Croissant 2015).

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

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

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

The data are structured so that there are varying numbers of clusters G; we examine cases with $G \in \{3, 6, 15, 21, 30, 60, 75, 90, 120\}$. 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).[16]

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 percent value of the test when using the CRSE replicates, except for very small numbers of clusters (where the rejection rate is lower than 5 percent). 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 percent 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 percent 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 percent power rates more quickly than any of the clustering alternatives.

The best performance is achieved by RE models, which maintain a 5 percent 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 three or six clusters the RE model's true positive performance is substantially better, ≈ 78 percent even with only three clusters.

[16] 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.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

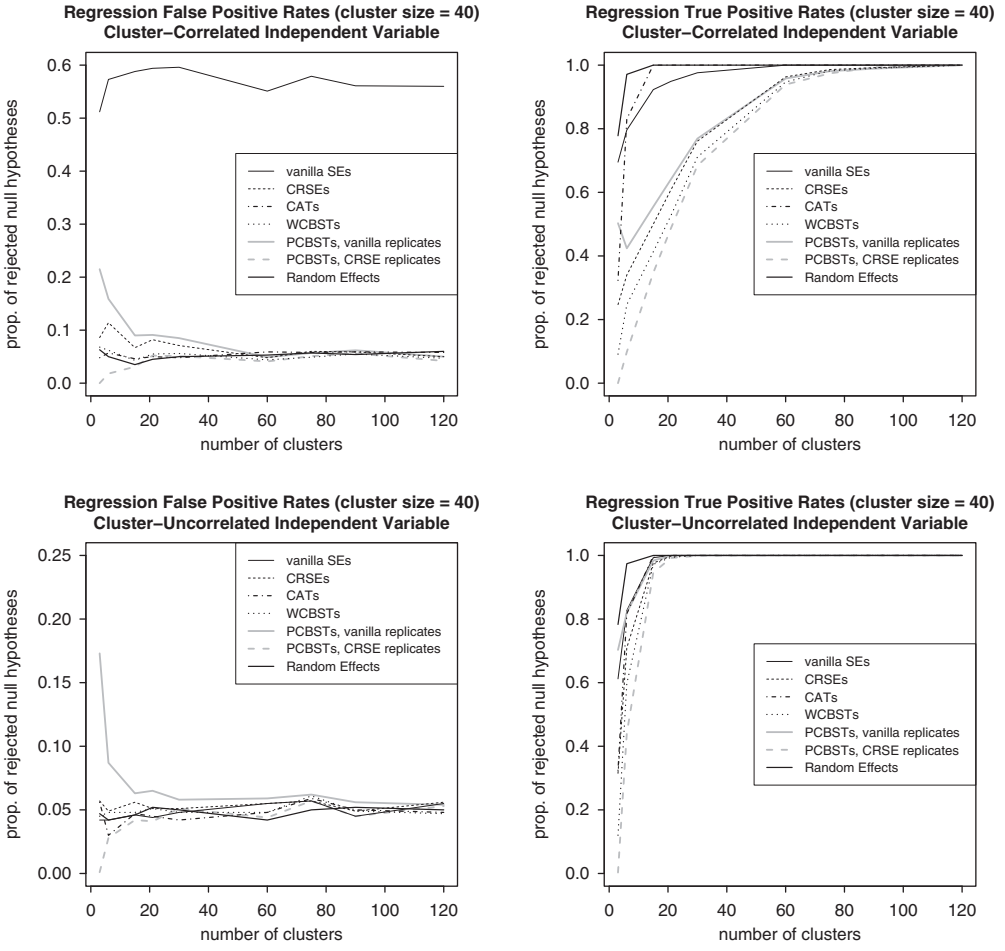


Fig. 1. Size and power assessment for linear dependent variables

Note: 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 generalized linear model (estimated using glm) with a different method of calculating statistical significance, as indicated in the legend; random effects models are correctly specified linear random effects 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.

SEs = standard error; CRSE = cluster-robust standard error; CATs = cluster-adjusted t-statistics; WCBSTs = wild cluster bootstrapped t-statistics; PCBSTs = pairs cluster bootstrapped t-statistics.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

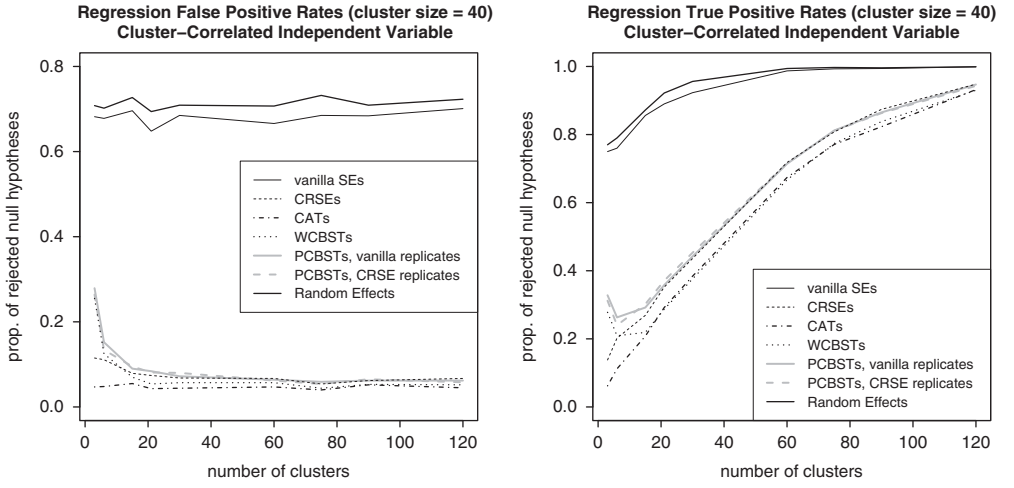


Fig. 2. Size and power assessment for linear dependent variables with fixed effects and serial dependence
Note: The graph on the left shows the proportion of rejected null hypotheses ($\beta = 0$) out of 1000 simulations for parameters whose true values is $\beta_x = 0$ in the linear model $y_i = \beta_x x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i$ 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 random effects 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).
SEs = standard error; CRSE = cluster-robust standard error; CATs = cluster-adjusted t-statistics; WCBSTs = wild cluster bootstrapped t-statistics; PCBSTs = pairs cluster bootstrapped t-statistics.

Simulation Results for Linear Models With μ_g/γ_g Correlation and Serial Dependence

The performance of the RE estimator (shown in Figure 2) is substantially worsened when the data-generating process includes (a) correlation between the average value of x and the group effect γ_g and (b) serial dependence in x and y. In this environment, the RE model produces a false positive for the cluster-correlated independent variable ≈ 70 percent 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 percent 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 substantially when the signal is made stronger (relative to noise[17]) so that $\beta_x = \beta_z = 0.5$. In this alternative environment, CATs achieve 78 percent detection of true positives with 21 clusters.

[17] 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.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

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 data sets with ≤ 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, for example, 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.[18] We provide an analytic argument in the Online Appendix that dropping these clusters will not distort inferences in many cases, particularly if the number of dropped clusters is relatively small. By dropping the clusters with missing and/or outlying estimates, we are able to estimate CATs on every data set with more than six clusters in our simulation[19] and our false positive rate remains very close to the nominal value of 5 percent 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.

Political Quid Pro Quo Agreements (Grosser, Reuben and Tymula 2013)

In a recent AJPS 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

[18] See the Online Appendix for more information about how our software handles missing and outlying estimates.

[19] 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.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

non-voting "candidate" subjects are also assigned to the group. In the experimental condition that we study in this replication, groups are fixed[20] 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 $\tau^\star(e-\overline{e})$ if rich, and gain $\tau^\star(\overline{e}-e)$ if poor; $\overline{e}$ 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' 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 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 page 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 $\left(D^{+}_{ij}=\max\left(\tau_{i(t-1)}-\tau_{j(t-1)},0\right)\right)$ and one for a negative difference $\left(D^{-}_{ij}=\max\left(\tau_{j(t-1)}-\tau_{i(t-1)},0\right)\right)$; these variables are designed to study how candidates react to one another's' proposals.

The regression analysis of Grosser, Reuben and Tymula (2013) uses CRSE 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)" (Grosser, Reuben, and Tymula, 2013; 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 (590–591).[21]

It is possible that uncertainty in the relationship between changes in transfers and changes in proposals might be understated by CRSE on data with only 17 groups. Therefore, we reproduce

[20] 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, for example, figure 2 on page 590 of their paper for details.

[21] See the Online Appendix for more details.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

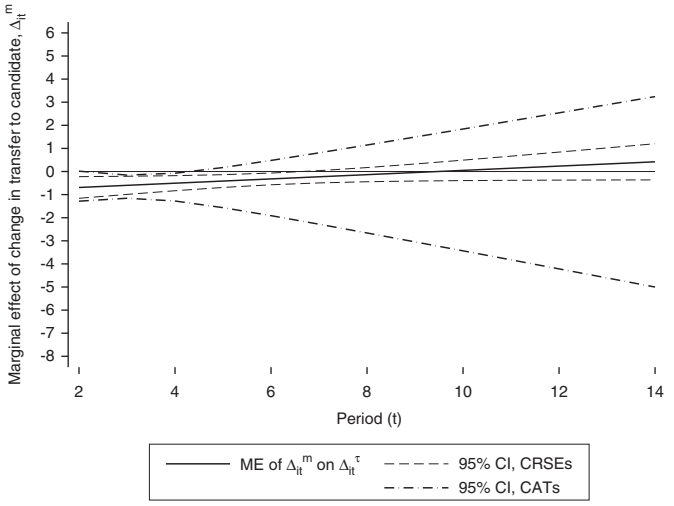


Fig. 3. Marginal effects (ME) plot, effect of changed in received transfer (Δ_{it}^{m}) on changes in proposed tax rate (Δ_{it}^{τ}) by period

Note: Dependent variable: change in candidate's proposed tax rate (Δ_{it}^{τ}). ME are calculated using the results in Online Appendix Table 2.

CI = confidence interval; CRSE = cluster-robust standard error; CATs = cluster-adjusted *t*-statistics.

the regression analysis of Grosser, Reuben and Tymula (2013) using their original CRSEs as well as PCBSTs and CATs. A marginal effect plot (see Brambor, Clark and Golder 2006) (Figure 3) for the marginal effect of changes in transfers on changes in tax policy in each period using these results[22] 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, 290) also separately examine the behavior of "high tax" groups, which have high winning tax policies (between 90 and 100 percent), and "low tax" groups, which have much lower winning tax policies (between 33.3 and 83.7 percent). 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.[23] 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.[24] 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.

[22] A coefficient table is shown in the Online Appendix Table 2.

[23] Coefficient tables are shown in the Online Appendix Tables 3 and 4.

[24] 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.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

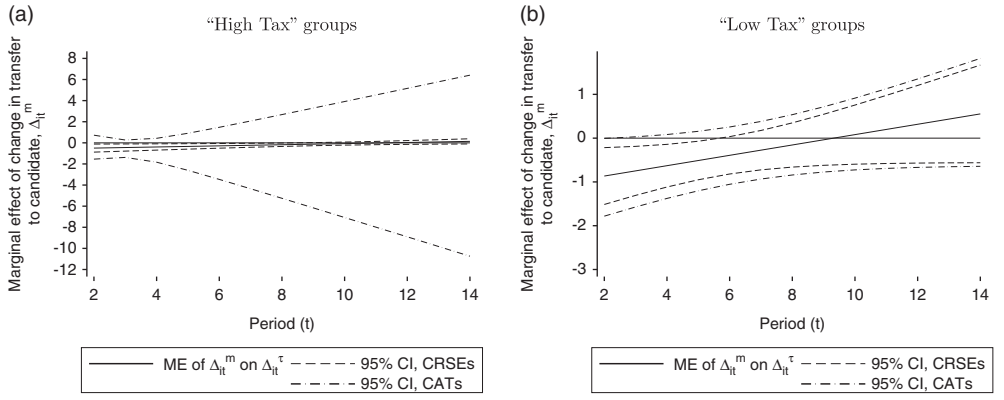


Fig. 4. Marginal effects (ME) plot, effect of changed in received transfer (Δ_{it}^{m}) on changes in proposed tax rate (Δ_{it}^{τ}) by period in high (a) and low tax groups (b)

Note: Dependent variable: change in candidate's proposed tax rate (Δ_{it}^{τ}). ME are calculated using the results in Online Appendix Table 3 for panel 4 and Online Appendix Table 4 for panel 4.
CI = confidence interval; CRSE = cluster-robust standard error; CATs = cluster-adjusted *t*-statistics.

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 is not robust to alternative methods of cluster adjustment for the SEs. 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 percent confidence intervals are 20 percent wider when using PCBSTs and 46 percent wider when using CATs compared with the original results using CRSEs.

CONCLUSION

Political scientists often use CRSE 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 SEs 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 CATs (based on the work of Ibragimov and Müller 2010) are the best choice among the options we examine for correcting SEs 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,

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

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), PCBSTs with CRSE replicates and WCBST generally provide better performance than vanilla SEs or CRSEs. Finally, in our simulations an accurate RE 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 SEs. Political scientists would be wary of deriving strong conclusions from a data set with $N = 15$ or 20 observations and five 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, Theodore W. 2003. *An Introduction to Multivariate Statistical Analysis* 3rd ed. New York, NY: Wiley.

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

Arellano, Manuel. 1987. 'Computing Robust Standard Errors for Within-Groups Estimators'. *Oxford Bulletin of Economics and Statistics* 49(4):431–34.

Bafumi, Joseph, and Andrew Gelman. 2006. 'Fitting Multilevel Models When Predictors and Group Effects Correlate'. Available at http://goo.gl/usvQsn, accessed 21 December 2017.

Bakirov, Nail K., and Gabor J. Szekely. 2006. 'Student's t-Test for Gaussian Scale Mixtures'. *Journal of Mathematical Sciences* 139(3):6497–505.

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. Available at http://CRAN.R-project.org/package=lme4, accessed 21 December 2017.

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):634–47.

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):274–78.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. 'How Much Should We Trust Differences-In-Differences Estimates?'. *The Quarterly Journal of Economics* 119(1):249–75.

Brambor, Thomas, William Roberts Clark, and Matthew Golder. 2006. 'Understanding Interaction Models: Improving Empirical Analyses'. *Political Analysis* 14(1):63–82.

Cameron, A. Colin, and Douglas L. Miller. 2015. 'A Practitioner's Guide to Cluster-Robust Inference'. *Journal of Human Resources* 50(2):317–72.

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):414–27.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge, UK: 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). Available at https://goo.gl/TUEQee, accessed 29 January 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.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

Croissant, Yves. 2015. 'Package "mlogit".' CRAN. Available at http://cran.r-project.org/web/packages/mlogit/mlogit.pdf, accessed 21 December 2017.

Croissant, Yves, and Giovanni Millo. 2008. 'Panel Data Econometrics in R: The plm Package'. *Journal of Statistical Software* 27(2):1–43.

Donald, Stephen G., and Kevin Lang. 2007. 'Inference With Difference-in-Differences and Other Panel Data'. *The Review of Economics and Statistics* 89(2):221–33.

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, Chris A., and Alan H. Welsh. 2007. 'Bootstrapping Clustered Data'. *Journal of the Royal Statistical Society: Series B* 69(3):369–90.

Gaibulloev, Khusrav, Todd Sandler, and Donggyu Sul. 2014. 'Dynamic Panel Analysis Under Cross-Sectional Dependence'. *Political Analysis* 22:258–73.

Green, Donald P., and Lynn Vavreck. 2008. 'Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches'. *Political Analysis* 16(2):138–52.

Grosser, Jens, Ernesto Reuben, and Agnieszka Tymula. 2013. 'Political Quid Pro Quo Agreements: An Experimental Study'. *American Journal of Political Science* 57:582–97.

Hainmueller, Jens, Michael Hiscox, and Sandra Sequeira. 2015. 'Consumer Demand for the Fair Trade Label: Evidence from a Multistore Field Experiment'. *Review of Economics and Statistics* 97(2):242–56.

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):223–46.

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

Horowitz, Joel L. 1997. 'Bootstrap Methods in Econometrics: Theory and Numerical Performance'. In David M. Kreps and Kenneth F. Wallis (eds), *Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress*, 189–222. Cambridge, UK: Cambridge University Press.

Hu, Feifang, and John D. Kalbeisch. 2000. 'The Estimating Function Bootstrap'. *Canadian Journal of Statistics* 28(3):449–81.

Ibragimov, Rustam, and Ulrich K. Müller. 2010. 't-Statistic Based Correlation and Heterogeneity Robust Inference'. *Journal of Business & Economic Statistics* 28(4):453–68.

Imbens, Guido W., and Michal Kolesar. 2012. 'Robust Standard Errors in Small Samples: Some Practical Advice' 98(4):701–12.

Judge, George G., R. Carter Hill, William E. Griffths, Helmut Lutkepohl, and Tsoung-Chao Lee. 1988. *Introduction to the Theory and Practice of Econometrics*. New York, NY: 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* 23:159–79.

Klar, Neil, and Allan Donner. 2001. 'Current and Future Challenges in the Design and Analysis of Cluster Randomization Trials'. *Statistics in Medicine* 20(24):3729–740.

Lacina, Bethany. 2014. 'How Governments Shape the Risk of Civil Violence: India's Federal Reorganization, 1950–56'. *American Journal of Political Science* 58(3):720–38.

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):1696–708.

Liu, Regina Y., and Kesar Singh. 1987. 'On a Partial Correction by the Bootstrap'. *The Annals of Statistics* 15(4):1713–718.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press

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):233–54.

Mancl, Lloyd A., and Timothy A. DeRouen. 2001. 'A Covariance Estimator for GEE With Improved Small-Sample Properties'. *Biometrics* 57(1):126–34.

Moulton, Brent R. 1986. 'Random Group Effects and the Precision of Regression Estimates'. *Journal of Econometrics* 32(3):385–97.

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* 72(2):334–38.

Nickell, Stephen. 1981. 'Biases in Dynamic Models With Fixed Effects'. *Econometrica* 49:1417–426.

Rogers, William. 1993. 'Regression Standard Errors in Clustered Samples'. *Stata Technical Bulletin* 13:19–23.

van der Vaart, Aad W. 1998. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

White, Halbert. 1980. 'A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity'. *Econometrica* 48(4):817–38.

Williams, Rick L. 2000. 'A Note on Robust Variance Estimation for Cluster-Correlated Data'. *Biometrics* 56(2):645–46.

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

Wu, C. F. Jeff. 1986. 'Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis'. *The Annals of Statistics* 14(4):1261–295.

https://doi.org/10.1017/psrm.2017.42 Published online by Cambridge University Press