Appendix A: Detailed descriptions of alternative cluster-robust uncertainty calculation procedures and analytic description of performance in data with few clusters

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). We also provide an analytic demonstration of how we expect each of these techniques, along with CRSEs, to perform in data with few clusters.

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, $|t_k|$.[27]

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.[28]

   (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.

[27] 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.

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

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

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.[29]

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 $\hat{\beta}_g$.

[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. 27) to make more efficient use of a smaller number of draws.

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

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 $\hat{\beta}$.

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 regression using a constant only:

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

where g indexes clusters $g = 1 \ldots G$ and i indexes individual observations within a cluster $i = 1 \ldots N_g$ 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.[30] 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:

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

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).

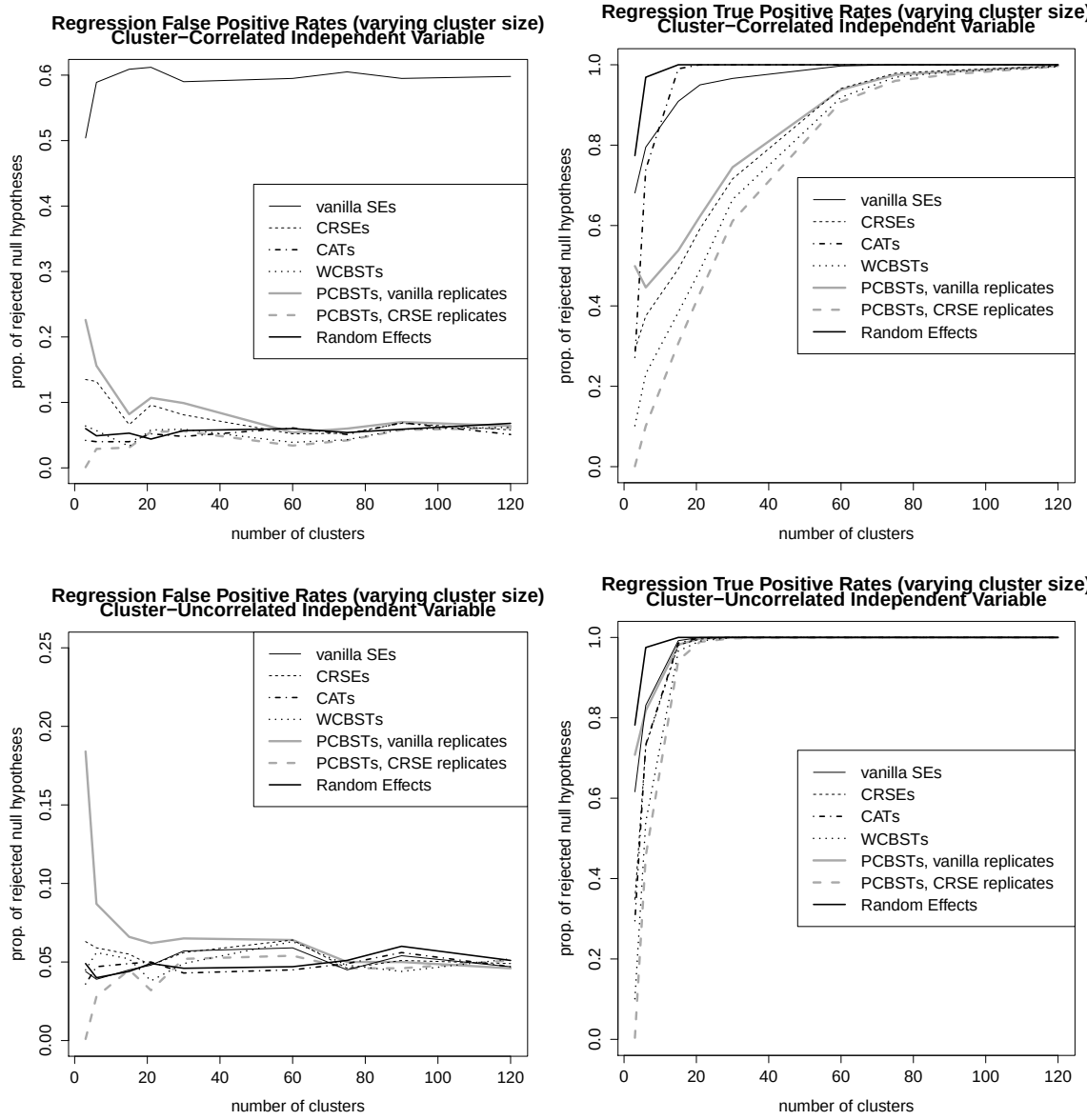
CATs are distinct from both CRSEs and cluster bootstrap procedures when the number of clusters is small because the small-sample characteristics of the underlying test statistic are *known*; this knowledge is thanks to analytical work from Bakirov and Szekely (2006) that is utilized by Ibragimov and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

[30] 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.

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

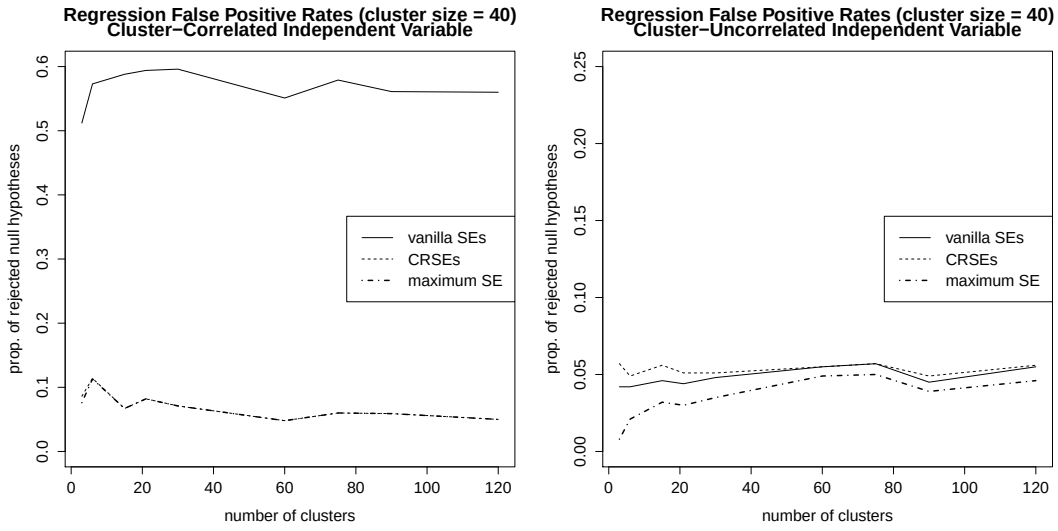


The graphs on the left show the proportion of rejected null hypotheses out of 1000 simulations for parameters whose true values are $\beta_x = \beta_z = 0$ in the linear model $y_i = \beta x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i$ with cluster dependency; this is a measure of the false 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



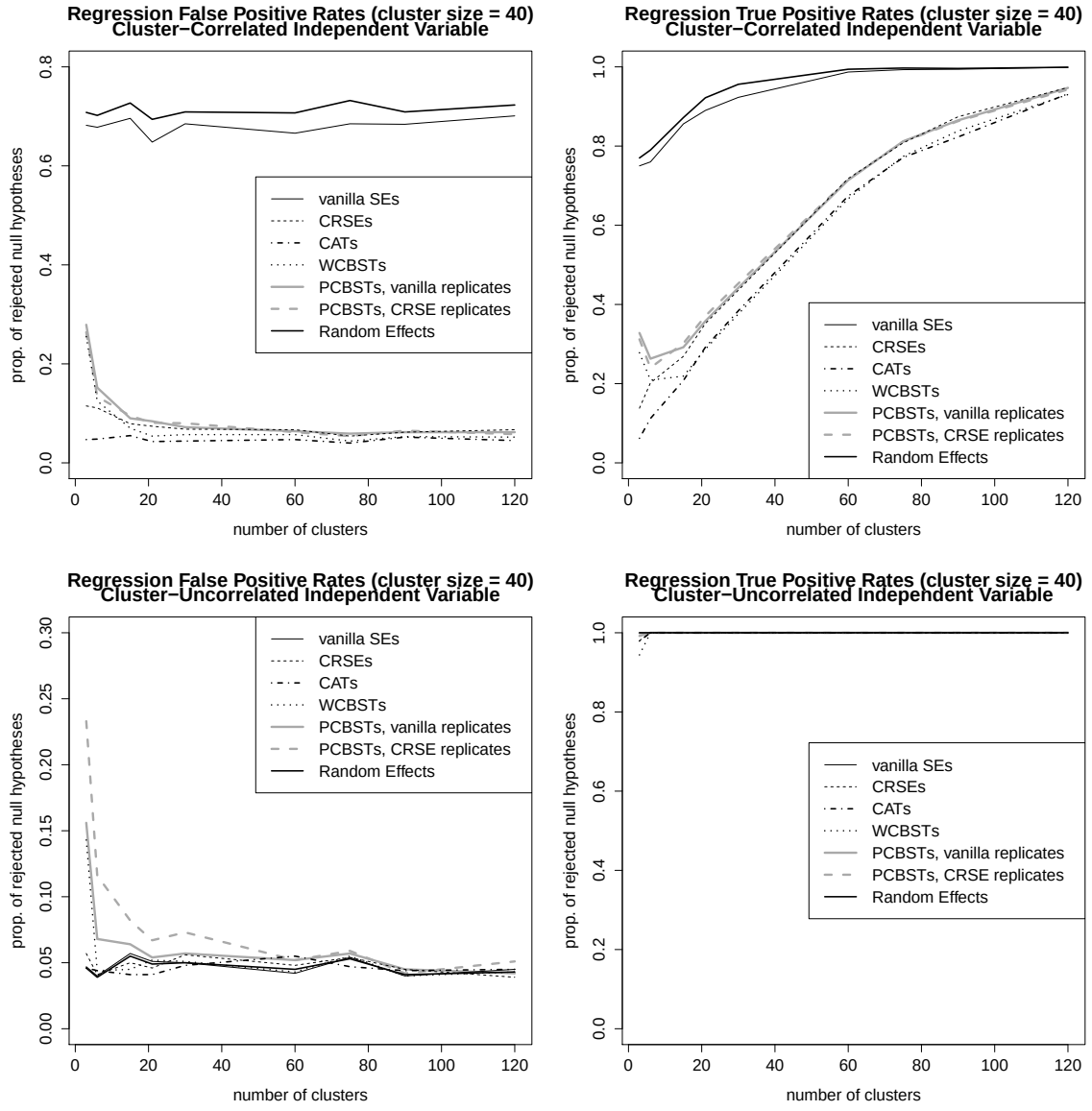
These graphs show the proportion of rejected null hypotheses out of 1000 simulations for parameters whose true values are $\beta_x = \beta_z = 0$ in the linear model $y_i = \beta_x x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i$ with cluster dependency; this is a measure of the false positive 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.

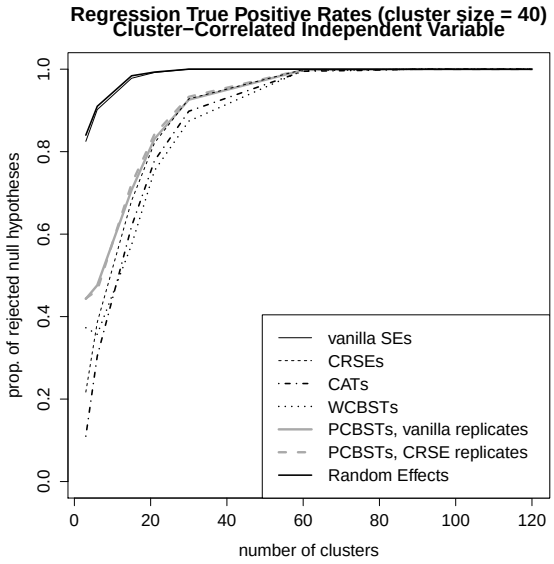
Figure 8 presents the results of simulations identical to those from Figure 7 with β_x and β_z increased to 0.50 from their original setting of 0.25. The detection of true positives for fixed effects models with CATs, WCBSTs, PCBSTs, and CRSEs are all improved relative to the original simulation.

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

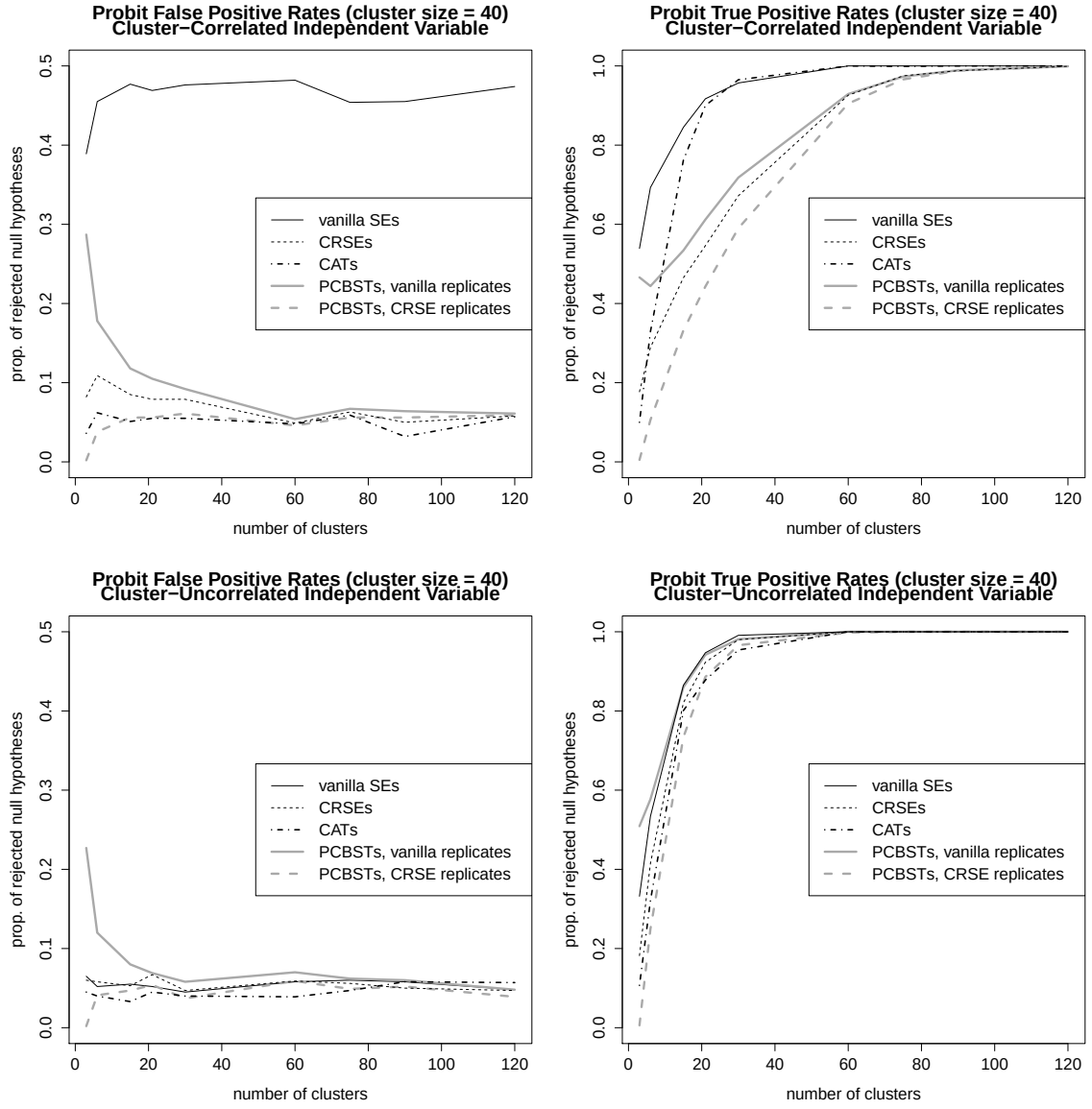


The graph shows the proportion of rejected null hypotheses ($\beta = 0$) out of 1000 simulations (with true values $\beta_x = \beta_z = 0.5$) for the x parameter whose mean is correlated with the group-level effect in the linear model $y_i = \beta_x x_i + \beta_z z_i + \beta_w w_i + \gamma_g + \varepsilon_i$ 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

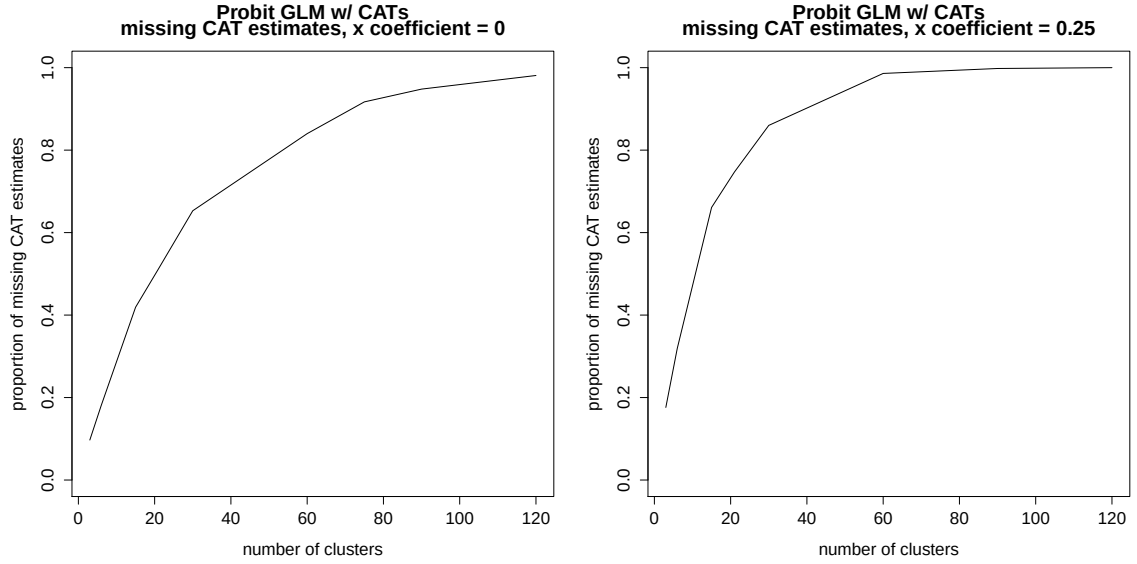
We note an anomaly: CATs often cannot be estimated in the probit model with a large number of clusters. For example, CATs cannot be calculated whenever any cluster has no variation on the dependent variable because the probit model is unidentified in this cluster and will thus fail to converge to an 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.[31]

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

[31] 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

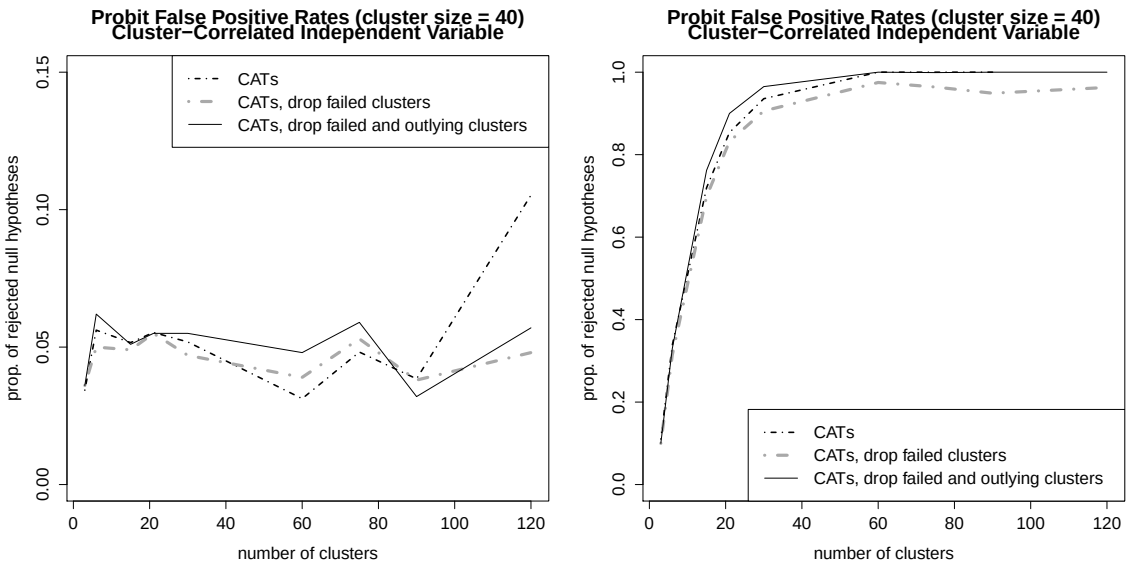
We note another anomaly: even if the model is technically identified in every cluster, extreme outlier β estimates can be produced in cases where perfect or near-perfect separation of the outcomes is predicted by one or more independent variables. The result is that the cluster-level distribution of 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.

How our software handles missing or outlying estimates

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.

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),[32] 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 $\sqrt{G_i}\left(D_i\left(\hat{\beta}_{\mathbf{G}} - \beta\right)\right) = \sqrt{G_i}\left(\hat{\beta}_{\mathbf{G_i}} - \beta\right) \overset{asym}{\sim} \Phi(\mathbf{0}, \mathbf{\Sigma_{G_i}})$. □

[32] 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:

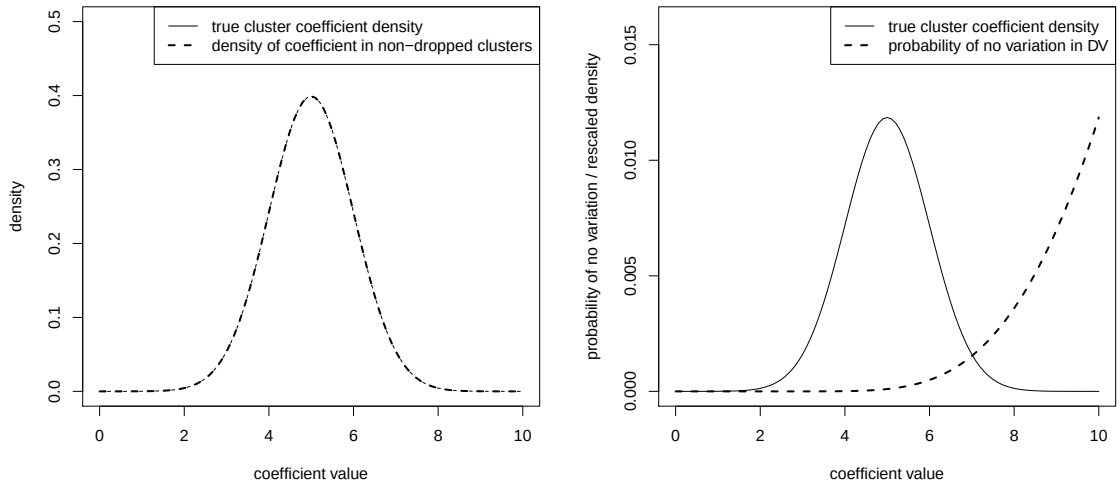
2. the degree of heterogeneity in the probability of dropping the cluster for different values of $\hat{\beta}_g$ is small, or $\pi(\hat{\beta}_g^a, X) - \pi(\hat{\beta}_g^b, X) \approx 0$ for all values of a and b so that $\pi(\hat{\beta}_g, X) \approx \int \pi(\hat{\beta}_g, X) f(\hat{\beta}_g) d\hat{\beta}_g$.

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

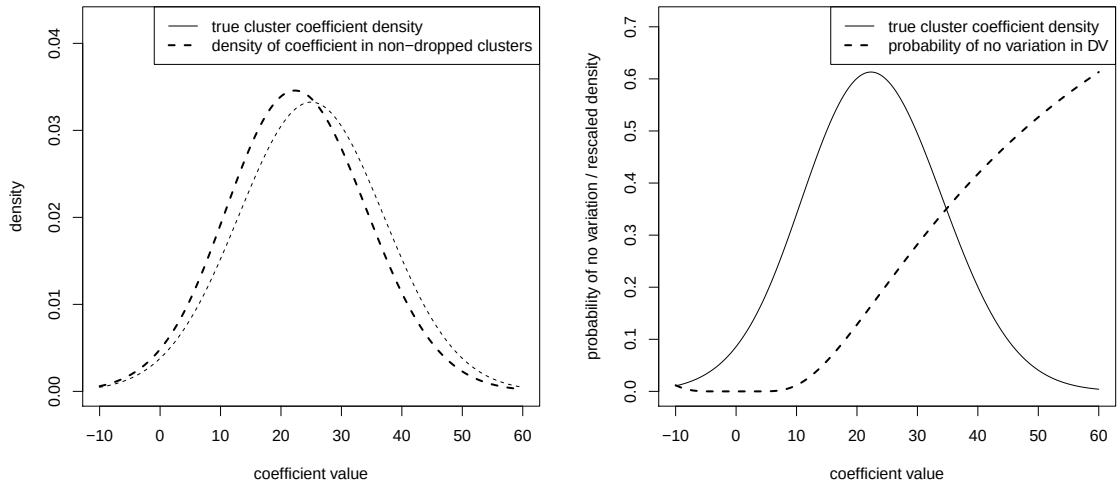


These graphs depict the result of calculating the probability of no variation in the dependent variable y under a model $y = \Phi\left(X\hat{\beta}\right)$ with cluster-level estimates $\hat{\beta}_g$ distributed according to $f(\hat{\beta}_g) = \phi(\mu = 5, \sigma = 1)$. The dataset X is a sequence 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



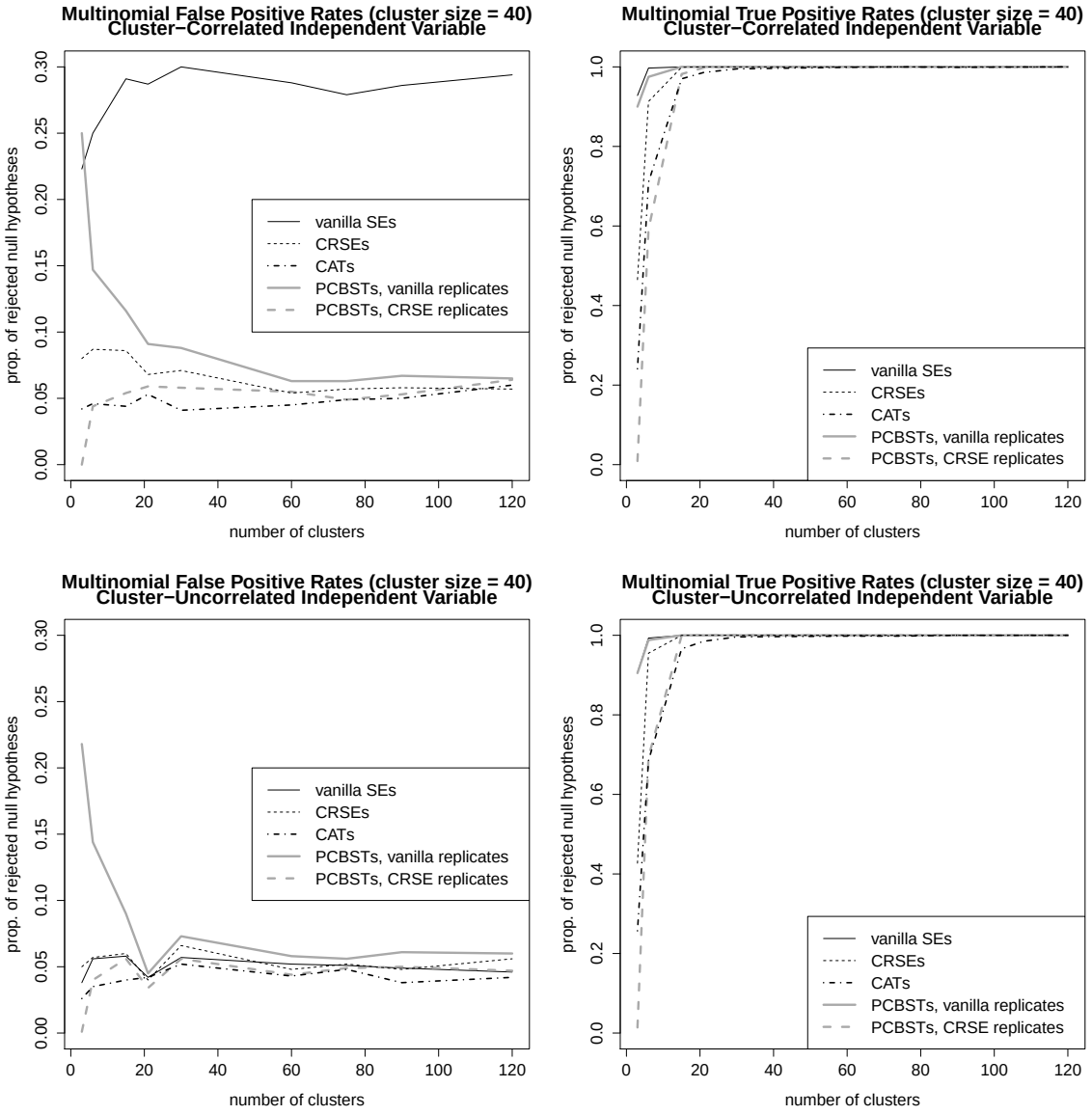
These graphs depict the result of calculating the probability of no variation in the dependent variable y under a model $y = \Phi\left(X\hat{\beta}\right)$ with cluster-level estimates $\hat{\beta}_g$ distributed according to $f(\hat{\beta}_g) = \Phi(\mu = 25, \sigma = 12)$. The dataset X is a sequence 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

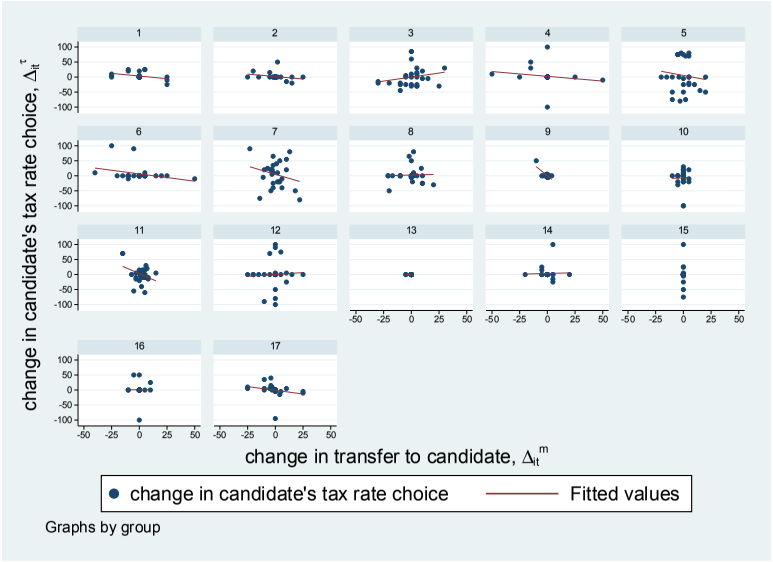
Figure 14: Size and power assessment for multinomial dependent variables



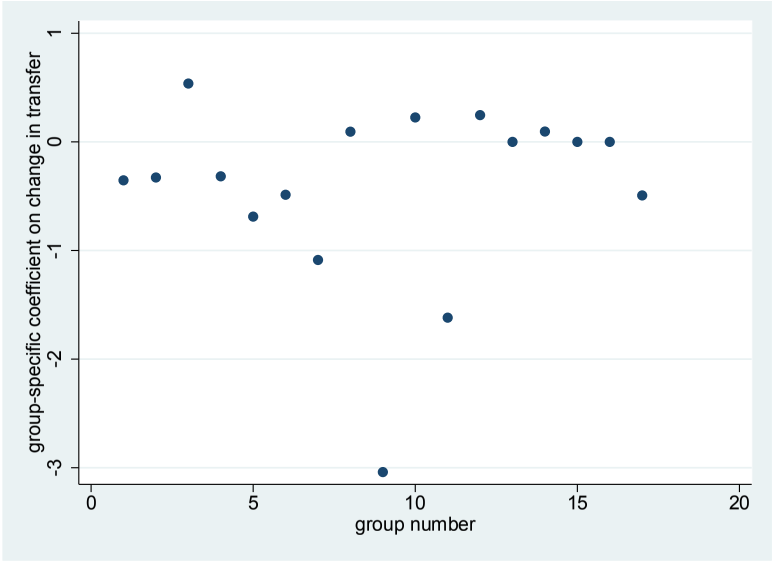
The graphs on the left show the proportion of rejected null hypotheses ($\beta = 0$) out of 1000 simulations for parameters whose true values are $\beta_{x2} = \beta_{z2} = 0$ in the multinomial logit model with cluster dependency; this is a measure of the false positive rate. Each model is a correctly 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)

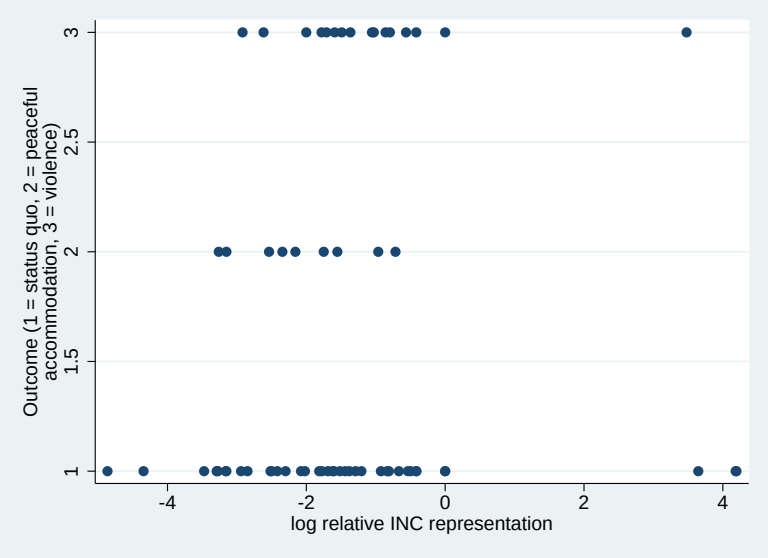
In the 2014 volume of the *American Journal of Political Science*, Lacina (2014) uses data from India to argue that "representation in the ruling party conditioned the likelihood of a violent statehood movement" (p. 720). Her primary empirical evidence comes from a data set consisting of 63 "language 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.

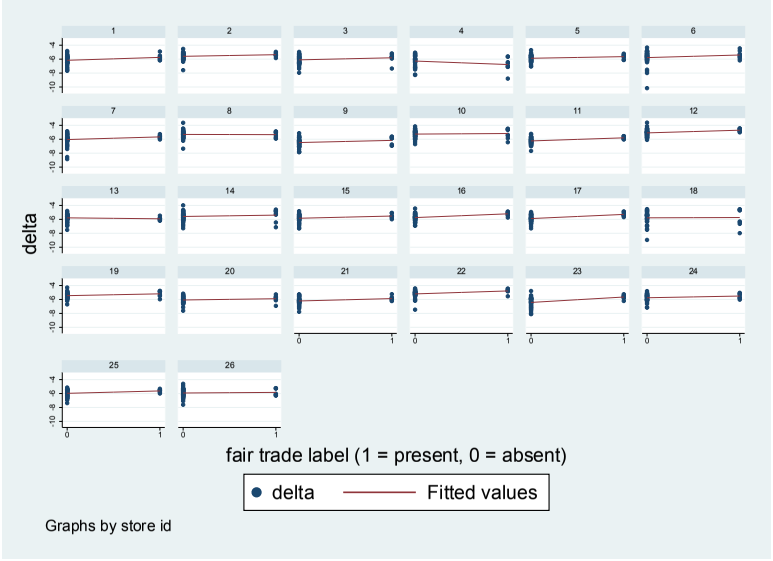
Visual assessment of the bivariate relationship between market share and fair trade labeling (in Figure 17a) seems to suggest that there is a small, positive relationship between fair trade labeling and market share in 23 out of 26 stores. This is confirmed in a plot of store-specific regression coefficients 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.

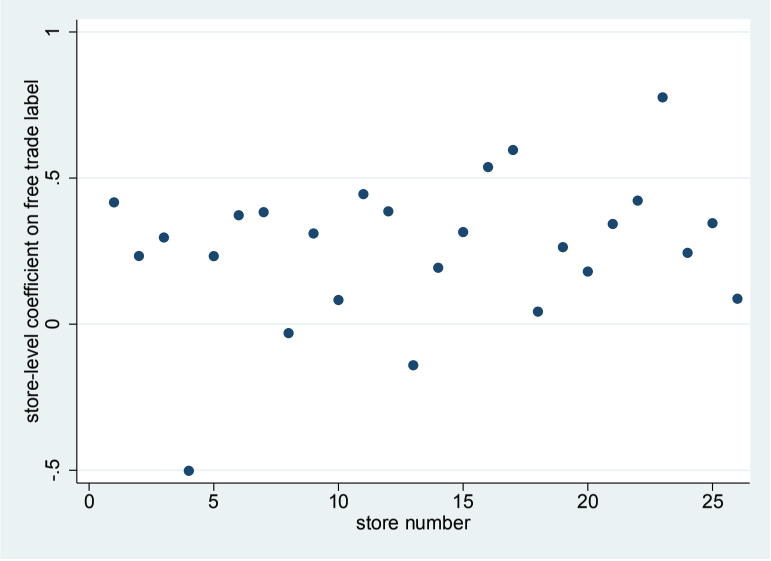
On the basis of this evidence, we conclude that the data collected in Hainmueller, Hiscox and Sequeira (2015) are generally supportive of their claim that fair trade labeling increases market share. There is somewhat greater uncertainty associated with the substantive

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.