A Modified Randomization Test for the Level of Clustering

Yong Cai
Department of Economics, Northwestern University, Evanston, IL

ABSTRACT
Suppose a researcher observes individuals within a county within a state. Given concerns about correlation across individuals, it is common to group observations into clusters and conduct inference treating observations across clusters as independent. However, a researcher that has chosen to cluster at the county level may be unsure of their decision, given knowledge that observations are independent across states. This article proposes a modified randomization test as a robustness check for the chosen level of clustering in a linear regression setting. Existing tests require either the number of states or number of counties to be large. Our method is designed for settings with few states and few counties. While the method is conservative, it has competitive power in settings that may be relevant to empirical work.

1. Introduction
Consider the following regression:

\[ Y_i = X_i'\beta + U_i, \quad E[X_iU_i] = 0, \]

where a researcher wants to perform inference on \( \beta \). If the researcher is concerned about correlation between \( U_i \) and \( U_j \), it is frequently helpful to group observations into independent clusters. These independent clusters can then be used to construct cluster-robust covariance estimators (CCE) as in Liang and Zeger (1986), or for approximate randomization tests as in Canay, Romano, and Shaikh (2017), and Cai et al. (2021).

However, these procedures require the assignment of units to clusters be known ex ante. In practice, researchers often have some freedom in choosing the level at which to cluster their standard errors. For example, those working with the American Community Survey (ACS) can cluster their data either at the individual, county or state level. Alternatively, those working with firm data from COMPSTAT have the option to cluster at either the 4-digit, 3-digit, or 2-digit Standard Industrial Classification (SIC) level, or even the firm level.

Clustering at the correct level is important for valid inference. A large body of simulation evidence shows that ignoring cluster dependence—in other words, clustering at too fine a level—leads to Type I errors that exceed the nominal error by as much as 10 times (Bertrand, Duflo, and Sendhil 2004; Cameron, Gelbach, and Miller 2008). On the other hand, clustering at excessively coarse levels can also lead to problems. For one, coarse clusters tend to be few in number. It is well-known that confidence intervals based on the cluster-robust standard errors tend to under-cover when the number of clusters is small (see Angrist and Pischke 2008 for instance), leading to poor size control. In the absence of under-coverage issues, unnecessarily coarse levels of clustering can also lead to tests with poor power since the researcher assumes less information than they actually have. Abadie et al. (2017) demonstrate via simulations, in a many-cluster setting, that CCEs based on coarse clusters can be too large. They also provide theoretical results in this vein, though they do so in the context of their “design-based” asymptotics that differ from those traditionally used to analyze clustered standard errors. Nonetheless, the problems with tests based on excessively coarse-clustering arise even with few clusters—the setting of interest for our article. We present a simple simulation to demonstrate these issues in Appendix B.

Given the above considerations, a researcher may choose to cluster at a fine level (e.g., individual or county) even when a coarse level of clustering (e.g., state), which is known to be valid, is also available. Nonetheless, they may be unsure if the fine level is appropriate. That is, whether observations across the fine clusters are approximately independent.

To help researchers assess the validity of their chosen clusters, we propose a modified randomization test that can be used as a robustness check for a given clustering specification. Our test requires large (fine) sub-clusters, but is justified under asymptotics that take the number of (coarse) clusters and (fine) sub-clusters as fixed. Inference is difficult in this setting because scores are not independent across sub-clusters even asymptotically, as we will explain in Section 2.2. Randomization tests, which typically require some type of asymptotic independence, thus, cannot be directly applied. We get around this problem by searching for worst-case values of the unobserved parameters to guard against over-rejection. We describe a simple method to search for this value, so that the computational complexity of the test is of the same order as the number of sub-clusters. This
is reasonable since our test is targeted toward applications with few sub-clusters. Our test has no power against negative correlation. However, ignoring negative correlation leads to variance estimators that are too large, and is thus less of an issue if the researcher is concerned about size control when performing inference on $\beta$.

To our knowledge, there are two other tests for the level of clustering. MacKinnon, Nielsen, and Webb (2020) proposes a test based on having large number of coarse clusters, relying on the wild bootstrap to improve finite sample performance. Meanwhile, Ibragimov and Müller (2016) proposes a test for the case when there are many sub-clusters. Our test, which takes the number of clusters and sub-clusters to be fixed, handles a more challenging situation, though this comes at the cost of being conservative, especially in settings with homogeneous clusters. However, as our simulations in Section 3 show, it has competitive power given heterogeneous clusters—a setting that could be relevant for empirical work. Indeed, our test detects correlation in the clusters chosen by Gneezy et al. (2019), demonstrating its potential usefulness in applied work (see Section 4). Finally, we note that the test of Ibragimov and Müller (2016) also has no power against negative correlation, although that limitation.

Abadie et al. (2017) takes a different approach to this issue. They argue for a “design-based” perspective on clustering, requiring researchers to determine ex ante the uncertainty that they face in either sampling or treatment assignment. For example, if the researcher believes that in their specific context, treatment assignment occurs at the sub-cluster level, then sub-clusters should be used for computing standard errors, regardless of whether or not residuals are correlated across the sub-clusters. While insightful, this approach requires researchers to answer an alternative question on which there is equally little theoretically guidance. We therefore develop our method under the “model-based” framework, in which the researcher has in mind some data-generating process that entails dependent clusters.

The remainder of this article is organized as follows. Section 2 describes our proposed test. Section 3 presents Monte Carlo simulations. Section 4 demonstrates an application to Gneezy et al. (2019). Section 5 concludes. Proofs are collected in Appendix A.

2. The Proposed Test

2.1. Model and Assumptions

In the following, we assume that the researcher has conducted inference on $\beta \in \mathbb{R}$, and seeks a robustness check for the level of clustering used for said inference. As will become clear in Section 2.3, using a scalar $\beta$ yields computational advantages, though the test can be feasibly computed for moderate dimensions of $\beta$. For this reason and for ease of exposition we limit our discussion to the scalar case.

Consider the linear regression:

$$Y_i = X_i \beta + W_i' \gamma + U_i, \quad E[X_i U_i] = 0, \quad E[W_i U_i] = 0,$$

where $\beta \in \mathbb{R}$ is the parameter of interest and $\gamma \in \mathbb{R}^d$ is a nuisance parameter. Suppose there are $r$ clusters, indexed by $k \in \mathcal{K}$. Within each cluster $k$, there are $q_k$ sub-clusters, indexed by $j \in \mathcal{J}_k$. Let $\mathcal{F} = \bigcup_{k \in \mathcal{K}} \mathcal{J}_k$ and $\mathcal{I} = \bigcup_{j \in \mathcal{J}} \mathcal{I}_j$. Further, let $n = \sum_{j \in \mathcal{J}} n_j$ and $q = \sum_{k \in \mathcal{K}} q_k = |\mathcal{F}|$. We also write $i \in \mathcal{I}_k$ when $i \in \mathcal{I}_j$ and $j \in \mathcal{J}_k$. In the following, we suppress dependence on $j$ and $k$ whenever this does not cause confusion.

**Assumption 1.** Suppose that for every cluster $j$, there exists a vector $\Pi_j$, with a consistent estimator $\tilde{\Pi}_j$, such that for all $i \in \mathcal{I}_j$:

$$X_i = W_i' \Pi_j + e_i, \quad E[W_i e_i] = 0. \quad (2)$$

Suppose that within a sub-cluster, $W$ has full rank. Then $\tilde{\Pi}_j$ can be chosen as the sub-cluster level OLS estimator of $X$ on $W$. Otherwise, we can just drop variables until we obtain a linearly independent subset $\tilde{W}$. The entries of $\tilde{\Pi}_j$ corresponding to the dropped variables can then be set to 0 while the remaining entries are chosen to be the corresponding coefficients from the sub-cluster level regression of $X$ on $\tilde{W}$. Alternatively, if the researcher is willing to assume that $\Pi_j$ is identical across clusters, $\tilde{\Pi}_j$ can also be obtained from the full sample regression of $X$ on $W$. Now define

$$Z_i = (X_i - W_i \tilde{\Pi}_j)U_i \quad \text{and} \quad \tilde{Z}_i = (X_i - W_i \tilde{\Pi}_j)\tilde{U}_i, \quad (3)$$

where $\tilde{U}_i$ is the full-sample OLS residual using (1). Suppose we know that clusters are independent, so that $E[Z_i Z_i'] = 0$ when $i \in \mathcal{I}_k, i' \in \mathcal{I}_k'$ and $k \neq k'$. Under this assumption, we test the null hypothesis that sub-clusters are uncorrelated:

$$H_0 : E[Z_i Z_i'] = 0 \quad \text{for all } i \in \mathcal{I}_j, i' \in \mathcal{I}_j', j \neq j' \quad (4)$$

against the alternative hypothesis that there exists sub-clusters within at least one cluster that exhibit correlation:

$$H_A : E[Z_i Z_i'] \neq 0 \quad \text{for some } i \in \mathcal{I}_j, i' \in \mathcal{I}_j' \quad \text{such that } j, j' \in \mathcal{J}_k, j \neq j'. \quad (4)$$

Note that changing the choice of $X_i$ and $W_i$ corresponds to testing different null hypotheses and could lead to differing outcomes. If a researcher wants to test the level of clustering used for inference on $\beta$, $X_i$ should be projected onto $W_i$. Similarly, if inference was conducted on $\gamma$, then $W_i$ should take the place of $X_i$ in (3).

**Remark 1.** A researcher interested in inference on $\beta$ only has to test the residualized hypothesis of (4). This is because to the first order, the asymptotic distribution of

$$\sqrt{n} \left( \hat{\beta} - \beta \right) \quad \text{and} \quad \sqrt{n_j} \left( \hat{\beta}_j - \beta \right)$$

depends only on

$$\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}} Z_i \quad \text{and} \quad \frac{1}{\sqrt{n_j}} \sum_{i \in \mathcal{I}_j} Z_i,$$

respectively. If the $Z_i$’s exhibit no correlation across clusters, then conducting inference using the sub-clusters is appropriate. We flesh out this argument in Appendix C.
Remark 2. Note that tests for different coefficients require different adjustments for clustering. This is unsurprising since our null hypothesis concerns dependence between $Z_i$'s. For intuition, consider a setting with two covariates, $X_1$ and $X_2$, which are mean zero random variables that are independent of each other and independent of $U$. Consider the null hypothesis when $X_1$ is our covariate of interest. Now,

$$Z_{1,i} = (X_{1,i} - X_{2,i} \Pi_{1,i}) U_i = X_{1,i} U_i \text{ since } X_1 \perp \perp X_2,$$

and similarly for $X_2$. Hence,

$$E[Z_{1,i} Z_{1,f}] = E[X_{1,i} X_{1,f}] E[U_i U_f],$$

$$E[Z_{2,i} Z_{2,f}] = E[X_{2,i} X_{2,f}] E[U_i U_f].$$

If $X_{1,i}$ is independent across sub-clusters, then $E[Z_{1,i} Z_{1,f}] = 0$. This is true even if $X_{2,i}$ is dependent within clusters, so that $E[Z_{2,i} Z_{2,f}] \neq 0$. A similar phenomenon arises in methods employing degrees of freedom correction for inference with a small number of clusters. Here, each slope parameter in a regression may require a test with different degrees of freedom approximation.

Remark 3. As with Ibragimov and Müller (2016) and MacKinnon, Nielsen, and Webb (2020), we require the researcher to specify independent clusters which nest the potentially correlated sub-clusters. For a concrete example, these methods cannot test the null of commuting zone level clustering against the alternative of state level clustering. However, they can be used to test the null of county level clustering against the alternative of state level clustering. Researchers interested in testing the null of commuting zone level clustering might instead consider the alternative of clustering at the level of labor market areas.

We further assume the following:

Assumption 2. Suppose $q$ and $r$ are fixed, but $n_j \to \infty$ for all $j \in J$. Let $Z_i$ be defined as in (3). Suppose there exists $\Omega \in \mathbb{R}^{q \times q}$ such that the $q$-vector $S_n$ converges in distribution to $S$, where

$$S_n := \left( \frac{1}{\sqrt{n_j}} \sum_{i \in T_j} Z_i \right)_{j \in J}, \quad S := N (0, \Omega). \quad (5)$$

Further, let $\hat{\beta}$ and $\hat{\gamma}$ be the (joint) respective OLS estimators of $\beta$ and $\gamma$, as defined in (1) and $\hat{\Pi}_j$ be the estimator of $\Pi_j$ as defined in (2). Suppose:

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\gamma} \xrightarrow{p} \gamma, \quad \sqrt{n_j} \left( \hat{\Pi}_j - \Pi_j \right) = O_p(1) \quad \text{for all } j \in J.$$

In other words, we assume that the errors are weakly correlated within each sub-cluster $j$. Imposing weak dependence within a (sub-)cluster is not an uncommon assumption (see for instance the discussion in Canay, Romano, and Shaikh 2017; Bester, Conley, and Hansen 2011). We note that under $H_0$, $\Omega$ is a diagonal matrix. On the other hand, under the alternative, it has a block diagonal structure due to correlation between sub-clusters.

Remark 4. Our assumption that $S_n \to S$ does not implicitly assume that (sub-)clusters have similar sizes. Intuitively, this is because our randomization test assigns “equal weight” to each sub-cluster: each sub-cluster is normalized by its own $n_j$, and the sign of each $S_{nj}$ contribute equally to the sign mismatch within its parent cluster. As such, heterogeneous sub-cluster sizes pose no issue for our test. Nonetheless, the quality of the asymptotic approximation is determined by $\min_{j \in J} n_j$, so the smallest cluster has to be large. We expand on this point in Appendix D and explain how the restricted heterogeneity assumptions that are required for inference with clustered data are not needed in our case.

Remark 5. Without further assumptions on $Z_i$, our test requires large sub-clusters. This rules out testing the null of no clustering where there is only one observation in each sub-cluster. However, the test is valid for the null of no clustering if we are willing to assume that each $Z_i$ is symmetrically distributed around 0. This might obtain, for example, if $U$ is symmetric around 0 conditional on $X$ and $W$. Such assumptions can be found in the econometrics literature. For example, Davidson and Flachaire (2008) use it to justify a wild-bootstrapped based $F$-test for the linear regression model. Nonetheless, we consider this assumption to be highly restrictive and hence justify our test via large sub-cluster asymptotics.

2.2. Test Statistic and Critical Value

In this section, we define the test statistic and explain the need to search over the worst case critical value. Before doing so, we first consider the infeasible test in which the true parameters -- $\beta$, $\gamma$ and $\Pi$ as defined in (1) and (2) -- are observed. Readers who are only interested in the details of implementation can skip to the end of Section 2.3.

2.2.1. Infeasible Test

Suppose we know $\beta$, $\gamma$, and $\Pi$. Given $Y_i$ and $X_n$, we can back out $U_i$ and construct the vector $S_n^*$, whose $j$th entry is

$$S_{nj}^* = \frac{1}{\sqrt{n_j}} \sum_{i \in T_j} Z_i = \frac{1}{\sqrt{n_j}} \sum_{i \in T_j} (X_i - W_i' \Pi_j) U_i. \quad (6)$$

Given $S_n^*$, we can then define the infeasible test statistic:

$$T(S_n^*) = \frac{1}{r} \sum_{k \in K} \sum_{j \in J_k} \left( I(S_{nj}^* \geq 0) - I(S_{nj}^* < 0) \right). \quad (7)$$

The inner sum is the net number of positive $S_{nj}^*$ within each cluster $k$. Intuitively, if the sub-clusters are independent, the net number of positive $S_{nj}^*$ should be close to 0. Conversely, if they are positively correlated, this number will be large in absolute value, since many sub-clusters will have $S_{nj}^*$ of the same sign. On the other hand, if they are negatively correlated, this number will be more concentrated around 0 than in the independent case. As will become clear below, our test interprets large absolute values of $T(S_n^*)$ as violation of the null. For this reason, it will not have power against negative correlation.
Remark 6. There are two advantages to having a test statistic that depends only on the sign of the $S_n^*$'s. First, large and small realizations of $S_n^*$ contribute the same amount to $T(S_n^*)$. As such, the performance of our test is not affected even if subclusters have wildly differing variances, a source of heterogeneity that may be important in applied work. We demonstrate this robustness property via simulations in Section 3.2. Second, the feasible version of this test requires searching over the worst case values of the test statistic. As will become clear in Section 2.3, this search is simplified by our choice of test statistic.

Next, denote by $G$ the set of sign changes. $G$ can be identified with the set $\{-1, 1\}^q$. That is, we can construct $G$ by enumerating all vectors of length $q$ with either 1 or $-1$ in each component. Then for each $g \in G$,

$$gS_n^* = \left( g_1 \cdot S_n^{*1}, \ldots, g_q \cdot S_n^{*q} \right).$$

Now let $p^*(S_n^*)$ be the proportion of $T\left(gS_n^*\right)$ that are no smaller than $T\left(S_n^*\right)$:

$$p(S_n^*) = \frac{1}{|G|} \sum_{g \in G} 1\{T\left(gS_n^*\right) \geq T\left(S_n^*\right)\}. \tag{8}$$

The test rejects the null hypothesis when $p(S_n^*)$ is small—that is, when $T\left(S_n^*\right)$ is extreme relative to $T\left(gS_n^*\right)$:

$$\phi_n = \begin{cases} 1 & \text{if } p(S_n^*) \leq \alpha \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The intuition for the randomization test is as follows. Since $S_n^*_{n_{ij}}$ involves only units within the same sub-cluster, under the null hypothesis, $S_n^*$ converges to a mean-zero normal distribution with independent components. Independence, together with symmetry of normal random variables about their means, implies that for any $g \in G$, $gS_n^*$, has the same distribution as $S_n^*$. Hence, the randomization distribution $\{T\left(gS_n^*\right)\}_{g \in G}$ is in fact the distribution of $T\left(S_n^*\right)$ conditional on the values of $|S_n^*|$, where $| \cdot |$ is applied component-wise. Rejecting the null hypothesis when we observe values of $T\left(S_n^*\right)$ that are extreme relative to $\{T\left(gS_n^*\right)\}_{g \in G}$ therefore leads to a test with the correct size.

Note that the randomization test defined above is non-randomized. Randomization tests can also employ a randomized rejection rule for the situation when

$$\frac{1}{|G|} \sum_{g \in G} 1\{T\left(gS_n^*\right) > T\left(S_n^*\right)\} < \alpha \quad \text{but} \quad \frac{1}{|G|} \sum_{g \in G} 1\{T\left(gS_n^*\right) \geq T\left(S_n^*\right)\} > \alpha .$$

Using a randomized rejection rule, the randomization test will have size equal to or exactly if the necessary symmetry properties hold in finite sample. The test defined in (9) is conservative since it never rejects when the above situation occurs. However, we present the deterministic version since the test that we propose is based on it.

### 2.2.2. Naive Test

Tests based on $S_n^*$ are infeasible since $\beta$, $\gamma$ and the $\Pi_j$’s are unknown. Suppose we simply replaced $Z_j$ with $\hat{Z}_j$ and performed the randomization test with the estimated scores. It turns out that this procedure is incorrect. To see this, let $\hat{S}_n$ be $S_n^*$ but with $\hat{Z}_j$ replacing $Z_j$. Then we can write each component of $\hat{S}_n$ as

$$\hat{s}_{n_{ij}} = \frac{1}{\sqrt{m_j}} \sum_{i \in I_j} \hat{Z}_i = \frac{1}{\sqrt{m_j}} \sum_{i \in I_j} (X_i - W_i'\hat{\Pi}) \hat{U}_i \tag{10}$$

$$\approx \frac{1}{\sqrt{m_j}} \sum_{i \in I_j} (X_i - W_i'\Pi_j) U_i - (\hat{\beta} - \beta) \frac{1}{\sqrt{m_j}} \sum_{i \in I_j} (X_i - W_i'\Pi_j)^2 .$$

In the above equation, $S_{n_{ij}}^*$ is the part that is informative about cluster structure. However, each component now has an additional nuisance term $A_j$ that does not go away under asymptotics that take the number of sub-clusters to be fixed. Because $\hat{\beta} - \beta$ is common across the $A_j$’s, it induces correlation across $\hat{s}_{n_{ij}}$ even when the $S_{n_{ij}}^*$’s are independent, leading potentially to over-rejection. Addressing this complication which does not arise in frameworks taking $q \to \infty$ results in the conservativeness of our test.

### 2.2.3. Feasible Test

If we knew $\hat{\beta} - \beta$, we could back out $S_{n_{ij}}^*$ for the randomization test using (11). Since that is not possible, we propose to search over values of $\hat{\beta} - \beta$ to ensure that the test controls size when the unobserved term takes on extreme values.

For a given $\lambda \in \mathbb{R}$, let $\hat{S}_n(\lambda)$ be $q \times 1$ vector whose $j$th entry is the following term:

$$\hat{s}_{n_{ij}}(\lambda) := \frac{1}{\sqrt{m_j}} \sum_{i \in I_j} (X_i - W_i'\hat{\Pi}) \hat{U}_i + \lambda \frac{1}{\sqrt{m_j}} \sum_{i \in I_j} (X_i - W_i'\hat{\Pi})^2 .$$

Note that $\hat{s}_{n_{ij}}(\hat{\beta} - \beta) = S_{n_{ij}}^*$ + $o_P(1)$. Define:

$$T(\hat{S}_n(\lambda)) = \frac{1}{p} \sum_{j \in K} \sum_{i \in I_j} \left( 1(\hat{s}_{n_{ij}}(\lambda) \geq 0) - 1(\hat{s}_{n_{ij}}(\lambda) < 0) \right).$$

For a given $\lambda$, this is just the test statistic in (7) but with $\hat{s}_{n_{ij}}(\lambda)$ taking the place of $S_{n_{ij}}^*$. As before, we denote by $G$ the set of sign changes and write:

$$g\hat{S}_n(\lambda) = \left( g_1 \cdot \hat{s}_{n_{ij}}(\lambda), \ldots, g_q \cdot \hat{s}_{n_{ij}}(\lambda) \right).$$

Now let $p(\hat{S}_n(\lambda))$ be the proportion of $T\left(g\hat{S}_n(\lambda)\right)$ that takes on extreme values relative to $T\left(\hat{S}_n(\lambda)\right)$:

$$p(\hat{S}_n(\lambda)) = \frac{1}{|G|} \sum_{g \in G} 1\{T\left(g\hat{S}_n(\lambda)\right) > T\left(\hat{S}_n(\lambda)\right)\} . \tag{12}$$

We can then define the randomization test as

$$\phi_n = \begin{cases} 1 & \text{if } \sup_{\lambda \in \mathbb{R}} p(\hat{S}_n(\lambda)) \leq \alpha \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

We can then prove the following result:
Remark 7. The worst-case test has no power if \( r = 1 \) since \( \lambda = \text{median}(\hat{S}_{n,j}) \) will set exactly half the signs of \( \hat{S}_{n,j}(\lambda) \) to be positive and half to be negative, so that the signs are completely balanced. However, this is no longer true with \( r > 1 \) since only a single value can be chosen to balance signs across multiple clusters. The implementation procedure provides further intuition for power in this test. See the next section.

Remark 8. As with standard randomization tests, \(|G|\) may sometimes be too large so that computation of \( p(\hat{S}_n(\lambda)) \) becomes onerous. In these instances, it is possible to replace \( p(\hat{S}_n(\lambda)) \) with a stochastic approximation. Formally, let \( \tilde{G} = \{g_1, \ldots, g^q\} \), where \( g^1 \) is the identity transformation and \( g^2, \ldots, g^q \) are iid Uniform\((G)\). Using \( \tilde{G} \) instead of \( G \) in equation (12) does not affect validity of Theorem 1. For implementation, we follow Canay, Romano, and Shaikh (2017) in evaluating the \( p(\hat{S}_n(\lambda)) \) completely when \( q \leq 10 \) and approximating it with \( B = 1000 \) when \( q > 10 \).

Remark 9. We advocate the use of our test as a robustness check, after a researcher has chosen a level of clustering for inference, in the same spirit that manipulation tests are routinely used in studies with regression discontinuity designs, or in tests for pretrends in studies involving difference-in-differences. In particular, the original inference results should be presented with results of the current test, regardless of the outcome. Conceptually, this is different from using the test as a pretest to select the level of clustering prior to inference. The distinction is important as pretesting is known to induce uniformity issues, where inference in the second stage (on \( \beta \)) suffers from distortion due to mistakes in the pretest (that happen with positive probability). These same concerns are articulated by Ibragimov and Müller (2016), who argue that their test “merely provides empirical evidence on the plausibility of one particular clustering assumption.” We take exactly the same view of our test.

2.3. Implementation

In this section we describe an efficient way of searching for \( \lambda \in \mathbb{R} \). This search is simplified by the fact that \( p(\hat{S}_n(\lambda)) \) depends only on the sign of \( \hat{S}_{n,j}(\lambda) \)'s. As such, to find \( \sup_{\lambda \in \mathbb{R}} \), we only need to search over sign combinations of \( \hat{S}_{n,j}(\lambda) \)'s. When \( \beta \) is scalar, the search can be completed in \( O(q) \) time. This is reasonable since the test is designed for use when \( q \) is small.

Suppose for now that \( \sum_{i \in I_j} \left( X_i - W_i \hat{\Pi}_j \right)^2 > 0 \) for all \( j \in \mathcal{J} \). Define:

\[
R_j = \frac{\sum_{i \in I_j} \left( X_i - W_i \hat{\Pi}_j \right) \hat{U}_i}{\sum_{i \in I_j} \left( X_i - W_i \hat{\Pi}_j \right)^2}.
\]

Then, \( \hat{S}_{n,j}(\lambda) \geq 0 \iff R_j + \lambda \geq 0 \). Sort the values of \( R_j \)'s so that \( R^{(1)} \geq R^{(2)} \geq \cdots \geq R^{(q)} \). We must have that \( R^{(j)} + \lambda \geq 0 \iff R^{(j')} + \lambda \geq 0 \) for all \( j' \leq j \). Let \( \hat{S}_{n,(1)}(\lambda), \ldots, \hat{S}_{n,(q)}(\lambda) \) denote the values of \( \hat{S}_{n,j}(\lambda) \) corresponding to \( R^{(1)}, \ldots, R^{(q)} \). Therefore, we only need to consider sequences of the form

\[
\hat{S}_{n,(1)} > \cdots > \hat{S}_{n,(j)} > 0, \quad \hat{S}_{n,(j+1)} < 0, \ldots, \hat{S}_{n,(q)} < 0,
\]

for some cutoff \( j \). Since the \( p \)-value, as defined in (12), depends only on the sign of \( \hat{S}_{n,j} \), we can compute it using \( \hat{S}_{n,j} \) in the place of \( \hat{S}_{n,j}(\lambda) \):

\[
\hat{S}_{n,(1)} = \cdots = \hat{S}_{n,(j)} = 1, \quad \hat{S}_{n,(j+1)} = \cdots = \hat{S}_{n,(q)} = -1.
\]

Here, we see that even when we are searching over the worst case \( \lambda \), we are only allowed to choose the cutoff point at which the signs change. We can therefore complete the search with no more than \( q \) randomization tests. Assuming that the time it takes for each test is \( O(1) \), the procedure takes \( O(q) \) time. The restriction that \( \hat{S}_{n,(j)} \geq \hat{S}_{n,(j')} \) for all \( j \leq j' \) also gives the test power. If all combinations of signs for the \( \hat{S}_{n,j} \)'s were allowed, the test will always return a \( p \)-value of 1 and will have no power.

Finally, suppose there are sub-clusters such that \( \sum_{i \in I_j} \left( X_i - W_i \hat{\Pi}_j \right)^2 = 0 \). We can repeat the above procedure excluding these sub-clusters. In the final step, we set \( \hat{S}_{n,j} \) corresponding to these clusters to 0. Hence,

Remark 10. We can further reduce computation time by the following. Let

\[
R_k^+ = \min_{f \in \mathcal{J}_k} \left\{ R_{j'} \text{ greater than or equal to } 0.5 \text{ of } \{R_{j}, j \in \mathcal{J}_k\} \right\}
\]

be the “upward-conservative” median. Also define the “downward-conservative” median:

\[
R_k^- = \max_{f \in \mathcal{J}_k} \left\{ R_{j} \text{ less than or equal to } 0.5 \text{ of } \{R_{j}, j \in \mathcal{J}_k\} \right\}.
\]

Now let \( R^+ = \max_{k \in \mathcal{K}} R_k^+ \) and \( R^- = \min_{k \in \mathcal{K}} R_k^- \). We only need to consider cutoffs below \( R^+ \). Setting the sign cutoff at the argmax of \( R^+ \) results in situation in which all clusters have at least half of their entries being \(-1\). If we now set the extreme \( \hat{S}_{n,j} \)'s to \(-1\), this will increase the net number of \(-1\)'s in all clusters. Since our test is based on sign imbalance within clusters, such sequences will lead to a strictly larger test statistic and smaller \( p \)-values than if they were set to 1. For the same reason, we only need to consider cutoffs above \( R^- \).

We summarize the implementation procedure in Algorithm 1.
Algorithm 1: Worst-Case Randomization Test

1. Perform full sample OLS to obtain residuals \( \hat{U}_i \). Compute \( \hat{\theta}_j \), for each \( j \in J \).

2. for \( j \in [q]\) do

3. if \( \sum_{i \in I_j} \left( X_i - W_i' \hat{\theta}_j \right)^2 > 0 \) then compute: \( R_j = \frac{\sum_{i \in I_j} \left( X_i - W_i' \hat{\theta}_j \right) \hat{U}_i}{\sum_{i \in I_j} \left( X_i - W_i' \hat{\theta}_j \right)^2} \)

4. else set \( R_j = 0 \).

5. Sort the values of \( R_j \)'s such that \( R^{(1)} \geq R^{(2)} \geq \ldots \geq R^{(q)} \).

6. for \( j \in [q] \), \( R_j \neq 0 \), \( R_j^* \leq R_j \leq R_j^{**} \) do

7. Set \( \hat{S}_{n(1)} = \ldots = \hat{S}_{n(q)} = 0 \), \( \hat{S}_{n(1)} = 1 \), \( \hat{S}_{n(1)} = \ldots = \hat{S}_{n(q)} = -1 \).

8. if \( R_j = 0 \) then replace \( \hat{S}_{n(j)} \) with 0.

9. Compute \( \rho(\hat{S}_n) \). This is as defined in (12), except with \( \hat{S}_n \) in place of \( \hat{S}_{n,j}(\lambda) \). Save this value as \( \hat{p}_j \).

10. if \( \max_{j \in [q]} R_j \neq 0 \hat{p}_j \leq \alpha \) then return 1. Reject the null hypothesis.

11. else return 0. Do not reject the null hypothesis.

2.4. Comparison with Existing Tests

To our knowledge, two other tests have been proposed for the level of clustering. They take either the number of sub-clusters in each cluster to infinity or the number of clusters to infinity. We assume both to be fixed. For ease of exposition, we restrict our discussion of these tests to the univariate case.

Ibragimov and Müller (2016) (IM hereafter) adopts an asymptotic framework that takes \( q_k \rightarrow \infty \) for all \( k \in K \). Consider estimating a regression coefficient cluster-by-cluster. Let \( \hat{\beta}_k \) denote coefficients estimated using only cluster \( k \). The IM test is based on the asymptotic distribution of an estimator for the variance of \( \frac{1}{r} \sum_{k=1}^r \hat{\beta}_k \). Let this variance be denoted by \( V \) and let \( \hat{\Omega}_{CCE}^k \) be the cluster-robust variance estimator for \( \hat{\beta}_k \), where the clustering is done at the sub-cluster level using \( j \in J_k \). Under the null hypothesis, \( \hat{\Omega}_{CCE}^k \) consistently estimates the variance of each \( \hat{\beta}_k \).

Under either the null or the alternative, but maintaining the assumption that coarse clusters are independent, consider estimating \( V \) by

\[
\hat{V} = \frac{1}{r - 1} \sum_{k=1}^r (\hat{\beta}_k - \bar{\beta})^2, \quad \bar{\beta} = \frac{1}{r} \sum_{k=1}^r \hat{\beta}_k.
\]

IM show that under the null, \( \hat{V} \xrightarrow{d} V^W \), where \( V^W = \frac{1}{r - 1} \sum_{k=1}^r (W_k - \bar{W})^2 \) and

\[
W \sim N(0, \text{diag}(\hat{\Omega}_{CCE}^1, \hat{\Omega}_{CCE}^2, \ldots, \hat{\Omega}_{CCE}^r)).
\]

The IM test constructs a reference distribution \( \hat{V}^W \) by drawing \( W \) from

\[
N(0, \text{diag}((\hat{\Omega}_{CCE}^1, \hat{\Omega}_{CCE}^2, \ldots, \hat{\Omega}_{CCE}^r))).
\]

and seeing if \( \hat{V} \) is larger than the \((1 - \alpha)\)th quantile of \( \hat{V}^W \).

There are two limitations to the IM test that our test does not share. First, they require the regression to be estimated cluster-by-cluster. This would be infeasible in, for example, differences-in-differences set ups where treatment varies at the cluster level. Second, since their asymptotics take \( q_k \rightarrow \infty \), we expect the test to have poor properties when \( q_k \) is small. Instead, our test is expected to have good properties even when \( q_k \) is small as long as \( n_j \) is large. These benefits come at a cost. We expect our test to perform worse if observations within sub-clusters are highly correlated, whereas the IM test allows unrestricted covariance within sub-clusters. Our test is also conservative under the null hypothesis. We note also that neither test has power against negative correlations. This is because both tests use test statistics that take on large value relative to their reference distributions only when there is positive correlation.

MacKinnon, Nielsen, and Webb (2020) (MNW hereafter) considers an asymptotic framework that takes \( r \rightarrow \infty \). In the same spirit as IM, the MNW test is a Hausman-type test based on the variance of regression coefficients. Consider the full sample regression coefficient \( \hat{\beta} \). Under the null hypothesis, the (full-sample) cluster-robust covariance estimator at the sub-cluster level, denoted, \( \hat{\Omega}_{CCE}^k \), is consistent for the asymptotic variance-covariance matrix.

Under either the null or the alternative, but maintaining the assumption that coarse clusters are independent, the (full-sample) cluster-robust covariance estimator at the cluster level, denoted, \( \hat{\Omega}_{CCE}^r \), is consistent for the asymptotic variance-covariance matrix. Under the null hypothesis, the authors show that their test statistic converges to a standard normal distribution

\[
\frac{\hat{\Omega}_{CCE}^k - \hat{\Omega}_{CCE}^r}{\sqrt{\hat{\Omega}_{MNW}^r}} \xrightarrow{d} N(0, 1) \text{ for an appropriately defined } \hat{V}^{MNW}.
\]

It is well known that the cluster-robust covariance estimator can be severely biased when \( r \) is small. In order to deal with such situations, the authors propose to conduct the test using the wild (sub-)cluster bootstrap. This procedure imposes the cluster structure specified in the null hypothesis when generating bootstrap samples. Under the null of no clustering, it reduces to the simple wild bootstrap. They prove the consistency of this approach in their large-\( r \) framework, showing power even against alternatives with negative correlations.

Compared to the MNW test, our test is theoretically justified when both \( r \) and \( q \) are small, provided that \( n_j \)'s are large. Our test could therefore be preferable in such applications since it is presently not known if the MNW test remains valid once we take \( r \) and \( q \) to be fixed. However, as with the IM test, the MNW test allows unrestricted covariance within sub-clusters, whereas our test is expected to have poor performance if observations
within sub-clusters are highly correlated. Our test is also conservative relative to the MNW test. On the other hand, simulation evidence suggests that it has comparable performance with the MNW test when clusters have differing variances (see Section 3).

3. Monte Carlo Simulations

In this section, we examine the finite sample performance of our worst-case randomization test (WCR) together with the IM and bootstrap version of the MNW tests via Monte Carlo simulations. We also study the naive randomization test (NR) as described in Section 2.2.2. Section 3.1 considers the effects of varying $r, q_k,$ and $n_j.$ Section 3.2 investigates the effects of cluster heterogeneity. Our data generating processes are as follows:

Model 1: Model 1 is defined by the following:

$$Y_{t,j,k} = X'_{t,j,k}\beta + \sigma_{j,k} \left( \rho V_{t,j,k} + \frac{1}{\sqrt{1-\phi^2}} U_{t,j,k} \right),$$

$$V_{t,j,k} \sim \text{N}(0,1), \quad U_{t,j,k} = \phi U_{t-1,j,k} + \varepsilon_{t,j,k}, \quad \varepsilon_{t,j,k} \sim \text{N}(0,1),$$

In particular, we set $X_{t,j,k} = \beta = 1$ and $\phi = 0.25.$ Errors are correlated within a sub-cluster, according to an AR(1) process, with autocorrelation coefficient $\phi.$ $\rho$ captures the importance of cluster level shock. Since $\frac{1}{\sqrt{1-\phi^2}} U_{t,j,k}$ has unit variance, $\rho$ is exactly the relative variance of cluster-to-sub-cluster-level shocks. $\sigma_{j,k}$ controls the variance of the unobserved term in each cluster $k.$ Here in Section 3.1, we set $\sigma_{j,k} = 1$ for all $j \in J, k \in K.$ In Section 3.2, we explore the consequences of cluster heterogeneity by varying $\sigma_{j,k}.$

Model 2: This is the model used in the simulations of MacKinnon, Nielsen, and Webb (2020), with the constant omitted. Let $m_k = \sum_{j \in J} n_j$ be the total number of observations in cluster $k.$ Let $U_k$ be the $m_k \times 1$ vector of $U_{t,j,k}$ for all observations in cluster $k.$ Then

$$U_k = \rho W_k \varepsilon_k + \sqrt{1-\rho^2} \varepsilon_k, \quad \varepsilon_k \sim \text{N}(0,I_{m_k}),$$

where $\varepsilon_k$ is a $10 \times 1$ vector distributed as

$$\varepsilon_{k,1} \sim \text{N}(0,1), \quad \varepsilon_{k,l} = \phi \varepsilon_{k,l-1} + \varepsilon_{k,l},$$

$$\varepsilon_{k,l} \sim \text{N}(0,1-\phi^2), \quad l \in \{2, \ldots, 10\}$$

and $W_k$ is the $m_k \times 10$ loading matrix with the $(i,j)$th entry $1 \{j = \lfloor (i - 1)10/m_k \rfloor + 1\}.$ Under this model, $\frac{1}{m_k}$ of the observations in each cluster are correlated because they depend directly on the same $\varepsilon_{k,j}.$ In addition, there is correlation between the $\varepsilon_{k,l}$'s since it is generated according to an AR(1) process. Observations are then ordered so that every sub-cluster contains the same number of observations that depend on each $\varepsilon_{k,l}.$ Finally, $\beta = (1,1)'$ and the two covariates are independent and generated in the same way as $U.$ This model features more complex correlations between and within the sub-clusters. Clusters are independent and identically distributed. As in Section 5.2 of MacKinnon, Nielsen, and Webb (2020), we set $\phi = 0.5.$ $\rho$ here is directly comparable to $w_k$ in their simulations.

For our simulations, we perform the test at the 5% level. 10,000 Monte Carlo simulations were drawn for each combination of the parameters. The nonstandard reference distribution in IM is evaluated using 1000 Monte Carlo draws. Wild bootstrap in MNW is evaluated using 399 draws as in their simulations.

3.1. Performance Over Values of $r, q_k,$ and $n_j$

To understand the size and power of each of our tests in scenarios with few clusters and few sub-clusters, we consider equal-sized clusters and sub-clusters, with $r \in \{4, 8, 12\}, q_k \in \{4, 8, 12\}$ and $n_j \in \{25, 50, 100\}.$ We consider $\rho \in \{0, 0.5\}.$

Table 1 presents results under the null hypothesis ($\rho = 0$). Across the two models, we see that regardless of $r,$ the IM test performs poorly when $q_k$ is small. With $q_k = 4,$ Type I error is between 15% and 20%. By $q_k = 12,$ however, the size is between 6% and 7%. Comparatively, our test, which is highly conservative, has Type I error less than 2% across all values of $q_k.$ The MNW and NR tests perform well across the board. Table 2 presents results under the alternative $\rho = 0.5.$ Relative to the IM and MNW tests, our test has power that is consistently lower. In particular, our test does poorly when $q_k$ is small. This is the weakness of the worst-case approach.

Figure 1 presents power of the tests for $r = 8, q_k = 12, n_j = 100$ as we vary $\rho$ from 0 to 2 in model 1 and 0 to 1 in model 2. Across the two models, we see that the IM and MNW tests have greater power than our test. However, as $\rho$ increases, our test quickly catches up in power.

3.2. Cluster Heterogeneity

As a growing body of papers document, heterogeneity across clusters can pose challenges for cluster-robust inference (see for instance Carter, Schnepel, and Steigerwald (2017), Djogbenou, MacKinnon, and Nielsen (2019), and Hu and Spamann (2020)). Three sources of heterogeneity are of particular concern: (i) distribution of errors, (ii) distribution of covariates, and (iii) cluster sizes. In the following, we investigate in turn how each of these issues affect tests for the level of clustering. Whereas the previous section suggests that our test has poor performance compared to other tests, a different picture emerges once we consider heterogeneous clusters.

3.2.1. Distribution of Errors

We first consider what happens when clusters differ in the distribution of regression errors. Specifically, we are interested in the case in which some clusters have much larger variances in their errors than others. This might happen if some clusters are exposed to more shocks than the others, or because clusters systematically differ in certain covariates and errors are heteroscedastic. We return to model 1 but with $\sigma_{j,1} \in \{5, 10, 15\}.$ That is, when all sub-clusters in cluster 1 are much noisier than the rest. Figure 2 plots power curves with $r = 8, q_k = 12, n_j = 100$ for $\sigma_{j,1} \in \{5, 10, 15\}.$ These curves are directly comparable with Figure 1. Starting from the within test comparison, we see that the performance of our test is unaffected by $\sigma_{j,1}.$ However, power of IM and MNW quickly degrade as $\sigma_{j,1}$ increases. Turning to the across test comparison, we see that the tests perform similarly when $\sigma_{j,1} = 5.$ As $\sigma_{j,1}$ increases to 10, our test starts to have more power than the IM and MNW tests for $\rho \geq 1.$ The across test comparison also shows how the NR test fails to control size. In particular, when $\sigma_{j,1}$ an NR test with nominal size 5% could wrongly reject over 40% of the time.

We see the same patterns when sub-clusters are heterogeneous. Consider again model 1 but with $\sigma_{j,k} \in \{5, 10, 15\}.$ That
Table 1. Monte Carlo rejection rates under the null hypothesis $\rho = 0$ at 5% level of significance.

| $r$ | $q_k$ | $n_j$ | Model 1 | Model 2 |
|-----|-------|-------|---------|---------|
|     |       |       | NR     | WCR    | IM     | MNW    | NR     | WCR    | IM     | MNW    |
| 4   | 25    | 0.019 | 0.000  | 0.145  | 0.053  | 0.019  | 0.000  | 0.161  | 0.053  |
|     | 50    | 0.022 | 0.000  | 0.155  | 0.059  | 0.021  | 0.000  | 0.160  | 0.055  |
|     | 100   | 0.020 | 0.000  | 0.154  | 0.058  | 0.019  | 0.000  | 0.157  | 0.055  |
| 8   | 25    | 0.024 | 0.000  | 0.099  | 0.050  | 0.025  | 0.000  | 0.103  | 0.055  |
|     | 50    | 0.022 | 0.000  | 0.095  | 0.054  | 0.022  | 0.000  | 0.098  | 0.052  |
|     | 100   | 0.022 | 0.000  | 0.094  | 0.050  | 0.026  | 0.000  | 0.098  | 0.053  |
| 12  | 25    | 0.024 | 0.001  | 0.075  | 0.050  | 0.026  | 0.001  | 0.083  | 0.053  |
|     | 50    | 0.022 | 0.000  | 0.075  | 0.050  | 0.026  | 0.000  | 0.086  | 0.057  |
|     | 100   | 0.025 | 0.001  | 0.079  | 0.052  | 0.026  | 0.000  | 0.081  | 0.054  |

NOTE: WCR refers to our worst-case randomization test. IM is the test from Ibragimov and Müller (2016). MNW is the bootstrap version of the test in MacKinnon, Nielsen, and Webb (2020). NR is the naïve randomization test. $r$ is the number of clusters, $q_k$ is the number of sub-clusters in each cluster, and $n_j$ is the number of individuals in each sub-cluster.

Figure 1. Power of various tests for level of clustering when $r = 8, q_k = 12, n_j = 100$. The black line indicates the nominal size of the tests (5%).

Figure 3 presents the results. Again, our test is not affected by changing $\sigma_{1,k}$. The power of the IM test falls by a large extent as $\sigma_{1,k}$ increases. The MNW test is also negatively affected by $\sigma_{1,k}$, though less so than the IM test.

3.2.2. Distribution of Covariates

We next consider the case when clusters differ in the distribution of covariates. Specifically, we are interested in the effects of having covariates that are perfectly correlated either within sub-clusters or clusters. These situations arise commonly in empirical work, when treatment assignment occurs at the sub-cluster or cluster level.

Our simulation is based on model 1 but with the following modification. When treatment is assigned at the individual level,

$$X_{t,j,k} := \begin{cases} 1 & \text{w.p. 0.5} \\ 2 & \text{otherwise} \end{cases}$$

is, when the first sub-cluster in each cluster is much noisier than the rest. Figure 3 presents the results. Again, our test is not affected by changing $\sigma_{1,k}$. The power of the IM test falls by a large extent as $\sigma_{1,k}$ increases. The MNW test is also negatively affected by $\sigma_{1,k}$, though less so than the IM test.

3.2.2. Distribution of Covariates

We next consider the case when clusters differ in the distribution of covariates. Specifically, we are interested in the effects of having covariates that are perfectly correlated either within sub-clusters or clusters. These situations arise commonly in empirical work, when treatment assignment occurs at the sub-cluster or cluster level.

Our simulation is based on model 1 but with the following modification. When treatment is assigned at the individual level,

$$X_{t,j,k} := \begin{cases} 1 & \text{w.p. 0.5} \\ 2 & \text{otherwise} \end{cases}$$

where $X_{t,j,k}$ is independent across observations. When treatment is assigned at the sub-cluster level, $X_{t,j,k} = X_{j,k}$ has the same distribution as above, but is identical within sub-clusters and
Table 2. Monte Carlo rejection rates under the alternative hypothesis $\rho = 0.5$ at 5% level of significance.

|   | Power – homogeneous clusters |   |   |
|---|---|---|---|
|   | Model 1 | Model 2 |   |
|   | NR | WCR | IM | MNW | NR | WCR | IM | MNW |
| 4 | 0.052 | 0.000 | 0.315 | 0.152 | 0.068 | 0.000 | 0.372 | 0.204 |
| 50 | 0.061 | 0.000 | 0.313 | 0.157 | 0.130 | 0.000 | 0.529 | 0.336 |
| 100 | 0.053 | 0.000 | 0.316 | 0.156 | 0.230 | 0.001 | 0.687 | 0.520 |
| 8 | 0.122 | 0.006 | 0.383 | 0.291 | 0.175 | 0.010 | 0.458 | 0.380 |
| 50 | 0.118 | 0.006 | 0.379 | 0.282 | 0.317 | 0.044 | 0.648 | 0.576 |
| 100 | 0.110 | 0.005 | 0.363 | 0.273 | 0.508 | 0.126 | 0.800 | 0.754 |
| 12 | 0.188 | 0.028 | 0.471 | 0.392 | 0.687 | 0.177 | 0.735 | 0.700 |
| 50 | 0.187 | 0.030 | 0.455 | 0.399 | 0.662 | 0.357 | 0.868 | 0.850 |
| 100 | 0.186 | 0.026 | 0.467 | 0.392 | 0.662 | 0.357 | 0.868 | 0.850 |

NOTE: WCR refers to our worst-case randomization test. IM is the test from Ibragimov and Müller (2016). MNW is the bootstrap version of the test in MacKinnon, Nielsen, and Webb (2020). NR is the naive randomization test. $r$ is the number of clusters, $q_k$ is the number of sub-clusters in each cluster, and $n_j$ is the number of individuals in each sub-cluster.

3.2.3. Cluster Sizes

Finally, we consider the effects of imbalanced cluster sizes. Specifically, we are interested in situations where each cluster has a sub-cluster that is much larger than the others. Our baseline model for this section follows that of Section 3.2.2. Inspired by Hu and Spamann’s (2020) study of clustering in the state corporate law literature, we set each cluster to contain a sub-cluster comprising 50% of all observations in that cluster. The large sub-cluster therefore contains 1100 observations, while the remaining 11 sub-clusters continue to have size 100. The case with individual treatment is presented in Figure 4. Comparing the case with equally sized sub-clusters (“Indiv”) and with imbalanced sub-clusters (“Indiv, 50%”), we see that imbalanced cluster sizes has negligible effects on the WCR test. It does, however, affect size control of the IM test, leading to greater over-rejection.

The MNW test becomes more slightly more sensitive to imbalanced sub-clusters once we allow treatment to be correlated within sub-cluster or clusters. Carter, Schnepel, and Steigerwald (2017) notes a similar phenomenon: cluster imbalance and correlated covariates interact to worsen size control in cluster-robust inference using the $t$-test. In our setting, this effect is most pronounced with cluster-level treatment assignment, presented in Figure 5. Here we see that increasing the size of the outlier sub-cluster from 33% to 67% of all observations in a cluster reduces power of MNW at
larger values of $\rho$. The performance of the WCR test remains unchanged, so that it starts to have higher power than the MNW test at larger values of $\rho$. As we saw in Figure 4, imbalanced sub-clusters increases the IM test’s Type I error. This continues to be true in Figure 5 with cluster level treatment assignment.

All in all, the simulation evidence suggests that our test manages to maintain Type I error below $\alpha$ when $q$ is small, whereas the IM and NR tests may see size distortion in such a setting. The cost of size control in a fixed $q$ setting is that the procedure is very conservative. This conservativeness limits the power of our test. However, the performance of our test is less sensitive to common sources of heterogeneity within and across clusters, such that it could become more powerful than the IM and MNW tests when clusters are heterogeneous. Indeed, as we will see in the next section, our test detects dependence in the clusters of Gneezy et al. (2019), demonstrating its potential relevance for empirical work.
4. Application: Gneezy et al. (2019)

In recent years, the poor performance of American students in assessment tests such as the Programme for International Student Assessment (PISA) has raised concerns among policymakers. Gneezy et al. (2019) argue that the testing gap reflects, among other things, the low effort that American students put in on tests, especially when compared to their higher scoring counterparts in other countries.

The authors test their hypothesis by a randomized controlled experiment in which students were rewarded with cash for correct answers in a 25-question test. Those assigned to the treatment group were offered roughly $1 USD per correct answer, while the control group received no payment. Students were
informed right before the test started to prevent them from changing their effort in test preparation. The experiments were conducted at four schools in Shanghai and two schools in the United States. Due to logistical reasons, the authors randomized treatment at the class level for some schools and individual level for others.

Various regression analyses were conducted to study the effect of treatment on test-taking effort and test performance. Panel A in their Table 3 examines whether monetary incentive increased the probability that students attempt a given question—a proxy for effort. It does so by estimating the following equation:

$$Y_{qi} = \beta Z_i + \gamma' W_i + U_{qi}.$$  

Here, the unit of analysis is a question and $Y_{qi}$ is an indicator for whether student $i$ attempted question $q$. $Z_i$ is the treatment indicator and $W_i$ is a vector of control variables, which include terms such as gender, ethnicity as well as question number fixed effects. We focus on Column 1 in Panel A, which looks at U.S. students’ responses to all 25 questions in the test, and Column 4, which looks at Shanghai students’ responses to the same test.

The authors present their linear regression estimate of $\beta$, together with standard errors clustered at the level of randomization. However, other levels of clustering are plausible:

- **G**: Group Level, that is, the level of randomization.
- **S**: School Level.
- **SY**: Experiments in Shanghai schools were conducted in 2016 and then 2018. We could plausibly interact school and year of experiment.
- **ST**: Schools in the United States separate students into tracks (Honors, Regular, Others). We could plausibly interact school and track.

We will refer to these levels of clustering by their initials hereafter. More information on the sizes of clusters can be found in Appendix E.

While the authors chose to cluster their standard errors by $G$, it seems reasonable to be concerned about correlation across individuals within the same school or among those who took the test in the same year. If these clusters were not independent, $t$-tests using the presented standard errors could lead to the wrong conclusions.

Table 3 presents the OLS estimates from Gneezy et al. (2019) as well as the $p$-values that would be obtained from testing the null hypothesis that $\beta = 0$ using several methods. Specifically, we consider the wild cluster bootstrap (Cameron, Gelbach, and Miller 2008), approximate randomization tests (Canay, Romano, and Shaikh 2017) and the t-distribution based procedure of Ibragimov and Müller (2010), denoted IM2010. We perform these tests using the various plausible levels of clustering. For Column 1, we consider the increasingly coarse levels of clustering $G$, $ST$ and $S$. For the United States, there are no schools sampled over multiple years, so $SY$ is the same as $S$. For Column 4, we consider the increasingly coarse levels of clustering $G$, $SY$ and $S$. In Shanghai schools, students are not separated by track, so $ST$ is the same as $S$.

**Remark 11.** Gneezy et al. (2019) present clustered standard errors but do not use them for inference. Instead, they conduct randomization inference by permuting treatment status as in Young (2019). This procedure tests the null hypothesis that the distribution of the $Y_{qi}$’s are the same with and without treatment. This is a stronger null hypothesis than the null of 0 average treatment effect ($\beta = 0$). We believe that the latter hypothesis is typically the one of interest and test it in our Table 3.

Turning to the results, for column 1, we see that CCE SE’s decrease as we move to increasingly coarse levels of clustering. Correspondingly, $p$-values from CCE-based $t$-tests decrease as we coarsen the clusters. Such a pattern is typically interpreted as arising from the downward bias of CCEs with few clusters (Angrist and Pischke 2008), so that these $p$-values would be considered unreliable. Faced with downward bias, practitioners commonly turn to the wild cluster bootstrap. With this method, the $p$-values increase as we coarsen the clusters. While clustering at $G$ and $ST$ may lead one to conclude that there is strong evidence that $\beta \neq 0$, the $p$-value at $S$ suggests the absence of strong
evidence. The same phenomenon arises with approximate randomization tests: at $ST$ there appears to be strong evidence that $\beta \neq 0$. At $S$, this is no longer true. With IM2010, the test does not reject in either case. We note that ART and IM2010 cannot be applied with $G$ as the chosen level of clustering, since both methods require $\beta$ to be estimated cluster-by-cluster. The results for column 4 are qualitatively similar. At $G$, CCE-based $t$-test and the wild cluster bootstrap find strong evidence that $\beta \neq 0$. This conclusion is overturned once we cluster at either $SY$ or $S$.

To assess the validity of the above specifications, we apply our WCR test, the IM test and the MNW tests. Table 4 presents the resulting $p$-values. The notation $G \to S$ means that the null hypothesis involves sub-clusters $G$ and coarse clusters $S$. For Column 1, clustering at $G$ appears to be appropriate, as all three tests fail to reject the null hypotheses $G \to ST$ and $G \to S$. For Column 4, all 3 tests find strong evidence that sub-clusters $G$ are inappropriate. The WCR test has higher $p$-values than the IM and MNW tests, likely due to its lower power. Nonetheless, they are close to 5%. The WCR and MNW tests do not reject the null hypothesis for $SY \to S$, whereas the IM test does. Given the that there are at most 2 school-year per school, the IM test is likely to over-reject. As such, we consider the conclusion of the WCR and MNW test to be more reliable in this instance. Thus, results based on clustering at $SY$ are plausible.

All in all, we see that settings with varying numbers of clusters and sub-clusters arise in empirical work. Our test, designed for applications with few clusters and sub-clusters is relevant and appears to work well in practical settings.

5. Conclusion

We propose to test for the level of clustering in a regression by means of a modified randomization test. We show that the test controls size even when the number of clusters and sub-clusters are small, provided that the size of sub-clusters are relatively large. This is a challenging situation not accommodated by existing tests. To ensure size control, our procedure may be conservative when clusters are homogeneous. However, in settings with heterogeneous clusters, it has power that is comparable with other tests. As such, our test can be useful when the researcher faces an application with few sub-clusters, particularly when these clusters are likely to be heterogeneous. Finally, we note that the test is easy to implement and could serve as a helpful robustness check to researchers working with clustered data. An R package is available from the author’s website.

Supplementary Materials

Extended Appendices Technical details including proof of Theorem 1, details concerning the application and additional Monte Carlo simulations.

Acknowledgments

I thank the editor Jianqing Fan, an associate editor, and two anonymous referees for comments and suggestions that greatly improved the article. I am grateful to Ivan Canay for extensive guidance on this project. I have also benefited from conversations with Eric Auerbach, Eduardo Campillo Betancourt, Grant Goehring and Joel Horowitz.

Disclosure Statement

The author reports there are no competing interests to declare.

ORCID

Yong Cai http://orcid.org/0000-0002-9735-0247

References

Abadie, A., Athey, S., Imbens, G., and Wooldridge, J. (2017), "When Should You Adjust Standard Errors for Clustering?" The Quarterly Journal of Economics, 138, 1–35. [933,934]

Angrist, J. D., and Pischke, J.-S. (2008), Mostly Harmless Econometrics, Princeton, NJ: Princeton University Press. [933,944]

Bell, R., and McCaffrey, D. E. (2002), "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," Survey Methodology, 28, 169–181. [935]

Bertrand, M., Duflo, E., and Sendhil, M. (2004), "How Much Should We Trust Differences-in-Differences Estimates?" Quarterly Journal of Economics, 119, 249–275. [933]

Bester, C. A., Conley, T. G., and Hansen, C. B. (2011), "Inference with Dependent Data Using Cluster Covariance Estimators," Journal of Econometrics, 165, 137–151. [935]

Cai, Y., Canay, I. A., Kim, D., and Shaikh, A. M. (2021), "A User’s Guide to Approximate Randomization Tests with a Small Number of Clusters," working paper. [933]

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008), "Bootstrap-Based Improvements for Inference with Clustered Errors," The Review of Economics and Statistics, 90, 414–427. [933,944]

Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017), "Randomization Tests under an Approximate Symmetry Assumption," Econometrica, 85, 1013–1030. [933,935,937,944]

Carter, A. V., Schnepel, K. T., and Steigerwald, D. G. (2017), "Asymptotic Behavior of At-test Robust to Cluster Heterogeneity," Review of Economics and Statistics, 99, 698–709. [935,939,941]

Davidson, R., and Flachaire, E. (2008), "The Wild Bootstrap, Tamed at Last," Journal of Econometrics, 146, 162–169. [935]

Djogbenou, A. A., MacKinnon, J. G., and Nielsen, M. Ø. (2019), "Asymptotic Theory and Wild Bootstrap Inference with Clustered Errors," Journal of Econometrics, 212, 393–412. [939]

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., and Xu, Y. (2019), "Measuring Success in Education: The Role of Effort on the Test Itself," American Economic Review: Insights, 1, 291–308. [934,937,942,943,944]

Hu, A., and Spamann, H. (2020), "Inference with Cluster Imbalance: The Case of State Corporate Laws," AEA Law & Corporate Governance working paper. [939,941]

Ibragimov, R., and Müller, U. K. (2010), "t-Statistic Based Correlation and Heterogeneity Robust Inference," Journal of Business & Economic Statistics, 28, 453–468. [944]

(2016), "Inference with Few Heterogeneous Clusters," The Review of Economics and Statistics, 98, 83–96. [934,935,937,938,940,941]

Imbens, G. W., and Kolesár, M. (2016), "Robust Standard Errors in Small Samples: Some Practical Advice," The Review of Economics and Statistics, 98, 701–712. [935]

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis for Generalized Linear Models, Biometrika, 73, 13–22. [933]

MacKinnon, J. G., Nielsen, M. A., and Webb, M. D. (2020), "Testing for the Appropriate Level of Clustering in Linear Regression Models," Queens Economics Department working paper no. 1428. [934,935,938,939,940,941]

Young, A. (2019), "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," The Quarterly Journal of Economics, 134, 557–598. [944]