Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity

Andrew V. Carter
Department of Statistics
University of California, Santa Barbara

Kevin T. Schnepel
School of Economics
University of Sydney

Douglas G. Steigerwald*
Department of Economics
University of California, Santa Barbara

February 1, 2016

Abstract

Abstract We study the behavior of a cluster-robust t statistic and make two principle contributions. First, we relax the restriction of previous asymptotic theory that clusters have identical size, and establish that the cluster-robust t statistic continues to have a Gaussian asymptotic null distribution. Second, we show how cluster heterogeneity governs the behavior of the test statistic. To do so, we develop the *effective* number of clusters, which scales down the actual number of clusters by a measure of three quantities that vary over clusters: cluster size, the cluster specific error covariance matrix and the actual value of the covariates. The implications for hypothesis testing in applied work are: 1) the number of clusters, rather than the number of observations, should be reported as the sample size, and 2) for data sets in which there is variation in the cluster sizes, or where a cluster-level covariate shows little variation across clusters, the effective number of clusters should be reported. If the effective number of clusters is large, then testing based on critical values from a normal distribution is appropriate.

KEYWORDS: Cluster robust, heteroskedasticity, t test

*We thank Dick Startz, together with members of the Econometrics Research Group at UC Santa Barbara, Colin Cameron, and Ulrich Müller, who served as discussant at the 2013 Econometric Society Winter Meetings, Mark Watson, and 2 referees for helpful comments. Corresponding author: doug@ucsb.edu

1 Introduction

In conducting inference with a cluster-robust t statistic, researchers often rely on the result that the statistic has a Gaussian asymptotic null distribution. The existing result is derived for the specific case in which clusters are equal in size. Because in many applications clusters are unequal in size, there is a gap between the existing result and empirical practice. We fill this gap by establishing that the conventional cluster-robust t statistic has a Gaussian asymptotic null distribution for the more general case in which clusters can vary in size. In so doing, we determine a sample specific measure of cluster heterogeneity that governs the behavior of this cluster-robust t statistic. From the sample specific measure we construct the effective number of clusters, which scales down the actual number of clusters by the measure of cluster heterogeneity. It is the effective number of clusters that governs inference: If the effective number of clusters is large, then Gaussian critical values are appropriate.

The conventional cluster-robust t statistic is based on the ordinary least squares coefficient estimator from the entire sample, together with a cluster-robust variance estimator based on the outer product of the residuals.[1] The original asymptotic theory, due to White (1984, Theorem 6.3, p. 136), applies to clusters of equal size that satisfy a further assumption of cluster homogeneity. Under cluster homogeneity White establishes two principle results. First, that the cluster-robust t statistic has a Gaussian asymptotic null distribution. Second, that the variance component, which appears in the denominator of the test statistic, is consistently estimated through the use of the cluster-robust variance estimator. Consistent estimation of the variance component is also established in Hansen (2007), who maintains the assumption that clusters have equal size while relaxing White's further assumption of cluster homogeneity. We allow both for unequal cluster size and for heterogeneity of clusters. Under these more general assumptions we establish that the cluster-robust variance estimator can be used to consistently estimate the variance component that appears in the denominator of the test statistic. We further establish that the cluster-robust t statistic has a Gaussian asymptotic null distribution.

To understand why variation in cluster sizes impacts the behavior of the cluster-robust t statistic, consider a sample of 20 observations divided into two clusters. Because observations are assumed to be independent across clusters, the number of nonzero elements of the error covariance matrix are all contained within the diagonal blocks that capture the correlation within clusters. If the clusters are equally sized, there are 110 potentially unique terms. As the size of one cluster grows, the number of elements of the error covariance matrix grows and reaches a maximum of 191 when one group contains 19 observations. Variation in cluster size, keeping fixed the total number of observations, alters the number of non-zero error covariance terms. Because each of these non-zero terms must be accounted for to avoid upward bias in the test statistic, as Kloek (1981) was among the first to show, the behavior of the cluster-robust t

[1] In what follows we refer to this test statistic simply as the cluster-robust t statistic.

statistic is impacted by variation in cluster size. Cameron, Gelbach and Miller (2008) find via simulation that for a small number of clusters, allowing clusters to have differing numbers of observations can substantially increase the size of a cluster-robust t test.

As we will show, use of the outer product of the residuals implies that the cluster-robust variance estimator is a function only of between cluster variation and, hence, that consistency of the variance estimator requires that the number of clusters grows without bound. Thus the number of clusters is the appropriate measure of the sample size. One immediate consequence is that it is not possible to conduct valid inference on cluster fixed effects with the cluster-robust t statistic. With a fixed effect limited to a single cluster the variance of the fixed effect is estimated from a sample of size 1, so the estimator of the variance is undetermined.

Because estimation and inference in practice are conditioned on the observed value of the covariates, the measure of cluster heterogeneity we derive is specific to each sample. The measure depends on how three quantities vary over clusters: cluster size, the cluster specific error covariance matrix and the observed value of the covariates. The measure of cluster heterogeneity scales down the number of clusters to produce the effective number of clusters. A low effective number of clusters leads to a higher mean-squared error for the cluster-robust variance estimator, which in turn affects the behavior of the cluster-robust t statistic.

The effective number of clusters can be thought of as a generalization of the correction formula reported in Moulton (1986). The correction formula, which requires that all observations be equally correlated within clusters, indicates how to increase standard error estimators to account for neglected cluster correlation. The effective number of clusters, which does not require equal correlation of all observations within clusters, does not alter the cluster-robust standard error estimator but rather alerts the researcher to the need for conservative critical values. The need to report the effective number of clusters is not restricted to data sets with unequal cluster sizes. For example, data sets with equal cluster sizes but where most clusters have the same value for a cluster-level covariate can have an effective number of clusters that is dramatically smaller than the actual number of clusters, which emphasizes the need to report the effective number of clusters when reporting a cluster-robust t statistic.

Through simulation we demonstrate this point and find that in many settings, while cluster heterogeneity reduces the effective number of clusters, the reduction results in only a moderate increase in the rejection rate for the test. In these cases, a researcher can report the effective number of clusters and proceed with Gaussian critical values. For settings with severe heterogeneity and substantial cluster correlation, the effective number of clusters can fall well below 20. When this is the case we find a downward bias in the cluster-robust standard errors, which in turn leads to rejection rates of up to 30 percent for a nominal size of 5 percent. In practice calculation of the effective number of clusters depends on the unknown error correlations. We show how to overcome this difficulty through use of an approximate measure that depends only on the

observed covariates and cluster sizes. The simulations reveal that the approximate measure, while conservative, closely tracks the effective number of clusters in precisely the situations where the calculation is of most importance, namely where correlation within clusters is substantial.

The paper is organized as follows. In Section 2 we define the general class of models under study and define the measure of cluster heterogeneity. We relate the measure to the mean-squared error of the cluster-robust variance estimator, establish that the asymptotic null distribution of the cluster-robust t statistic is Gaussian and show that consistent testing of fixed effects is not possible. In Section 3, we define the effective number of clusters and emphasize, through simulation, that the effective number of clusters is a sample specific measure that varies with the coefficient under test. For several empirical settings we report an effective number of clusters for the key hypotheses under test and discuss appropriate inference, in Section 4. While not our principle focus, we discuss how to select conservative critical values in Section 5.

2 Asymptotic Behavior

We consider a set of n observations from the linear model

$$y = X\beta + u, \tag{1}$$

where the covariate matrix X consists of k linearly independent columns. The key feature of the model is that the observations can be sorted into G clusters, where the errors are independent between clusters. Hence the covariance matrix of u, given X, Ω is a block-diagonal matrix where each diagonal block Ω_g is the covariance matrix for cluster g. Because Ω is block diagonal, the variance of the ordinary least squares estimator $\hat{\beta}$ can be written as the sum of the G cluster specific variance components. We have

$$V := Var\left[\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}u\middle| X\right] = \sum_{g=1}^{G} Var\left[\left(X^{\mathrm{T}}X\right)^{-1}X_g^{\mathrm{T}}u_g\middle| X\right],$$

where X_g and u_g are the covariate matrix and error vector for cluster g, respectively.

The hypotheses under test are formed from subsets of the coefficients in (1). The general form of null hypothesis is $H_0 : a^{\mathrm{T}}\beta = a^{\mathrm{T}}\beta_0$, where a is a selection vector of dimension k. Because any factor that multiplies the selection vector cancels out of the test statistic, we assume without loss of generality that $\|a\|^2 = 1$, where $\|a\|$ is the Euclidean norm of the vector a. The cluster-robust t statistic is

$$Z = \frac{a^{\mathrm{T}}\left(\hat{\beta} - \beta_0\right)}{\sqrt{\widehat{Var}\left(a^{\mathrm{T}}\hat{\beta}\right)}}, \tag{2}$$

where the variance component is $\widehat{Var}\left(a^{\mathrm{T}}\hat{\beta}\right) = a^{\mathrm{T}}\widehat{V}a$ and $\widehat{V}$ is the cluster-robust variance estimator. The cluster-robust variance estimator, which Shah, Holt and Folsom (1977) are among the first to use, is the sample analog for V where the observed residuals $\hat{u}_g$ replace the errors u_g:

$$\widehat{V} = \left(X^{\mathrm{T}}X\right)^{-1}\sum_{g=1}^{G} X_g^{\mathrm{T}}\hat{u}_g\hat{u}_g^{\mathrm{T}}X_g\left(X^{\mathrm{T}}X\right)^{-1}. \tag{3}$$

White establishes asymptotic results for the cluster-robust t statistic and for the variance component $\widehat{V}_a := a^{\mathrm{T}}\widehat{V}a$. White's proof has two key assumptions: 1) that all clusters have an identical, fixed, number of observations and 2) that $\mathbb{E}\left(X_g^{\mathrm{T}}\Omega_g X_g\right)$ not vary over g. He then proves that, if $G \to \infty$ as $n \to \infty$ then Z has a Gaussian asymptotic null distribution and $\widehat{V}_a$ is a consistent estimator of V_a. We relax both of White's key, cluster homogeneity, assumptions. We allow the cluster size, n_g, to vary: over clusters, so that clusters need not be of identical size, and to vary with the sample size, so that cluster sizes need not be fixed. We also allow $\mathbb{E}\left(X_g^{\mathrm{T}}\Omega_g X_g\right)$ to vary over g. We then prove that, if $G \to \infty$ as $n \to \infty$, then Z has a Gaussian asymptotic null distribution and $\widehat{V}_a$ is a consistent estimator of V_a.

Because Ω_g is restricted only by the requirements of a positive definite matrix, the test statistic Z is robust to a wide range of correlated processes. But this general robustness has an important implication: $\widehat{V}$ is a function only of between cluster variation. It immediately follows that first, consistency of $\widehat{V}$ requires that the number of clusters grow without bound, and second, that the behavior of Z, even for hypothesis tests of coefficients on covariates that vary within clusters, is governed by the number of clusters, not the total number of observations.[2]

To establish these facts, we first show that the variance of $\hat{\beta}$ can be expressed as a weighted sum of the variances for the ordinary least squares estimators based only on the observations for cluster g, $\hat{\beta}_g$. We then show that it follows that $\widehat{V}$ is a function only of between cluster variation, where between cluster variation corresponds to the difference between the cluster specific means $\left(X_1^{\mathrm{T}}\hat{\beta}_1, \ldots, X_G^{\mathrm{T}}\hat{\beta}_G\right)$ and the overall mean $X^{\mathrm{T}}\hat{\beta}$. We collect these findings in the following result (algebraic details that verify the result are contained in the Appendix).

Result 1:
a) The covariance matrix V, together with the estimator $\widehat{V}$, can be expressed as functions of $\hat{\beta}_g$:

$$V = \sum_g A_g Var\left(\hat{\beta}_g \middle| X\right) A_g^{\mathrm{T}}, \tag{4}$$

[2] If the researcher groups observations into clusters to allow for the possibility of cluster correlation, then, even if the observations are independent, the number of clusters must grow to infinity for consistency of $\widehat{V}$.

$$\widehat{V} = \sum_g A_g \left(\hat{\beta}_g - \hat{\beta}\right)\left(\hat{\beta}_g - \hat{\beta}\right)^{\mathrm{T}} A_g^{\mathrm{T}},$$

where $A_g = \left(X^{\mathrm{T}}X\right)^{-1} X_g^{\mathrm{T}} X_g$.

b) Furthermore, $\left(\hat{\beta}_g - \hat{\beta}\right)$ *isolates the variation between clusters from the variation within clusters, so the estimator* $\widehat{V}$ *is not a function of within cluster variation.*

Remarks: The cost of the general robustness of Z, even under cluster homogeneity, is reflected in Result 1b. Because $\widehat{V}$ is a function only of between cluster variation (and the design through A_g), consistency of $\widehat{V}$ requires that the number of clusters grow without bound. Thus, to ensure we have a consistent test, we require that the selection vector a include only covariates for which the number of clusters in which the covariate takes non-zero values grows without bound. Corollary 1, below, formalizes this remark.

Importantly, we establish consistency of $\widehat{V}_a/V_a$ rather than $\widehat{V}_a - V_a$. We do so because if $\hat{\beta}$ is a consistent estimator of β, then the elements of V converge to zero and do so at a rate that depends on the behavior of the cluster sizes. The rate of convergence of V to zero must be explicitly accounted for in $\widehat{V}_a - V_a$, while it is implicitly controlled in $\widehat{V}_a/V_a$. This point is clearly revealed in Hansen, who studies $\widehat{V}_a - V_a$ and so must establish separate results depending on the rate at which n_g grows with the sample size. Under the assumption that clusters have an identical number of observations, but where $\mathbb{E}\left(X_g^{\mathrm{T}}\Omega_g X_g\right)$ is allowed to vary over g, Hansen establishes that $\widehat{V}_a$ is a consistent estimator of V_a for two rates of growth of n_g. The situation becomes more complex if n_g varies over g, as the appropriate result depends on assumptions governing the growth of specific cluster sizes. Through study of $\widehat{V}_a/V_a$ we avoid the need for rate-specific results and our theorem accommodates a wide range of behavior for n_g.

The estimator $\widehat{V}_a$ can be decomposed into two parts, one of which contains no bias, so that

$$\frac{\widehat{V}_a - V_a}{V_a} = \frac{\widetilde{V}_a - V_a}{V_a} + \frac{\widehat{V}_a - \widetilde{V}_a}{V_a},$$

where $\widetilde{V}_a$ is constructed from the unbiased function

$$\widetilde{V} = \sum_g A_g \left(\hat{\beta}_g - \beta\right)\left(\hat{\beta}_g - \beta\right)^{\mathrm{T}} A_g^{\mathrm{T}}.$$

(We note that $\widetilde{V}$ is equivalently represented as the right side of (3) with u_g in place of $\hat{u}_g$.) One heuristic for understanding the decomposition of the error in the estimator is that $\left(\hat{\beta}_g - \beta\right)$ is likely to be much larger than $\left(\hat{\beta} - \beta\right)$. As a result, our estimator that is a function of $\left(\hat{\beta}_g - \hat{\beta}\right)$ has an error that is mostly dependent on $\left(\hat{\beta}_g - \beta\right)$, as in $\widetilde{V}$. However, the bias of the estimator is

$\mathbb{E}\left(\widehat{V}_a - \widetilde{V}_a\right)$, which is the bias of the second term in the decomposition.

To establish consistency for $\widehat{V}_a$ we will show that both $\left|\frac{\widetilde{V}_a - V_a}{V_a}\right|$ and $\left|\frac{\widehat{V}_a - \widetilde{V}_a}{V_a}\right|$ converge to 0. We do so in a way that allows us to determine the sample specific features that govern the performance of the cluster robust variance estimator. In Lemma 1 we bound the mean-squared error of $\frac{\widetilde{V}_a - V_a}{V_a}$ conditionally on X, which captures the main contribution to the variance of $\widehat{V}_a$. In Lemma 2 we bound the expectation of $\left|\frac{\widehat{V}_a - \widetilde{V}_a}{V_a}\right|$ conditionally on X, which captures the bias of $\widehat{V}_a$. We use these bounds in Theorem 1 to derive the unconditional asymptotic null distribution of the test statistic.

We prove the results under moment assumptions on the (conditional) distribution of the error. In Lemma 1 we show how the results simplify if the error has a conditionally normal distribution.

ASSUMPTION 1: *Conditional on the covariate matrix X, the distribution of the error vector u satisfies:*
(i) u has mean zero.
(ii) u_g satisfies a fourth-order moment condition; specifically there exists an Ω_g such that $u_g = \Omega_g^{1/2} Z_g$ with $\{Z_g\}$ a sequence of uncorrelated random variables that satisfy $\mathbb{E}\left(Z_{gi} Z_{gj} Z_{gk} Z_{gl}\right) = 0$, $\mathbb{E}\left(Z_{gi} Z_{gj} Z_{gk}^2\right) = 0$, $\mathbb{E}\left(Z_{gi} Z_{gj}^3\right) = 0$, $\mathbb{E}\left(Z_{gi}^2 Z_{gj}^2\right) = 1$, and $\mathbb{E} Z_{gi}^4 \leq M_4$. This implies $\mathbb{E} u_i^4 < \infty$.
(iii) u is independent across clusters and has a block diagonal covariance matrix Ω. Specifically, the error vector can be heteroskedastic and have cluster correlation that varies both within and across clusters.

LEMMA 1: *Under Assumption 1,*

$$\mathbb{E}\left\{\left[\frac{\widetilde{V}_a - V_a}{V_a}\right]^2 \middle| X\right\} \leq \frac{1 + \Gamma(\Omega, X)}{G}\left(2 + \frac{M_4 - 3}{n^*}\right),$$

where n^* is defined in the appendix (under cluster homogeneity $n^* = n/G$) and the quantity Γ is defined by

$$\begin{aligned} \gamma_g(\Omega, X) &= a^{\mathrm{T}} A_g Var\left(\hat{\beta}_g \middle| X\right) A_g^{\mathrm{T}} a, \\ \Gamma(\Omega, X) &= \frac{1}{G}\sum_{g=1}^{G} \frac{\left(\gamma_g - \bar{\gamma}\right)^2}{\bar{\gamma}^2}, \end{aligned}$$

with $\bar{\gamma} := \bar{\gamma}(\Omega, X) = \frac{1}{G}\sum \gamma_g(\Omega, X)$.[3] *If Assumption 1(ii) is strengthened to u is normally distributed, then*

$$\mathbb{E}\left\{\left[\frac{\widetilde{V}_a - V_a}{V_a}\right]^2 \middle| X\right\} = \frac{2}{G}\left(1 + \Gamma(\Omega, X)\right).$$

[3] Because the researcher selects a through specification of the null hypothesis, we do not explicitly include a as an argument in $\Gamma(\Omega, X)$.

PROOF: See Appendix.

Remarks: Because $\widetilde{V}_a$ is unbiased for V_a, the (relative) mean-squared error in Lemma 1 consists entirely of the variation in $\widetilde{V}_a$. The quantity $\Gamma(\Omega, X)$, which is the squared coefficient of variation for $\gamma(\Omega, X)$, is the measure of cluster heterogeneity that drives the variation in $\widetilde{V}_a$. To see this, for u normally distributed if $\Gamma(\Omega, X) = 0$, then $\widetilde{V}_a \sim \chi^2_{(G)}$ and $\mathbb{E}\left\{ \left[\frac{\widetilde{V}_a - V_a}{V_a} \right]^2 \middle| X \right\} = \frac{2}{G}$. If $\Gamma(\Omega, X) \neq 0$, then $\widetilde{V}_a \nsim \chi^2_{(G)}$ and the mean-squared error increases by the factor $(1 + \Gamma(\Omega, X))$.

We next bound the bias of $\frac{\widehat{V}_a}{V_a}$; below we will establish that $\frac{\widehat{V}_a - \widetilde{V}_a}{V_a}$ is $o_{\mathbb{P}}(1)$ and so the bias vanishes asymptotically.

LEMMA 2: *Under Assumption 1,*

$$\mathbb{E}\left\{ \left[\left| \frac{\widehat{V}_a - \widetilde{V}_a}{V_a} \right| \right] \middle| X \right\} \leq \frac{1}{G} + \frac{1}{V_a} a^{\mathrm{T}} \sum_{g=1}^{G} \left[\left(A_g - \frac{1}{G} I \right) V \left(A_g - \frac{1}{G} I \right)^{\mathrm{T}} \right] a + \\ + 2 \left(\frac{1}{V_a} a^{\mathrm{T}} \sum_{g=1}^{G} \left[\left(A_g - \frac{1}{G} I \right) V \left(A_g - \frac{1}{G} I \right)^{\mathrm{T}} \right] a \right)^{\frac{1}{2}}.$$

PROOF: See Appendix.

We are now able to establish our principle asymptotic result that the cluster-robust test statistic has an (unconditional) Gaussian asymptotic null distribution.

ASSUMPTION 2:
(i) As $n \to \infty$ the number of clusters is increasing, $G \to \infty$.
(ii) As $G \to \infty$, $\frac{\mathbb{E}[\Gamma(\Omega, X)]}{G} \to 0$.
(iii) As $n \to \infty$, $\frac{1}{V_a} a^{\mathrm{T}} \sum_{g=1}^{G} \left[\left(A_g - \frac{1}{G} I \right) V \left(A_g - \frac{1}{G} I \right)^{\mathrm{T}} \right] a \xrightarrow{\mathbb{P}} 0$.

Let W be the class of error distributions that satisfy Assumptions 1 and 2. The null hypothesis is $H_0 : a^{\mathrm{T}} \beta = a^{\mathrm{T}} \beta_0$, where the error distribution belongs to the class W.

THEOREM 1: *If Assumptions 1-2 hold, then $\widehat{V}_a$ is a consistent estimator of V_a and, under H_0:*

$$Z \rightsquigarrow N(0, 1),$$

where $\rightsquigarrow$ denotes convergence in distribution.

PROOF: See Appendix.

Remarks: Assumption 2 governs the heterogeneity across clusters as well as the growth rate of cluster sizes. The possible growth rates of cluster sizes are governed by Assumption 2(i)-(ii). The allowable heterogeneity across clusters is contained Assumption 2(ii)-(iii).

For the growth rate of cluster sizes, Assumption 2(i) rules out the case in which all clusters remain a constant proportion of the sample as n grows,

because the number of clusters must go to infinity. Assumption 2(ii) rules out the case in which any of the clusters remains a constant proportion of the sample as n grows, but does allow cluster sizes to grow with n. Because, in general, $Var\left(\hat{\beta}_g \middle| X\right) = O_{\mathbb{P}}\left(\frac{1}{n_g}\right)$ and $Var\left(\hat{\beta} \middle| X\right) = O_{\mathbb{P}}\left(\frac{1}{n}\right)$, the quantity $\left\|\left(\gamma_g - \bar{\gamma}\right)\right\|^2 = O_{\mathbb{P}}\left(\frac{n_g^2}{n^2}\right)$ and $\frac{\mathbb{E}[\Gamma(\Omega,X)]}{G} = O\left(\frac{n_g^{\max}}{n}\right)$, where $n_g^{\max}$ is the size of the largest cluster. Thus, if $n_g^{\max} = o(n)$, then Assumption 2(ii) is satisfied and, hence, Theorem 1 encompasses both the case in which cluster sizes are fixed as the number of clusters grows and cases in which the cluster sizes and the number of clusters go to infinity jointly.

Assumption 2(ii) governs the heterogeneity arising from Ω while Assumption 2(iii) governs the heterogeneity arising from variation in the covariate matrix X. It may be helpful to relate Assumption 2(ii)-(iii) to earlier work in which cluster heterogeneity is considered. While it is difficult to relate these conditions to the work of Hansen, who does not have an explicit condition controlling cluster heterogeneity, it is possible to relate these conditions to the work of Rogers (1993). Although he does not derive an asymptotic null distribution, Rogers conjectured that a Gaussian approximation would be adequate for Z if $\max \frac{n_g}{n} < .05$. To link the conjecture to Assumption 2(ii), consider a model with only an intercept and common intracluster correlation, so that $\gamma_g = \sigma^2 \left(\frac{n_g}{n}\right)^2$. We see that the adequacy of a Gaussian approximation does depend on $\frac{n_g}{n}$, albeit through the squared coefficient of variation, rather than the maximal value.

Under Assumption 2(iii) problematic designs, in which $X^{\mathrm{T}}X$ is (nearly) singular, occur with negligible probability. Assumption 2(iii) principally governs heterogeneity arising from the covariate matrix X. Observe that if all the elements of β are consistently estimated, it is useful to write Assumption 2(iii) as

$$\frac{\lambda_V^*}{V_a} \sum_{g=1}^{G} \left\| a^{\mathrm{T}} \left(A_g - \frac{1}{G} I \right) \right\|^2 \xrightarrow{\mathbb{P}} 0 \quad \text{as } n \to \infty,$$

where λ_V^* is the largest eigenvalue of V. From this expression it is clear that heterogeneity enters only through the covariate matrix. Assumption 2(iii) also allows for models with coefficients that are not consistently estimated (e.g. cluster fixed-effect coefficients where n_g does not grow with n, so the variance of the estimators does not converge to zero). Assumption 2(iii) requires that the selection vector a assign zero weight to these coefficients, so that Theorem 1 applies to models that contain nuisance coefficients that are not consistently estimated.

Moreover, it is not possible to consistently estimate the variance of cluster fixed-effect coefficients with $\widehat{V}$ as defined in (3), even if the coefficients are consistently estimated.

Corollary 1:

If Assumption 1(ii) is strengthend to u is normally distributed, then for coefficient estimators that depend only on a fixed subset of clusters, the elements

of $\widehat{V}$ that correspond to these estimators are inconsistent.

PROOF: Because $\widehat{V}$ is a function only of between cluster variation, consistency of $\widehat{V}$ requires information from a growing number of clusters. If a coefficient estimator depends only on a finite set of clusters, the requirement is not met. Consider a covariate that takes non-zero values for a fixed subset m of the clusters. (For a cluster specific control, $m = 1$.) The element of A_g that corresponds to this covariate is zero for all clusters other than the set of m, so γ_g is nonzero on m elements. Hence $\bar{\gamma}$ is $O\left(\frac{m}{G}\right)$ and Γ is $O\left(\frac{G}{m}\right)$, so that $\frac{1}{G}\Gamma = O\left(\frac{1}{m}\right)$ which does not tend to zero as $G \to \infty$. *Q.E.D.*

We expect that Corollary 1 generally holds for non-normal errors as well. Leading examples of such covariates are cluster specific controls (most often termed cluster fixed effects), controls that correspond to a group of clusters, and, for a model in which only one cluster is treated, the coefficient on the treatment covariate.

3 Effective Number of Clusters

We have established conditions under which the cluster-robust variance estimator $\widehat{V}$ is consistent and the test statistic Z has a Gaussian asymptotic null distribution. We now turn to the question: How should a researcher use the results to inform empirical analysis? An important component to the answer for this question is contained in Lemma 1, where we establish that the relative mean-squared error of $\widetilde{V}_a$ is inversely proportional to

$$G^* = \frac{G}{1 + \Gamma(\Omega, X)}.$$

From the proof of Theorem 1 the leading term that governs the asymptotic behavior of $\widehat{V}_a/V_a$ corresponds to the relative mean-squared error of $\widetilde{V}_a$, so G^* is the key measure of the adequacy of the asymptotic results. Further, because $\widetilde{V}_a$ is unbiased, this analysis reveals that the variance of $\widehat{V}_a$ plays an important role in the finite sample behavior of $\widehat{V}_a$.

We refer to G^* as the *effective number of clusters* to reflect the fact that the results in Section 2 extend the conventional analysis (under cluster homogeneity) in which the number of clusters is the measure of the adequacy of the asymptotic results. To calculate G^*, recall from Lemma 1 that

$$\Gamma(\Omega, X) = \frac{\hat{\sigma}_\gamma^2}{\bar{\gamma}^2},$$

where $\hat{\sigma}_\gamma^2 = \frac{1}{G}\sum_{g=1}^{G}\left(\gamma_g - \bar{\gamma}\right)^2$. Because $\Gamma(\Omega, X) \geq 0$, $G^* \leq G$ so that the effective number of clusters is no larger than the actual number of clusters. Importantly the magnitude of the difference between G^* and G increases nonlinearly in the measure of cluster heterogeneity $\Gamma(\Omega, X)$. To construct $\Gamma(\Omega, X)$ note that the cluster-specific component γ_g, defined in Lemma 1, can also be

written as

$$\gamma_g = a^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} X_g^{\mathrm{T}} \Omega_g X_g \left(X^{\mathrm{T}} X\right)^{-1} a. \tag{5}$$

From this expression we see that $X_g^{\mathrm{T}} \Omega_g X_g$ is the quantity that drives cluster heterogeneity, so variation in cluster size is not required for cluster heterogeneity. Cluster heterogeneity can arise with clusters of equal size, but where the cluster error covariance matrix differs over clusters. Moreover, even if Ω_g is identical across clusters, the fact that the covariates differ over clusters induces heterogeneity. For this reason the vast majority of empirical analyses with cluster-robust inference are characterized by heterogeneous clusters.

To construct G^* in practice, one must approximate the unknown error covariance matrix Ω. It is tempting to use the estimated residuals to replace Ω_g with $\widehat{\Omega}_g = \widehat{u}_g \widehat{u}_g^{\mathrm{T}}$. Yet $\widehat{\Omega}_g$ has already been used to construct the test statistic Z, so using the same data to approximate G^* would lead to dependence between Z and the critical value for Z. Dependence between a test statistic and the critical value used in the test is difficult to account for when determining the size of a test. To avoid this dependence we approximate Ω_g with a matrix that is not constructed from the data. We replace Ω_g with a matrix of ones, which corresponds to perfect correlation between all observations within a cluster. If all observations are perfectly correlated within a cluster, then the information contained in the cluster is reduced to the information contained in any one observation and so our approximation may be conservative for G^*. Let G^{*A} be a feasible version of G^* with this approximation, so G^{*A} is constructed by replacing γ_g with

$$\gamma_g^A = a^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} X_g^{\mathrm{T}} \left(\iota_g \iota_g^{\mathrm{T}}\right) X_g \left(X^{\mathrm{T}} X\right)^{-1} a,$$

where ι_g is a vector of length n_g with each element equal to one.

To illustrate how variation across clusters affects hypothesis testing in empirical settings we turn to simulations. The simulations reveal to what degree certain characteristics in the data cause the size of the test to rise above the nominal level. Moreover, we are able to suggest a threshold for the feasible effective number of clusters, such that if the computed feasible effective number of clusters is above the threshold then it is appropriate to use critical values from the normal distribution. The simulations also provide further insight into how cluster sizes, the distribution of the covariates, and the properties of the error all translate into cluster heterogeneity as reflected in the effective number of clusters.

The data generating process is

$$y_{gi} = \beta_0 + \beta_1 x_{gi} + u_{gi}, \tag{6}$$

together with the error-components model

$$u_{gi} = \varepsilon_g + v_{gi}, \tag{7}$$

where the cluster component $\varepsilon_g | X \sim i.i.d.\ \mathcal{N}(0,1)$ is independent of the individual component $v_{gi} | X \sim \mathcal{N}\left(0, cx_{gi}^2\right)$.[4]

A useful way to capture cluster heterogeneity is to allow the cluster sizes to vary. This allows one to compare the simulated designs to data sets used in empirical research. In each of our experimental designs there are 2500 observations divided into 100 clusters. The first design places 25 observations in each cluster. In each succeeding design the size of the first cluster grows, as observations are moved from clusters 2 through 100 into the first cluster. To keep track of the growing cluster heterogeneity, we calculate the coefficient of variation for the cluster sizes and use this to index the designs in our graphs. (In the Appendix we describe the designs, together with the other settings of the simulations, in detail.)

While the feasible effective number of clusters depends only on the design matrix X, the test size depends on the specification of the error. Perhaps the most important feature of the specification is the value of c. Consider an arbitrary pair of observations, i and j, that are in the same cluster. Because the correlation between these two observations is

$$Corr\left(u_{gi}, u_{gj} | X\right) = \left(1 + cx_{gi}^2\right)^{-1/2} \left(1 + cx_{gj}^2\right)^{-1/2}, \tag{8}$$

if $c = 0$ then the correlation does not vary over i, j, or g. With constant correlation the correction proposed by Moulton would be correct and there would be no need to compute cluster-robust standard errors. A second important feature of the specification is the distribution of v_{gi}. As any distributional assumption is arbitrary and unverifiable, we do not want our findings to be specific to the selection of a normal distribution. To capture the richness of empirical settings in which $\widehat{V}$ is typically employed, we allow c to vary over a range of values and the distribution of v_{gi} to be non-normal.

We initially focus on hypothesis testing for a cluster-level covariate, to reflect the understanding that the effect of cluster correlation on hypothesis testing is most pronounced when the covariate under test is highly correlated within clusters. We then go on to explore the effect on hypothesis testing when the covariate is not as highly correlated within clusters.[5]

Cluster-Level Covariate

To capture the effect of cluster variation on hypothesis testing for a cluster-level covariate, let

$$x_{gi} = x_g,$$

with $\{x_g\}$ a sequence of independent Bernoulli random variables with equal probability of 0 or 1. As written this is a pure treatment model, but for a model with multiple covariates this would correspond to testing for the impact

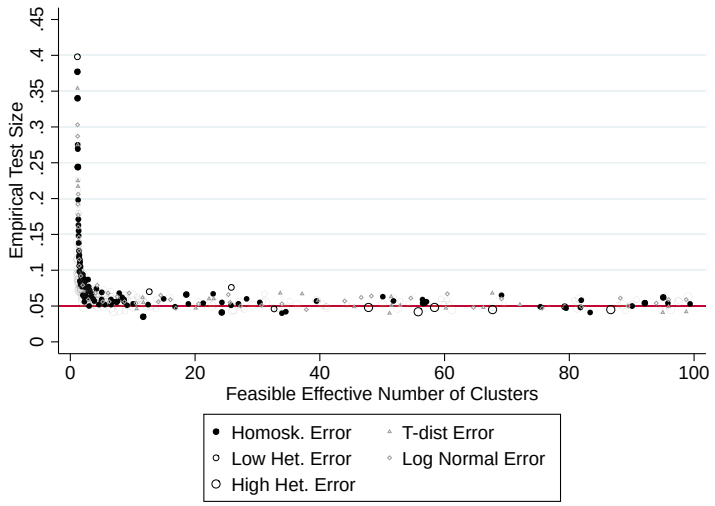
[4] For models with multiple covariates, the effective number of clusters may vary depending on which coefficient is selected for testing (γ_g in (5) depends on the selection vector a).

[5] Carter, Schnepel and Steigerwald (2013), which contains simulation results for a more exhaustive set of models, compares the actual and feasible effective number of clusters and also analyzes the bias, as a proportion of the MSE, for $\hat{V}_a$.

of class size on student test scores in a data set with equal numbers of each of two class sizes. Importantly, because the number of clusters in which x_g takes non-zero values grows with the sample size, the cluster-invariant covariate is distinct from a cluster-specific fixed effect and the statistic Z is consistent for hypothesis testing on β_1.

We display in Figure 1 the test size as a function of the feasible effective number of clusters. The test size depends on three distinct components of the data (the design of cluster sizes, the value of c, and the error distribution) and this dependence could prevent the emergence of a clear relation between the feasible effective number of clusters and the test size. For example, if the level of average dependence between the error terms in a cluster affects the behavior of the test statistic, then results with $c = 0.9$, for which the average dependence is 0.75, would differ from the results with $c = 9$, for which the average dependence is 0.44.[6] Happily, there is a clear relation between the feasible effective number of clusters and the test size. We see that the empirical test size rises sharply above the nominal size of 5%, but does so only when the feasible effective number of clusters falls below 10. This result is robust to the degree of heteroskedasticity c and to the underlying error distribution.[7]

Figure 1



We saw in Figure 1 the dramatic increase in the test size as the feasible effective number of clusters declined. What observable features of the data lead to such a sharp increase in the test size? We display in Figure 2 the effective

[6] Larger values of c reduce the influence of ε_g, thereby reducing the within-cluster error correlation.

[7] The innovation v_{gi} is allowed to be non-normal. For the elements in Figure 2 labeled: 't-dist' the innovation is $t_{(4)}$, and 'log norm' the innovation is $\log(\mathcal{N}(0,1))$, both standardized to have mean 0 and variance 1.

number of clusters as a function of the coefficient of variation of cluster sizes. The plot is quite revealing. From the pattern represented by the squares, which indicate the median value (over 1000 simulations) for each design, we see that the effective number of clusters declines sharply in cluster size variation, nearly falling to the minimum size of 1 when the variation mirrors the population distribution across US states. From the length of the vertical lines, which indicate the maximum and minimum values (over 1000 simulations) for each design, we see that the effective number of clusters also depends on the pattern of values for the covariate, and can fall sharply even for a design with no variation in cluster size.

Figure 2

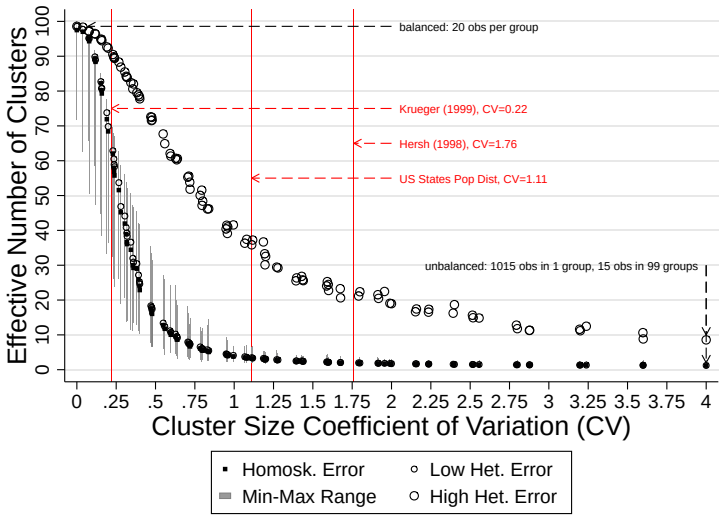
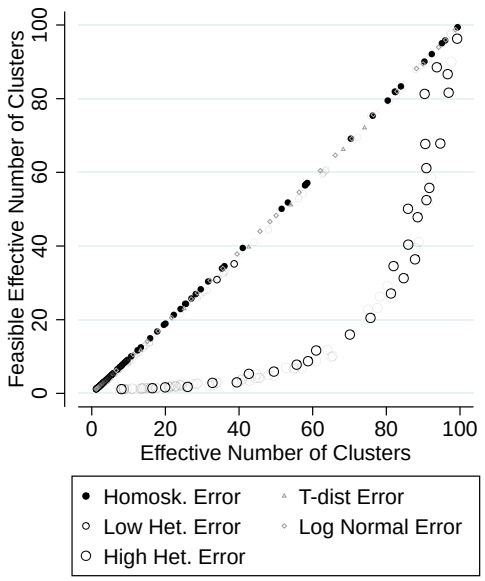


Figure 2 reveals that observable features of the data can indicate a substantial reduction in the effective number of clusters. Does the feasible effective number of clusters show a similar pattern? In Figure 3 we display the feasible effective number of clusters as a function of the number of clusters. The figure reveals a clear pattern. The majority of simulation settings fall near the 45 degree line, indicating a near match between the effective number of clusters and the feasible counterpart we suggest. For simulation settings in which there is a much lower degree of correlation within clusters, the consequence of setting the correlation to 1 when constructing the feasible measure is revealed. For these settings, the feasible measure lies below the effective number of clusters, indicating that the feasible measure is a conservative bound. A conservative bound can be useful: If a researcher finds the feasible effective number of clusters is relatively large, then there is strong evidence that critical values from a normal distribution are appropriate.

Figure 3



Individual-Level Covariate

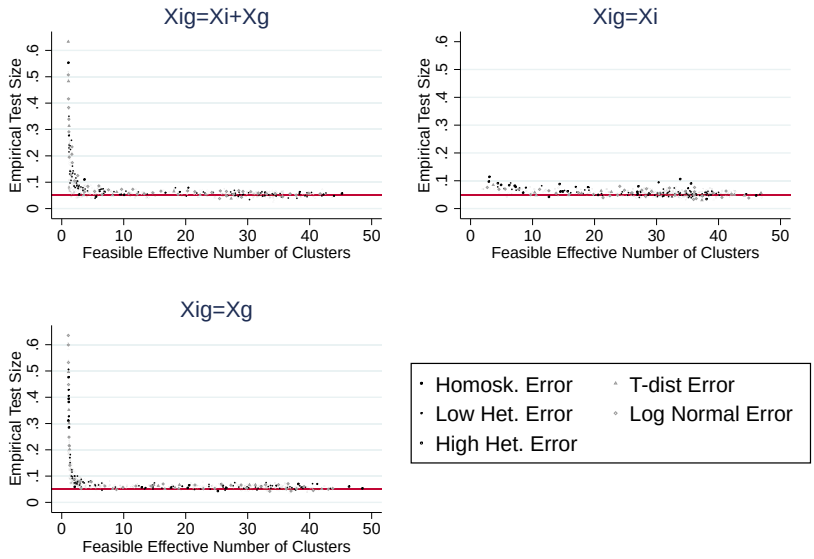
To capture the effect of cluster variation on hypothesis testing for a continuous, individual-level covariate, we consider both

$$\text{(a)} \quad x_{gi} = z_g + z_{gi} \qquad \text{(b)} \quad x_{gi} = \sqrt{2} \cdot z_{gi},$$

where $\{z_g\}$ and $\{z_{gi}\}$ are sequences of independent $\mathcal{N}(0,1)$ random variables. This would correspond to testing the effect of parental income on test scores. The two equations for x_{gi} represent two levels of correlation within clusters: in (a) the correlation is .5 while in (b) the correlation is 0, which would reflect the presence (or absence) of sorting into classes by parental income. We also consider $x_{gi} = z_g$, to show that the results in Figure 1 are not specific to a binary covariate.

Figure 4 displays the test size as a function of the feasible effective number of clusters. What emerges clearly is the importance of the degree of cluster correlation in the covariate under test. The left panels, in which the covariate exhibits substantial cluster correlation, reveal the striking pattern observed in Figure 1. The test size can far exceed the nominal size, but does so only when the feasible effective number of clusters falls below 10. Again the result is robust to the degree of heteroskedasticity and the underlying error distribution. For the right panel, in which the covariate is uncorrelated within clusters, there is no evidence of inflated test size.

Figure 4



4 Empirical Settings

To illustrate how the research design impacts the effective number of clusters, we calculate the effective number of clusters for two empirical settings in which unobserved shocks that are common within a cluster naturally arise: data on children grouped by classroom and workers grouped by industry. Importantly, growth of the sample size can occur through the addition of classrooms or industries, so that each of these settings accommodates the assumption that the number of clusters grows with the sample size.

The first setting corresponds to measurement of the impact of class size on student achievement. Krueger (1999) analyzes data from the STAR experiment in which students were randomly assigned to classrooms of different sizes, identifying the class size effect using the following regression model

$$a_{gi} = \beta_0 + \beta_1 s_g + z_{gi}^{\mathrm{T}}\gamma + u_{gi},$$

where a_{gi} is the test score of student i in classroom g, s_g is the number of students in classroom g and z_{gi} captures other observed determinants of student performance, including the race, gender and socioeconomic status of student i. For kindergarten students, the public use version of the data employed by Krueger contains 5,743 students grouped into 318 classrooms. In describing regression results Krueger reports a sample size corresponding to the number of children (Table V, p. 513). Yet for the purpose of inference, even regarding a

coefficient on a cluster-varying covariate, the appropriate sample size is based on the number of classrooms.

As classrooms form the clusters, the data set has $G = 318$, which appears to be well in excess of the number needed to use Gaussian critical values. Yet the number of students varies across classrooms, from a low of 9 to a high of 27. The mean number of students per classroom is 18 with a variance of 15.7. To determine how the variation in cluster sizes, together with other sources of variation in the design, impacts inference, we compute the effective number of clusters for test of hypotheses on β_1 and find $G^{*A} = 192$. While the variation in the design across clusters has reduced the effective number of clusters to 60 percent of the actual number of clusters, the initial large number of clusters leaves the effective number of clusters sufficiently large that Gaussian inference is reliable.

The second setting corresponds to measurement of the impact of injury risk on wages. Hersch (1998) analyzes data on individual wages from the Current Population Survey, together with injury rates for workers by industry:

$$w_{gi} = \beta_0 + \beta_1 r_g + z_{gi}^{\mathrm{T}}\gamma + u_{gi},$$

where w_{gi} is the (logarithm of the) wage for individual i working in industry g, r_g is the industry-specific injury rate and z_{gi} captures other observed determinants of individual wages. For male workers, the Hersch data set (Table 3, Panel B, column 1) contains 5,960 workers grouped into 211 industries.[8]

As industries form clusters, the data set has $G = 211$, which again appears to be well in excess of the number needed to use Gaussian critical values. The number of workers varies dramatically across industries, ranging from a low of 1 to a high of 517. The mean number of workers per industry is 28 with a variance of 2,474. For test of hypotheses on β_1, we compute $G^{*A} = 19$, which indicates caution in using Gaussian critical values. In this setting the degree of variation in cluster sizes, together with other sources of variation in the design, is large enough to drive the effective number of clusters into a warning area, even though the actual number of clusters is quite large.[9]

The effective number of clusters calculated in these empirical examples is in line with our simulation results presented in Section 3. The Krueger setting (the coefficient of variation for cluster sizes is $cv = 22$) contains less cluster heterogeneity than the first unbalanced simulation design of one large group with 124 observations and 99 groups with 24 observations ($cv = 40$). The cluster size heterogeneity in Hersch ($cv = 176$) is similar to the variation in the designs including one large group of more than 420 observations and 99 groups with 21 or less observations. For these designs, the effective number of clusters is very small compared to the actual number of clusters. Hersch provides an empirical

[8] We thank Colin Cameron for providing the data needed to replicate the Hersch results.

[9] In some specifications Hersch (1998) also includes occupation-specific injury rates and clusters by either occupation or industry. Cameron, Gelbach and Miller (2011) replicate Hersch's results for men and compare clustering on industry and occupation with clustering by industry or occupation. The impact of cluster heterogeneity in multi-way clustering scenarios is left for future research.

setting in which the degree of cluster heterogeneity can lead to large increases in the mean squared error of the conventional cluster-robust variance estimator and a downward bias of test statistics. Along with the simulation results, these examples help emphasize the importance of calculating the effective number of clusters–*even when the number of clusters is large*–to gauge whether inference using the cluster-robust t statistic is appropriate.

5 Remarks

Consistency of the cluster-robust variance estimator, together with a null distribution for the resultant t test statistic as the number of clusters grows large, have previously been established under the assumption of equally sized clusters. We allow the size of clusters to vary and establish conditions under which parallel asymptotic results hold. Our theory yields a sample specific adjustment to the number of clusters, which we term the effective number of clusters. The key innovation is that it is the effective number of clusters that must grow without bound. The effective number of clusters replaces the number of clusters; if the effective number of clusters is large, then the asymptotic theory provides a reliable guide to inference.

Use of the effective number of clusters as a measure of the adequacy of the asymptotic approximation is related to degrees of freedom corrections in related testing problems. For data with error covariance matrices that are not block diagonal, in which a bandwidth parameter mirrors the role of cluster sizes, Sun (2014) derives an "equivalent degrees of freedom", where the adjustment to the degrees of freedom is a function of the bandwidth.

The effective number of clusters depends on two sample specific measures in addition to variation in cluster sizes. First, the measure depends on the cluster-specific error covariance matrices. As these matrices are latent, direct calculation of the effective number of clusters is infeasible. The assumption of perfect within-cluster error correlation provides a useful lower bound on the effective number of clusters. When this feasible measure of the effective number of clusters is large, Gaussian critical values can be used with the cluster-robust t test statistic.

Second, the effective number of clusters depends on how the realized values of the covariates are distributed across clusters. This is the essence of the sample specific nature of the effective number of clusters. Because in virtually all data sets the realized values of the covariates are not identical across clusters, the effective number of clusters will be less than the number of clusters. In consequence, the effective number of clusters should be measured in virtually all studies that use cluster-robust inference.

A researcher should calculate the effective number of clusters to determine if the measure obtained from their sample is large enough to use Gaussian critical values. A natural question arises: If the effective number of clusters is not large, then how should critical values be obtained? Use of critical values from a t distribution is argued for by Kott (1994). Although his analysis does not

contain formal asymptotic results, he suggests that the degrees of freedom should be selected to mirror the variation of the cluster-robust variance estimator. In a related analysis, Imbens and Kolesar (2012) argue for the use of critical values that match the first two moments of the distribution of the variance ratio to the distribution of a χ^2 random variable. As we establish that the variation of the cluster-robust variance estimator depends on the effective number of clusters, the logical implication would be to set the degrees of freedom for the t distribution equal to the effective number of clusters.

The appeal of this approach to the problem at hand would be enhanced by the ability to bound the error introduced by use of the t distribution to approximate the finite sample distribution of Z. To understand the difficulty in constructing such a bound, consider the behavior of $\widetilde{Z} = \frac{a^{\mathrm{T}}\left(\hat{\beta}-\beta_0\right)}{\sqrt{\widetilde{V}_a}}$, which uses the (infeasible) unbiased estimator $\widetilde{V}_a$. Even under homogeneous clusters, for which $G\frac{\widetilde{V}_a}{V_a} \sim \chi^2_G$, $\widetilde{Z} \nsim t_{(G)}$ because the numerator and denominator of $\widetilde{Z}$ are correlated. The error from approximating $\widetilde{Z}$ by a t distribution is magnified under cluster heterogeneity because $G^* \frac{\widetilde{V}_a}{V_a}$ is not a $\chi^2_{(G^*)}$ random variable. A further source of approximation error is introduced by use of $\widehat{V}_a$, rather than $\widetilde{V}_a$, to construct the test statistic Z. Because it is difficult to bound the approximation error that these three sources induce, use of critical values from a $t_{(G^*)}$ distribution could lead to difficulty in controlling the size of the test.

An alternative approach is to use critical values from a re-sampling method, as Cameron, Gelbach and Miller recommend when clusters are equal in size and G is small. MacKinnon and Webb (2013) compare a re-sampling method with inference based on a $t_{(G^*)}$ distribution. They find that, over a range of simulations in which clusters are unequal in size, the re-sampling method often yields an empirical test size that is closer to the nominal level. Analytic treatment of these approaches under a full range of cluster heterogeneity remains a topic for further study.

6 Appendix

6.1 Technical Proofs

VERIFICATION OF RESULT 1: Let X_g^* be the $n \times k$ covariate matrix with all rows that do not correspond to cluster g set to zero.
Part a: The cluster specific estimator $\hat{\beta}_g$ is constructed with a generalized inverse to allow both for cluster invariant covariates and for clusters with $n_g < k$. Observe that because $X^{\mathrm{T}} y = \sum_g X_g^{*\mathrm{T}} y$,

$$\hat{\beta} = \sum_g \left(X^{\mathrm{T}} X\right)^{-1} X_g^{\mathrm{T}} X_g \left(X_g^{\mathrm{T}} X_g\right)^{-} X_g^{*\mathrm{T}} y \equiv \sum_g A_g \hat{\beta}_g, \tag{9}$$

where $\left(X_g^{\mathrm{T}} X_g\right)^{-}$ is a generalized inverse.[10] As $V \equiv Var\left(\hat{\beta}|X\right)$, the cluster representation of V in (4) follows directly from (9). To derive the cluster representation of $\widehat{V}$ in (4), note that $X_g^{\mathrm{T}} X_g = X_g^{*\mathrm{T}} X_g^* = X_g^{*\mathrm{T}} X$. Hence

$$\begin{aligned} X_g^{*\mathrm{T}} \left(y - X\hat{\beta}\right) &= \left[X_g^{*\mathrm{T}} - X_g^{*\mathrm{T}} X \left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}}\right] y \\ &= X_g^{\mathrm{T}} X_g \left[\left(X_g^{\mathrm{T}} X_g\right)^{-} X_g^{*\mathrm{T}} - \left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}}\right] y \\ &= X_g^{\mathrm{T}} X_g \left(\hat{\beta}_g - \hat{\beta}\right). \end{aligned}$$

Thus

$$A_g \left(\hat{\beta}_g - \hat{\beta}\right) = \left(X^{\mathrm{T}} X\right)^{-1} X_g^{*\mathrm{T}} \left(y - X\hat{\beta}\right) = \left(X^{\mathrm{T}} X\right)^{-1} X_g^{\mathrm{T}} \hat{u}_g, \tag{10}$$

because $\left(y - X\hat{\beta}\right) = \hat{u}$ and $X_g^{*\mathrm{T}} \hat{u} = X_g^{\mathrm{T}} \hat{u}_g$. Hence the cluster representation of $\widehat{V}$ in (4) follows directly from (10).
Part b: The estimator $\widehat{V}$ is a function of the residuals

$$y - X\left[X^{\mathrm{T}} X\right]^{-1} X^{\mathrm{T}} y = \left(I_n - \Pi_X\right) y = \hat{u},$$

where $\Pi_X = X\left[X^{\mathrm{T}} X\right]^{-1} X^{\mathrm{T}}$. These residuals can be decomposed into two components

$$\hat{u} = \left(I_n - \Pi_G + \Pi_G - \Pi_X\right) y = \left(I_n - \Pi_G\right) y + \left(\Pi_G - \Pi_X\right) y = \hat{u}_W + \hat{u}_B,$$

where $\Pi_G = \sum_g X_g^* \left[X_g^{*\mathrm{T}} X_g^*\right]^{-} X_g^{*\mathrm{T}}$ is the projection operator onto the cluster specific models. The residual component $\hat{u}_W$ captures the within cluster variation while the residual component $\hat{u}_B$ captures the between cluster variation.

[10] Because the covariate matrix may not be of full column rank within cluster g, we use the generalized inverse $\left(X_g^{\mathrm{T}} X_g\right)^{-}$ defined such that $\left(X_g^{\mathrm{T}} X_g\right)\left(X_g^{\mathrm{T}} X_g\right)^{-} X_g^{\mathrm{T}} = X_g^{\mathrm{T}}$ (Harville 1997, Theorem 12.3.4 part (5), p. 167). The generalized inverse, for which $\left(X_g^{\mathrm{T}} X_g\right)\left(X_g^{\mathrm{T}} X_g\right)^{-} X_g^{*\mathrm{T}} = X_g^{*\mathrm{T}}$ also holds, presents the issue that $\hat{\beta}_g$ is not uniquely defined, but any convenient choice of generalized inverse results in an identical variance estimator.

The quantity $\widehat{V}$ depends on the residuals through the linear function $X_g^{\mathrm{T}}\hat{u}_g = X_g^{*\mathrm{T}}\hat{u}$. Hence,

$$X_g^{\mathrm{T}}\hat{u}_g = X_g^{*\mathrm{T}}\hat{u}_W + X_g^{*\mathrm{T}}\hat{u}_B.$$

Because the least squares residuals are orthogonal to the corresponding model space,

$$\begin{aligned} X_g^{*\mathrm{T}}\hat{u}_W &= X_g^{*\mathrm{T}}\left(I_n - \Pi_G\right)y \\ &= \left(X_g^{*\mathrm{T}} - X_g^{*\mathrm{T}}\right)y = 0, \end{aligned}$$

and

$$\begin{aligned} X_g^{*\mathrm{T}}\hat{u}_B &= X_g^{*\mathrm{T}}\left(\Pi_G - \Pi_X\right)y \\ &= \left(X_g^{*\mathrm{T}} - A_g^{\mathrm{T}}X^{\mathrm{T}}\right)y \neq 0. \end{aligned}$$

Thus, $\widehat{V}$ is only a function of between cluster variation.

PROOF OF LEMMA 1: Let $Q_g := a^{\mathrm{T}}A_g\left(\hat{\beta}_g - \beta\right)$. Because the components $\left(\hat{\beta}_g - \beta\right)$ are independent across clusters, $\mathbb{E}\left(\widetilde{V}_a - V_a\right)^2 = \sum_g Var\left(Q_g^2\right)$. Let c_g be defined such that $Q_g = c_g^{\mathrm{T}}Z_g$, where $u_g = \Omega_g^{1/2}Z_g$ with $\{Z_g\}$ a sequence of uncorrelated random variables as in Assumption 1(ii). We then have

$$\begin{aligned} \mathbb{E}\left[Q_g^2\right] &= \sum_i c_{gi}^2 \\ \mathbb{E}\left[Q_g^4\right] &= \sum_i c_{gi}^4\mathbb{E}Z_{gi}^4 + 3\sum_{i\neq j} c_{gi}^2c_{gj}^2\mathbb{E}\left(Z_{gi}^2Z_{gj}^2\right) \\ &\leq M_4\sum_i c_{gi}^4 + 3\sum_{i\neq j} c_{gi}^2c_{gj}^2 \\ &= 3\left(\sum_i c_{gi}^2\right)^2 + (M_4 - 3)\sum_i c_{gi}^4. \end{aligned}$$

Thus,

$$\mathbb{E}\left[\left(\frac{\widetilde{V}_a - V_a}{V_a}\right)^2\middle| X\right] \leq \left[2\sum_g\left(\sum_i c_{gi}^2\right)^2 + (M_4 - 3)\sum_i c_{gi}^4\right]\left[\sum_{g,i} c_{gi}^2\right]^{-2}.$$

Note that $\gamma_g = \sum_i c_{gi}^2$, so that $\sum_g\left(\sum_i c_{gi}^2\right)^2 = G\overline{\gamma}^2 + \sum_g\left(\gamma_g - \overline{\gamma}\right)^2$ and $\sum_{g,i} c_{gi}^4 = \sum_g\sum_{i=1}^{n_g}\left(c_{gi}^2 - \frac{\gamma_g}{n_g}\right)^2 + \sum_g\frac{\gamma_g^2}{n_g}$, hence

$$\mathbb{E}\left[\left(\frac{\widetilde{V}_a - V_a}{V_a}\right)^2\middle| X\right] \leq \frac{1 + \Gamma\left(\Omega, X\right)}{G}\left(2 + \frac{M_4 - 3}{n^*}\right),$$

where $n^* = \frac{n}{G}\left[1+\frac{\sum_g \gamma_g^2\left(\frac{n/G-n_g}{n_g}\right)}{\sum_g \gamma_g^2}\right]^{-1}\left[1+\frac{\sum_{g,i}\left(c_{gi}^2-\gamma_g/n_g\right)^2}{\sum_g \gamma_g^2/n_g}\right]^{-1}$.

If we replace the finite fourth moment assumption with the normality assumption, then $M_4 = 3$ and

$$\mathbb{E}\left\{\left[\frac{\widetilde{V}_a - V_a}{V_a}\right]^2 \middle| X\right\} = \frac{2}{G}\left(1+\Gamma\left(\Omega, X\right)\right).$$

Q.E.D.

Proof of Lemma 2: The setting of the problem follows from the expansion

$$a^{\mathrm{T}}\left(\widehat{V}-\widetilde{V}\right)a = \sum_g a^{\mathrm{T}} A_g\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}} - 2\left(\widehat{\beta}_g-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\right]A_g^{\mathrm{T}} a.$$

We use the fact that $\sum_g A_g = I$, to introduce the matrix $\left(A_g - \frac{1}{G}I\right)$, together with the fact $\sum_g A_g\widehat{\beta}_g = \widehat{\beta}$ to obtain

$$\begin{aligned} a^{\mathrm{T}}\left(\widehat{V}-\widetilde{V}\right)a &= \frac{1}{G}a^{\mathrm{T}}\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}} a + \sum_g a^{\mathrm{T}} A_g\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g - \frac{1}{G}I\right]^{\mathrm{T}} a + \\ &\quad -\frac{2}{G}a^{\mathrm{T}}\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}} a - 2\sum_g a^{\mathrm{T}} A_g\left(\widehat{\beta}_g-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g - \frac{1}{G}I\right]^{\mathrm{T}} a. \end{aligned}$$

Combining terms on the right side yields

$$\begin{aligned} a^{\mathrm{T}}\left(\widehat{V}-\widetilde{V}\right)a &= -\frac{1}{G}a^{\mathrm{T}}\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}} a + \sum_g a^{\mathrm{T}} A_g\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g - \frac{1}{G}I\right]^{\mathrm{T}} a + \\ &\quad -2\sum_g a^{\mathrm{T}} A_g\left(\widehat{\beta}_g-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g - \frac{1}{G}I\right]^{\mathrm{T}} a. \end{aligned} \tag{11}$$

A bound for $\mathbb{E}_X\left|a^{\mathrm{T}}\left(\widehat{V}-\widetilde{V}\right)a\right|$, where $\mathbb{E}_X$ denotes expectation conditional on X, follows directly from the expansion (11) as

$$\begin{aligned} \mathbb{E}_X\left|a^{\mathrm{T}}\left(\widehat{V}-\widetilde{V}\right)a\right| &\leq \mathbb{E}_X\left|\frac{1}{G}a^{\mathrm{T}}\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}} a\right| + \\ &\quad +\mathbb{E}_X\left|\sum_g a^{\mathrm{T}} A_g\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g - \frac{1}{G}I\right]^{\mathrm{T}} a\right| + \\ &\quad +2\mathbb{E}_X\left|\sum_g a^{\mathrm{T}} A_g\left(\widehat{\beta}_g-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g - \frac{1}{G}I\right]^{\mathrm{T}} a\right|. \end{aligned} \tag{12}$$

As the first two terms on the right side are squared norms of vectors (we show details for the second term below), we can ignore the absolute value for these terms.

The first term in (12) is

$$\mathbb{E}_X\left[\frac{1}{G}a^{\mathrm{T}}\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}a\right]=\frac{V_a}{G}, \tag{13}$$

which is the magnitude of the downward bias present even when clusters are homogeneous.

For the second term in (12) first note

$$\begin{aligned}&\sum_g a^{\mathrm{T}}A_g\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]^{\mathrm{T}}a\\ =\;&\sum_g a^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]^{\mathrm{T}}a,\end{aligned}$$

where the second line follows from the fact that $\sum_g A_g = I$. Because it is the squared norm of the a vector $\mathbb{E}_X\left|\sum_g a^{\mathrm{T}}A_g\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]^{\mathrm{T}}a\right|$ equals

$$\begin{aligned}&\mathbb{E}_X\left[\sum_g a^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]^{\mathrm{T}}a\right] \qquad (14)\\ =\;&\sum_g a^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]V\left[A_g-\frac{1}{G}I\right]^{\mathrm{T}}a,\end{aligned}$$

which is an upward bias due to the heterogeneity of the covariate matrices across clusters.

For the third term in (12), we have

$$\begin{aligned}&\left|\sum_g a^{\mathrm{T}}A_g\left(\widehat{\beta}_g-\beta\right)\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]^{\mathrm{T}}a\right|\leq\sum_g\left|a^{\mathrm{T}}\left[A_g\left(\widehat{\beta}_g-\beta\right)\right]\right|\left|\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]^{\mathrm{T}}a\right|\\ &\leq\left(\sum_g\left|a^{\mathrm{T}}\left[A_g\left(\widehat{\beta}_g-\beta\right)\right]\right|^2\right)^{1/2}\left(\sum_g\left|\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]^{\mathrm{T}}a\right|^2\right)^{1/2},\end{aligned}$$

where the first inequality follows from the Triangle Inequality. Then by the Cauchy-Schwarz Inequality

$$\begin{aligned}&\mathbb{E}_X\left|\sum_g a^{\mathrm{T}}\left[A_g\left(\widehat{\beta}_g-\beta\right)\right]\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]^{\mathrm{T}}a\right|\\ \leq\;&\mathbb{E}_X\left[\left(\sum_g\left|a^{\mathrm{T}}\left[A_g\left(\widehat{\beta}_g-\beta\right)\right]\right|^2\right)^{1/2}\left(\sum_g\left|\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]^{\mathrm{T}}a\right|^2\right)^{1/2}\right]\\ \leq\;&\left(\sum_g\mathbb{E}_X\left|a^{\mathrm{T}}\left[A_g\left(\widehat{\beta}_g-\beta\right)\right]\right|^2\right)^{1/2}\left(\sum_g\mathbb{E}_X\left|\left(\widehat{\beta}-\beta\right)^{\mathrm{T}}\left[A_g-\frac{1}{G}I\right]^{\mathrm{T}}a\right|^2\right)^{1/2}.\end{aligned}$$

Let $B_g := A_g Var\left(\hat{\beta}_g \middle| X\right) A_g^{\mathrm{T}}$, so $\sum_g B_g = V$ (because $\sum_g A_g \hat{\beta}_g = \hat{\beta}$ and the $\hat{\beta}_g$ are uncorrelated. Hence

$$\mathbb{E}_X \left| a^{\mathrm{T}} \left[A_g \left(\widehat{\beta}_g - \beta \right) \right] \right|^2 = a^{\mathrm{T}} B_g a,$$

and

$$\mathbb{E}_X \left| \left(\widehat{\beta} - \beta \right)^{\mathrm{T}} \left[A_g - \frac{1}{G} I \right]^{\mathrm{T}} a \right|^2 = a^{\mathrm{T}} \left[A_g - \frac{1}{G} I \right] V \left[A_g - \frac{1}{G} I \right]^{\mathrm{T}} a.$$

Now

$$\begin{aligned} & \mathbb{E}_X \left| \sum_g a^{\mathrm{T}} \left[A_g \left(\widehat{\beta}_g - \beta \right) \right] \left(\widehat{\beta} - \beta \right)^{\mathrm{T}} \left[A_g - \frac{1}{G} I \right]^{\mathrm{T}} a \right| \qquad (15) \\ \leq \quad & \left((a^{\mathrm{T}} V a) \sum_{g=1}^{G} a^{\mathrm{T}} \left[A_g - \frac{1}{G} I \right] V \left[A_g - \frac{1}{G} I \right]^{\mathrm{T}} a \right)^{1/2}. \end{aligned}$$

From (13), (14) and (15) we have

$$\begin{aligned} \mathbb{E}\left\{ \left| \frac{\widehat{V}_a - \widetilde{V}_a}{V_a} \right| \middle| X \right\} \quad \leq \quad & \frac{1}{G} + \frac{1}{V_a} a^{\mathrm{T}} \sum_{g=1}^{G} \left[\left(A_g - \frac{1}{G} I \right) V \left(A_g - \frac{1}{G} I \right)^{\mathrm{T}} \right] a + \\ & + 2 \left(\frac{1}{V_a} a^{\mathrm{T}} \sum_{g=1}^{G} \left[\left(A_g - \frac{1}{G} I \right) V \left(A_g - \frac{1}{G} I \right)^{\mathrm{T}} \right] a \right)^{\frac{1}{2}}. \end{aligned}$$

Q.E.D.

Proof of Theorem 1: The first step is to establish that

$$T = a^{\mathrm{T}} \left(\widehat{\beta} - \beta_0 \right) / \sqrt{a^{\mathrm{T}} V a} \rightsquigarrow N(0,1).$$

We have

$$T = \sum_{g=1}^{G} D_g,$$

where $D_g | X := a^{\mathrm{T}} A_g \left(\widehat{\beta}_g - \beta_0 \right) / \sqrt{a^{\mathrm{T}} V a}$ forms a sequence of independent random variables that satisfy $\mathbb{E}(D_g | X) = 0$.

$$\mathbb{E}(D_g | X) = 0 \text{ and } Var(D_g | X) = a^{\mathrm{T}} A_g Var\left(\widehat{\beta}_g \middle| X \right) A_g^{\mathrm{T}} a.$$

For $s_G^2 := \sum_{g=1}^{G} Var(D_g | X) = a^{\mathrm{T}} V a$, under Assumption 1(ii) $\mathbb{E}\left(D_g^4\right) < \infty$, so there exists a $\delta > 0$ for which

$$\lim_{G \to \infty} \frac{1}{s_G^{2+\delta}} \sum_{g=1}^{G} \mathbb{E}\left[|D_g | X|^{2+\delta} \right] = 0,$$

hence by the Lyapunov Central Limit Theorem the distribution function of $D_g | X$ converges to a standard normal. The convergence is almost surely over X, so

$$T \rightsquigarrow N(0,1).$$

The test statistic

$$Z = T\left[\frac{a^{\mathrm{T}} V a}{a^{\mathrm{T}} \widehat{V} a}\right]^{\frac{1}{2}},$$

will converge in distribution to T if $\frac{a^{\mathrm{T}} \widehat{V} a}{a^{\mathrm{T}} V a} \xrightarrow{\mathbb{P}} 1$, by Slutsky's lemma.

It is enough to show that $\left|\frac{\widetilde{V}_a - V_a}{V_a}\right|$ and $\left|\frac{\widetilde{V}_a - \widehat{V}_a}{V_a}\right|$ are each $o_{\mathbb{P}}(1)$. Lemma 1 and Chebyshev's Inequality imply that

$$\mathbb{P}\left\{\left|\frac{\widetilde{V}_a - V_a}{V_a}\right| > \varepsilon \middle| X\right\} \leq \frac{1}{\varepsilon^2} \frac{1 + \Gamma(\Omega, X)}{G}\left(2 + \frac{M_4 - 3}{n^*}\right).$$

Under Assumption 2 (i)-(ii) the expected value of this bound goes to 0, so $\mathbb{P}\left\{\left|\frac{\widetilde{V}_a - V_a}{V_a}\right| > \varepsilon\right\} \to 0$.

In order to show that $\left|\frac{\widetilde{V}_a - \widehat{V}_a}{V_a}\right|$ is $o_{\mathbb{P}}(1)$, it is sufficent to show that $\mathbb{P}\left\{\left|\frac{\widetilde{V}_a - \widehat{V}_a}{V_a}\right| > \varepsilon \middle| X\right\} \xrightarrow{\mathbb{P}} 0$. This follows because $\mathbb{P}\left\{\left|\frac{\widetilde{V}_a - \widehat{V}_a}{V_a}\right| > \varepsilon \middle| X\right\}$ is bounded and, hence, is uniformly integrable as a function of X. Under uniform integrability, convergence in probability implies convergence in expectation so $\mathbb{P}\left\{\left|\frac{\widetilde{V}_a - \widehat{V}_a}{V_a}\right| > \varepsilon\right\} = \mathbb{E}\left[\mathbb{P}\left\{\left|\frac{\widetilde{V}_a - \widehat{V}_a}{V_a}\right| > \varepsilon \middle| X\right\}\right] \to 0$.

By Lemma 2 and Markov's inequality

$$\begin{aligned}\mathbb{P}\left\{\left|\frac{\widetilde{V}_a - \widehat{V}_a}{V_a}\right| > \varepsilon \middle| X\right\} \quad \leq \quad & \frac{1}{\varepsilon}\left(\frac{1}{G} + \frac{1}{V_a} a^{\mathrm{T}} \sum_{g=1}^{G}\left[\left(A_g - \frac{1}{G} I\right) V \left(A_g - \frac{1}{G} I\right)^{\mathrm{T}}\right] a + \right. \\ & \left. + 2\left(\frac{1}{V_a} a^{\mathrm{T}} \sum_{g=1}^{G}\left[\left(A_g - \frac{1}{G} I\right) V \left(A_g - \frac{1}{G} I\right)^{\mathrm{T}}\right] a\right)^{\frac{1}{2}}\right).\end{aligned}$$

Under Assumption 2 (i) and (iii) the bound goes to 0 in probability. As noted above, because the probability is bounded, $\mathbb{P}\left\{\left|\frac{\widetilde{V}_a - \widehat{V}_a}{V_a}\right| > \varepsilon \middle| X\right\} \xrightarrow{\mathbb{P}} 0$ implies $\mathbb{P}\left\{\left|\frac{\widetilde{V}_a - \widehat{V}_a}{V_a}\right| > \varepsilon\right\} \to 0$.

Because Ω is unknown in practice it is useful to note that this result holds for any error distribution in the set W. Over this set

$$\lim_{n \to \infty} \mathbb{P}\left\{\left|\frac{\widehat{V}_a}{V_a} - 1\right| > \varepsilon\right\} = 0,$$

which implies convergence in distribution. *Q.E.D.*

6.2 Simulation Details

We construct a sequence of 101 cluster-size designs, in which the proportion of the sample in the first cluster grows monotonically from 1 percent to 37 percent. The full description of design variation is contained in Table 1.

Table 1: Cluster-Size Designs

Design 1	$n_1 = 25$	$n_2 = \cdots = n_{100} = 25$	
Design 2	$n_1 = 34$	$n_2 = \cdots = n_{10} = 24$	$n_{11} = \cdots = n_{100} = 25$
Design 3	$n_1 = 44$	$n_2 = \cdots = n_{20} = 24$	$n_{21} = \cdots = n_{100} = 25$
Design 11	$n_1 = 124$	$n_2 = \cdots = n_{100} = 24$	
Design 12	$n_1 = 133$	$n_2 = \cdots = n_{10} = 23$	$n_{11} = \cdots = n_{100} = 25$
Design 101	$n_1 = 1015$	$n_2 = \cdots = n_{100} = 15$	

To construct the figures that display the effective number of clusters as a function of cluster size variation, for each cluster-size design the covariate matrix is simulated 1000 times.

To construct the empirical test size as a function of the effective number of clusters G^* we first generate 5 covariate matrices X from each cluster-size design simulation, yielding 505 distinct values of X (and so 505 distinct values for G^*). For each value of X we then perform the following procedure. Select the first error specification (detailed in Table 2) and simulate 1000 values of u. (Each simulated error vector u has length 2500.) The empirical test size is then the rejection probability over the 1000 data sets $\{X, u\}$ that share a common X. Repeat the procedure for error specifications 2 through 9.

Table 2: Error Specifications[11] $v_{gi} = cx_{gi} \cdot \eta_{gi}$

Specification 1a	$c = 0$	$\eta_{gi} \sim \mathcal{N}(0,1)$
Specification 1b	$c = 0.9$	$\eta_{gi} \sim \mathcal{N}(0,1)$
Specification 1c	$c = 9$	$\eta_{gi} \sim \mathcal{N}(0,1)$
Specification 2a	$c = 0$	$\eta_{gi} = \frac{1}{\sqrt{2}}\delta_{gi} \quad \delta_{gi} \sim t_{(4)}$
$\vdots$		
Specification 3c	$c = 9$	$\eta_{gi} = \frac{1}{\sqrt{4.67}}(\delta_{gi} - 1.65) \quad \delta_{gi} \sim \log\mathcal{N}(0,1)$

[11] For each specification, η_{gi} has mean 0 and variance 1.

References

[1] Cameron, A., J. Gelbach and D. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors", *Review of Economics and Statistics*, 2008, **90**, 414-427.

[2] Carter, A., K. Schnepel and D. Steigerwald, "Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity", *Economics Department Working Papers, UC Santa Barbara*, 2013, URL www.econ.ucsb.edu/~doug/researchpapers/Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity.pdf.

[3] Harville, D., *Matrix Algebra from a Statistician's Perspective*, 1997, Springer: New York.

[4] Hansen, C.,"Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large", *Journal of Econometrics*, 2007, **141**, 597-620.

[5] Hersch, J., "Compensating Differential for Gender-Specific Job Injury Risks", *American Economic Review*, 1998, **88**, 598-607.

[6] Imbens, G. and M. Kolesar, "Robust Standard Errors in Small Samples: Some Practical Advice", *Graduate School of Business Working Papers, Stanford University*, 2012.

[7] Kloek, T., "OLS Estimation in a Model where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated", *Econometrica*, 1981, **49**, 205-207.

[8] Kott, P., "A Hypothesis Test of Linear Regression Coefficients with Survey Data", *Survey Methodology*, 1994, **20**, 159-164.

[9] Krueger, A., "Experimental Estimates of Educational Production Functions", *Quarterly Journal of Economics*, 1999, **114**, 497-532.

[10] MacKinnon, J. and M. Webb, "Wild Bootstrap Inference for Wildly Different Cluster Sizes", *Economics Department Working Papers, Queens University*, 2015.

[11] Moulton, B., "Random Group Effects and the Precision of Regression Estimates", *Journal of Econometrics*, 1986, **32**, 385-397.

[12] Rogers, W., "Regression Standard Errors in Clustered Samples", *Stata Technical Bulletin 13*, 1993, **3**, 19-23.

[13] Shah, B., M. Holt and R. Folsom, "Inference About Regression Models from Sample Survey Data", *Bulletin of the International Statistical Institute Proceedings of the 41st Session*, 1977, **47:3**, 43-57.

[14] Sun, Y., "Let's Fix It: Fixed$-b$ Asymptotics versus Small-b Asymptotics in Heteroskedasticity and Autocorrelation Robust Inference", *Journal of Econometrics*, 2014, **178**, 659-677.

[15] White, H., *Asymptotic Theory for Econometricians*, 1984, Academic Press: San Diego.