When Should You Adjust Standard Errors for Clustering?

Alberto Abadie
MIT

Susan Athey
Stanford

Guido W. Imbens
Stanford

Jeffrey M. Wooldridge
MSU

September 19, 2022

Abstract

Clustered standard errors, with clusters defined by factors such as geography, are widespread in empirical research in economics and many other disciplines. Formally, clustered standard errors adjust for the correlations induced by sampling the outcome variable from a data-generating process with unobserved cluster-level components. However, the standard econometric framework for clustering leaves important questions unanswered: *(i)* Why do we adjust standard errors for clustering in some ways but not others, e.g., by state but not by gender, and in observational studies, but not in completely randomized experiments? *(ii)* Why is conventional clustering an "all-or-nothing" adjustment, while within-cluster correlations can be strong or extremely weak? *(iii)* In what settings does the choice of whether and how to cluster make a difference? We address these and other questions using a novel framework for clustered inference on average treatment effects. In addition to the common sampling component, the new framework incorporates a design component that accounts for the variability induced on the estimator by the treatment assignment mechanism. We show that, when the number of clusters in the sample is a non-negligible fraction of the number of clusters in the population, conventional cluster standard errors can be severely inflated, and propose new variance estimators that correct for this bias.

The questions addressed in this article partly originated in discussions with Gary Chamberlain. We are grateful for questions raised by Chris Blattman and seminar audiences, and for insightful comments by Colin Cameron, Vicente Guerra, four reviewers, Larry Katz, and Jesse Shapiro. Jaume Vives-i-Bastida provided expert research assistance. This work was supported by the Office of Naval Research under grants N00014-17-1-2131 and N00014-19-1-2468.

1. Introduction

Imagine you estimated the effect of attending college on labor earnings using linear regression on a cross-section of U.S. workers. How should you calculate the standard error? Empirical studies in economics often report heteroskedasticity-robust standard errors (henceforth "robust") associated with the work by Eicker [1963], Huber [1967], and White [1980]. A common alternative is to report cluster-robust standard errors (henceforth "cluster") associated with the work by Liang and Zeger [1986] and Arellano [1987], with clustering often applied within geographic units such as states or counties. Moulton [1986, 1987] and Bertrand, Duflo, and Mullainathan [2004] have shown that clustering adjustments can make a substantial difference, and since the 1980s cluster standard errors have become commonplace in empirical economics.

Later in this section, we estimate a log-linear regression of earnings on an indicator for some college using data from the 2000 U.S. Census. We find that standard errors clustered at the state level are more than 20 times larger than robust standard errors. Which ones should a researcher report? The conventional framework for clustering [see Cameron and Miller, 2015, MacKinnon, Nielsen, and Webb, 2021, for recent reviews] suggests that if the clustering adjustment matters, in the sense that the cluster standard errors are substantially larger than the robust standard errors, one should use the cluster standard errors. In this article, we develop a new framework for cluster adjustments to standard errors that nests the conventional framework as a limiting case. The new framework suggests novel standard error formulas that can substantially improve over robust and cluster standard errors in settings like the earnings regression described above.

Our proposed clustering framework differs from the standard one in that it includes a design component that accounts for between-clusters variation in treatment assignments. We argue that the new design component is important because between-cluster variation in treatment assignments often motivates the use of clustered standard errors in empirical studies [see, e.g., Gentzkow and Shapiro, 2008, Cohen and Dupas, 2010]. In addition, our framework shifts the focus of interest from features of infinite super-populations/data-

generating processes to average treatment effects defined for the finite (but potentially large) population at hand. As a result of this shift, it is the sampling process and the treatment assignment mechanism that solely determine the correct level of clustering; the presence of cluster-level unobserved components of the outcome variable becomes irrelevant for the choice of clustering level. Moreover, by focusing on finite populations (which could be entirely or substantially sampled in the data) we obtain standard errors smaller than those aiming to measure uncertainty with respect to features of infinite super-populations. We derive the large sample variances for the least squares and fixed effect estimators under our proposed framework and show that they differ in general from both the robust and the cluster variances. We also propose two estimators for the large sample variances, one analytic and one based on a re-sampling (bootstrap) approach. For the U.S. earnings application, our proposals produce standard errors that are substantially larger than the robust standard errors, but also substantially smaller than the conventional version of cluster standard errors.

We use our framework to highlight three common misconceptions surrounding clustering adjustments. The first misconception is that the need for clustering hinges on the presence of a non-zero correlation between residuals for units belonging to the same cluster. We show that the presence of such correlation does not imply the need to use cluster adjustments, and that the absence of such correlation does not imply that clustering is not required. The second misconception is that there is no harm in using clustering adjustments when they are not required, with the implication that if clustering the standard errors makes a difference, one should do so. To see that both of these claims are in fact incorrect, consider the following simple example. Suppose that, based on a random sample from the population of interest, we use the sample average of a variable to estimate its population mean. Suppose also that the population can be partitioned into clusters such as geographical units. If outcomes are positively correlated within clusters, the cluster variance will be larger than the robust variance. However, standard sampling theory directly implies that if the units are sampled randomly from the population there is no need to cluster. The harm in clustering in this case is that confidence intervals will be unnecessarily conservative, possibly by a wide margin.

A third misconception is that researchers have only two choices: either fully adjust for clustering and use the cluster standard errors, or not adjust the standard errors at all and use the robust standard errors. We show that a combination of the robust and the cluster variance estimators can substantially improve accuracy over its two components.

The new clustering framework in this article has the advantage of providing actionable guidance on a question of substantial consequence for empirical practice in econometrics: When should standard errors be clustered, and at what level? In the conventional model-based econometric framework, the researcher takes a stand on the error component structure of a model for the outcome variable. For example, suppose that, following Moulton [1986, 1987], the researcher posits a random effects model, with random effects at the state level. In this setting, a repeated sampling thought experiment entails that, for each sample, different values of the state random effects are drawn from their distributions. This model-based approach implies that if we are estimating a population mean using a sample average one needs to cluster the standard errors at the state level *even if the sample is a random sample of individuals* and not a clustered sample. A drawback of the model-based econometric framework for clustering is that empirical researchers need to take a stand on the structure of the error components of their models.

A second, closely related, framework for clustering that is often invoked in the econometrics literature is motivated by a sampling mechanism that in a first stage selects clusters at random from an infinite population, followed by a second stage of random sampling of units from the sampled clusters (or keeping all units in a cluster). Although this framework is appropriate for some applications in the analyses of surveys, where it originated [Kish, 1995, Thompson, 2012], we argue that it is not appropriate for many of the data sets economists and other social scientists analyze. In many applications in economics, researchers do observe units from *all* the clusters they are interested in, e.g., all the states in the U.S., and a framework based on randomly sampling a small fraction of a large population of clusters does not apply.

Neither of the two conventional frameworks for clustered inference described above fully

incorporates the design aspect of clustering. And it is the lack of a design component that makes them inappropriate for inference on treatment effects. To gain insight on the importance of the assignment mechanism for the standard errors of treatment effects estimators, consider a setting with individuals sampled at random from a population, but where treatment is assigned at the cluster level, with the same treatment value for all the individuals in the same cluster. Assume that the quantity of interest is the population average treatment effect. Clustered assignment to treatment is equivalent to clustered sampling of potential outcomes. Because the parameter of interest depends on averages of potential outcomes, which are sampled in a clustered manner, clustering of the standard errors is required in this setting, even when the individual observations are sampled at random. Our framework for clustered inference in this setting is close in spirit to the sampling framework described in the previous paragraph, but it incorporates a design component.

By shifting the attention from parameters of a data generating process for the outcomes to the average treatment effect for the population at hand, a researcher applying the proposals in this article does not need to take a stand on the error component structure of a model for the outcome variable to calculate standard errors. Instead, all the relevant variability of the estimator with respect to the average treatment effect is generated by the sampling mechanism, which extracts the sample from the population, and the assignment mechanism, which determines which units are exposed to the treatment. We see this as an intrinsic advantage of the framework proposed in this article in settings where it is difficult to justify a particular error component structure.

In this article we make three contributions. The first one is a novel framework for clustering, building on the one developed by Abadie et al. [2020] for the analysis of regression estimators from a design perspective. We allow for clustering both in the sampling process and in the assignment process. As a result, the framework nests both the traditional case of clustered sampling and the case of clustered treatment assignment in experiments as special cases. It also allows for intermediate cases. In particular, treatment assignment may depend on cluster but not perfectly so, and there remains variation in treatments within-clusters.

This framework clarifies the separate roles of clustering in the sampling process and clustering in the assignment process. It also clarifies what we can learn from the data about the need to adjust standard errors for clustering. In our framework, the data are *not* informative about the need to adjust for clustering in the sampling process, but they *are* informative about the need to adjust for clustering in the assignment process.

In our second contribution, we derive central limit theorems and large sample variances for the least squares and the fixed effect estimators of average treatment effects that take into account variation both from sampling and assignment. Comparing these variances to limit versions of the robust and cluster variances shows that the robust standard errors are generally too small, and the cluster standard errors are unnecessarily conservative. These comparisons also highlight how heterogeneity in treatment effects affects inference in the estimation of average treatment effects. Often researchers specify models that implicitly assume constant treatment effects without appreciating the implications for inference. We show, however, that heterogeneity in treatment effects introduces additional variance components that affect the need for clustering adjustments.

In our third contribution, we propose new variance formulas and bootstrap procedures for treatment effects estimators in the presence of clustering. We use the term Causal Cluster Variance (CCV) for the analytic variance formulas. For the case of a least squares estimator of average treatment effects, the intuition for the CCV variance formula is as follows. The error of the least squares estimator is approximately equal to a sum, over all units, of an expression involving products of regression errors and regressors values. The robust variance is approximately equal to a sum, over all units, of the squares of these products. In contrast, the conventional cluster variance estimator is approximately equal to a sum, over all clusters, of squares of within-cluster sums of the same products. Although the sum over all clusters of the expectation of the within-cluster sums of these products is zero, the expectation for each cluster separately is not. For each cluster in the sample, it is possible to estimate the expectation of the sum of the products between regression errors and regressors values. The CCV formula uses these estimates to correct the bias of the conventional cluster

variance. The CCV correction does not help much if only a small fraction of clusters are sampled. However, when a large fraction of the clusters are represented in the sample, the CCV correction can lead to substantial improvements. This adjustment relies on estimates of cluster-level treatment effects, and thus requires within-cluster variation in treatment assignment. In addition, we propose a bootstrap version of the variance estimator., which we compare to two benchmarks. In contrast to conventional bootstrap procedures, which are based on resampling individual units or entire clusters of units, our proposed Two-Stage-Cluster-Bootstrap (TSCB) conducts resampling in two stages. In the first stage, the fraction treated for each cluster is drawn from the empirical distribution of cluster-specific treatment fractions. In the second stage, the researcher samples the treated and control units from each cluster, with their number of units determined in the first stage. The CCV and TSCB variance estimators are designed for applications with large number of observations and substantial variation in treatment assignment within clusters.

To illustrate the empirical relevance of our results, we analyze a sample from the 2000 U.S. Decennial Census, which includes 2,632,838 individuals. We define 52 clusters according to residency in the 50 states, Puerto Rico, and the District of Columbia. We consider two log-linear regressions of individual earnings on a treatment variable that encodes information on college attendance. In the first specification, the treatment variable is measured as an average, at the state level. In a second specification, we measure college attendance at the individual level.

In Panel A of Table 1, we report results for a regression where the only explanatory variable is a binary treatment that takes value one if the fraction of individuals with at least some college residing in the state is 0.55 or higher, and value zero otherwise (we chose the 0.55 value to ensure sufficient variation in the treatment over the 52 clusters). Notice that the treatment is constant within states. We report the ordinary least squares (OLS) estimate, as well as robust and cluster standard errors. Since the late 1980s, it has been common practice to report cluster standard errors in settings where the regressors are constant within a cluster. Clustering at the state level makes a substantial difference relative to using robust standard

Table 1: College effects in the Census sample

Dependent variable: Log labor earnings		
Panel A		
Treatment: State indicator for share of some college greater than 0.55		
	OLS	
coefficient	0.1022	
standard error:		
robust	(0.0012)	
cluster	(0.0312)	
Panel B		
Treatment: Individual indicator for some college		
	OLS	FE
coefficient	0.4656	0.4570
standard error:		
robust	(0.0012)	(0.0012)
cluster	(0.0269)	(0.0276)
causal cluster variance (CCV)	(0.0035)	(0.0014)
two-stage cluster bootstrap (TSCB)	(0.0036)	(0.0014)

errors, with the cluster standard errors approximately twenty-six times larger than the robust standard errors.

In Panel B of Table 1, the sole regressor is an individual-level indicator for at least some college. In addition to OLS, we report the fixed effects (FE) estimate (with fixed effects for the 50 states, plus Washington DC and Puerto Rico) and robust, cluster, CCV, and TSCB standard errors in parentheses. Like for the regression of the first panel, clustering at the state level makes a substantial difference in the standard errors, with the cluster standard errors approximately twenty-three times larger than the robust standard errors, both for the OLS and the FE regressions. In Panel B, our proposed CCV and TSCB standard errors for the OLS estimate are 0.0035 and 0.0036 respectively, in between the robust standard errors (0.0012) and the cluster standard errors (0.0269), and substantially different from both. The same holds for the FE estimator. The cluster standard error is 0.0276, quite different from the robust standard errors, 0.0012. The CCV and TSCB standard errors are 0.0014, in

between robust and cluster but much closer to robust.

2. A Framework for Clustering

In this section, we describe in detail the framework for our analysis. There are multiple components to our set-up that are not explicitly modeled in the usual analysis of the variance of econometric estimators. In general, quantifying the uncertainty of parameter estimates requires describing the population and articulating the assumptions that describe how the sample was generated from that population (that is, building a model for the data generating process). In our framework, there are three distinct sources of sampling variation that lead to variation in the estimates. First, there is variation across samples in which units are observed in each cluster. Second, there is potentially variation in which clusters are observed (which leads to different units being observed). Third, there is variation in the treatment assignment across units. Whereas the standard framework for clustering focuses solely on the first two (sampling) sources of uncertainty, our proposed framework allows for all three. How much these three components matter for the variance of the least squares and fixed effects estimators of the average treatment effect depends on *(i)* the sampling process, *(ii)* the assignment process, and *(iii)* the heterogeneity in the treatment effects across clusters. To facilitate the calculation of asymptotic approximations in a range of relevant settings for empirical practice, it is convenient to formally consider a sequence of populations where we can separately control the fraction of units in the population that are sampled and the fraction of clusters in the population that is sampled, as well as the assignment mechanism.

2.1. A Sequence of Populations

We have a sequence of populations indexed by k. The k-th population has n_k units, indexed by $i = 1, \ldots, n_k$. The population is partitioned into m_k clusters. Let $m_{k,i} \in \{1, \ldots, m_k\}$ denote the cluster that unit i of population k belongs to. The number of units in cluster m of population k is $n_{k,m} \geqslant 1$. For each unit, i, there are two potential outcomes, $y_{k,i}(1)$ and $y_{k,i}(0)$, corresponding to treatment and no treatment. Thus the population is characterized by the set of triples $(m_{k,i}, y_{k,i}(0), y_{k,i}(1))$, for units $1, \ldots, n_k$ and clusters $1, \ldots, m_k$. The

object of interest is the population average treatment effect

$$\tau_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\big(y_{k,i}(1) - y_{k,i}(0)\big).$$

The population average treatment effect by cluster is

$$\tau_{k,m} = \frac{1}{n_{k,m}}\sum_{i=1}^{n_k} 1\{m_{k,i} = m\}(y_{k,i}(1) - y_{k,i}(0)).$$

Therefore,

$$\tau_k = \sum_{m=1}^{m_k} \frac{n_{k,m}}{n_k}\tau_{k,m}.$$

We assume that potential outcomes, $y_{k,i}(1)$ and $y_{k,i}(0)$, are bounded in absolute value, uniformly for all (k, i).

For each unit in the population, we define the stochastic treatment indicator, $W_{k,i} \in \{0, 1\}$. The realized outcome for unit i in population k is $Y_{k,i} = y_{k,i}(W_{k,i})$. For a random sample of the population, we observe the triple $(Y_{k,i}, W_{k,i}, m_{k,i})$. Inclusion in the sample is represented by the random variable $R_{k,i}$, which takes value one if unit i belongs to the sample, and value zero if not. We next describe the two components of the stochastic nature of the sample: the sampling process that determines the values of $R_{k,i}$, and the assignment process that determines the values of $W_{k,i}$.

2.2. The Sampling Process

The sampling process that determines the values of $R_{k,i}$ is independent of the potential outcomes and the assignments. It consists of two stages. First, clusters are sampled with cluster sampling probability $q_k \in (0, 1]$. Second, units are sampled from the subpopulation consisting of all the sampled clusters, with unit sampling probability equal to $p_k \in (0, 1]$. Both q_k and p_k may be equal to one, or close to zero. If $q_k = 1$, we sample all clusters. If $p_k = 1$, we sample all units from the sampled clusters. If $q_k = p_k = 1$, all units in the population are sampled. The standard framework for analyzing clustering focuses on the special case where $q_k \to 0$, so only a small fraction of the clusters in the population are sampled. The case $q_k = 1$ and $p_k \to 0$ corresponds to taking a relatively small random

sample of units from the population. While this is an important special case, there are also many applications where the sampled clusters comprise a large fraction of the overall set of clusters. We refer to the case of $q_k = 1$ as *random sampling* and to the case of $q_k < 1$ as *clustered sampling*.

2.3. The Assignment Process

The assignment process that determines the values of $W_{k,i}$ also consists of two stages. In the first stage of the assignment process, for cluster m in population k, an assignment probability $A_{k,m} \in [0,1]$ is drawn randomly from a distribution with mean μ_k, bounded away from zero and one uniformly in k, and variance σ_k^2, independently for each cluster. The variance σ_k^2 is key. If σ_k^2 is zero, then $A_{k,m}$ is the same for all m, and $W_{k,i}$ is randomly assigned across clusters. We refer to this case as *random assignment*. For positive values of σ_k^2 assignment probabilities depend on cluster. Because $A_{k,m}^2 \leqslant A_{k,m}$, it follows that σ_k^2 is bounded above by $\mu_k(1-\mu_k)$ and that the bound is attained when $A_{k,m}$ can only take values zero or one, so all units within a cluster have the same values for the treatment. We use the term *clustered assignment* to refer to the case $\sigma_k^2 = \mu_k(1-\mu_k)$, when there is no within-cluster variation in $W_{k,i}$. We use the term *partially clustered assignment* to refer to the case $0 < \sigma_k^2 < \mu_k(1-\mu_k)$, where assignment depends on cluster but not all units in the same cluster necessarily have the same value of $W_{k,i}$. In the second stage of the assignment process, each unit in cluster m is assigned to the treatment independently, with cluster-specific probability $A_{k,m}$.

3. The Least Squares Estimator and its Variance

Let

$$N_{k,1} = \sum_{i=1}^{n_k} R_{k,i} W_{k,i} \quad \text{and} \quad N_{k,0} = \sum_{i=1}^{n_k} R_{k,i}(1 - W_{k,i})$$

be the number of treated and untreated units in the sample, respectively; these are random variables. The total sample size is $N_k = N_{k,1} + N_{k,0}$.

We first analyze the OLS estimator of a regression of the outcome $Y_{k,i}$ on an intercept and the treatment indicator $W_{k,i}$. The OLS estimator (modified so it is well-defined even

when $N_{k,1} = 0$ or $N_{k,0} = 0$) is equal to the difference in means:

$$\widehat{\tau}_k = \frac{1}{N_{k,1} \vee 1} \sum_{i=1}^{n_k} R_{k,i} W_{k,i} Y_{k,i} - \frac{1}{N_{k,0} \vee 1} \sum_{i=1}^{n_k} R_{k,i}(1 - W_{k,i}) Y_{k,i}, \tag{1}$$

where $N_{k,1} \vee 1$ and $N_{k,0} \vee 1$ are the maxima of $N_{k,1}$ and 1 and of $N_{k,0}$ and 1, respectively.

We make the following assumptions about the sampling process and the cluster sizes: *(i)* $m_k q_k \to \infty$, *(ii)* $\liminf_{k\to\infty} p_k \min_m n_{k,m} > 0$, and *(iii)* $\limsup_{k\to\infty} \max_m n_{k,m} / \min_m n_{k,m} < \infty$. The first assumption implies that the expected number of sampled clusters goes to infinity as k increases. The second assumption implies that the average number of observations sampled per cluster, conditional on the cluster being sampled, does not go to zero. The third assumption restricts the imbalance between the number of units across clusters. Notice that assumptions *(i)* and *(ii)* imply $n_k p_k q_k \to \infty$, so the sample size becomes larger in expectation as k increases.

3.1. Large k Distribution of the Least Squares Estimator

Our first main result derives the large k distribution of $\widehat{\tau}_k$. Let $\alpha_k = (1/n_k) \sum_{i=1}^{n_k} y_{k,i}(0)$, $u_{k,i}(1) = y_{k,i}(1) - (\alpha_k + \tau_k)$, and $u_{k,i}(0) = y_{k,i}(0) - \alpha_k$. Under additional regularity conditions in the Appendix,

$$\sqrt{N_k}(\widehat{\tau}_k - \tau_k) / v_k^{1/2} \xrightarrow{d} N(0,1),$$

where

$$\begin{aligned} v_k = & \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\frac{u_{k,i}^2(1)}{\mu_k} + \frac{u_{k,i}^2(0)}{1-\mu_k} \right) \\ & - p_k \frac{1}{n_k} \sum_{i=1}^{n_k} \big(u_{k,i}(1) - u_{k,i}(0)\big)^2 - p_k \sigma_k^2 \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k} \right)^2 \\ & + p_k(1-q_k) \frac{1}{n_k} \sum_{m=1}^{m_k} \left(\sum_{i=1}^{n_k} 1\{m_{k,i} = m\} \big(u_{k,i}(1) - u_{k,i}(0)\big) \right)^2 \\ & + p_k \sigma_k^2 \frac{1}{n_k} \sum_{m=1}^{m_k} \left(\sum_{i=1}^{n_k} 1\{m_{k,i} = m\} \left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k} \right) \right)^2. \end{aligned} \tag{2}$$

The expression for the variance v_k has multiple terms that make its interpretation challenging. We first interpret v_k in some special cases to highlight the implications of clustered

sampling and clustered assignment. In Section 3.3, we compare v_k to the large-k form of the robust and cluster variance estimators.

For the case of random sampling ($q_k = 1$) and random assignment ($\sigma_k^2 = 0$), the variance simplifies to

$$\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}^2(1)}{\mu_k} + \frac{u_{k,i}^2(0)}{1-\mu_k}\right) - p_k\frac{1}{n_k}\sum_{i=1}^{n_k}\big(u_{k,i}(1) - u_{k,i}(0)\big)^2.$$

As we show in Section 3.2 below, the first term in this variance is estimated by the robust variance estimator. The second term is a finite sample correction that is familiar from the literature on randomized experiments [e.g., Neyman, 1923/1990, Imbens and Rubin, 2015, Abadie et al., 2020]. This finite sample correction vanishes if there is either no heterogeneity in the treatment effects (so $u_{k,i}(1) - u_{k,i}(0) = y_{k,i}(1) - y_{k,i}(0) - \tau_k = 0$), or if the sample is a small fraction of the population ($p_k \approx 0$).

Adding clustered sampling, $q_k < 1$, increases the variance by

$$p_k(1-q_k)\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k} 1\{m_{k,i} = m\}\big(u_{k,i}(1) - u_{k,i}(0)\big)\right)^2,$$

which is the same as

$$p_k(1-q_k)\frac{1}{n_k}\sum_{m=1}^{m_k} n_{k,m}^2(\tau_{k,m} - \tau_k)^2.$$

This term vanishes if there is no heterogeneity in the average treatment effect across clusters. Although the sample is informative about heterogeneity in cluster average treatment effects, it is not informative about the value of q_k. Information about the need to adjust for clustered sampling ($q_k < 1$) must come from outside the sample.

Clustered assignment, $\sigma_k^2 > 0$, adds two terms to the variance,

$$-p_k\sigma_k^2\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k}\right)^2 + p_k\sigma_k^2\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k} 1\{m_{k,i} = m\}\left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k}\right)\right)^2.$$

As we explain in more detail in section 3.3, the sign of this expression depends on the amount of variation in potential outcomes that can be explained by the clusters. Note that in contrast to the lack of sample information about the need to adjust for clustered sampling, the sample is potentially informative about the need to account for clustered assignment.

The five terms making up the asymptotic variance v_k can be of different order. The first term is an average of bounded terms, and so under our assumptions will be of order $\mathcal{O}(1)$. The second and third terms will be at most of the same order as the first one. If $p_k \approx 0$ so we can think of the sample as small relative to the population of sampled clusters, the first term dominates the second and third terms. If cluster sizes are bounded as k increases, the fourth and fifth terms in are also order $\mathcal{O}(1)$. If, on the other hand, cluster sizes increase with k, these terms can be of higher order and dominate the variance. Whether they do so or not depends on the *(i)* magnitude of p_k, *(ii)* presence of clustering in sampling, *(iii)* presence of clustering in assignment, and *(iv)* heterogeneity in potential outcomes.

3.2. The Robust and Cluster Robust Variance Estimators

Let $\widehat{U}_{k,i} = Y_{k,i} - \widehat{\alpha}_k - \widehat{\tau}_k W_{k,i}$ be the residuals from the regression of $Y_{k,i}$ or a constant and $W_{k,i}$. Here, $\widehat{\alpha}_k$ is the intercept of the regression and $\widehat{\tau}_k$ is the coefficient on $W_{k,i}$ (equal to the expression in (1) with probability approaching one).

There are two common estimators of the variance of $\sqrt{N_k}(\widehat{\tau}_k - \tau_k)$. First, the conventional robust variance estimator (Eicker [1963], Huber [1967], White [1980]):

$$\widehat{V}_k^{\text{robust}} = \frac{1}{\overline{W}_k^2(1-\overline{W}_k)^2}\left\{\frac{1}{N_k}\sum_{i=1}^{n_k} R_{k,i}\widehat{U}_{k,i}^2(W_{k,i}-\overline{W}_k)^2\right\}, \tag{3}$$

where

$$\overline{W}_k = \frac{1}{N_k \vee 1}\sum_{i=1}^{n_k} R_{k,i}W_{k,i}.$$

Let

$$v_k^{\text{robust}} = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}^2(1)}{\mu_k} + \frac{u_{k,i}^2(0)}{1-\mu_k}\right).$$

Under regularity conditions (see appendix), $\widehat{V}_k^{\text{robust}}$ and v_k^{robust} are close in the following sense,

$$\frac{\widehat{V}_k^{\text{robust}}}{v_k} = \frac{v_k^{\text{robust}}}{v_k} + o_p(1),$$

motivating our focus on the comparison of v_k^{robust} and v_k. In general the difference $v_k^{\text{robust}} - v_k$ can be positive or negative, so the robust variance estimator can be invalid in large samples.

The second common variance estimator is the cluster variance [Liang and Zeger, 1986, Arellano, 1987],

$$\widehat{V}_k^{\text{cluster}} = \frac{1}{\overline{W}_k^2(1-\overline{W}_k)^2}\left\{\frac{1}{N_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k} 1\{m_{k,i}=m\}R_{k,i}\widehat{U}_{k,i}(W_{k,i}-\overline{W}_k)\right)^2\right\}. \tag{4}$$

Define

$$\begin{aligned} v_k^{\text{cluster}} &= \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}^2(1)}{\mu_k}+\frac{u_{k,i}^2(0)}{1-\mu_k}\right) \\ &\quad - p_k\frac{1}{n_k}\sum_{i=1}^{n_k}\big(u_{k,i}(1)-u_{k,i}(0)\big)^2 - p_k\sigma_k^2\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}(1)}{\mu_k}+\frac{u_{k,i}(0)}{1-\mu_k}\right)^2 \\ &\quad + p_k\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}1\{m_{k,i}=m\}\big(u_{k,i}(1)-u_{k,i}(0)\big)\right)^2 \\ &\quad + p_k\sigma_k^2\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}1\{m_{k,i}=m\}\left(\frac{u_{k,i}(1)}{\mu_k}+\frac{u_{k,i}(0)}{1-\mu_k}\right)\right)^2. \end{aligned}$$

Then, $\widehat{V}_k^{\text{cluster}}$ is close to v_k^{cluster} in the sense that

$$\frac{\widehat{V}_k^{\text{cluster}}}{v_k} = \frac{v_k^{\text{cluster}}}{v_k} + o_p(1).$$

The difference $v_k^{\text{cluster}} - v_k$ is always nonnegative. Therefore, for large k, the cluster variance estimator can be conservative but cannot underestimate the variance of $\widehat{\tau}_k$.

3.3. Discussion

From the formulas for v_k, v_k^{robust}, and v_k^{cluster} it follows that if p_k is small enough, then v_k^{robust} and v_k^{cluster} are approximately equal to v_k. In this case, clustered sampling and clustered assignment do not matter much because the probability that two sample units belong to the same cluster is small.

The difference $v_k^{\text{robust}} - v_k$ depends on two terms. The first term,

$$p_k\frac{1}{n_k}\left[\sum_{i=1}^{n_k}\big(u_{k,i}(1)-u_{k,i}(0)\big)^2 - (1-q_k)\sum_{m=1}^{m_k} n_{k,m}^2(\tau_{k,m}-\tau_k)^2\right], \tag{5}$$

is equal to zero when treatment effects are constant (in which case, $u_{k,i}(1)-u_{k,i}(0)=0$ for $i=1,\ldots,n_k$ and $\tau_{k,m}-\tau_k=0$ for all $m=1,\ldots,m_k$). If all clusters are sampled, so $q_k=1$,

and treatment effects are heterogeneous, (5) is positive. When only a fraction of the clusters are sampled, $q_k < 1$, the sign of (5) depends on the extent to which heterogeneity in treatment effects can be explained by the clusters. If there is no variation in average treatment effects across clusters, the expression in (5) is non-negative. However, when clusters explain much of the variation in treatment effects, the expression in (5) can be negative and very large in magnitude because of the factor $n_{k,m}^2$. The second term of $v_k^{\text{robust}} - v_k$ is equal to

$$p_k \sigma_k^2 \sum_{m=1}^{m_k} \frac{n_{k,m}}{n_k} \Bigg[\frac{1}{n_{k,m}} \sum_{i=1}^{n_k} 1\{m_{k,i} = m\} \left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k} \right)^2 - n_{k,m} \left(\frac{1}{n_{k,m}} \sum_{i=1}^{n_k} 1\{m_{k,i} = m\} \left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k} \right) \right)^2 \Bigg]. \quad (6)$$

This term is equal to zero if there is no clustered assignment, that is, $\sigma_k^2 = 0$. If $\sigma_k^2 > 0$, the sign of (6) depends on how much of the heterogeneity in potential outcomes is explained by the clusters. The expression in (6) is close to zero when there is little heterogeneity in potential outcomes, so $u_{k,i}(1)$ and $u_{k,i}(0)$ are close to zero. If there is heterogeneity in potential outcomes but average potential outcomes are nearly constant across clusters, (6) is positive. When the clusters explain enough heterogeneity in potential outcomes (6) can be negative and potentially very large in magnitude because of the factor $n_{k,m}$ multiplying the second term of the sum in (6). That is, the robust variance formula can severely underestimate the variance of $\widehat{\tau}_k$.

Clustered standard errors are conservative in general, that is, $v_k^{\text{cluster}} \geqslant v_k$. In particular, the difference $v_k^{\text{cluster}} - v_k$ is

$$v_k^{\text{cluster}} - v_k = p_k q_k \frac{1}{n_k} \sum_{m=1}^{m_k} \left(\sum_{i=1}^{n_k} 1\{m_{k,i} = m\} \big(u_{k,i}(1) - u_{k,i}(0) \big) \right)^2,$$

which can be rewritten as

$$v_k^{\text{cluster}} - v_k = \left(\frac{p_k n_k}{m_k} \right) q_k \left\{ \frac{1}{m_k} \sum_{m=1}^{m_k} \left(\frac{n_{k,m} m_k}{n_k} \right)^2 (\tau_{k,m} - \tau_k)^2 \right\}. \quad (7)$$

When the expected fraction of clusters in the sample, q_k, is small, or when the average treatment effect is nearly constant between clusters, then $v_k^{\text{cluster}} \approx v_k$. Aside from these

special cases, the $p_k n_k/m_k$ factor in the formula above indicates that cluster standard errors can be extremely conservative in general.

4. Two New Variance Estimators

Estimation of the variance of $\widehat{\tau}_k$ is challenging because the different terms in v_k can be of different orders of magnitude. In this section, we propose two estimators of the variance of $\widehat{\tau}_k$ that allow us to correct the bias of the cluster variance estimator, one analytic, and one resampling-based. As the expression for the bias of the cluster variance in (7) shows, the cluster variance is heavily biased if the fraction of the sampled clusters is large and there is substantial variation in the cluster-specific treatment effects. Although the proposed analytic variance estimator is defined irrespective of the value of σ_k^2, in order to for the correction to be effective we need to be able to estimate the cluster-specific treatment effects, and thus we need σ_k^2 to be less than its maximum value of $\mu_k(1-\mu_k)$ to ensure that there is variation in the treatment assignment within clusters. One of the proposed variance estimators is based on a correction to $\widehat{V}_k^{\text{cluster}}$, and the other is based on resampling methods. An alternative would be to directly estimate the bias term in (7) and subtract that from the cluster variance. A challenge with this approach is that the estimation error for the adjustment term is large (often leading to negative variances estimates) because the order of magnitude of the correction is itself large and this approach did not work well in our simulations. We do not report formal results for the variance estimators in the current paper. We demonstrate their performance in the simulations in Section 6. There may well be further refinements possible.

If q_k is close to zero, the proposed variance estimators are close to $\widehat{V}_k^{\text{cluster}}$, which has little bias in that case. If $\sigma_k^2 = \mu_k(1-\mu_k)$ (that is, when $W_{k,i}$ is constant within clusters), the proposed resampling variance estimator is not defined. To be effective both variance estimators rely on estimating the variation in treatment effects across clusters, and therefore require a substantial number of both treated and control observations per cluster. The proposed variance estimators lead to substantial improvements over $\widehat{V}_k^{\text{cluster}}$ in cases where $\widehat{V}_k^{\text{cluster}}$ has a large upward bias. The downside of the proposed variance estimators is that they can be conservative when there is no need to cluster because there is no heterogeneity

in treatment effects, or when there are too few treated and control observations per cluster to estimate the heterogeneity in the treatment effects precisely.

We first consider in Section 4.1 the case with $q_k = 1$ so we have random sampling. Next we consider in Section 4.2 the case with clustered sampling $q_k < 1$. In Section 4.3 we propose a bootstrap procedure for estimating the variance. The proposed variance estimators perform very well in the simulation study of Section 6. The derivation of their formal properties is left for future work.

4.1. The Case with All Clusters Observed

First we focus on the case with $q_k = 1$ (all clusters observed), but allowing for general p_k. Let $U_{k,i} = W_{k,i}u_{k,i}(1) + (1 - W_{k,i})u_{k,i}(0)$. The first step is to approximate the normalized error of the least squares estimator $\widehat{\tau}_k$ by a normalized sample average over clusters,

$$\sqrt{N_k}(\widehat{\tau}_k - \tau_k)/v_k^{1/2} = \frac{1}{\sqrt{n_k p_k v_k}\mu_k(1-\mu_k)} \sum_{m=1}^{m_k} C_{k,m} + o_p(1), \tag{8}$$

where the terms

$$C_{k,m} = \sum_{i=1}^{n_k} 1\{m_{k,i} = m\} R_{k,i}(W_{k,i} - \mu_k)U_{k,i}$$

are independent across clusters. In the appendix, we show

$$\widehat{V}_k^{\text{cluster}}/v_k = \frac{1}{n_k p_k v_k}\left(\frac{1}{\mu_k(1-\mu_k)}\right)^2 \sum_{m=1}^{m_k} C_{k,m}^2 + o_p(1). \tag{9}$$

The expectation of $C_{m,k}$ is

$$E[C_{k,m}] = n_{k,m} p_k \mu_k(1-\mu_k)(\tau_{k,m} - \tau_k),$$

with sum over clusters

$$\sum_{m=1}^{m_k} E[C_{k,m}] = p_k\mu_k(1-\mu_k) \sum_{m=1}^{m_k} n_{k,m}(\tau_{k,m} - \tau_k) = 0. \tag{10}$$

That is, although the sum of the expectations of $C_{k,m}$ over clusters is equal to zero, these expectations are not equal to zero in general for each cluster separately. Because $\text{var}(C_{k,m}) \leqslant E[C_{k,m}^2]$, the first term on the right-hand side of (9) is conservative on expectation relative to the variance of $\sqrt{N_k}(\widehat{\tau}_k - \tau_k)/v_k^{1/2}$, which explains the conservativeness of $\widehat{V}_k^{\text{cluster}}$.

Because of (10), we can replace the terms $C_{k,m}$ in (8) by $C_{k,m} - E[C_{k,m}] = C_{k,m,1} + C_{k,m,2}$, where

$$C_{k,m,1} = \sum_{i=1}^{n_k} 1\{m_{k,i} = m\}(R_{k,i} - p_k)(\tau_{k,m} - \tau_k)\mu_k(1-\mu_k),$$

and

$$C_{k,m,2} = \sum_{i=1}^{n_k} 1\{m_{k,i} = m\}R_{k,i}\Big((W_{k,i} - \mu_k)U_{k,i} - (\tau_{k,m} - \tau_k)\mu_k(1-\mu_k)\Big).$$

Therefore,

$$\sqrt{N_k}(\widehat{\tau}_k - \tau_k)/v_k^{1/2} = \frac{1}{\sqrt{n_k p_k v_k}\mu_k(1-\mu_k)}\left(\sum_{m=1}^{m_k} C_{k,m,1} + \sum_{m=1}^{m_k} C_{k,m,2}\right) + o_p(1). \tag{11}$$

It can be shown that $C_{k,m,1}$ and $C_{k,m,2}$ have means equal to zero and are uncorrelated. In addition, $C_{k,m,1}$ and $C_{k,m,2}$ are uncorrelated across clusters. The variance of $\sum_{m=1}^{m_k} C_{k,m,1}/(\sqrt{n_k p_k}\mu_k(1-\mu_k))$ is

$$(1-p_k)\sum_{m=1}^{m_k}\frac{n_{k,m}}{n_k}(\tau_{k,m} - \tau_k)^2.$$

Let $\widehat{\tau}_{k,m}$ be difference between the sample average of the outcome for treated and nontreated units in cluster m. A direct estimator the variance of $\sum_{m=1}^{m_k} C_{k,m,2}$ is

$$\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k} 1\{m_{k,i} = m\}R_{k,i}\Big((W_{k,i} - \overline{W}_k)\widehat{U}_{k,i} - (\widehat{\tau}_{k,m} - \widehat{\tau}_k)\overline{W}_k(1-\overline{W}_k)\Big)\right)^2, \tag{12}$$

In practice, the estimator in (12) is biased from the correlations between the estimation errors of its components. We apply sampling splitting to address this bias. We first split the sample randomly into two subsamples. Let $Z_{k,i} \in \{0,1\}$ be the indicator that unit i belongs to the second subsample, and let $\overline{Z}_k$ be the mean of $Z_{k,i}$. Using the subsample with $Z_{k,i} = 0$, we obtain estimates $\widehat{\tau}^*_{k,m}$, $\widehat{\alpha}^*_k$, and $\widehat{\tau}^*_k$ of $\tau_{k,m}$, α_k, and τ_k, respectively. Next, for observations with $Z_{k,i} = 1$, we calculate the residuals $\widehat{U}^*_{k,i} = Y_{k,i} - \widehat{\alpha}^*_k - \widehat{\tau}^*_k W_{k,i}$. Finally, we estimate the normalized variance for the case with $q_k = 1$ as

$$\widehat{V}_k^{\mathrm{CCV}}(1) = \frac{1}{N_k\overline{W}_k^2(1-\overline{W}_k)^2}\sum_{m=1}^{m_k}\left[\frac{1}{\overline{Z}_k^2}\left(\sum_{i=1}^{n_k} 1\{m_{k,i} = m\}R_{k,i}Z_{k,i}\Big((W_{k,i} - \overline{W}_k)\widehat{U}^*_{k,i} - (\widehat{\tau}^*_{k,m} - \widehat{\tau}^*_k)\overline{W}_k(1-\overline{W}_k)\Big)\right)^2\right.$$

$$- \frac{1-\overline{Z}_k}{\overline{Z}_k^2} \sum_{i=1}^{n_k} 1\{m_{k,i}=m\} R_{k,i} Z_{k,i} \Big((W_{k,i} - \overline{W}_k)\widehat{U}^*_{k,i} - (\widehat{\tau}^*_{k,m} - \widehat{\tau}^*_k)\overline{W}_k(1-\overline{W}_k) \Big)^2 \Bigg]$$
$$+ (1-p_k) \sum_{m=1}^{m_k} \frac{\overline{N}_{k,m}}{N_k} (\widehat{\tau}_{k,m} - \widehat{\tau}_k)^2, \tag{13}$$

where $\overline{N}_{k,m}$ is the size of the sample in cluster m. For clusters with no variation in the treatment variable, we replace $\widehat{\tau}_{k,m}$ in (13) with $\widehat{\tau}_k$. For clusters with no variation in the treatment variable for a particular subsample, we replace $\widehat{\tau}^*_{k,m}$ in (13) with $\widehat{\tau}^*_k$. We derive the form of the CCV estimator in the appendix. To improve the precision of $\widehat{V}_k^{\text{CCV}}(1)$, we re-estimate it multiple times with new sample splits (new values for $Z_{k,i}$) and then average the corresponding variance estimators. In our simulations of section 6, we re-estimate the variance estimator four times, and use sample splits with in expectation an equal number of units in each subsample, so $E[\overline{Z}_k] = 1/2$.

4.2. The Case When Not All Clusters Are Sampled

To motivate the modification of the variance estimator $\widehat{V}_k^{\text{CCV}}(1)$ for the $q_k < 1$ case, notice that

$$v_k(q_k) - v_k^{\text{cluster}} = q_k \times (v_k(1) - v_k^{\text{cluster}}),$$

where $v_k(q_k)$ denotes the value of the true variance v_k evaluated at q_k. That is, the variance for the general q_k case is a convex combination of the true variance at $q_k = 1$ and the cluster variance,

$$v_k(q_k) = q_k \times v_k(1) + (1-q_k) \times v_k^{\text{cluster}}.$$

Let $\widehat{q}_k$ be the ratio between the number of sampled clusters and the total number of clusters in the population. The proposed variance estimator, $\widehat{V}_k^{\text{CCV}}$, is a convex combination of $\widehat{V}_k^{\text{CCV}}(1)$ and $\widehat{V}_k^{\text{cluster}}$ with weights $\widehat{q}_k$ and $1-\widehat{q}_k$,

$$\widehat{V}_k^{\text{CCV}} = \widehat{q}_k \times \widehat{V}_k^{\text{CCV}}(1) + (1-\widehat{q}_k) \times \widehat{V}_k^{\text{cluster}}. \tag{14}$$

Computation of $\widehat{q}_k$ requires knowledge of m_k, the total number of clusters in the population.

4.3. A Bootstrap Variance Estimator

In the previous sections, we have discussed an analytic variance estimator. Here we suggest a resampling-based variance estimator, initially for the case with $q_k = 1$. Like the causal bootstrap in Imbens and Menzel [2021], the proposed bootstrap procedure takes into account the causal nature of the estimand and creates bootstrap samples where units (in this case clusters) have different assignments and assignment probabilities than they have in the original sample. It differs from earlier bootstrap variance estimators for clustered settings [e.g., Cameron and Miller, 2015, Menzel, 2021] in that it allows for the possibility that a large fraction of clusters are observed.

The specific resampling procedure, which we call the two-stage-cluster-bootstrap (TSCB), consists of two stages. For each of the clusters, let $\overline{N}_{k,m}$ be the cluster-level sample size and $\overline{W}_{k,m} = N_{k,m,1}/(\overline{N}_{k,m} \vee 1)$ the cluster-level fraction of treated units. In the first stage of the bootstrap procedure, for each cluster we draw $\overline{W}^b_{k,m}$ with replacement from the empirical distribution of the cluster-level fractions of treated units, that is with probability $1/m_k$ from the set $\{\overline{W}_{k,1}, \ldots, \overline{W}_{k,m_k}\}$. In the second stage, we draw $\overline{N}_{k,m}\overline{W}^b_{k,m}$ units with replacement from the set of treated units in cluster m and $\overline{N}_{k,m}(1 - \overline{W}^b_{k,m})$ units with replacement from the set of untreated units in cluster m. In order for this to be well-defined we do need all the $\overline{W}_{k,1}$ to be strictly between zero and one. We do this for all clusters to create the bootstrap sample, and calculate the bootstrap standard errors as the standard deviation of the treatment effect estimates across bootstrap iterations.

Next, consider the case with $q_k < 1$. In this case, we need to take into account the fact that we see a fraction of the clusters in the population. We follow the approach proposed in Chao and Lo [1985]. Suppose $q = 1/2$, so we observe half the clusters in the population. The bootstrap procedure first creates a pseudo population consisting of the original population of clusters, plus one additional replica of each cluster. Then, to get a bootstrap sample, we sample randomly, without replacement, from the clusters in this pseudo population. Given the clusters in the bootstrap sample, we proceed as before, and ultimately calculate the bootstrap variance as the variance of the estimator over the bootstrap samples. Chao and

Lo [1985] provide details and extensions to the case for the case where $1/q_k$ is not an integer.

The algorithm for the TSCB is summarized here.

Algorithm 1 Two Stage Cluster Bootstrap

Input:

Sample $(Y_{k,i}, W_{k,i}, m_{k,i})$

Fraction sampled clusters q_k

Number of bootstrap replications B

Stage 1:

1a: Create pseudo population by replicating each cluster $1/q_k$ times

1b: For each cluster in the pseudo population, calculate the assignment probability $\overline{W}_{k,m}$

1c: Create a bootstrap sample of clusters by randomly drawing clusters from the pseudo population from Stage 1a, where cluster k is sampled with probability q_k

1d: For each sampled cluster, draw an assignment probability $A_{k,m}$ from the empirical distribution of the $\overline{W}_{k,m}$ from Stage 1b

Stage 2:

2a: Randomly draw from the set of treated units in cluster m, $\lfloor N_{k,m} A_{k,m} \rfloor$ units

2b: Randomly draw from the set of control units in cluster m, $\lfloor N_{k,m}(1 - A_{k,m}) \rfloor$ units

Calculations:

For the units in the bootstrap sample constructed in Stage 2, collect the values for $(Y_{k,i}, W_{k,i}, m_{k,i})$ and calculate the least squares or fixed effect estimator

Calculate the standard deviation of the least squares or fixed effect estimator over the B bootstrap samples

5. The Fixed Effect Estimator

In this section, we report results for the fixed effect estimator often used in empirical research in economics. Arellano [1987], Bertrand, Duflo, and Mullainathan [2004], Cameron and Miller [2015] and MacKinnon, Nielsen, and Webb [2021] have pointed out that cluster adjustments may still be necessary in fixed effects regressions. However, a view of clustering based on models with cluster-specific variance components creates ambiguity in the role of clustered standard errors for estimators with cluster fixed effects, which are specifically aimed to absorb cluster-level variation.

We first characterize the fixed effect estimator and derive its large k distribution. Then, we discuss the properties of the two conventional variance estimators, the robust and cluster

robust variance estimators. As in the least squares case, we find that the robust standard errors may be too small and the cluster standard errors may be unnecessarily large, especially in cases when the number of observations per cluster is large. We propose CCV and TSCB variance estimators. The CCV estimator for fixed effects has a different form than the one for least squares in section 4.

The fixed effect estimator is based on a regression of the outcome on the treatment indicator and indicators for each of the clusters in the sample. It can be written as the least squares estimate for a regression of the outcome on the treatment, with both variables measured in deviation from cluster means,

$$\widehat{\tau}_k^{\text{fixed}} = \frac{\sum_{m=1}^{m_k}\sum_{i=1}^{n_k} 1\{m_{k,i}=m\} R_{k,i} Y_{k,i}(W_{k,i} - \overline{W}_{k,m})}{\sum_{m=1}^{m_k}\sum_{i=1}^{n_k} 1\{m_{k,i}=m\} R_{k,i} W_{k,i}(W_{k,i} - \overline{W}_{k,m})}. \tag{15}$$

Like in section 3, we assume that that potential outcomes are bounded, $m_k q_k \to \infty$, and $\limsup_{k\to\infty} \max_m n_{k,m}/\min_m n_{k,m} < \infty$. In addition, we assume *(i)* $(m_k q_k)/((p_k n_k)/m_k) \to 0$, and *(ii)* the supports of the cluster probabilities, $A_{k,m}$, are bounded away from zero and one (uniformly in k and m). Assumption *(i)* restricts the focus of our analysis in this section to settings where the expected number of sampled clusters is small relative to the expected number of sampled observations per sampled cluster. Together with the previous assumptions, assumption *(i)* implies $(p_k n_k)/m_k \to \infty$, $n_k p_k q_k \to \infty$, and $p_k \min_m n_{k,m} \to \infty$. This last result, along with assumption *(ii)*, ensures that $\widehat{\tau}_k^{\text{fixed}}$ in (15) is well-defined with probability approaching one.

Let $\alpha_{k,m} = (1/n_{k,m}) \sum_{i=1}^{n_k} 1\{m_{k,i} = m\} y_{k,i}(0)$. For an observation, i, with $m_{k,i} = m$, we define the within-cluster residuals $e_{k,i}(0) = y_{k,i}(0) - \alpha_{k,m}$ and $e_{k,i}(1) = y_{k,i}(1) - \tau_{k,m} - \alpha_{k,m}$. Let

$$\tilde{v}_k = f_k/(\mu_k(1-\mu_k) - \sigma_k^2)^2 \tag{16}$$

where

$$f_k = E[A_{k,m}(1-A_{k,m})^2]\frac{1}{n_k}\sum_{i=1}^{n_k} e_{k,i}^2(1) + E[A_{k,m}^2(1-A_{k,m})]\frac{1}{n_k}\sum_{i=1}^{n_k} e_{k,i}^2(0)$$

$$
\begin{aligned}
&- p_k E[A_{k,m}^2(1-A_{k,m})^2]\frac{1}{n_k}\sum_{i=1}^{n_k}(e_{k,i}(1)-e_{k,i}(0))^2 \\
&+ \Big(E[A_{k,m}(1-A_{k,m})] - (5+p_k)E[A_{k,m}^2(1-A_{k,m})^2] \\
&\qquad + 2q_k(E[A_{k,m}(1-A_{k,m})])^2\Big)\sum_{m=1}^{m_k}\frac{n_{k,m}}{n_k}(\tau_{k,m}-\tau_k)^2 \\
&+ \Big(p_k E[A_{k,m}^2(1-A_{k,m})^2] - p_k q_k(E[A_{k,m}(1-A_{k,m})])^2\Big)\sum_{m=1}^{m_k}\frac{n_{k,m}^2}{n_k}(\tau_{k,m}-\tau_k)^2.
\end{aligned}
$$

Under additional regularity conditions, which are described in the Appendix, we obtain the large k distribution of the fixed effects estimator,

$$
\sqrt{N_k}(\widehat{\tau}_k^{\text{fixed}}-\tau_k)/\tilde{v}_k^{1/2} \xrightarrow{d} N(0,1). \tag{17}
$$

Let $\widetilde{U}_{k,i} = \widetilde{Y}_{k,i} - \widehat{\tau}_k^{\text{fixed}}\widetilde{W}_{k,i}$, where $\widetilde{Y}_{k,i} = Y_{k,i} - \overline{Y}_{k,m_{k,i}}$, $\widetilde{W}_{k,i} = (W_{k,i} - \overline{W}_{k,m_{k,i}})$. The robust estimator of the variance of $\sqrt{N_k}(\widehat{\tau}_k^{\text{fixed}}-\tau_k)$ is

$$
\widetilde{V}_k^{\text{robust}} = \frac{1}{N_k}\sum_{i=1}^{n_k}R_{k,i}\widetilde{W}_{k,i}^2\widetilde{U}_{k,i}^2 \Bigg/ \left(\frac{1}{N_k}\sum_{i=1}^{n_k}R_{k,i}\widetilde{W}_{k,i}^2\right)^2. \tag{18}
$$

Now let,

$$
\tilde{v}_k^{\text{robust}} = f_k^{\text{robust}}/(\mu_k(1-\mu_k)-\sigma_k^2)^2.
$$

with

$$
\begin{aligned}
f_k^{\text{robust}} &= E[A_{k,m}(1-A_{k,m})^2]\frac{1}{n_k}\sum_{i=1}^{n_k}e_{k,i}^2(1) + E[A_{k,m}^2(1-A_{k,m})]\frac{1}{n_k}\sum_{i=1}^{n_k}e_{k,i}^2(0) \\
&\quad + E[A_{k,m}(1-A_{k,m})(1-3A_{k,m}(1-A_{k,m}))]\sum_{m=1}^{m_k}\frac{n_{k,m}}{n_k}(\tau_{k,m}-\tau_k)^2.
\end{aligned}
$$

Notice that all terms of f_k^{robust} are bounded. In the appendix, we show that

$$
\widetilde{V}_k^{\text{robust}} = \tilde{v}_k^{\text{robust}} + o_p(1).
$$

The cluster variance estimator for fixed effects is

$$
\widetilde{V}_k^{\text{cluster}} = \frac{1}{N_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}1\{m_{k,i}=m\}R_{k,i}\widetilde{W}_{k,i}\widetilde{U}_{k,i}\right)^2 \Bigg/ \left(\frac{1}{N_k}\sum_{i=1}^{n_k}R_{k,i}\widetilde{W}_{k,i}^2\right)^2. \tag{19}
$$

Let,

$$\tilde{v}_k^{\text{cluster}} = f_k^{\text{cluster}}/(\mu_k(1-\mu_k) - \sigma_k^2)^2.$$

with

$$\begin{aligned} f_k^{\text{cluster}} &= E[A_{k,m}(1-A_{k,m})^2]\frac{1}{n_k}\sum_{i=1}^{n_k} e_{k,i}^2(1) + E[A_{k,m}^2(1-A_{k,m})]\frac{1}{n_k}\sum_{i=1}^{n_k} e_{k,i}^2(0) \\ &- p_k E[A_{k,m}^2(1-A_{k,m})^2]\frac{1}{n_k}\sum_{i=1}^{n_k}(e_{k,i}(1) - e_{k,i}(0))^2 \\ &+ (E[A_{k,m}(1-A_{k,m})] - (5+p_k)E[A_{k,m}^2(1-A_{k,m})^2])\sum_{m=1}^{m_k}\frac{n_{k,m}}{n_k}(\tau_{k,m}-\tau_k)^2 \\ &+ p_k E[A_{k,m}^2(1-A_{k,m})^2]\sum_{m=1}^{m_k}\frac{n_{k,m}^2}{n_k}(\tau_{k,m}-\tau_k)^2. \end{aligned}$$

We obtain in the appendix,

$$\frac{\widetilde{V}_k^{\text{cluster}}}{\tilde{v}_k} = \frac{\tilde{v}_k^{\text{cluster}}}{\tilde{v}_k} + o_p(1).$$

Similar to the least squares case, the robust variance can underestimate the true variance, and the cluster variance is generally too large. Our proposed variance estimator is a convex combination of $\widetilde{V}_k^{\text{cluster}}$ and and $\widetilde{V}_k^{\text{robust}}$, with the weights selected to correct the bias of the cluster variance estimator as k increases (see appendix for details).

$$\widetilde{V}_k^{\text{CCV}} = \widehat{\lambda}_k \widetilde{V}_k^{\text{cluster}} + (1-\widehat{\lambda}_k)\widetilde{V}_k^{\text{robust}}. \tag{20}$$

where the estimated weight for the cluster variance is

$$\widehat{\lambda}_k = 1 - q_k \frac{\left(\frac{1}{M_k}\sum_{m=1}^{m_k} Q_{k,m}\overline{W}_{k,m}(1-\overline{W}_{k,m})\right)^2}{\frac{1}{M_k}\sum_{m=1}^{m_k} Q_{k,m}\overline{W}_{k,m}^2(1-\overline{W}_{k,m})^2},$$

where $Q_{k,m}$ is an indicator that takes value one if cluster m of population k is sampled, and $M_k = \sum_{m=1}^{m_k} Q_{k,m}$ is the total number of sampled clusters. The second factor in the second term approximately (that is, ignoring the variance of $\overline{W}_{k,m}$ conditional on $A_{k,m}$]) estimates the variance of $A_{k,m}(1-A_{k,m})$ divided by its second moment, so that

$$\tilde{\lambda} \approx 1 - q_k \frac{V(A_{k,m}(1-A_{k,m}))}{E[(A_{k,m}(1-A_{k,m}))^2]}.$$

If there is no variation in $W_{k,i}$ within any of the clusters the fixed effect estimator is not defined, and neither is this variance estimator. In all other cases the variance estimator is well-defined.

We also consider a bootstrap standard error, based on the same resampling procedure described in Section 4.3.

6. Simulations

We next report simulation results that illustrate the performance of the proposed variance estimators relative to existing alternatives. To operate in an empirically relevant setting, we create an artificial population based on the Census data briefly described in the introduction, which contains information on log earnings, an indicator for college attendance, and an indicator for state of residence for 2,632,838 individuals.

For each individual in this population of 2,632,838 individuals, we define $m_{k,i}$ using state of residence (plus Washington, DC, and Puerto Rico), for a total of 52 clusters. We assign potential outcomes as $y_{k,i}(0) = Y_{k,i} - \widehat{\tau}_{k,m} W_{k,i}$ and $y_{k,i}(1) = Y_{k,i} + \widehat{\tau}_{k,m}(1 - W_{k,i})$, so treatment effects are constant within clusters. We then repeatedly create samples from this population. Creating a sample requires fixing p_k, q_k, and fixing the distribution of $A_{k,m}$ and then drawing from the implied distribution for $R_{k,i}$ and $W_{k,i}$ to generate outcomes for all sampled units. In the baseline design, we set $p_k = q_k = 1$, so we sample all $m_k = 52$ clusters and all $n_k = 2{,}632{,}838$ individuals in the population. For the assignment mechanism in the baseline design, we convert cluster means of the treatment variable into log-odds, $\widehat{\ell}_{k,m} = \ln(\overline{W}_{k,m}/(1 - \overline{W}_{k,m}))$. Let $(\widehat{\mu}_\ell, \widehat{\sigma}_\ell)$ be the average and the sample standard deviation of $\widehat{\ell}_{k,m}$. We then draw $\ln(A_{k,m}/(1 - A_{k,m}))$ for cluster m from a normal distribution with expected value $\widehat{\mu}_\ell$ and standard deviation $\widehat{\sigma}_\ell$. Given the cluster assignment probability $A_{k,m}$, we assign the treatment in cluster m by drawing from a binomial distribution with parameter $A_{k,m}$.

We calculate the standard deviation of the least squares and fixed effect estimators, normalized by the square root of the sample size, $N_k^{1/2}$s.d., across 10,000 samples drawn according to the procedure outlined above. This is the benchmark against which we compare

Table 2: Average standard errors across simulations

					normalized standard error			
		$N_k^{1/2}$s.d.	$v_k^{1/2}$	$\widetilde{v}_k^{1/2}$	robust	cluster	CCV	TSCB
Baseline design:								
$p_k = 1$, $q_k = 1$,	OLS	5.91	5.90		1.90	44.86	6.32	5.80
$\sigma_{\tau_k} = .120$, $\sigma_k = .057$	FE	2.34		2.32	1.90	44.63	2.31	2.29
Second Design:								
$p_k = .1$, $q_k = 1$,	OLS	2.61	2.59		1.90	14.28	3.78	2.60
$\sigma_{\tau_k} = .120$, $\sigma_k = .057$	FE	1.95		1.95	1.90	14.21	1.95	1.94
Third Design:								
$p_k = .1$, $q_k = 1$,	OLS	14.50	14.17		1.98	56.46	13.70	14.33
$\sigma_{\tau_k} = .480$, $\sigma_k = .206$	FE	12.14		11.89	2.13	56.79	11.61	12.07
Fourth design:								
$p_k = .1$, $q_k = 1$,	OLS	9.39	9.39		1.90	8.20	9.19	9.37
$\sigma_{\tau_k} = 0$, $\sigma_k = .206$	FE	2.04		2.04	2.04	1.97	2.04	2.09
Fifth design:								
$p_k = .1$, $q_k = 1$,	OLS	1.95	1.97		1.97	56.42	4.53	2.04
$\sigma_{\tau_k} = .480$, $\sigma_k = 0$	FE	1.91		1.94	1.94	56.42	1.96	1.90

Notes: $N_k^{1/2}$s.d. is the standard deviation of the estimators over the simulations, multiplied by the square root of the sample size. $v_k^{1/2}$ is the square root of the asymptotic variance in equation (2). $\tilde{v}_k^{1/2}$ is the square root of the asymptotic variance of the fixed effect estimator in (16). The remaining four columns report average values of robust, cluster, CCV, and TSCB standard errors across simulations (multiplied by $N_k^{1/2}$). p_k and q_k are the unit and cluster sampling probabilities, respectively. σ_{τ_k} is the standard deviation of the cluster average treatment effect. σ_k is the standard deviation across clusters of the treatment assignment probabilities.

the various estimates of standard errors. For the least squares and the fixed effects estimators, respectively, we first calculate the (infeasible) asymptotic standard errors $v_k^{1/2}$ and $\widetilde{v}_k^{1/2}$ to benchmark the performance of the feasible variance estimators. Next, we calculate the averages across 10,000 simulations of the robust, cluster, CCV, and TCSB standard errors, where we use 100 bootstrap replications in each simulation. Table 2 reports the results. Table 3 reports coverage rates for 95 percent confidence intervals. In the design column of the two tables σ_{τ_k} is the standard deviation of the cluster average treatment effect.

For the baseline design, the normalized standard deviation of the least squares estimator is

Table 3: Coverage rates across simulations

		coverage of 95 percent confidence interval					
		$v_k^{1/2}$	$\widetilde{v}_k^{1/2}$	robust	cluster	CCV	TSCB
Baseline design:							
$p_k = 1$, $q_k = 1$,	OLS	0.949		0.467	1.000	0.971	0.947
$\sigma_{\tau_k} = .120$, $\sigma_k = .057$	FE		0.950	0.893	1.000	0.947	0.942
Second design:							
$p_k = .1$, $q_k = 1$,	OLS	0.951		0.846	1.000	0.996	0.952
$\sigma_{\tau_k} = .120$, $\sigma_k = .057$	FE		0.950	0.944	1.000	0.950	0.948
Third design:							
$p_k = .1$, $q_k = 1$,	OLS	0.947		0.208	1.000	0.960	0.950
$\sigma_{\tau_k} = .480$, $\sigma_k = .206$	FE		0.941	0.284	1.000	0.918	0.948
Fourth design:							
$p_k = .1$, $q_k = 1$,	OLS	0.952		0.308	0.905	0.966	0.952
$\sigma_{\tau_k} = 0$, $\sigma_k = .206$	FE		0.952	0.951	0.932	0.951	0.955
Fifth design:							
$p_k = .1$, $q_k = 1$,	OLS	0.952		0.953	1.000	1.000	0.959
$\sigma_{\tau_k} = .480$, $\sigma_k = 0$	FE		0.954	0.955	1.000	0.957	0.949

Notes: See notes of Table 2.

5.91. This is well approximated by the asymptotic standard error, 5.90. The robust standard error is on average over the simulations 1.90, less than one-third of the normalized standard deviation of the estimator. The cluster standard error is far too large, on average 44.86, more than seven times the value of the normalized standard deviation. CCV improves considerably over robust and cluster. The average CCV standard error is 6.32, about 7 percent higher than the normalized standard deviation. The TSCB standard error is the most accurate, on average equal to 5.80. For the fixed effect estimator, the asymptotic standard error is again accurate. The robust standard error is about 16 percent too small, leading to a coverage rate for the nominal 95 percent confidence interval of 0.89 in Table 3. The cluster standard error is too large by a factor of 20. CCV and TSCB standard errors closely approximate the normalized standard error.

It is also interesting to consider the variation in the different variance estimators over the repeated samples relative to the true value of the standard deviation of the estimator. In the baseline design the normalized standard deviation is 5.91. The robust standard error is very precisely estimated, with a standard deviation of the normalized robust standard over the 10,000 simulations equal to 0.005. The standard deviation of the cluster standard error is much larger, 1.48. For the CCV standard error the standard deviation is 1.21, and for the resampling-based TSCB the standard deviation is consideralby lower at 0.69.

We vary the design from the baseline case by changing (i) the fraction of sampled units p_k, (ii) the amount of treatment effect heterogeneity across clusters, σ_{τ_k}, and (iii) the cross-cluster standard deviation of the assignment probability, σ_k. In the second design, $p_k = 0.1$ is the only change relative to the baseline design. This makes the robust standard errors less biased downward, and the cluster standard errors less biased upward. The result of decreasing the fraction of sampled units (and thus decreasing the sample size) is that the performance of the analytic CCV variance estimator declines, whereas the bootstraping vaiance estimator TSCB continues to perform well. We keep $p_k = 0.1$ for the remaining three designs. In the third design, we increase both the treatment effect heterogeneity and the within-cluster correlation of the treatment by increasing the differences in treatment effects $\tau_{k,m} - \tau_k$ and the differences of the logs odds ratio $\ell_{k,m} - \ell_k$ by a factor of four. The resulting increase in σ_k^2 makes the performance of the robust standard error substantially worse, consistent with equation (6). In this design, the bias of the robust standard error is substantial, also for the fixed effect estimator. The difference between the cluster variance and the true variance for the least squares estimator is proportional to the variation in the cluster average treatment effects, implying that the bias will increase for this design relative to the second design. In the fourth design, we remove the heterogeneity in the treatment effect but keep the correlation in the treatment assignment the same as in the third design. Now, the cluster variance performs well, but the robust variance remains poor. In the fifth design, the assignment probabilities are identical in all clusters, and the treatment effect heterogeneity is the same as in the third and fourth designs. In this case the robust

standard errors perform well, but the cluster standard errors substantially over-estimate the uncertainty, as expected. In all designs, the CCV and especially the TSCB standard errors outperform the robust and cluster standard errors.

7. Implications for Practice

The analysis in this article has several implications for how to compute and, most importantly, interpret, standard errors in a variety of empirical settings. Some settings are clear cut and others are more subtle. First, we discuss the case where there is no cluster sampling. If one has a random sample of units from a large population with randomized treatment assignment at the unit level, there is no reason to cluster the standard errors of the least squares estimator. Doing so can be harmful, resulting in unnecessarily wide confidence intervals. In this case, clustering is not appropriate even if there is within-cluster correlation in outcomes (however those clusters are defined), and thus even if clustering makes a substantial difference in the magnitude of the standard errors. For example, if workers are sampled at random from a some population of interest and then randomly assigned to a job training program, clustering the standard errors at, say, the industry, county, or state level can result in standard errors that are unnecessarily conservative, often by a wide margin. Similarly, in a judge-leniency design—where defendants are randomly assigned to judges—standard errors should not be clustered at the level of the judge [Chyn, Frandsen, and Leslie, 2022]. If the sample represents a large fraction of the population and treatment effects are heterogeneous across units, robust standard errors are also conservative. If the data contains information on attributes of the units that are correlated with unit-level treatment effects, the methods in Abadie et al. [2020] can be applied to obtain less conservative standard errors.

Next, consider the case of clustered assignment, and where we either have random sampling or we observe the entire population. This is one case where clustering becomes relevant, although conventional cluster standard errors can be extremely conservative. If assignment is perfectly clustered so that units that belong to the same cluster have the same treatment assignment, there is no improvement from using the CCV variance and the TSCB variance estimator is not applicable. If assignment is partially clustered—so there is variation in

treatment assignment within clusters—and cluster sizes are large, the CCV and TSCB can be applied and can produce standard errors considerably smaller than the usual clustered standard errors.

Another reason to cluster standard errors is cluster sampling. The case with q_k close to zero is sometimes relevant, especially when the sample is a panel data on individuals or a cross-section of families, and the individuals or families in the sample are a small fraction of the population. Then, the clustered variance estimator of the least squares estimator is asymptotically correct regardless of whether the treatment assignment is clustered or not. The same result holds when clusters are large (*e.g.*, states), q_k is a substantial fraction of the clusters in the population, but p_k is small—so the sample includes only a small number of units from each cluster. In other cases, cluster standard errors can be considerably larger than necessary. If cluster sizes are large and there is treatment variation within clusters, CCV and TSCB can substantially reduce the magnitude of standard errors.

The insights in this article are relevant in other common settings of empirical economics. Consider a setting with unit-level panel data on outcomes and a treatment that is implemented on the same period for all units in the treatment group. The difference-in-difference estimator is in this case equal to the coefficient on the treatment variable in a cross-sectional regression of the change in unit-level average outcomes between the post-treatment and the pre-treatment periods on a constant and a treatment indicator that takes value one if the unit belongs to the treatment group. If treatment assignment is random across units, and the sample includes a random subset of the population or the entire population, robust standard error provide inference that is generally conservative if the sample is large relative to the population and treatment effects are heterogeneous. Here too, the methods in Abadie et al. [2020] can be applied to correct the bias of robust standard errors. With clustered assignment, one should cluster the standard errors at the level of assignment—for example, cluster at the village level if all farmers are assigned the same treatment status. Adding group-level fixed effects to this regression allows for group-specific linear trends in the underlying potential outcomes series but does not change the answer to the question whether

one needs to adjust for clustering. Under partially clustered assignment, CCV and TSCB standard errors can continue to provide substantial improvements over conventional cluster standard errors for the fixed effect estimator.

8. Conclusion

This article proposes a research framework aimed to address a question of central relevance for empirical practice: when and how we should cluster standard errors. Like in Abadie et al. [2020], we shift the attention from estimation of features of a data-generating process (i.e., infinite superpopulation) to estimation of average treatment effects of the finite population at hand. We show that, in this framework, the decision on when and how to cluster standard errors depends on the nature of the sampling and the assignment processes only, and not on the presence of within-cluster error components in the outcome variable. We derive expressions of the large sample variances of the OLS and FE estimators of the average treatment effect for a setting with clustered sampling and where assignment is random within clusters with assignment probabilities that may vary across clusters. For this setting, we demonstrate that robust standard errors can be too small and conventional cluster standard errors can be unnecessarily large. We propose two novel procedures, CCV and TSCB, that can be used to calculate more precise standard errors in settings with large clusters and where there is enough variation in treatment assignment within cluster (so that average treatment effects within clusters can be precisely estimated). While CCV and TSCB are designed for this particular setting, the general principles of the framework remain valid for other settings and estimators. If sampling is not clustered, standard errors should be clustered at the treatment assignment level because the estimand of interest depends on potential outcomes and the sampling of potential outcomes is determined only by the assignment mechanism. When the fraction of sampled clusters is non-negligible and there is variation in average treatment effects across clusters, conventional clustered standard errors may be off, and we provide an analytical framework that can be applied to derive appropriate standard errors. When sampling and assignment are random, clustering standard errors is not appropriate regardless of the structure of the covariance of the outcomes across the units in the population. In

this setting, if there is substantial treatment effect heterogeneity and the sample represents a large fraction of the population of interest, robust standard errors are conservative in large samples. This bias can be corrected using the methods in Abadie et al. [2020]. Deriving standard error formulas for sampling and assignment processes other than the ones featured in this article is an important avenue for future research. Rambachan and Roth [2022] is a recent contribution in this direction. In addition, in the present article we have restricted the analysis to linear estimators (least squares and fixed effects). Xu [2019] extends the ideas and framework of this article to analyze the distribution of non-linear estimators.

References

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020.

Manuel Arellano. Practitioners corner: Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434, 1987.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004.

A Colin Cameron and Douglas L Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015.

Min-Te Chao and Shaw-Hwa Lo. A bootstrap method for finite population. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 399–405, 1985.

Eric Chyn, Brigham Frandsen, and Emily Leslie. Examiner and judge designs in economics: A practitioner's guide. 2022. Working paper.

Jessica Cohen and Pascaline Dupas. Free distribution or cost-sharing? evidence from a randomized malaria prevention experiment. *The Quarterly Journal of Economics*, 125(1): 1–45, 2010.

F. Eicker. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statististics*, 34(2):447–456, 06 1963. doi: 10.1214/aoms/1177704156.

Matthew Gentzkow and Jesse M. Shapiro. Preschool television viewing and adolescent test scores: Historical evidence from the coleman study. *The Quarterly Journal of Economics*, 123(1):279–323, 2008.

Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press, 1967.

Guido Imbens and Konrad Menzel. A causal bootstrap. *Annals of Statistics*, 49(3):1460–1488, 2021.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Leslie Kish. *Survey Sampling*. Wiley-Interscience, 1995.

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

James G MacKinnon, Morten Ørregaard Nielsen, and Matthew Webb. Cluster-robust inference: A guide to empirical practice. 2021. Working paper.

Konrad Menzel. Bootstrap with cluster-dependence in two or more dimensions. *Econometrica*, 89(5):2143–2188, 2021.

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

Brent R Moulton. Diagnostics for group effects in regression analysis. *Journal of Business & Economic Statistics*, 5(2):275–282, 1987.

Jerzey Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472, 1923/1990.

Ashesh Rambachan and Jonathan Roth. Design-based uncertainty for quasi-experiments. 2022. arXiv:2008.00602.

Steven K Thompson. *Sampling*. John Wiley & Sons, 2012.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, pages 817–838, 1980.

Ruonan Xu. Asymptotic properties of M-estimators with finite populations under cluster sampling and cluster assignment. 2019. Working paper.