# Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed-effect models

James E. Pustejovsky*
Department of Educational Psychology
University of Texas at Austin

and

Elizabeth Tipton
Department of Human Development
Teachers College, Columbia University

October 16, 2015

**Abstract**

The text of your abstract. 200 or fewer words.

*Keywords:* 3 to 6 keywords, that do not appear in the title

*The authors thank Dan Knopf for helpful discussions about the linear algebra behind the cluster-robust variance estimator. Coady Wing,...

# 1 INTRODUCTION

The use of cluster-robust variance estimation (CRVE Arellano, 1987; Liang and Zeger, 1986; White, 1984) is now common across a wide range of economic analyses. Standard errors are routinely "clustered" to account for correlations arising from the sampling of aggregate units (e.g., countries, regions, states, villages), each containing multiple individuals. Likewise, CRVE is now routinly used in analysis of panel data to account for correlation of measurements on the same individual unit across time periods. The method is an extension to another economic mainstay, heteroscedasticity-robust standard errors (Eicker, 1967; Huber, 1967; White, 1980), which are used to account for non-constant variance in regression models. Although CRVE has existed for over 30 years, it has become standard practice only in the last decade, as illustrated by its coverage in major textbooks and review articles (e.g., Angrist and Pischke, 2009; Cameron and Miller, 2015; Wooldridge, 2010).

CRVE allows analysts to estimate causal and structural models using ordinary least squares (OLS) or weighted least squares (WLS), while adjusting standard errors thereafter. Cluster-robust variance estimators can also be used as the basis for both single- or multiple-parameter hypothesis tests. For example, an analyst may wish to understand the effects on employment outcomes of several state-level policy shifts, where the policies were implemented at different time-points in each state. A standard approach to estimating such effects is via a regression model that includes indicator variables for each policy shift, as well as fixed effects for states and time-periods in order to control for unobserved confounding in each of these dimensions. The model might be estimated by OLS, with the fixed effects included as indicator variables; more commonly, the effects of the policy indicators might be estimated after absorbing the fixed effects, a computational technique that is also known as the fixed effects or within transformation (Wooldridge, 2010). Standard errors would then be clustered by state to account for dependence in the residual errors from a given state, and these clustered standard errors would be used to test hypotheses regarding each policy (i.e., a t-test) or across policies (i.e., an F-test). The need to cluster the standard errors by state, even when including state fixed effects, was highlighted by Bertrand, Duflo and Mullainathan (2004), who showed that to do otherwise can lead to inappropriately small standard errors and hypothesis tests with incorrect rejection rates.

In CRVE, standard errors are estimated empirically, thus not requiring analysts to assume a particular correlation structure. The standard errors produced are consistent estimates of the true standard errors, leading to appropriate hypotheses tests when the number of clusters is large. As the method has become more common, however, researchers have turned attention to the performance of these tests in small and moderate samples. Cameron and Miller (2015) provide a thorough review of this literature, including a discussion of current practice, possible solutions, and open problems. They highlight the well-known phenomenon that in small samples, CRVE has a downward bias and that hypothesis tests based on CRVEs can have Type-I error rates that are considerably larger the nominal level of the test. Moreover, they review recent research showing that the small-sample corrections for t-tests typically found in software such as Stata and SAS are inadequate.

Cameron and Miller also point to a potentially promising solution to these problems, the bias-reduced linearization (BRL) method, which was introduced by McCaffrey, Bell and Botts (2001) and Bell and McCaffrey (2002). BRL entails correcting the bias of the CRVE so that it is exactly unbiased under a working model specified by the analyst, while also remaining asymptotically consistent under arbitrary true variance structures. Simulations reported by Bell and McCaffrey (2002) demonstrate that the BRL correction serves to reduce the bias of the CRVE even when the working model is mis-specified. The same authors also proposed and studied small-sample corrections to single-parameter hypothesis tests using the BRL variance estimator, based on Satterthwaite (Bell and McCaffrey, 2002) or saddlepoint approximations (McCaffrey and Bell, 2006).

Despite promising simulation evidence that BRL performs well (e.g., Imbens and Kolesar, 2012), several problems arise making it difficult to implement in the kinds of analyses common in economics. First, as Angrist and Pischke (2009) argue, when clustering is doubly accounted for (both through fixed effects and clustered standard errors) the BRL adjustment breaks down and cannot be implemented. Second, however, as Cameron and Miller (2015) highlight, even when the BRL adjustment can be computed, the standard errors that result from absorption can be substantially different than those produced using dummy fixed effects. This problem also plagues the small sample corrections more

commonly implemented, and leads analysts to have to choose the appropriate method for accounting for fixed effects on an ad hoc basis, with little guidance regarding the best strategy. Third—and more generally—although Bell and McCaffrey (2002) provide a method for conducting single parameter tests, small-sample methods for multiple-parameter tests are lacking. These tests occur commonly in the broader economics literature and are found not only in panel data (e.g., the Hausman test), but also more broadly in seemingly unrelated regression models, and when analyzing experimental data (e.g., baseline equivalence), particularly when there are multiple treatment groups.

Briefly review alternative small-sample stuff.

In this paper, we address each of these three concerns, in the end articulating a BRL methodology that is suitable for everyday econometric practice. In the remainder of this section, we introduce our econometric framework and review the standard CRVE methods, as implemented in most software applications. In Section 2, we review the original BRL correction, and propose modifications that make it possible to implement BRL in a broad class of models with fixed effects. In Section 3, we propose a method for testing multiple-constraint hypotheses (i.e., F-tests) based on CRVE with the BRL adjustments, and show that the t-test proposed by Bell and McCaffrey (2002) is a special case. We then provide simulation evidence that this small-sample F-test offers drastic improvements over commonly implemented alternatives. In Section 4, we illustrate the use of CRVE in small samples, implementing the proposed hypothesis tests in three examples that cover a variety of contexts where CRVE is commonly used. We conclude the paper with a discussion (Section 5), where we argue that the BRL approach given here is not only superior to CRVE more generally, but also that it is potentially more useful in practice than other resampling based methods.

Bold!

## 1.1 Econometric framework

We begin by considering a generic model of the form,

$$y_{ij} = \mathbf{r}_{ij}'\boldsymbol{\beta} + \mathbf{s}_{ij}'\boldsymbol{\gamma} + \mathbf{t}_{ij}'\boldsymbol{\mu} + \epsilon_{ij} \tag{1}$$

where for observation $j$ in cluster $i$, $\mathbf{r}_{ij}$ is a vector of $r$ predictors of primary interest in an analysis (e.g., policy variables) as well as additional control variables, $\mathbf{s}_{ij}$ is a vector of $s$

fixed effects that vary across clusters, and $\mathbf{t}_{ij}$ is a vector of $t$ fixed effects that are identified within clusters. In the state-policy example described in the introduction, the $\mathbf{r}_{ij}$ would include indicators for each policy under study and additional demographic controls, $\mathbf{s}_{ij}$ would include year fixed effects, and $\mathbf{t}_{ij}$ would indicate state fixed effects. Interest would focus on testing hypotheses regarding the coefficients in $\boldsymbol{\beta}$ that correspond to the policy indicators, while $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ would be considered incidental.

In developing theory, it is often easier to work with the matrix version of this model, where now

$$\mathbf{y}_i = \mathbf{R}_i\boldsymbol{\beta} + \mathbf{S}_i\boldsymbol{\gamma} + \mathbf{T}_i\boldsymbol{\mu} + \boldsymbol{\epsilon}_i, \tag{2}$$

where for cluster $j$, $\mathbf{R}_i$ is an $n_i \times r$ matrix of focal predictors and controls; $\mathbf{S}_i$ is an $n_i \times s$ matrix describing fixed effects that vary across clusters, and $\mathbf{T}_i$ is an $n_i \times t$ matrix describing fixed effects that are identified only within clusters.

We assume that $\mathrm{E}\left(\boldsymbol{\epsilon}_i \,|\mathbf{R}_i, \mathbf{S}_i, \mathbf{T}_i\right) = \mathbf{0}$ and $\mathrm{Var}\left(\boldsymbol{\epsilon}_i \,|\mathbf{R}_i, \mathbf{S}_i, \mathbf{T}_i\right) = \boldsymbol{\Sigma}_i$, for $i = 1, ..., m$, where the form of $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_m$ may be unknown but the errors are independent across clusters. For notational convenience, let $\mathbf{U}_i = [\mathbf{R}_i\ \mathbf{S}_i]$ denote the set of predictors that vary across clusters, $\mathbf{X}_i = [\mathbf{R}_i\ \mathbf{S}_i\ \mathbf{T}_i]$ denote the full set of predictors, $\boldsymbol{\alpha} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\mu}')'$, and $p = r + s + t$. Denote the total number of individual observations by $N = \sum_{i=1}^{m} n_i$. Let $\mathbf{y}$, $\mathbf{R}$, $\mathbf{S}$, $\mathbf{T}$, $\mathbf{U}$, and $\mathbf{X}$ denote the matrices obtained by stacking their corresponding components, as in $\mathbf{R} = (\mathbf{R}_1'\ \mathbf{R}_2'\ \cdots\ \mathbf{R}_m')'$.

In this model, inferential interest is confined to $\boldsymbol{\beta}$ and the fixed effects $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ are treated as nuisance parameters. The distinction between the covariates $\mathbf{R}_i$ versus the fixed effects $[\mathbf{S}_i\ \mathbf{T}_i]$ thus depends on context and the analyst's inferential goals. However, the distinction between the two fixed effect matrices $\mathbf{S}_i$ and $\mathbf{T}_i$ is unambiguous, in that the within-cluster fixed effects satisfy $\mathbf{T}_h\mathbf{T}_i' = \mathbf{0}$ for $h \neq i$. We further assume that $(\mathbf{U}'\mathbf{U} - \mathbf{U}_i'\mathbf{U}_i)$ is of full rank for $i = 1, ..., m$.

We can estimate the $\boldsymbol{\beta}$ using weighted least squares (WLS), where for cluster $i$ we define $\mathbf{W}_i$ to be a symmetric, $n_i \times n_i$ weighting matrix of full rank. Importantly, the WLS framework includes the unweighted case (where $\mathbf{W}_i = \mathbf{I}_i$, an identity matrix), as well as feasible GLS.[1] In the latter case, it is assumed that $\mathrm{Var}\left(\mathbf{e}_i \,|\mathbf{X}_i\right) = \boldsymbol{\Phi}_i$, where $\boldsymbol{\Phi}_i$ is a known

[1]The WLS estimator also encompasses the estimator proposed by Ibragimov and Müller (2010) for

function of a low-dimensional parameter. For example, an auto-regressive error structure might be posited to describe repeated measures on an individual over time. The weighting matrices are then taken to be $\mathbf{W}_i = \hat{\mathbf{\Phi}}_i^{-1}$, where the $\hat{\mathbf{\Phi}}_i$ are constructed from estimates of the variance parameter. Finally, for analysis of data from complex survey designs, WLS may be used with sampling weights in order to account for unequal selection probabilities.

## 1.2 Absorption

In most analyses, the goal is to estimate and test hypotheses regarding the parameters in $\boldsymbol{\beta}$. This means that the values of $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ are nuisance parameters and are not of inferential interest. Estimating each of these fixed effects - as occurs if the fixed effects are included as dummy variables in the model - be computationally intensive and numerically inaccurate if the number of clusters is large (i.e., $s + t$ large). In the policy example given above, for example, this could easily result in over 70 parameters (i.e., 50 states, 20 time periods). An alternative that is commonly implemented, therefore, is to first absorb the fixed effects. This amounts to demeaning the data by substracting the cluster-mean value from both the outcomes and covariates; this results in the "within" estimator, which is commonly implemented in panel data analyses. By aborbing the fixed effects, only the $r$ parameters in $\boldsymbol{\beta}$ need to be estimated, which results in a more computationally efficient and numerically accurate procedure.

In Section 2 of this paper, we will discuss more fully comparisons between the dummy variable and aborption approaches to fixed effects. In order to do, we now formalize the absorption method. To begin, denote the full block-diagonal weighting matrix as $\mathbf{W} = \text{diag}\left(\mathbf{W}_1, ..., \mathbf{W}_m\right)$. Let $\mathbf{K}$ be the $x \times r$ matrix that selects the covariates of interest, so that $\mathbf{XK} = \mathbf{R}$ and $\mathbf{K}'\boldsymbol{\alpha} = \boldsymbol{\beta}$. For a generic matrix $\mathbf{Z}$ of full column rank, let $\mathbf{M_Z} = \left(\mathbf{Z}'\mathbf{WZ}\right)^{-1}$ and $\mathbf{H_Z} = \mathbf{ZM_ZZ}'\mathbf{W}$.

The absorption technique involves obtaining the residuals from the regression of $\mathbf{y}$ on $\mathbf{T}$ and from the multivariate regressions of $\mathbf{U} = [\mathbf{R}\ \mathbf{S}]$ on $\mathbf{T}$. The $\mathbf{y}$ residuals and $\mathbf{R}$

clustered data. Assuming that $\mathbf{X}_i$ has rank $p$ for $i = 1, ..., m$, their proposed approach involves estimating $\boldsymbol{\beta}$ separately within each cluster and taking the simple average of these estimates. The resulting average is equivalent to the WLS estimator with weights $\mathbf{W}_i = \mathbf{X}_i \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-2} \mathbf{X}_i$.

residuals are then regressed on the $\mathbf{S}$ residuals. Finally, these twice-regressed $\mathbf{y}$ residuals are regressed on the twice-regressed $\mathbf{R}$ residuals to obtain the WLS estimates of $\boldsymbol{\beta}$. Let $\ddot{\mathbf{S}} = (\mathbf{I} - \mathbf{H_T})\,\mathbf{S}$, $\ddot{\mathbf{R}} = (\mathbf{I} - \mathbf{H_{\ddot{S}}})\,(\mathbf{I} - \mathbf{H_T})\,\mathbf{R}$, and $\ddot{\mathbf{y}} = (\mathbf{I} - \mathbf{H_{\ddot{S}}})\,(\mathbf{I} - \mathbf{H_T})\,\mathbf{y}$. In what follows, subscripts on $\ddot{\mathbf{R}}$, $\ddot{\mathbf{S}}$, $\ddot{\mathbf{U}}$, and $\ddot{\mathbf{y}}$ refer to the rows of these matrices corresponding to a specific cluster. The WLS estimator of $\boldsymbol{\beta}$ can then be written as

$$\hat{\boldsymbol{\beta}} = \mathbf{M}_{\ddot{\mathbf{R}}} \sum_{i=1}^{m} \ddot{\mathbf{R}}_i' \mathbf{W}_i \ddot{\mathbf{y}}_i. \tag{3}$$

This estimator is algebraically identical to the direct WLS estimator based on the full set of predictors,

$$\hat{\boldsymbol{\beta}} = \mathbf{K}' \mathbf{M}_{\mathbf{X}} \sum_{i=1}^{m} \mathbf{X}_i' \mathbf{W}_i \mathbf{y}_i,$$

but avoids the need to solve a system of $x$ linear equations.

## 1.3 Standard CRVE

In the remainder of this paper, we focus on the general case in which fixed effects are either included as dummy variables or absorbed. In either case, the goal of the analysis is test hypotheses regarding $\boldsymbol{\beta}$ using the WLS estimator $\hat{\boldsymbol{\beta}}$, which has true variance,

$$\text{Var}\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{M}_{\ddot{\mathbf{R}}} \left( \sum_{i=1}^{m} \ddot{\mathbf{R}}_i' \mathbf{W}_i \boldsymbol{\Sigma}_i \mathbf{W}_i \ddot{\mathbf{R}}_i \right) \mathbf{M}_{\ddot{\mathbf{R}}}, \tag{4}$$

which depends upon the unknown variance matrices $\boldsymbol{\Sigma}_i$. A model-based approach to estimating this variance would involve assuming a structure to this $\boldsymbol{\Sigma}_i$; for example, it may be assumed that the structure was hierarchical or auto-regressive. However, if the model is mis-specified, the model-based variance estimator will be inconsistent and inferences based upon it will be invalid.

The CRVE approach is to instead estimate $\text{Var}\left(\hat{\boldsymbol{\beta}}\right)$ empirically. While there are several versions of this approach, all can be written in the form

$$\mathbf{V}^{CR} = \mathbf{M}_{\ddot{\mathbf{R}}} \left( \sum_{i=1}^{m} \ddot{\mathbf{R}}_i' \mathbf{W}_i \mathbf{A}_i \mathbf{e}_i \mathbf{e}_i' \mathbf{A}_i' \mathbf{W}_i \ddot{\mathbf{R}}_i \right) \mathbf{M}_{\ddot{\mathbf{R}}}, \tag{5}$$

for some $n_i$ by $n_i$ adjustment matrix $\mathbf{A}_i$. Note that this estimator replaces the unknown $\boldsymbol{\Sigma}_i$ with the cross-product of the the residuals, $\mathbf{e}_i \mathbf{e}_i'$.

The form of these adjustments parallels those of the heteroscedastity-consistent (HC) variance estimators proposed by MacKinnon and White (1985). Setting $\mathbf{A}_i = \mathbf{I}_i$, an $n_i \times n_i$ identity matrix, results in the most basic form, described by Liang and Zeger (1986). Following Cameron and Miller (2015), we refer to this estimator as $\mathbf{V}^{CR0}$. Setting $\mathbf{A}_i = c\mathbf{I}_i$, where $c = \sqrt{(m/(m-1))(N/(N-p))}$, results in a slightly larger estimator, denoted $\mathbf{V}^{CR1}$. Note that when $N >> p$, $c \approx \sqrt{m/(m-1)}$, and some software uses the latter approximation (e.g., SAS). Both the CR0 and CR1 estimators rely on asymptotic properties of the residuals in order to consistently estimate $\mathbf{\Sigma}_i$. The CR1 estimator is now standard in most analyses in economics.

In addition to CR1, two other estimators are also currently available for improving small sample properties. Unlike the CR1 estimator, these approaches result in adjustments that take into account features of the covariates in $\mathbf{X}_i$. In the next section, we describe in detail the BRL approach, which is an extension of the HC2 estimator for regressions with heteroskedastic but uncorrelated errors; we therefore refer to it as CR2. A further alternative is CR3, which uses adjustment matrices given by $\mathbf{A}_i = \left(\mathbf{I} - \ddot{\mathbf{R}}_i \mathbf{M}_{\ddot{\mathbf{R}}} \ddot{\mathbf{R}}_i' \mathbf{W}_i\right)^{-1}$. The CR3 estimator closely approximates the jackknife re-sampling variance estimator (Bell and McCaffrey, 2002; Mancl and DeRouen, 2001). As they indicate, however, the CR3 estimator tends to over-correct the downward bias, while the CR1 estimator tends to under-correct. The CR2 estimator offers a solution in the middle, and for this reason we focus on it for the remainder of this paper.

# 2 BIAS REDUCED LINEARIZATION

The unadjusted CR0 estimator and, to a smaller degree, the CR1 estimator both tend to under-estimate the true variance of $\hat{\boldsymbol{\beta}}$ (Cameron and Miller, 2015). Simulation studies indicate that the degree of this bias, however, depends not only on the number of clusters $m$, but also on features of the covariates in $\mathbf{X}_i$. MacKinnon (2013) shows that this bias is largest when the covariate exhibits large imbalances, skew, or leverage. For this reason, it is desirable to develop an adjustment that takes into account features of the covariates. More formally, this amounts to developing a method for defining the adjustment matrices $\mathbf{A}_i \neq c\mathbf{I}_i$ for any constant $c$.As noted above, both the BRL approach (CR2) and the jacknife

I don't think this formality is necessary.

approach (CR3) meet these requirements.

Unlike CR1 and CR3, however, the BRL approach requires the analyst to specify a "working" model for the correlation structure. The BRL estimator then uses the adjustment matrices $\mathbf{A}_i$ chosen so that the CR2 estimator is exactly unbiased when this working model is correct. The idea of specifying a model may seem antithetical to the purpose of using CRVE, yet extensive simulation studies have illustrated that the method performs better in small samples than any of the other approaches, even when the working model is incorrect. Although the CR2 estimator is no longer exactly unbiased when the working model is mis-specified, its bias tends to be greatly reduced compared to CR1 or CR0 (thus the name "bias reduced linearization"). Furthermore, as the number of clusters increases, reliance on the working model diminishes. In a sense, CR2 provides necessary scaffolding in the small-sample case, which falls away when there is sufficient data.

Cite evidence.

Let $\mathbf{\Phi}_i$ denote a working model for the covariance of the errors in cluster $i$, with $\mathbf{\Phi} = \text{diag}\left(\mathbf{\Phi}_1, ..., \mathbf{\Phi}_m\right)$. In the original formulation of Bell and McCaffrey (2002), the BRL adjustment matrices were chosen to satisfy the criterion

$$\mathbf{A}_i \left(\mathbf{I} - \mathbf{H}_\mathbf{X}\right)_i \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_\mathbf{X}\right)_i' \mathbf{A}_i' = \mathbf{\Phi}_i \tag{6}$$

where $\left(\mathbf{I} - \mathbf{H}_\mathbf{X}\right)_i$ denotes the rows of $\mathbf{I} - \mathbf{H}_\mathbf{X}$ corresponding to cluster $i$. Calculation of the adjustment matrices $\mathbf{A}_i$ involves taking the inverse of the symmetric square-root of a matrix. For example, if the working model and weight matrices are both taken to be identity matrices, then $\mathbf{A}_i = \left(\mathbf{I}_i - \ddot{\mathbf{R}}_i \mathbf{M}_{\ddot{\mathbf{R}}} \ddot{\mathbf{R}}_i'\right)^{-1/2}$. However, this formulation of $\mathbf{A}_i$ is problematic for some fixed effects models that are common in economic applications. In the next two subsections, we address two problems that arise, thereby articulating a BRL methodology that is suitable for a wide range of applications.

## 2.1 Generalized Inverse

The equality defining the $\mathbf{A}_i$ matrices cannot always be solved because it is possible that some of the matrices involved are not of full rank, and thus cannot be inverted. For example, Angrist and Pischke (2009) note that this problem arises in balanced state-by-year panel models that include fixed effects for states and for years. In order to address this concern,

we provide an alternative criterion for the adjustment matrices that can always be satisfied. Instead of criterion (6), we seek adjustment matrices $\mathbf{A}_i$ that satisfy:

$$\ddot{\mathbf{R}}_i'\mathbf{W}_i\mathbf{A}_i\left(\mathbf{I}-\mathbf{H}_{\mathbf{X}}\right)_i\boldsymbol{\Phi}\left(\mathbf{I}-\mathbf{H}_{\mathbf{X}}\right)_i'\mathbf{A}_i'\mathbf{W}_i\ddot{\mathbf{R}}_i=\ddot{\mathbf{R}}_i'\mathbf{W}_i\boldsymbol{\Phi}_i\mathbf{W}_i\ddot{\mathbf{R}}_i. \tag{7}$$

A variance estimator that uses such adjustment matrices will be exactly unbiased when the working model is correctly specified.

The above criterion (7) does not uniquely define $\mathbf{A}_i$. Following McCaffrey et al. (2001), we propose to use a symmetric solution in which

$$\mathbf{A}_i=\mathbf{D}_i'\mathbf{B}_i^{+1/2}\mathbf{D}_i, \tag{8}$$

where $\mathbf{D}_i$ is the upper-right triangular Cholesky factorization of $\boldsymbol{\Phi}_i$,

$$\mathbf{B}_i=\mathbf{D}_i\left(\mathbf{I}-\mathbf{H}_{\ddot{\mathbf{R}}}\right)_i\left(\mathbf{I}-\mathbf{H}_{\ddot{\mathbf{S}}}\right)\left(\mathbf{I}-\mathbf{H}_{\mathbf{T}}\right)\boldsymbol{\Phi}\left(\mathbf{I}-\mathbf{H}_{\mathbf{T}}\right)'\left(\mathbf{I}-\mathbf{H}_{\ddot{\mathbf{S}}}\right)'\left(\mathbf{I}-\mathbf{H}_{\ddot{\mathbf{R}}}\right)_i'\mathbf{D}_i', \tag{9}$$

and $\mathbf{B}_i^{+1/2}$ is the symmetric square root of the Moore-Penrose inverse of $\mathbf{B}_i$. The Moore-Penrose inverse is well-defined and unique even when $\mathbf{B}_i$ is not of full rank (Banerjee and Roy, 2014, Thm. 9.18). These adjustment matrices satisfy criterion (7), as stated in the following theorem.

**Theorem 1.** *Let* $\mathbf{L}_i=\left(\ddot{\mathbf{U}}'\ddot{\mathbf{U}}-\ddot{\mathbf{U}}_i'\ddot{\mathbf{U}}_i\right)$ *and assume that* $\mathbf{L}_1,...,\mathbf{L}_m$ *have full rank* $r+s$*. Further assume that* $Var\left(\boldsymbol{\epsilon}_i\left|\mathbf{R}_i,\mathbf{S}_i,\mathbf{T}_i\right.\right)=\boldsymbol{\Phi}_i$*, for* $i=1,...,m$*. Then the adjustment matrix* $\mathbf{A}_i$ *defined in (8) and (9) satisfies criterion (7) and* $\mathbf{V}^{CR2}$ *is exactly unbiased.*

Proof is given in Appendix A. If $\mathbf{B}_i$ is of full rank, then the adjustment matrices also satisfy the original criterion (6). Furthermore, because the adjustment matrices are defined in terms of all three components of the predictors ($\mathbf{R}$, $\mathbf{S}$, and $\mathbf{T}$), they are invariant to whether the model is estimated by direct WLS estimation or after absorbing some or all of the fixed effects. Thus, these modified BRL adjustment matrices can be calculated in a wider range of applications.

## 2.2 Absorption and Dummy Equivalence

A second problem, highlighted by Cameron and Miller (2015), is that the small-sample CRVE approach can result in a different estimator depending upon if the fixed effects are

included in the model as dummies or absorbed. For example, this problem arises with the CR1 estimator, which has the form $\mathbf{A}_i = c\mathbf{I}_i$, where $c = \sqrt{(m/(m-1))(N/(N-p))}$. In this estimator, $p$ depends on the total number of covariates estimated in the model. When fixed effects are included as dummies, $p = r + s + t$, whereas when the fixed effects are absorbed, instead $p = r$. Cameron and Miller highlight that this can be particularly problematic if the clusters are small, as when they each include a pair of individuals; in these studies, the correction results in a variance that is over twice as large when using absorption compared to the use of dummies.

This non-equivalence problem given above can also arise when implementing the BRL method.

Explain how the non-equivalence stuff works—it's a matter of assuming a working model for the errors or the working model for the residuals (after absorption).

Below, we show that by defining the problem in a particular way, the adjustment matrices $\mathbf{A}_i$ are invariant to the method for adjusting for fixed effects. To see how, begin by noting that in many applications, it will make sense to choose weighting matrices that are the inverses of the working covariance model, so that $\mathbf{W}_i = \mathbf{\Phi}_i^{-1}$. In this case, the adjustment matrices can be calculated using $\tilde{\mathbf{B}}_i$ in place of $\mathbf{B}_i$, where

$$\tilde{\mathbf{B}}_i = \mathbf{D}_i \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_i \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}}\right) \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}}\right)' \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_i' \mathbf{D}_i'. \tag{10}$$

**Theorem 2.** *Let* $\tilde{\mathbf{A}}_i = \mathbf{D}_i' \tilde{\mathbf{B}}_i^{+1/2} \mathbf{D}_i$, *where* $\tilde{\mathbf{B}}_i$ *is given in (10). If* $\mathbf{T}_i \mathbf{T}_k' = \mathbf{0}$ *for* $j \neq k$ *and* $\mathbf{W} = \mathbf{\Phi}^{-1}$, *then* $\mathbf{A}_i = \tilde{\mathbf{A}}_i$.

Proof: See Appendix A.

As theorem 2 above demonstrates, using $\tilde{\mathbf{B}}_i$ rather than $\mathbf{B}_i$ leads to algebraically identical adjustment matrices; the form of $\tilde{\mathbf{B}}_i$ is simply more convenient for computation. Interestingly, in the simple case of ordinary (unweighted) least squares, in which the working variance model posits that the errors are all independent and homoskedastic and $\mathbf{W} = \mathbf{\Phi} = \mathbf{I}$, the adjustment matrices simplify further to

$$\mathbf{A}_i = \left(\mathbf{I}_i - \ddot{\mathbf{U}}_i \left(\ddot{\mathbf{U}}' \ddot{\mathbf{U}}\right)^{-1} \ddot{\mathbf{U}}_i'\right)^{+1/2},$$

where $\ddot{\mathbf{U}} = \left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right) \mathbf{U}$. Importantly, unlike the CR1 approach, this means that in the CR2 approach the analyst is not left to choose the method for accounting for fixed effects

in an ad hoc fashion. Together, these two reformulations to the BRL method provided here allow for the approach to be implemented in a broad range of economic applications. In the next section, we address a final set of concerns: how to implement the method in hypothesis testing.

# 3 HYPOTHESIS TESTING

Until now, we have focused on different approaches to estimating cluster-robust standard errors in small samples. However, standard errors are of limited inherent interest—rather, their main use is for the construction of hypothesis tests and confidence intervals. Cluster-robust Wald-type test statistics are a function of the parameter estimates $\hat{\boldsymbol{\beta}}$ and the corresponding CRVE matrix. Such tests are justified based on the asymptotic behavior of robust Wald statistics as the number of clusters grows large (i.e., as $m \to \infty$).

Like the research on the bias of the CRVE estimator, evidence from a wide variety of contexts indicates that the asymptotic limiting distribution of these statistics may be a poor approximation when the number of clusters is small, even if corrections such as CR2 are employed (Bell and McCaffrey, 2002; Bertrand et al., 2004; Cameron, Gelbach and Miller, 2008). Like the bias of the CRVE estimator itself, the accuracy of the asymptotic approximations depends on design features such as the degree of imbalance, skewness, and leverage in the covariates, and similarity of cluster sizes (McCaffrey et al., 2001; Tipton and Pustejovsky, forthcoming; Webb and MacKinnon, 2013). This provides motivation for development of general-purpose hypothesis testing procedures that have accurate rejection rates in small samples.

In this section, we develop a general method for conducting hypothesis tests based on CRVE. We consider linear constraints on $\boldsymbol{\beta}$, where the null hypothesis has the form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ for fixed $q \times r$ matrix $\mathbf{C}$ and $q \times 1$ vector $\mathbf{d}$. The cluster-robust Wald statistic is then

$$Q = \left(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}\right)' \left(\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'\right)^{-1} \left(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}\right), \tag{11}$$

where $\mathbf{V}^{CR}$ is any of the cluster-robust estimators described above. In large samples, it can be shown that this Wald test rejects $H_0$ at level $\alpha$ if $Q$ exceeds $\chi^2(\alpha; q)$, the $\alpha$ critical

value from a chi-squared distribution with $q$ degrees of freedom. When $q = 1$, an equivalent statement is that the test follows a standard normal distribution in large samples. Regardless of the value of $q$ however, in practice it is rarely clear how large samples need to be for Wald tests to be implemented. In the $q = 1$ case, the standard is therefore to instead use the t-distribution instead, with degrees of freedom $m-1$ when CR1 is employed. Similarly, when $q > 1$, this results in the test $F = Q/q$, which is compared to the $F(q, m-1)$ reference distribution.

Can we wait to get into the q = 1 case?

## 3.1 Small-sample corrections for t-tests

Consider testing the hypothesis $H_0 : \mathbf{c}'\boldsymbol{\beta} = 0$ for some fixed $r \times 1$ contrast vector. For this one-dimensional constraint, an equivalent to the Wald F test is to use the test statistic $Z = \mathbf{c}'\hat{\boldsymbol{\beta}}/\sqrt{\mathbf{c}'\mathbf{V}^{CR}\mathbf{c}}$, which follows a standard normal distribution in large samples. In small samples, following the CR1 approach, it is common to instead approximate the distribution of $Z$ by a $t(m-1)$ distribution. Hansen (2007) provided one justification for the use of this reference distribution by identifying conditions under which $Z$ converges in distribution to $t(m-1)$ as the within-cluster sample sizes grow large, with $m$ fixed (see also Donald and Lang, 2007). Ibragimov and Müller (2010) proposed a weighting technique derived so that that $t(m-1)$ critical values would be conservative (leading to rejection rates less than or equal to $\alpha$). However, both of these arguments require that $\mathbf{c}'\boldsymbol{\beta}$ be separately identified within each cluster. Outside of these circumstances, using $t(m-1)$ critical values can still lead to over-rejection (Cameron and Miller, 2015). Furthermore, using these critical values does not take into account that the distribution of $\mathbf{V}^{CR}$ is affected by the structure of the covariate matrix. An alternative, proposed by Bell and McCaffrey (2002), is to approximate the distribution of $Z$ by a $t$ distribution with degrees of freedom determined by a Satterthwaite approximation, under the working covariance model.

The t-test developed by Bell and McCaffrey (2002) involves using a $t(\nu)$ references distribution with degrees of freedom $\nu$, which are estimated by a Satterthwaite approximation. The Satterthwaite approximation (Satterthwaite, 1946) entails using degrees of freedom that are a function of the the first two moments of the sampling distribution of

$\mathbf{c}'\mathbf{V}^{CR}\mathbf{c}$. Theoretically, these degrees of freedom should be

$$\nu = \frac{2\left[\mathrm{E}\left(\mathbf{c}'\mathbf{V}^{CR2}\mathbf{c}\right)\right]^2}{\mathrm{Var}\left(\mathbf{c}'\mathbf{V}^{CR2}\mathbf{c}\right)}. \tag{12}$$

Expressions for the first two moments of $\mathbf{c}'\mathbf{V}^{CR2}\mathbf{c}$ can be derived under the assumption that the errors $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_m$ are normally distributed; see Appendix B.

In practice, both moments involve the variance structure $\boldsymbol{\Sigma}$, which is unknown. Bell and McCaffrey (2002) proposed to estimate the moments based on the same working model that is used to derive the adjustment matrices. This "model-assisted" estimate of the degrees of freedom is then calculated as

$$\nu_M = \frac{\left(\sum_{i=1}^m \mathbf{p}_i'\hat{\boldsymbol{\Phi}}\mathbf{p}_i\right)^2}{\sum_{i=1}^m \sum_{i=1}^m \left(\mathbf{p}_i'\hat{\boldsymbol{\Phi}}\mathbf{p}_i\right)^2}, \tag{13}$$

where $\mathbf{p}_i = (\mathbf{I} - \mathbf{H}_\mathbf{X})_i' \mathbf{A}_i \mathbf{W}_i \ddot{\mathbf{R}}_i \mathbf{M}_{\ddot{\mathbf{R}}} \mathbf{c}$. Alternately, for any of the CRVEs one could instead use an "empirical" estimate of the degrees of freedom, constructed by substituting $\mathbf{e}_i\mathbf{e}_i'$ in place of $\boldsymbol{\Sigma}_i$. However, Bell and McCaffrey (2002) found using simulation that this plug-in degrees of freedom estimate led to very conservative rejection rates.

The Bell and McCaffrey (2002) approach has been shown to perform well in a variety of conditions (CITE simulation studies). These studies encompass a variety of data generation processes, covariate types, and weighting procedures. A key finding is that the degrees of freedom depend not only on the number of clusters $m$, but also on features of the covariates. When the covariate is balanced across clusters—as occurs in balanced panels with a dichotomous covariate with the same proportion of ones in each cluster—the degrees of freedom are $m - 1$ even in small samples. However, when the covariate exhibits large imbalances—as occurs when the panel is not balanced or if the proportion of ones varies from cluster to cluster—the degrees of freedom can be considerably smaller. Similarly, covariates with large leverage points will tend to exhibit lower of degrees of freedom.

By adjusting the degrees of freedom to account for these features, the Type I error rate of the test is nearly-always less than or equal to nominal. This is in comparison to tests following the CR1 degrees of freedom (i.e., $m - 1$), wherein the test only performs well when in the cases in which the covariates are balanced. Importantly, because the degrees of freedom are covariate-dependent, it is not possible to assess whether a small-sample

correction is needed based solely on the total number of clusters in the data. Consequently, these studies argues that t-tests based on CRVE should routinely use the CR2 variance estimator and the Satterthwaite degrees of freedom, regardless even when $m$ appears to be large.

## 3.2 Small-sample corrections for F-tests

Little research has considered small-sample corrections for multiple-constraint hypothesis tests based on cluster-robust Wald statistics. Cameron and Miller highlight this problem, proposing a set of ad-hoc adjustments based on the BRL approach to t-tests, noting that some form of adjustment must be required given the extensive work on single-parameter tests. In this sub-section, we propose an approach to multi-paramter test that closely parallels the BRL method for t-tests. In this approach, we approximate the distribution of $Q/q$ by a multiple of an F distribution with estimated degrees of freedom. The sampling distribution of $Q$ is then approximated by Hotelling's $T^2$ distribution, a multiple of an $F$ distribution. Specifically, suppose that $\eta \mathbf{C}\mathbf{V}^{CR2}\mathbf{C}'$ approximately follows a Wishart distribution with $\eta$ degrees of freedom and scale matrix $\mathbf{C}\text{Var}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{C}'$, then

$$\left(\frac{\eta - q + 1}{\eta q}\right) Q \dot{\sim} F(q, \eta - q + 1). \tag{14}$$

We will refer to this as the approximate Hotelling's $T^2$ (AHT) test. We consider how to estimate $\eta$ below. This approach is conceptually similar to the Satterthwaite approximation for one-dimensional constraints and reduces to it if $q = 1$. For $q > 1$, however, the test depends on multivariate features of the covariates, including both CRVE estimates of variances and covariances.

Although not previously considered for Wald tests based on CRVE, the strategy of approximating the distribution of a robust variance estimator by a Wishart has been studied for some simpler models that are to special cases of CRVE. Zhang (2012*a*,1) described an AHT test for contrasts in analysis of variance models with unequal within-cell variance, which are particularly simple cases of linear models with heteroskedastic error terms. Zhang (2012*b*) extended the method to multivariate analysis of variance models where the covariance of the errors differs across cells, a special case of model (2) in which the CR2

variance estimator has a particularly simple form. In all of these cases, Zhang demonstrated that the robust variance estimator is a mixture of Wishart distributions that is well-approximated by a Wishart distribution with estimated degrees of freedom. Additionally, Pan and Wall (2002) described an F-test based on CR0 for use in GEE models, which also uses the Wishart approximation to the distribution of $\mathbf{V}^{CR}$ but estimates the degrees of freedom using a different method than the one we describe below.

The contribution of the present paper is to extend the AHT test to the more general setting of linear models with fixed effects. The remaining question is how to estimate the parameter $\eta$, which determines scalar multiplier and demoninator degrees of freedom of the F-test. To do so, we estimate the degrees of freedom of the Wishart distribution so that they match the mean and variance of $\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'$ under the working variance model $\mathbf{\Phi}$, just as in the degrees of freedom for the t-test. The problem that arises in doing so is that when $q > 1$ it is not possible to exactly match both moments. Pan and Wall (2002) proposed to choose $\eta$ to minimize the squared differences between the covariances among the entries of $\eta\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'$ and the covariances of the Wishart distribution with $\eta$ degrees of freedom and scale matrix $\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'$. Zhang (2012*b*) instead matches the mean and total variance of $\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'$ (i.e., the sum of the variances of its entries), which avoids the need to calculate any covariances. In what follows we focus on this latter approach, which we have found to perform better in practice (Tipton and Pustejovsky, forthcoming).

Let $\mathbf{c}_1, ..., \mathbf{c}_q$ denote the $p \times 1$ row-vectors of $\mathbf{C}$. Let $\mathbf{p}_{sh} = (\mathbf{I} - \mathbf{H})'_h \mathbf{A}'_h \mathbf{W}_h \mathbf{X}_h \mathbf{M} \mathbf{c}_s$ for $s = 1, ..., q$ and $h = 1, ..., m$. The degrees of freedom are then estimated under the working model as

$$\eta_M = \frac{\sum_{s,t=1}^{q} \sum_{h,i=1}^{m} b_{st} \mathbf{p}'_{sh} \hat{\mathbf{\Phi}} \mathbf{p}_{th} \mathbf{p}'_{si} \hat{\mathbf{\Phi}} \mathbf{p}_{ti}}{\sum_{s,t=1}^{q} \sum_{h,i=1}^{m} \mathbf{p}'_{sh} \hat{\mathbf{\Phi}} \mathbf{p}_{ti} \mathbf{p}'_{sh} \hat{\mathbf{\Phi}} \mathbf{p}_{ti} + \mathbf{p}'_{sh} \hat{\mathbf{\Phi}} \mathbf{p}_{si} \mathbf{p}'_{th} \hat{\mathbf{\Phi}} \mathbf{p}_{ti}}, \tag{15}$$

where $b_{st} = 1 + (s = t)$ for $s, t = 1, .., q$. Note that $\eta_M$ reduces to $\nu_M$ from Equation (13) if $q = 1$.

Note that this formulation is different from what we did in the JEBS paper. Are they equivalent? Are they both invariant to transformation of the constraint matrix?

This F-test shares features with the t-test developed by Bell and McCaffrey. Like the t-test, the degrees of freedom of this F-test depend not only on the number of clusters, but also on features of the covariates being tested. Again, these degrees of freedom can be much smaller than $m - 1$, and are particularly smaller when the covariates being tested exhibit high imbalances or leverage. Unlike the t-test case, however, in multi-parameter case, it

is often more difficult to diagnose the cause of these small degrees of freedom. In some situations, however, these are straightforward extensions to the findings in t-tests. For example, if the goal is to test if there are differences across a four-arm treatment study, the degrees of freedom are largest (and close to $m-1$) when the treatment is allocated equally across the four groups within each cluster. When the proportion varies across clusters, these degrees of freedom fall, often leading to degrees of freedom in the "small sample" territory even when the number of clusters is large.

# 4 Simulation Study

Throughout this paper, we have argued that the CR2 hypothesis tests (i.e., the BRL t-test and the AHT F-test) perform significantly better than the standard CR1 tests commonly used in economic applications. This argument is based on results from several large simulation studies. In this section, we review the design of these studies and their results, with particular emphasis on the role of covariate features and sample size on the Type I error rates of these tests. Throughought, we use "CR2" test to refer to that with both the BRL correction to the variance and the Satterthwaite degrees of freedom.

To date, four papers have examined the performance of the CR2 t-test; the results of these studies are indicated in TABLE 1. As Table 1 illustrates, these simulation studies have included a range of applications; for example, Cameron and Miller (2015) and Imbens and Kolesar (2015) have focused on conditions common in economics, while Bell and McCaffrey (2002) focused on those common in complex surveys, and Tipton (2015) on those in meta-analysis. Some of these studies have focused on policy dummies in the balanced case, while others have varied the degree of balance, and yet others have examined continous covariates that are symmetrically distributed, as well as those with high skew and leverage. In particular, these studies show that the CR1 t-test can have Type I error far above nominal when the covariate tested is unbalanced, skewed (i.e., high leverage), or when the number of observations per cluster varies. These studies have also examined the role of the number of clusters, with values ranging from 6 to 50, as well as the number of observations per clusters (from 1 to roughly 260). Results not shown in Table 1 indicate that differences between the CR1 and CR2 test performance have less to do with the number

of clusters, however, than the degrees of freedom, and that these degrees of freedom can be tremendously smaller than the number of clusters and depend on covariate features. Finally, each of the simulation studies tests how well the CR2 approach holds up to model misspecification, particularly since the approach requires specification of a "working" model (see columns for "error structure"). To do so, these studies have simulated data with various degress of heteroskedasticity and clustering, while using working models that are more simple.

Table 1 also highlights the range of Type I error values found in these studies for both the CR1 and CR2 tests. Looking over all the conditions studied, this ranges 0.001 - 0.13 for CR2, compared to 0.01 - 0.34 for CR1, with these differences being most pronounced for covariates that are unbalanced or have high leverage. Importantly, both Bell and McCaffrey (2002) and Tipton (2015) highlight that the t-distribution approximation used in the CR2 test holds only when the degrees of freedom are greater than about 4 or 5, and therefore break out their results separately for these two cases. These results (not shown here) indicate that the CR2 test has nominal Type I error whenever the degrees of freedom are greater than 4 or 5, but can have Type I error slightly above or below nominal for smaller degrees of freedom. This is important since the highest Type I error for the CR2 test is indicated in the Imbens and Kolesar (2015) paper (i.e., 0.13), and this paper does not separate results out by degrees of freedom. Given the conditions studied (i.e., only 3 out of 30 clusters have a policy), we would expect that the degrees of freedom here are very small, thus explaining the larger than typical Type I error.

In comparison to the t-test, the CR2 F-test has only been tested by fewer authors. Tipton and Pustejovsky (2015) initially proposed the test in the meta-analytic case, however, and conducted extensive simulations in the conditions encountered there commonly. These conditions are found in the final panel of Table 1. As these rows indicate, these simulations varied the number of clusters $m$ (10 - 100), as well as the number of within-cluster observations $n$ (from 1 to 10), the dimension $q$ of the test (from 2 - 5) and the number of covariates in the regression model $p$ from 2 - 5. In the meta-analytic context, it is typical to use inverse-variance weighting (akin to feasible-WLS) to account for heteroskedasticity, and these simulations thus base both the working model and the weights upon a simple

heteroscedastic structure. The true correlation structure is varied far from this by inclusion of both correlations between errors within studies and random effects between studies; this is similar to the structure used in Tipton (2015). As Table 2 indicates, over this range of conditions, the Type I error for the CR2 F-test ranges between XXX-XXX, while varying considerably from XX-XXX for the CR1 test.

Given the fact that the CR2 F-test is newer, in this section we conduct a new simulation study focused on the contexts found more commonly in the economics literature. This includes tests where $q$ is relatively small (e.g., 2 - 4) while $p$ is larger (since the model often includes demographic controls). Most importantly, we focus on the case in which the working model assumes homoscedasticity (what the literature refers to as the "BM" working model) or assumes random effects (suggested by Imbens-Kolesar). We also increase the within-study sample size to $n = 100$, to become more closely in line with conditions found in practice. Like previous studies, we then vary the true correlation structure so as to test the sensitivity of the CR2 approach to model misspecification.

### 4.1 Results

# 5 EXAMPLES

In this section we examine three short examples of the use of CRVE with small samples, spanning a variety of applied contexts. In the first example, the effects of substantive interest are identified within each cluster. In the second example, the effects involve between-cluster contrasts. The third example involves a cluster-robust Hausmann test for differences between within- and across-cluster information. In each example, we illustrate how the proposed small-sample t- and F-tests can be used and how they can differ from both the standard CR1 and Wild bootstrap tests. R code and data files are available for each analysis as an online supplement.

### 5.1 Tennessee STAR class-size experiment.

The Tennessee STAR class size experiment is one of the most intensively studied interventions in education. The experiment involved students in kindergarten through third grade

across 79 schools. Within each school, students and their teachers were randomized to one of three conditions: small class-size (targetted to have 13-17 students), regular class-size, or regular class-size with an aide (for a more detailed review, see Schanzenbach, 2006). Analyses of the original study and follow-up waves have found that being in a small class improves a variety of student outcomes, including higher test scores (Schanzenbach, 2006), increased likelihood of taking college entrance exams (Krueger and Whitmore, 2001), and increased rates of home ownership and earnings (Chetty, Friedman, Hilger, Saez, Schanzenbach and Yagan, 2011).

The class-size experiment consists of three treatment conditions and multiple, student-level outcomes of possible interest. Here, we examine the results for the subgroup of students who were in kindergarten during the first year of the study, focusing on reading, word recognition, and math scores measured at the end of kindergarten (Achilles, Bain, Bellott, Boyd-Zaharias, Finn, Folger, Johnston and Word, 2008). Following Krueger and Whitmore (2001), we put the three outcomes on comparable scales by converting the scores to percentile ranks based upon their distributions in the control condition. The analytic model is then

$$y_{ijk} = \mathbf{r}'_{ij}\boldsymbol{\beta}_k + \mathbf{s}'_{ij}\boldsymbol{\gamma}_0 + \gamma_k + \mu_i + \epsilon_{ijk}, \tag{16}$$

where $y_{ijk}$ is the percentile rank on outcome $k$ for student $j$ in school $i$; $\mathbf{r}_{ij}$ includes indicators for the small-class and regular-plus-aide conditions; $\mathbf{s}_{ij}$ includes student demographic covariates (i.e., free or reduced-price lunch status; race; gender; age); $\gamma_k$ is a fixed effect for outcome $k$; and $\mu_i$ is a fixed effect for school $i$.

We estimated the model in two ways. First, we estimated $\boldsymbol{\beta}_k$ separately for each outcome $k$ and tested the null hypothesis that $\boldsymbol{\beta}_k = \mathbf{0}$. Second, we use the seemingly unrelated regression (SUR) framework to simultaneously test for treatment effects across conditions and across outcomes. In the SUR model, separate treatment effects are estimated for each outcome, but the student demographic effects and school fixed effects are pooled across outcomes. An overall test of the differences between conditions thus amounts to testing the null hypothesis that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = \mathbf{0}$. In all models, we estimated $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}$ after absorbing the school fixed effects and clustered the errors by school.

Table 1 displays the results of the hypothesis tests. Across all outcomes, use of CR2

| Outcome | Correction | Test | F | df | p |
|---|---|---|---|---|---|
| Reading | CR1 | F | 9.073 | 78.0 | 0.00029 |
| | CR2 | F | 9.046 | 78.0 | 0.00029 |
| | CR2 | AHZ | 8.916 | 68.5 | 0.00036 |
| Math | CR1 | F | 6.838 | 78.0 | 0.00183 |
| | CR2 | F | 6.823 | 78.0 | 0.00186 |
| | CR2 | AHZ | 6.725 | 68.6 | 0.00215 |
| Word recognition | CR1 | F | 8.094 | 78.0 | 0.00064 |
| | CR2 | F | 8.073 | 78.0 | 0.00065 |
| | CR2 | AHZ | 7.957 | 68.5 | 0.00078 |
| Combined (SUR) | CR1 | F | 3.284 | 78.0 | 0.00622 |
| | CR2 | F | 3.276 | 78.0 | 0.00632 |
| | CR2 | AHZ | 3.042 | 64.9 | 0.01103 |

Table 1: Tests of treatment effects in the Tennessee STAR class size experiment

variance estimates with the AHZ test leads to slightly smaller F statistics, smaller degrees of freedom, and larger $p$-values compared to use of CR1 or CR2 with $J - 1$ degrees of freedom. In all cases, the reduction in degrees of freedom is relatively minor (i.e., less than 10 for the tests of individual outcomes and less than 15 for the joint test). This is because the class-size experiment used a randomized block design, and so the constraints being tested are identified within every cluster and are relatively balanced. Nonetheless, the use of the AHZ test still makes a bigger difference in the F statistic and $p$-values than does the use of CR2 standard errors.

Why do degrees of freedom change across outcomes? Is it just little changes in missing observations? Should we just show reading and SUR results?

Should we do any t-tests?

Why does SUR test have bigger p-value than individual tests?

## 5.2 Achievement Awards demonstration

Angrist and Lavy (2009) reported results from a randomized trial in Israel that aimed to

increase completion rates of the Bagrut, the national matriculation certificate for post-secondary education, among low-achieving high school students. In the Achievement Awards demonstration, 40 non-vocational high schools with low rates of Bagrut completion were selected from across Israel, including 10 Arab and 10 Jewish religious schools and 20 Jewish secular schools. The 40 schools were then pair-matched based on 1999 rates of Bagrut completion, and within each pair one school was randomized to receive a cash-transfer program. In these treatment schools, seniors who completed certification were eligible for payments of approximately $1,500. Student-level covariate and outcome data were drawn from administrative records, available for the school years ending in June of 2000, 2001, and 2002.

Our analysis focuses the effect of the cash incentive program on Bagrut completion rates. For sake of simplicity, we report results for the sub-sample of female students. Bagrut completion rates for June, 2000 constitute baseline measures. The incentive program was in effect for the group of seniors in treatment schools taking the Bagrut exams in Spring of 2001. However, the program was discontinued for the following year, and so we treat 2002 completion rates as being unaffected by treatment assignment.

We estimate two distinct models and tested several different hypotheses in order to demonstrate the performance of the small-sample hypothesis tests under a range of condition. First, we test for whether the average effect of the program differs from zero. Angrist and Lavy (2009) provided evidence that the program effects for female students are moderated by the prior academic performance. We therefore estimate separate average effects for students in the lower and upper halves of the distribution of prior test scores. Let $h = 1, 2, 3$ index the sector of each school (Arab religious, Jewish religious, or Jewish secular). We then use the model:

$$y_{hitj} = z_{hit}\mathbf{r}'_{hitj}\boldsymbol{\beta} + \mathbf{s}'_{hitj}\boldsymbol{\gamma} + \gamma_{ht} + \mu_{hi} + \epsilon_{hitj} \tag{17}$$

In this model for student $j$ in year $t$ in school $i$ in sector $h$, $z_{hit}$ is an indicator equal to one in the treatment schools for the 2001 school year and otherwise equal to zero; $\mathbf{r}_{hitj}$ is a vector of indicators for whether the student is in the lower or upper half of the distribution of prior academic performance; and $\boldsymbol{\beta} = (\beta_1, \beta_2)$ is a vector of average treatment effects. The vector $\mathbf{s}_{hitj}$ includes the following individual student demographic measures: mother's

and father's education, immigration status, number of siblings, and indicators for each quartile in the distribution of prior-year academic performance. The model also includes fixed effects $\gamma_{ht}$ for each sector in each year and $\mu_{hi}$ for each school.

Based on Model (17), we test for whether the average treatment effects in the lower and upper halves differ from zero, first as separate t-tests (i.e., $H_1 : \beta_1 = 0$, $H_2 : \beta_2 = 0$) and then using a joint F-test (i.e., $H_3 : \beta_1 = \beta_2 = 0$). In a modified model, we test for whether school sector moderates the program effects. To do so, we allow the coefficients on the average treatment effects to vary by sector, replacing $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_h$ in Model (17). We then test for the equality of the treatment effects (i.e., $H_4 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3$).

| Hypothesis | Correction | Test | F | df | p |
|---|---|---|---|---|---|
| ATE - lower half (q = 1) | CR1 | F | 0.079 | 34.00 | 0.78065 |
| | CR2 | F | 0.068 | 34.00 | 0.79573 |
| | CR2 | AHZ | 0.068 | 20.67 | 0.79674 |
| ATE - upper half (q = 1) | CR1 | F | 5.746 | 34.00 | 0.02217 |
| | CR2 | F | 5.169 | 34.00 | 0.02943 |
| | CR2 | AHZ | 5.169 | 15.86 | 0.03726 |
| ATE - joint (q = 2) | CR1 | F | 3.848 | 34.00 | 0.03116 |
| | CR2 | F | 3.589 | 34.00 | 0.03854 |
| | CR2 | AHZ | 3.371 | 15.46 | 0.06096 |
| Moderation by sector (q = 4) | CR1 | F | 8.213 | 34.00 | 0.00010 |
| | CR2 | F | 5.603 | 34.00 | 0.00142 |
| | CR2 | AHZ | 2.895 | 3.21 | 0.19446 |

Table 2: Tests of treatment effects in the Achievement Awards Demonstration

Table 2 reports the results of all four hypothesis tests, using the CR1 and CR2 adjustments for variance estimation and either the naivë degrees of freedom (for these data, $m - 1 = 34$) or the approximate Hotelling's $T^2$-Z degrees of freedom. Several features of these results are worth noting. First, for the tests of single-parameter hypotheses, the differences between the CR1 and CR2 test statistics are minor, although the AHZ (Sat-

terthwaite) degrees of freedom are substantially lower than the naivë ones. As a result, the $p$-value for the ATE in the upper half of the prior achievement distribution is 68% larger based on our preferred test than based on the standard F-test with CR1. Second, the joint test that the ATEs are zero in both halves of the prior achivement distribution is sensitive to both the CR adjustment and the degrees of freedom: the $p$-value based on our preferred test is 0.061, compared to the $p = 0.031$ for the standard test. Finally, the test of moderation by sector is strongly affected by the CR2 adjustment and AHZ degrees of freedom. In particular, the AHZ degrees of freedom are just 3.21, far lower than the number of clusters. The low degrees of freedom are a consequence of the small number of schools of each type in each treatment condition. Although the total sample includes 35 schools, there are only 9 Arab religious, 7 Jewish religious, and 19 Jewish secular schools, each split across two treatment conditions, and the treatment effects are estimated by making comparisons across clusters. As a result, the F statistic based on CR2 has a very large sampling variance under the null hypothesis.

Connect to simulation results–higher $q$, between-cluster

## 5.3 Robust Hausmann test

In this final example, we shift focus from analyses of experiments to panel data. Here we build off of an example first developed in Bertrand et al. (2004) using Current Population Survey (CPS) data to relate demographics to earnings. Following Cameron and Miller (2015), we aggregated the data from the individual level to the time period, producing a balanced panel with $T = 36$ time points within 51 states (including the District of Columbia). We focus on the model,

$$y_{it} = \mathbf{r}_{it}'\boldsymbol{\beta} + \gamma_t + \mu_i + \epsilon_{it}. \tag{18}$$

In this model, time-point $t$ is nested within state $i$; the outcome $y_{it}$ is log-earnings, which are reported in 1999 dollars; $\mathbf{r}_{it}$ is a vector of demographic covariates specific to the time point, including years of education, age, the square of age, and an indicator for female; $\gamma_t$ is a fixed effect for time point $t$; and $\mu_i$ is an effect for state $i$.

For sake of example, we focus here on a Hausman test to determine whether to use a fixed effects (FE) estimator or a random effects (RE) estimator for the four parameters

in $\boldsymbol{\beta}$. In an OLS model with uncorrelated, the Hausmann test directly compares the vectors of FE and RE estimates using a chi-squared test. Although this formulation of the Hausmann test is not available when cluster-robust standard errors are employed, it remains possible to use an artificial-Hausman test (Arellano, 1993; Wooldridge, 2002). This test instead amends the model to include within-cluster deviations (or cluster aggregates) of the variables of interest. In our example, this becomes,

$$y_{it} = \mathbf{r}_{it}'\boldsymbol{\beta} + \ddot{\mathbf{r}}_{it}'\boldsymbol{\delta} + \gamma_t + \mu_i + \epsilon_{it}, \tag{19}$$

where $\ddot{\mathbf{r}}_{it}$ denotes the vector of within-cluster deviations of the covariates (i.e., $\ddot{\mathbf{r}}_{it} = \mathbf{r}_{it} - \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_{it}$). The four parameters in $\boldsymbol{\delta}$ represent the differences between the between-panel and within-panel estimates of $\boldsymbol{\beta}$. The artificial Hausmann test therefore reduces to testing the null hypothesis that $\boldsymbol{\beta} = \mathbf{0}$ using an F test with $q = 4$. We estimate the model using WLS with weights derived under the assumption that $\mu_1, ..., \mu_m$ are mutually independent, normally distributed, and independent of $\epsilon_{it}$.

# 6 DISCUSSION

# A BRL adjustment matrices

This appendix states and provides proof of two theorems regarding the BRL adjustment matrices.

**Theorem 3.** *Let* $\mathbf{L} = \left(\ddot{\mathbf{U}}'\ddot{\mathbf{U}} - \ddot{\mathbf{U}}_i'\ddot{\mathbf{U}}_i\right)$ *and assume that* $\mathbf{L}$ *has full rank* $r + s$*, so that its inverse exists. Then the adjustment matrices* $\mathbf{A}_i$ *defined in (8) and (9) satisfy criterion (7) and* $\mathbf{V}^{CR2}$ *is exactly unbiased when the working covariance model* $\boldsymbol{\Phi}$ *is correctly specified.*

*Proof.* The Moore-Penrose inverse of $\mathbf{B}_i$ can be computed from its eigen-decomposition. Let $b \leq n_i$ denote the rank of $\mathbf{B}_i$. Let $\boldsymbol{\Lambda}$ be the $b \times b$ diagonal matrix of the positive eigenvalues of $\mathbf{B}_i$ and $\mathbf{V}$ be the $n_i \times b$ matrix of corresponding eigen-vectors, so that $\mathbf{B}_i = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'$. Then $\mathbf{B}_i^{+} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}'$ and $\mathbf{B}_i^{+1/2} = \mathbf{V}\boldsymbol{\Lambda}^{-1/2}\mathbf{V}'$.

Now, observe that $(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}})_i (\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}}) (\mathbf{I} - \mathbf{H}_{\mathbf{T}}) = (\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i$. Thus,

$$\begin{aligned}\ddot{\mathbf{R}}_i' \mathbf{W}_i \mathbf{A}_i (\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i \boldsymbol{\Phi} (\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i' \mathbf{A}_i' \mathbf{W}_i \ddot{\mathbf{R}}_i &= \ddot{\mathbf{R}}_i' \mathbf{W}_i \mathbf{D}_i \mathbf{B}_i^{+1/2} \mathbf{B}_i \mathbf{B}_i^{+1/2} \mathbf{D}_i' \mathbf{W}_i \ddot{\mathbf{R}}_i \\ &= \ddot{\mathbf{R}}_i' \mathbf{W}_i \mathbf{D}_i \mathbf{V}\mathbf{V}' \mathbf{D}_i' \mathbf{W}_i \ddot{\mathbf{R}}_i. \end{aligned} \tag{20}$$

Because $\mathbf{D}_i$, and $\boldsymbol{\Phi}$ are positive definite and $\mathbf{B}_i$ is symmetric, the eigenvectors $\mathbf{V}$ define an orthonormal basis for the column span of $(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}})_i$. We now show that $\ddot{\mathbf{U}}_i$ is in the column space of $(\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i$. Let $\mathbf{Z}_i$ be an $n_i \times (r+s)$ matrix of zeros. Let $\mathbf{Z}_k = -\ddot{\mathbf{U}}_k \mathbf{L}^{-1} \mathbf{M}_{\ddot{\mathbf{U}}}^{-1}$, for $k \neq j$ and take $\mathbf{Z} = (\mathbf{Z}_1', ..., \mathbf{Z}_m')'$. Now observe that $(\mathbf{I} - \mathbf{H}_{\mathbf{T}}) \mathbf{Z} = \mathbf{Z}$. It follows that

$$\begin{aligned}(\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i \mathbf{Z} &= (\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}})_i (\mathbf{I} - \mathbf{H}_{\mathbf{T}}) \mathbf{Z} = (\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}})_i \mathbf{Z} \\ &= \mathbf{Z}_i - \ddot{\mathbf{U}}_i \mathbf{M}_{\ddot{\mathbf{U}}} \sum_{k=1}^{m} \ddot{\mathbf{U}}_k' \mathbf{W}_k \mathbf{Z}_k = \ddot{\mathbf{U}}_i \mathbf{M}_{\ddot{\mathbf{U}}} \left( \sum_{k \neq j} \ddot{\mathbf{U}}_k' \mathbf{W}_k \ddot{\mathbf{U}} \right) \mathbf{L}^{-1} \mathbf{M}_{\ddot{\mathbf{U}}}^{-1} \\ &= \ddot{\mathbf{U}}_i.\end{aligned}$$

Thus, there exists an $N \times (r+s)$ matrix $\mathbf{Z}$ such that $(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}})_i \mathbf{Z} = \ddot{\mathbf{U}}_i$, i.e., $\ddot{\mathbf{U}}_i$ is in the column span of $(\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i$. Because $\mathbf{D}_i \mathbf{W}_i$ is positive definite and $\ddot{\mathbf{R}}_i$ is a sub-matrix of $\ddot{\mathbf{U}}_i$, $\mathbf{D}_i \mathbf{W}_i \ddot{\mathbf{R}}_i$ is also in the column span of $(\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i$. It follows that

$$\ddot{\mathbf{R}}_i' \mathbf{W}_i \mathbf{D}_i \mathbf{V}\mathbf{V}' \mathbf{D}_i' \mathbf{W}_i \ddot{\mathbf{R}}_i = \ddot{\mathbf{R}}_i' \mathbf{W}_i \boldsymbol{\Phi}_i \mathbf{W}_i \ddot{\mathbf{R}}_i. \tag{21}$$

Substituting (21) into (20) demonstrates that $\mathbf{A}_i$ satisfies criterion (7).

Under the working model, the residuals from cluster $i$ have mean $\mathbf{0}$ and variance

$$\text{Var}\left(\ddot{\mathbf{e}}_i\right) = (\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i \boldsymbol{\Phi} (\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i',$$

It follows that

$$\begin{aligned}\text{E}\left(\mathbf{V}^{CR2}\right) &= \mathbf{M}_{\ddot{\mathbf{R}}} \left[ \sum_{i=1}^{m} \ddot{\mathbf{R}}_i' \mathbf{W}_i \mathbf{A}_i (\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i \boldsymbol{\Phi} (\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i' \mathbf{A}_i \mathbf{W}_i \ddot{\mathbf{R}}_i \right] \mathbf{M}_{\ddot{\mathbf{R}}} \\ &= \mathbf{M}_{\ddot{\mathbf{R}}} \left[ \sum_{i=1}^{m} \ddot{\mathbf{R}}_i' \mathbf{W}_i \boldsymbol{\Phi}_i \mathbf{W}_i \ddot{\mathbf{R}}_i \right] \mathbf{M}_{\ddot{\mathbf{R}}} \\ &= \text{Var}\left(\hat{\boldsymbol{\beta}}\right)\end{aligned}$$

□

**Theorem 4.** *Let* $\tilde{\mathbf{A}}_i = \mathbf{D}_i'\tilde{\mathbf{B}}_i^{+1/2}\mathbf{D}_i$*, where* $\tilde{\mathbf{B}}_i$ *is given in (10). If* $\mathbf{T}_i\mathbf{T}_k' = \mathbf{0}$ *for* $j \neq k$ *and* $\mathbf{W} = \mathbf{\Phi}^{-1}$*, then* $\mathbf{A}_i = \tilde{\mathbf{A}}_i$*.*

*Proof.* From the fact that $\ddot{\mathbf{U}}_i'\mathbf{W}_i\mathbf{T}_i = \mathbf{0}$ for $i = 1, ..., m$, it follows that

$$\begin{aligned}
\mathbf{B}_i &= \mathbf{D}_i \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right)_i \left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right) \hat{\mathbf{\Phi}} \left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right)' \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right)_i' \mathbf{D}_i' \\
&= \mathbf{D}_i \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}} - \mathbf{H}_{\mathbf{T}}\right)_i \hat{\mathbf{\Phi}} \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}} - \mathbf{H}_{\mathbf{T}}\right)_i' \mathbf{D}_i' \\
&= \mathbf{D}_i \left(\mathbf{\Phi}_i - \ddot{\mathbf{U}}_i \mathbf{M}_{\ddot{\mathbf{U}}} \ddot{\mathbf{U}}_i' - \mathbf{T}_i \mathbf{M}_{\mathbf{T}} \mathbf{T}_i'\right) \mathbf{D}_i'
\end{aligned}$$

and

$$\mathbf{B}_i^+ = \left(\mathbf{D}_i'\right)^{-1} \left(\mathbf{\Phi}_i - \ddot{\mathbf{U}}_i \mathbf{M}_{\ddot{\mathbf{U}}} \ddot{\mathbf{U}}_i' - \mathbf{T}_i \mathbf{M}_{\mathbf{T}} \mathbf{T}_i'\right)^+ \mathbf{D}_i^{-1}. \tag{22}$$

Let $\mathbf{\Omega}_i = \left(\mathbf{\Phi}_i - \ddot{\mathbf{U}}_i \mathbf{M}_{\ddot{\mathbf{U}}} \ddot{\mathbf{U}}_i'\right)^+$. Using a generalized Woodbury identity (Henderson and Searle, 1981),

$$\mathbf{\Omega}_i = \mathbf{W}_i + \mathbf{W}_i \ddot{\mathbf{U}}_i \mathbf{M}_{\ddot{\mathbf{U}}} \left(\mathbf{M}_{\ddot{\mathbf{U}}} - \mathbf{M}_{\ddot{\mathbf{U}}} \ddot{\mathbf{U}}_i' \mathbf{W}_i \ddot{\mathbf{U}}_i \mathbf{M}_{\ddot{\mathbf{U}}}\right)^+ \mathbf{M}_{\ddot{\mathbf{U}}} \ddot{\mathbf{U}}_i' \mathbf{W}_i.$$

It follows that $\mathbf{\Omega}_i \mathbf{T}_i = \mathbf{W}_i \mathbf{T}_i$. Another application of the generalized Woodbury identity gives

$$\begin{aligned}
\left(\mathbf{\Phi}_i - \ddot{\mathbf{U}}_i \mathbf{M}_{\ddot{\mathbf{U}}} \ddot{\mathbf{U}}_i' - \mathbf{T}_i \mathbf{M}_{\mathbf{T}} \mathbf{T}_i'\right)^+ &= \mathbf{\Omega}_i + \mathbf{\Omega}_i \mathbf{T}_i \mathbf{M}_{\mathbf{T}} \left(\mathbf{M}_{\mathbf{T}} - \mathbf{M}_{\mathbf{T}} \mathbf{T}_i' \mathbf{\Omega}_i \mathbf{T}_i \mathbf{M}_{\mathbf{T}}\right)^+ \mathbf{M}_{\mathbf{T}} \mathbf{T}_i' \mathbf{\Omega}_i \\
&= \mathbf{\Omega}_i + \mathbf{W}_i \mathbf{T}_i \mathbf{M}_{\mathbf{T}} \left(\mathbf{M}_{\mathbf{T}} - \mathbf{M}_{\mathbf{T}} \mathbf{T}_i' \mathbf{W}_i \mathbf{T}_i \mathbf{M}_{\mathbf{T}}\right)^+ \mathbf{M}_{\mathbf{T}} \mathbf{T}_i' \mathbf{W}_i \\
&= \mathbf{\Omega}_i.
\end{aligned}$$

The last equality follows from the fact that $\mathbf{T}_i \mathbf{M}_{\mathbf{T}} \left(\mathbf{M}_{\mathbf{T}} - \mathbf{M}_{\mathbf{T}} \mathbf{T}_i' \mathbf{W}_i \mathbf{T}_i \mathbf{M}_{\mathbf{T}}\right)^- \mathbf{M}_{\mathbf{T}} \mathbf{T}_i' = \mathbf{0}$ because the fixed effects are nested within clusters. Substituting into (22), we then have that $\mathbf{B}_i^+ = \left(\mathbf{D}_i'\right)^{-1} \mathbf{\Omega}_i \mathbf{D}_i^{-1}$. But

$$\tilde{\mathbf{B}}_i = \mathbf{D}_i \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right)_i \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right)_i' \mathbf{D}_i' = \mathbf{D}_i \left(\mathbf{\Phi}_i - \ddot{\mathbf{U}}_i \mathbf{M}_{\ddot{\mathbf{U}}} \ddot{\mathbf{U}}_i'\right) \mathbf{D}_i' = \mathbf{D}_i \mathbf{\Omega}_i^+ \mathbf{D}_i',$$

and so $\mathbf{B}_i^+ = \tilde{\mathbf{B}}_i^+$. It follows that $\mathbf{A}_i = \tilde{\mathbf{A}}_i$ for $i = 1, ..., m$. □

# B DISTRIBUTION THEORY FOR $\mathbf{V}^{CR}$

The small-sample approximations for t-tests and F-tests both involve the distribution of the entries of $\mathbf{V}^{CR2}$. This appendix explains the relevant distribution theory.

Standard results regarding quadratic forms can be used to derive the moments of the linear combination (e.g., Searle, 2006, Sec. 13.5). We now assume that $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_m$ are multivariate normal with zero mean and variance $\boldsymbol{\Sigma}$. It follows that

$$\mathrm{E}\left(\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2\right) = \sum_{i=1}^{m} \mathbf{p}_{1j}'\boldsymbol{\Sigma}\mathbf{p}_{2j} \tag{23}$$

$$\mathrm{Var}\left(\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2\right) = \sum_{i=1}^{m}\sum_{i=1}^{m} \left(\mathbf{p}_{1i}'\boldsymbol{\Sigma}\mathbf{p}_{2j}\right)^2 + \mathbf{p}_{1i}'\boldsymbol{\Sigma}\mathbf{p}_{1j}\mathbf{p}_{2i}'\boldsymbol{\Sigma}\mathbf{p}_{2j} \tag{24}$$

$$\mathrm{Cov}\left(\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2, \mathbf{c}_3'\mathbf{V}^{CR}\mathbf{c}_4\right) = \sum_{i=1}^{m}\sum_{i=1}^{m} \mathbf{p}_{1i}'\boldsymbol{\Sigma}\mathbf{p}_{4j}\mathbf{p}_{2i}'\boldsymbol{\Sigma}\mathbf{p}_{3j} + \mathbf{p}_{1i}'\boldsymbol{\Sigma}\mathbf{p}_{3j}\mathbf{p}_{2i}'\boldsymbol{\Sigma}\mathbf{p}_{4j}. \tag{25}$$

Furthermore, the distribution of $\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2$ can be expressed as a weighted sum of $\chi_1^2$ distributions (Mathai and Provost, 1992), with weights given by the eigen-values of the $m \times m$ matrix with $(i,j)^{th}$ entry $\mathbf{p}_{1i}'\boldsymbol{\Sigma}\mathbf{p}_{2j}$, $i, i = 1, ..., m$.

# References


Achilles, C. M., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J. and Word, E. (2008), 'Tennessee's Student Teacher Achievement Ratio (STAR) project'.

**URL:** *http://hdl.handle.net/1902.1/10766*

Angrist, J. D. and Lavy, V. (2009), 'The effects of high stakes high school achievement awards : Evidence from a randomized trial', *American Economic Review* **99**(4), 1384–1414.

Angrist, J. D. and Pischke, J. (2009), *Mostly harmless econometrics: An empiricist's companion*, Princeton University Press, Princeton, NJ.

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

Arellano, M. (1993), 'On the testing of correlated effects with panel data', *Journal of Econometrics* **59**(1-2), 87–97.

Banerjee, S. and Roy, A. (2014), *Linear Algebra and Matrix Analysis for Statistics*, Taylor & Francis, Boca Raton, FL.

Bell, R. M. and McCaffrey, D. F. (2002), 'Bias reduction in standard errors for linear regression with multi-stage samples', *Survey Methodology* **28**(2), 169–181.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004), 'How much should we trust differences-in-differences estimates?', *Quarterly Journal of Economics* **119**(1), 249–275.

Cameron, A. C., Gelbach, J. B. and Miller, D. (2008), 'Bootstrap-based improvements for inference with clustered errors', *The Review of Economics and Statistics* **90**(3), 414–427.

Cameron, A. C. and Miller, D. L. (2015), A practitioner's guide to cluster-robust inference.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W. and Yagan, D. (2011), 'How does your kindergarten classroom affect your earnings? Evidence from Project STAR', *The Quarterly Journal of Economics* **126**(4), 1593–1660.

Donald, S. G. and Lang, K. (2007), 'Inference with difference-in-differences and other panel data', *Review of Economics and Statistics* **89**(2), 221–233.

Eicker, F. (1967), Limit theorems for regressions with unequal and dependent errors, *in* 'Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, pp. 59–82.

Hansen, C. B. (2007), 'Asymptotic properties of a robust variance matrix estimator for panel data when T is large', *Journal of Econometrics* **141**, 597–620.

Henderson, H. V. and Searle, S. R. (1981), 'On deriving the inverse of a sum of matrices', *Siam Review* **23**(1), 53–60.

Huber, P. J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions, *in* 'Proceedings of the fifth Berkeley symposium on Mathematical Statistics and Probability', University of California Press, Berkeley, CA, pp. 221–233.

Ibragimov, R. and Müller, U. K. (2010), 't-Statistic based correlation and heterogeneity robust inference', *Journal of Business & Economic Statistics* **28**(4), 453–468.

Imbens, G. W. and Kolesar, M. (2012), Robust standard errors in small samples: some practical advice.

**URL:** *http://www.nber.org/papers/w18478*

Krueger, A. and Whitmore, D. (2001), 'The effect of attending a small class in the early

grades on college-test taking and middle school test results: Evidence from Project STAR', *The Economic Journal* **111**(468), 1–28.

Liang, K.-Y. and Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

MacKinnon, J. G. and White, H. (1985), 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties', *Journal of Econometrics* **29**, 305–325.

Mancl, L. A. and DeRouen, T. A. (2001), 'A covariance estimator for GEE with improved small-sample properties', *Biometrics* **57**(1), 126–134.

Mathai, A. M. and Provost, S. B. (1992), *Quadratic forms in random variables: theory and applications*, M. Dekker, New York.

McCaffrey, D. F. and Bell, R. M. (2006), 'Improved hypothesis testing for coefficients in generalized estimating equations with small samples of clusters.', *Statistics in medicine* **25**(23), 4081–98.

McCaffrey, D. F., Bell, R. M. and Botts, C. H. (2001), Generalizations of biased reduced linearization, *in* 'Proceedings of the Annual Meeting of the American Statistical Association', number 1994.

Pan, W. and Wall, M. M. (2002), 'Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations.', *Statistics in medicine* **21**(10), 1429–41.

Satterthwaite, F. E. (1946), 'An approximate distribution of estimates of variance components', *Biometrics bulletin* **2**(6), 110–114.

Schanzenbach, D. W. (2006), 'What have researchers learned from Project STAR?', *Brookings Papers on Education Policy* **2006**(1), 205–228.

Searle, S. R. (2006), *Matrix Algebra Useful for Statistics*, john wiley edn, Hoboken, NJ.

Tipton, E. and Pustejovsky, J. E. (forthcoming), 'Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression', *Journal of Educational and Behavioral Statistics* .

Webb, M. and MacKinnon, J. G. (2013), Wild Bootstrap Inference for Wildly Different Cluster Sizes.

White, H. (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica* **48**(4), 817–838.

White, H. (1984), *Asymptotic theory for econometricians*, Academic Press, Inc., Orlando, FL.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edn, MIT Press, Cambridge, MA.

Zhang, J.-T. (2012*a*), 'An approximate degrees of freedom test for heteroscedastic two-way ANOVA', *Journal of Statistical Planning and Inference* **142**(1), 336–346.

Zhang, J.-T. (2012*b*), 'An approximate Hotelling T2 -test for heteroscedastic one-way MANOVA', *Open Journal of Statistics* **2**, 1–11.

Zhang, J.-T. (2013), 'Tests of linear hypotheses in the ANOVA under heteroscedasticity', *International Journal of Advanced Statistics and Probability* **1**(2), 9–24.