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

December 21, 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 observations. Likewise, CRVE is now routinly used in the analysis of panel data to account for correlation of measurements on the same individual units 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. These standard errors 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, demographic controls, and 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 us-

ing 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 $\text{Var}(\mathbf{e}_i | \mathbf{X}_i) = \mathbf{\Phi}_i$, where $\mathbf{\Phi}_i$ is a known 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—can 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 $\mathbf{W} = \text{diag}(\mathbf{W}_1, ..., \mathbf{W}_m)$. 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 $\mathbf{M}_\mathbf{Z} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}$ and $\mathbf{H}_\mathbf{Z} = \mathbf{Z}\mathbf{M}_\mathbf{Z}\mathbf{Z}'\mathbf{W}$.

The absorption technique involves obtaining the residuals from the regression of $\mathbf{y}$ on

[1] The WLS estimator also encompasses the estimator proposed by Ibragimov and Müller (2010) for 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 $\mathbf{W}_i = \mathbf{X}_i (\mathbf{X}_i'\mathbf{X}_i)^{-2} \mathbf{X}_i$.

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

James – it says here that this is general (dummies or absorbed), but the math indicates absorbtion. It may be better to start by saying our focus is on absorbing since it is more common, and then later discuss how it's important to show equivalence to the dummy case.

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 $\mathbf{\Sigma}_i$ 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. Note that when $N >> p$, $c \approx \sqrt{m/(m-1)}$, and software typically uses the latter approximation (e.g., SAS); for this reason we refer to this approximation as $\mathbf{V}^{CR1}$. 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. As

noted above, both the BRL approach (CR2) and the jacknife 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 defines the adjustment matrices $\mathbf{A}_i$ 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 (see Section 4 of this paper for a review). 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.

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:

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

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 $\mathbf{\Phi}_i$,

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

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* $Var\left(\boldsymbol{\epsilon}_i\left|\mathbf{R}_i,\mathbf{S}_i,\mathbf{T}_i\right.\right)=\mathbf{\Phi}_i$, *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.

Connect this to the IK case.

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 $\ddot{\mathbf{U}} = (\mathbf{I} - \mathbf{H_T})\,\mathbf{U}$. 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 α if Q exceeds $\chi^2(\alpha; q)$, the α critical value from a chi-squared distribution with q degrees of freedom. In practice it is rarely clear how large samples need to be for Wald tests to be implemented.

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 α). 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 t-test developed by Bell and McCaffrey (2002) involves using a $t(\nu)$ references distribution with degrees of freedom ν, 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 (see Section 4 of this paper). 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, so long as the degrees of freedom are larger than 4 or 5 (Bell and McCaffrey, 2002; Tipton, 2015). This is because when the degrees of freedom are smaller, the t-distribution approximation to the sampling distribution doesn't hold, and the Type I error can be higher than the stated α level. In comparison, the CR1 degrees of freedom (i.e., $m - 1$) are constant, and 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, 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-parameter testing 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 η 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 η 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.

Tipton and Pustejovsky (in press) recently introduced this test for use in a special case of CRVE used in meta-analysis. This test was developed in relation to other literature that focused on approximating the distribution of a robust variance estimator by a Wishart for some simpler models that are special cases of CRVE. Zhang (2012a,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 (2012b) 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 η, 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 η 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 η 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 Tipton and Pustejovsky (in press) have found to perform better in practice.

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 η_M reduces to ν_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 based hypothesis tests 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 "CR2S" test to refer to those tests employing both the BRL correction to the variance and the Satterthwaite degrees of freedom. For t-tests, CR2S thus refers to test described in Section 3.1, while for the F-test CR2S refers the that found in Section 3.2.

4.1 Review of previous simulation studies

To date, four papers have examined the performance of the CR2 t-test and these papers have included nearly 100 parameter combinations. 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. 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). Finally, the studies have examined a range of both true error structures—including various combinations of heteroscedasticity and clustering—and estimation strategies, with a focus in particular on different 'working' models. By varying the working model from the true error structure, the studies allow for the effect of model mispecification to be studied. Finally, while most of these studies focus on comparisons in OLS, one study (Tipton, 2015) focuses on the use of WLS.

The last column of Table 1 indicates the range of Type I error rates given over the conditions studied in each of the simulation studies, with values given for both the CR1 and CR2S tests. Across these studies, this Type I error for the CR1 t-test ranges from 0.01 to 0.34, when the stated α level was 0.05. These values are particularly far above nominal when the covariate tested is unbalanced, skewed (i.e., high leverage), or when the number of observations per cluster varies. While not illustrated by the table, importantly these high Type I error rates occurred not only when the number of clusters was very small, but also at moderate sizes if the covariate imbalance or skew was large.

In comparison, the CR2S t-test performs considerably better across the range of conditions studied, with Type I error rates ranging between 0.01 and 0.13. Notably, the largest value observed here is for the Imbens and Kolesar (2015) study, which does not break results out by degrees of freedom. Given the condition studied (30 clusters, with only 3 having a policy dummy), these degrees of freedom are possibly below the 4 or 5 cut-off at which others have shown the t-test approximation to fail. Putting this value aside, the maximum Type I error observed in these conditions is 0.06, only slightly higher than nominal. Importantly, these nearly nominal Type I error rates hold across a wide degree of model mispecification (i.e., when the working model is far from the true error structure) and covariate types. While not shown in the table, this is because the CR2S degrees of freedom take into account these covariate features in their degrees of freedom, which are often far below those in the CR1 test.

In comparison to the t-test, the CR2 F-test has only been recently developed and studied in a single paper focused on the meta-analytic case (Tipton and Pustejovsky, 2015). While focusing only on the use of CRVE with weights (WLS), however, the simulation

study was comprehensive in other regards. The simulation studies examined the effects of the number of covariates in the full model (up to $p = 5$) and the number of constraints being tested ($q = 2, 3, 4, 5$), including cases in which $p = q$ and in which $p > q$. These simulations examined models with various combinations of the 5 covariates found in Tipton (2015), found in the last line of Table 1. This set included both balanced and imbalanced dummies, as well as symmetric and skewed continuous covariates. Like Tipton (2015), these simulations focused on true correlation structures that included heteroscedasticity, clustering (i.e., a cluster specific random effect), and correlated errors. The working models for these were then varied far from the true error stucture; for example, based on models with no clustering at all. Finally, the number of clusters was varied from 10 to 100, each with between 1 and 10 observations. Type I error was then compared over a range of α levels for both the CR1 and CR2S F-tests.

The results of the simulations by Tipton and Pustejovsky indicate that the CR2S F-test always has Type I error less than or equal to the stated α level, except in cases with extreme model mispecification. Even in these cases, however, in the conditions studied the Type I error was in line with those observed for t-tests; for example, for $\alpha = 0.05$ the error was not above 0.06. In comparison, the Type I error of the CR1 test was often very high (between XX and XXX), with particularly large differences when the degrees of freedom of the two tests differed. Again, like the t-test, the degrees of freedom of the CR2S F-test here were driven by covariate features, with particularly low degrees of freedom resulting when the covariate set studied included imbalances or skew.

4.2 Simulation Design

While the simulation study produced by Tipton and Pustejovsky (2015) included a variety of conditions, the set studied focused largely on the types of data found in meta-analysis and regression more broadly. These differ from the economic context in two ways. First, in meta-analysis, it is common for there to be both heteroscedasticity assumed and for analysts to implement weights in the analysis (typically inverse-variance). This use of weights is less common in economic applications. Second, in meta-analysis, analysts are typically interested in a variety of covariate relationships. In comparison, in economics,

many applications focus on the effects of different policies (and thus dummies), particularly policies that may vary at the cluster and/or observation level. For example, a policy may be implemented in a portion of states (the cluster level), or in all states, but at different time points (the observation level), or a combination of the two. Given these differences, in this paper we provide an additional simulation study focused on designs more commonly found in economics.

In this simulation study, we focus on several research designs and analysis models, including the analysis of experiments, panel data (difference-in-difference models), and seemingly unrelated regression. In the analysis of experiments, two designs are common: the cluster randomized design (CRT), in which treatments are assigned randomly at the cluster level; and the randomized block design (RBD), in which treatments are randomly assigned within each cluster. For both designs we investigated the effect of a design with three treatment arms (2 policy indicators), with units randomized either equally across these arms or unequally. This resulted in two cluster randomized designs: one with exactly 1/3 of clusters in each of the three arms (CR-balanced) and the other with unequal allocations (.5, .3, .2; CR-unbalanced). For randomized block designs, this resulted in two designs as well: one with exactly 1/3 of units in each cluster in each of the three arms (RB-balanced), and the other with unequal allocations (n/2,n/3,n/6; RB-unbalanced). In the analysis of panel data, often policies are observed at all time points in some clusters, while changing across time points in others. Here we investigated a total of four cases. In the first case, XXXX. In the second case, XXXX.

The set-up described above for both the experimental and panel-data cases resulted in single-paramter ($q = 1$) and multi-parameter ($q = 2$) hypothesis tests. For $q = 1$, tests focused on the hypothesis that a single policy arm differed from the control, while for $q = 2$, tests focused on the hypothesis that there were no differences across all arms. For each of these tests, the data was generated ... MORE HERE ... (JAMES HELP!)

For the final set of simulations, the above structure was replicated across three outcome variables, generated such that the outcomes were correlated (e.g., reading, math, science test scores). This resulted in a seemingly unrelated regression model, allowing hypotheses of higher dimensions to be tested. The first of these tested if a single treatment arm differed

from the control arm across the three outcomes ($q = 3$), while the second tested if there were differences across all three outcomes and arms ($q = 6$). In order to generate this data, in addition to the above strucutre, the outcomes were generated to be either correlated weakly (0.20) or strongly (0.80). For all of these designs, Appendix C includes further details on the simulation study design.

4.3 Simulation Results

In this section, we provide an overview of the results from the previously described simulation study. We focus here on three trends, each depicted visually in a figure and discussed in the text. Importantly, these trends are similar to results found in Tipton and Pustejovsky (2015), providing further support that the CR2S F-test performs well in a wide range of data generating mechanisms and parameter combinations.

The first result is that the CR2S F-test has Type I error close to the stated α level for all parameter combinations studied, whereas the CR1 does not. These results are found in Figure 1, which provides evidence across dimension of the test q (the columns) and α levels (the rows). In each of these figures, the number of clusters is depicted from left to right, varying from 15 to 50, and the stated α level is given with a horizontal black line; the dashed line indicates the simulation error upper bound. In this figure, the CR2S test is seen to have Type I error near the stated α level, even with a small number of clusters. When the number of clusters is very small ($m = 15$) the Type I error can be samller than the stated α level, and while there exist situations in which the error is above the simulation bounds, at most the Type I error is XXX for $\alpha = 0.01$, XXX for $\alpha = 0.05$, and XXX for $\alpha = 0.10$. In comparison, the Type I error for the CR1 test can be markedly higher than the stated α level, particularly when the number of clusters is small. This is particularly true as the dimension of the test increases; for example, for $\alpha = 0.05$, when there are 15 clusters the maximum Type I error ranges from 0.12 ($q = 1$) to 0.65 ($q = 6$). Perhaps even more important for practice, even when there are 50 clusters, the Type I error for the CR1 test can be far above the stated α level. Again, focusing on the $\alpha = 0.05$ case, this error ranges from a maximum of 0.07 ($q = 1$) to 0.20 ($q = 6$).

In order to better understand the effects of different paramter combinations on the

performance of both tests, Figure 2 focuses on the $\alpha = 0.05$ case and divides the results out by number of clusters (columns) and test (rows). Within each graph, results are given by the policy design, with 6 combinations illustrated, and within each from left to right by the dimension of the test ($q = 1, 2, 3, 6$). The top panel focuses on the CR1 test, highlighting both that the Type I error climbs in relation to the dimension of the test (q) and in relation to the degree of imbalance in the policy indicators. Here differences between the balanced and unbalanced cases are largest for the cluster-randomized (CR) designs and the panel data (DD) designs, with the best performance arising for the randomized block designs (RB). In the bottom panel, results are given for the CR2S test; here we focus on three trends. First, as for the CR1 test, the Type I error increases as the dimension of the test increases, though never above 0.07. Second, whereas imbalance results in Type I error larger than nominal for the CR1 test, for the CR2S test, it results in Type I error below nominal. This trend is the strongest for the cluster-randomized (CR) design where Type I error can be close to 0 at its minimum. Third, for studies with at least 30 clusters, the Type I error is very close to nominal (between 0.03 and 0.055) for all conditions studied.

Finally, by simulating the errors across a variety of parameter combinations, we were also able to test the impact of misspecification of the working model on Type I error. Figure 3 depicts Type I error rates for $\alpha = 0.05$ for the CR2S test by the dimension q of the test (columns) and the number of clusters (rows). Within each, results are separated by the 9 combinations of treatment effect variance and ICCs, noting that the working model assumed values of zero for each. Within each of these 12 blocks, results are shown to be very similar across these 9 error structures, with no clear pattern to the small differences that emerge. These results follow closely those from Tipton and Pustejovsky (2015), which also found that even with extreme model misspecification the Type I error of the CR2S test was close to nominal.

Figure 1 shows that Naive-F = bad, AHZ = rad! Also, performance of the naive test degrades (becomes liberal) for larger q, whereas AHZ test becomes more conservative.

Figure 2 shows that rejection rates are strongly affected by imbalance of the clusters.

Figure 3 shows that model mis-specification doesn't matter for AHZ.

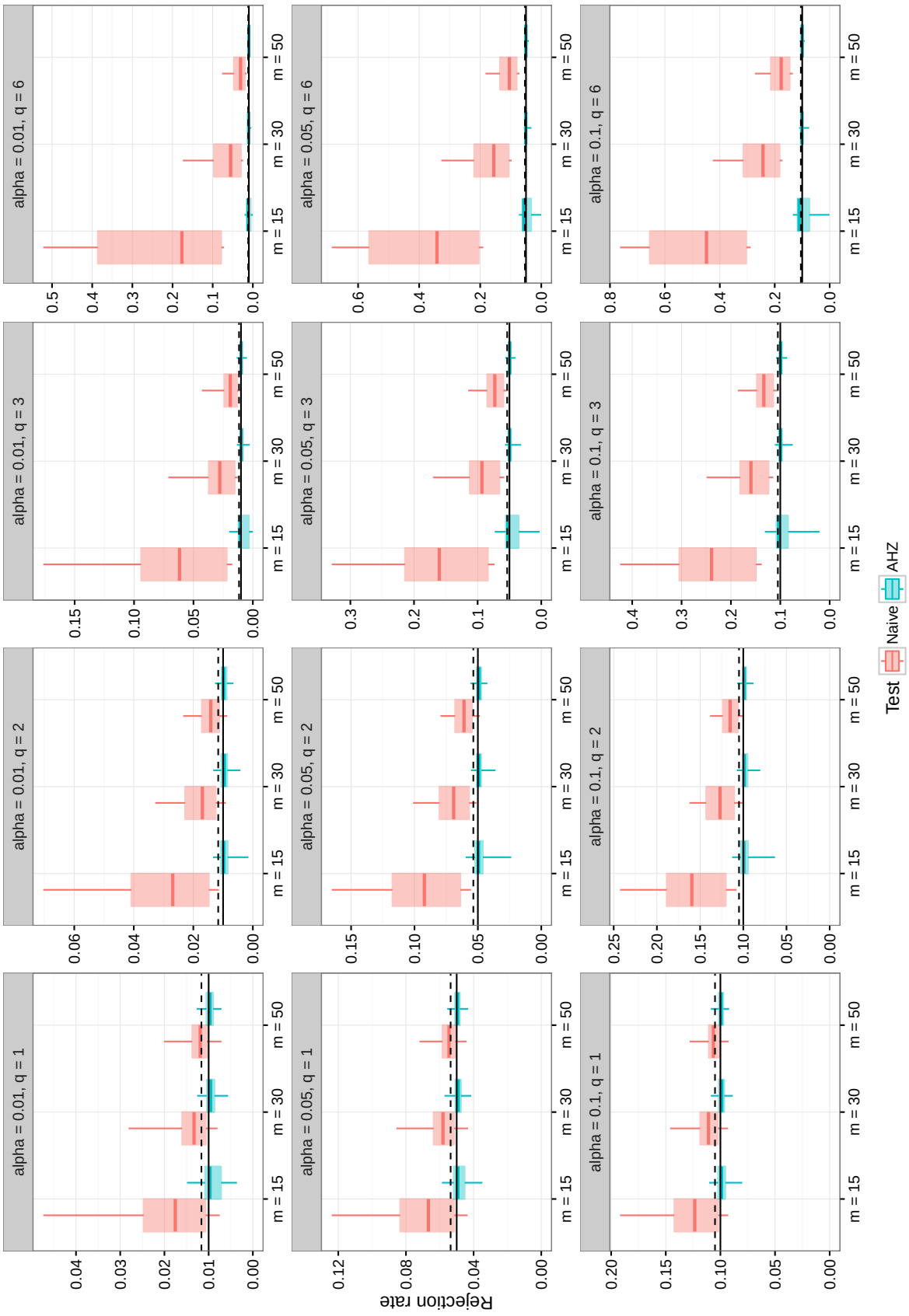

Figure 1: Rejection rates of Naive and AHZ tests, by dimension of hypothesis (q) and nominal type I error (α)

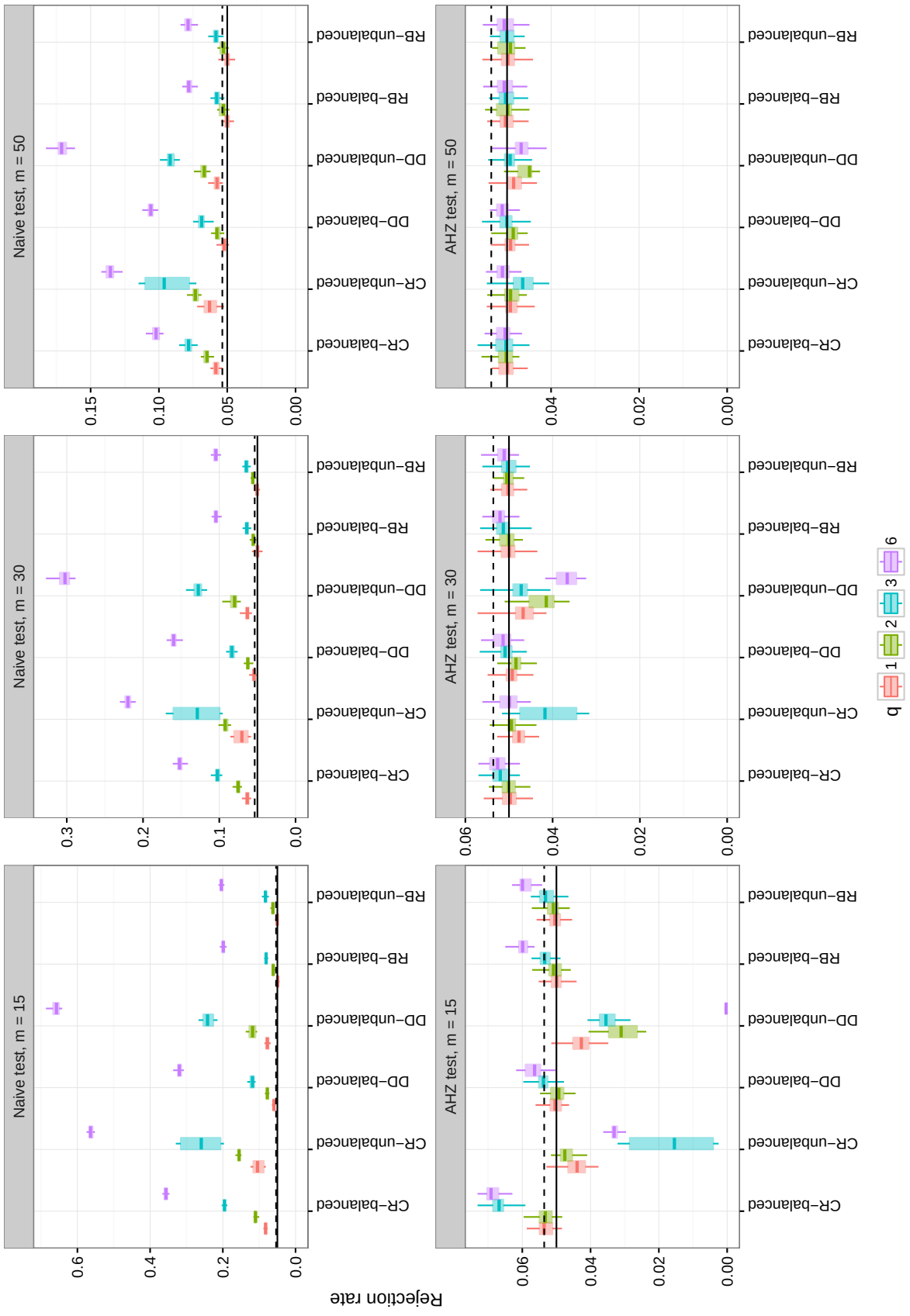

Figure 2: Rejection rates of Naive and AHZ tests, by study design and dimension of hypothesis (q). CR = cluster-randomized design; DD = difference-in-differences design; RB = randomized block design.

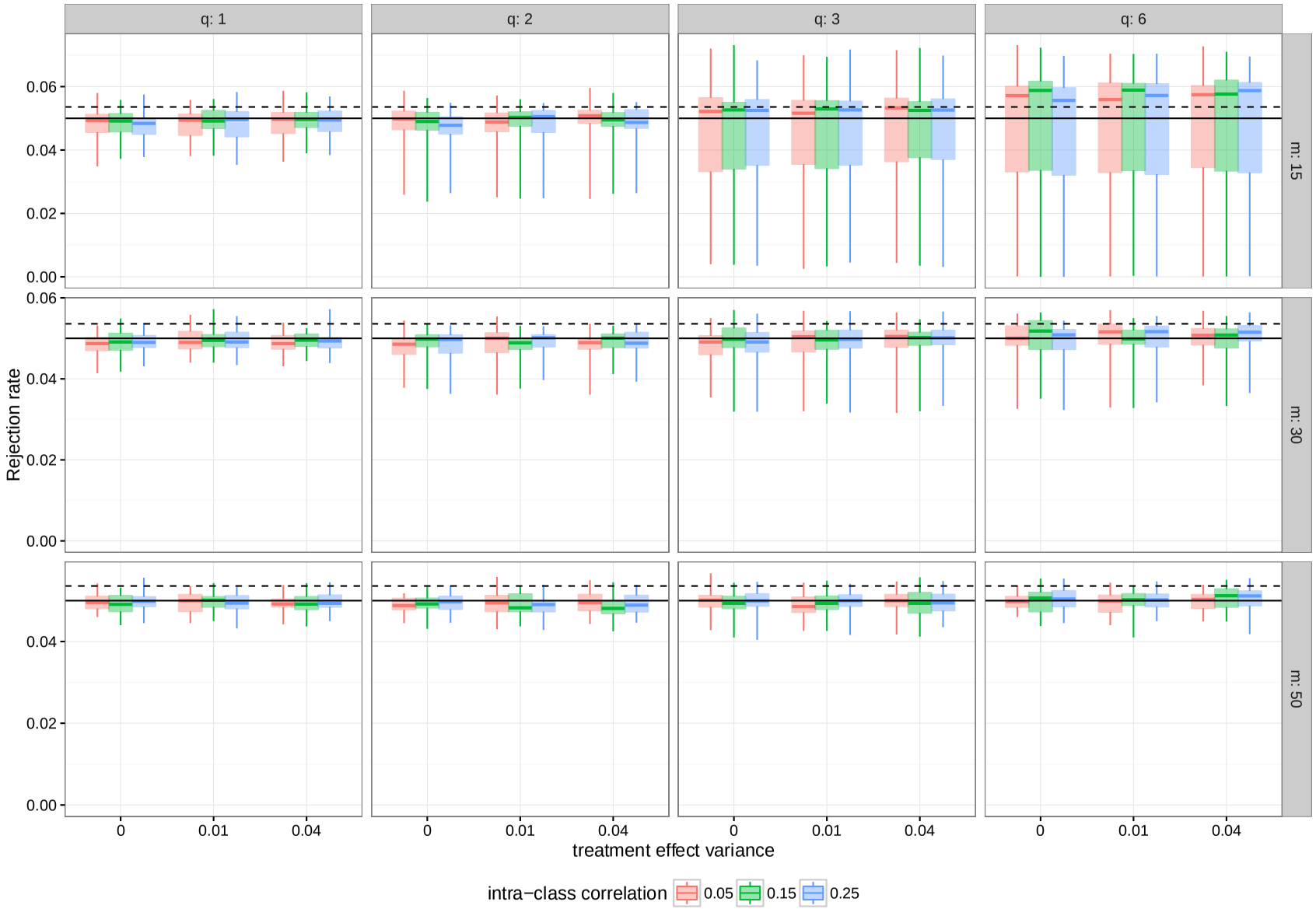

Figure 3: Rejection rates of AHZ test, by treatment effect variance and intra-class correlation.

5 EXAMPLES

This section presents three short examples of the use of CRVE with small or moderate samples of clusters, 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 demonstrate the proposed small-sample t- and F-tests and compare them to 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 (for a detailed review, see Schanzenbach, 2006). The experiment involved students in kindergarten through third grade across 79 schools. Within each school, students and their teachers were randomized equally to one of three conditions: small class-size (targetted to have 13-17 students), regular class-size, or regular class-size with an aide. Among other outcomes, subsequent research has focused on the effects of these conditions on kindergarden reading, math, and word recognition (Achilles, Bain, Bellott, Boyd-Zaharias, Finn, Folger, Johnston and Word, 2008); high school test scores (Schanzenbach, 2006); college entrance exam participation (Krueger and Whitmore, 2001); and home ownership and earnings (Chetty, Friedman, Hilger, Saez, Schanzenbach and Yagan, 2011).

The STAR experiment involved three treatment conditions and multiple outcomes, providing a scenario where both t-tests (with $q = 1$) and F-tests with varying constraint dimensions can be applied. For simplicity, we focus only on the subgroup of students who were in kindergarten during the first year of the study, and on the three outcomes measured at the end of the year, standardized to percentile ranks (Krueger and Whitmore, 2001): reading, word recognition, and math (Achilles et al., 2008). The analytic model is:

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

where y_{ijk} is the percentile rank on outcome k for student j in school i; $\mathbf{r}_{ij}$ includes indica-

tors 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); γ_k is a fixed effect for outcome k; and μ_i is a fixed effect for school i. In this model, β_{1k} represents the average effect of being in a small class and β_{2k} represents the average of effect of being in a regular class with an aid, in each case compared to a regular-size class without an aid.

Using this model, we test four distinct hypotheses that vary in dimension from $q = 1$ to $q = 6$. First, using only the data for outcome k, we test the effects of small class size ($H_0 : \beta_{1k} = 0$) while maintaining the assumption that the additional classroom aide has no effect on student achievement (i.e., constraining $\beta_{2k} = 0$). Second, again only using the data for outcome k, we test the hypothesis that there are no differences across the three class-size conditions (i.e., $H_0 : \boldsymbol{\beta}_k = \mathbf{0}$). Third, combining the data across all three outcomes, we test the hypothesis that small class size (vs regular and regular plus aide) had no effects on any outcome (i.e., $\beta_{11} = \beta_{12} = \beta_{13} = 0$). Finally, we test the hypothesis that there are no differences across the three class-size conditions on any outcome (i.e., $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = \mathbf{0}$). The third and fourth tests use the seemingly unrelated regression (SUR) framework, in which separate treatment effects are estimated for each outcome, but the student demographic effects and school fixed effects are pooled across outcomes. In all models, we estimated $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}$ after absorbing the school fixed effects and clustering the errors by school.

Outcome	Effect	Test	F	df	p
Math	Small class (q=1)	Standard	13.624	78.0	0.00041
		AHZ	13.590	69.0	0.00045
	Small class and classroom aide (q=2)	Standard	6.838	78.0	0.00183
		AHZ	6.725	68.6	0.00215
Combined	Small class (q=3)	Standard	6.408	78.0	0.00062
		AHZ	6.206	67.0	0.00088
	Small class and classroom aide (q=6)	Standard	3.284	78.0	0.00622
		AHZ	3.042	64.9	0.01103

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

Table 1 displays the results for a representative subset of these hypothesis tests, using either the standard CR1 hypothesis test or the CR2S test. These results illustrate two important points regarding the use of the AHZ test in practice. First, across all three analyses, the CR2 t- and F-tests are typically slightly smaller than those produced using CR1. This trend is common, since the BRL adjustments to the variance estimator are typically small in scope. Second, in the case in which treatment is randomly allocated in equal proportions within each cluster – as occurs in the TN STAR experiment – the degrees of freedom for the CR2 tests are only slightly smaller than those for the CR1 tests. In combination with the rather large sample size of 79 schools, these differences have only a minimal effect on the p-values for these tests. Importantly, these conditions do not always occur in practice, as the next examples will illustrate.

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 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. 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 completion rates for 2000 and 2002 as being unaffected by treatment assignment. The primary outcome of interest is Bagrut completion.

This study provides an opportunity to examine the CR2 tests in a situation in which the treatment was assigned at the cluster level, with a smaller number of clusters. For simplicity, we restrict our analysis to the sample of female students. Following the original

analysis of Angrist and Lavy (2009), we allow the program's effects to vary depending on whether a students was in the upper or lower half of the distribution of prior-year academic performance. Letting $h = 1, 2, 3$ index the sector of each school (Arab religious, Jewish religious, or Jewish secular), we consider the following analytic model:

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

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 $\boldsymbol{\beta}_h = (\beta_{1h}, \beta_{2h})$ is a vector of average treatment effects for schools in sector h. 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 γ_{ht} for each sector in each year and μ_{hi} for each school.

Based on Model (17), we test four hypotheses, again with the goal of exploring the use of the CR2S tests in a range of conditions. First, we assume that the program effects are constant across sector (i.e., $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = \boldsymbol{\beta}$) and test for whether the program affected completion rates for students in the upper half of the prior achievement distribution ($H_0 : \beta_2 = 0$, with $q = 1$). Second, we test for whether the program was effective in either half of the prior academic performance ($H_0 : \boldsymbol{\beta} = 0$, with $q = 2$), still assuming that program effects are constant across sector. Third, we test for whether program effects in the upper half of the prior achievement distribution vary by school sector ($H_0 : \beta_{21} = \beta_{22} = \beta_{23}$, with $q = 3$). Finally, we conduct a joint test for whether program effects differ across student prior achievement crossed with school sector ($H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3$, with $q = 4$).

Table 2 reports the results of all four hypothesis tests. These results indicate three important trends. First, in the case of the first two hypotheses, the CR2 test statistics are only slightly smaller than their CR1 counterparts, though the degrees of freedom are considerably smaller. These differences reflect the fact that the treatment indicator occurs at the cluster level, but that the subgroups varied within each cluster. Together these lead to a larger impact on results than in the STAR case, where treatment was assigned at the student level. Second, the third and fourth hypotheses tests – comparing treatment effects

Hypothesis	Test	F	df	p
ATE - upper half (q = 1)	Standard	5.746	34.00	0.02217
	AHZ	5.169	15.86	0.03726
ATE - joint (q = 2)	Standard	3.848	34.00	0.03116
	AHZ	3.371	15.46	0.06096
Moderation - upper half (q = 2)	Standard	3.186	34.00	0.05393
	AHZ	0.091	3.19	0.91520
Moderation - joint (q = 4)	Standard	8.213	34.00	0.00010
	AHZ	2.895	3.21	0.19446

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

across sectors and subgroups – illustrates cases in which the CR2S and CR1 tests greatly diverge. In these cases, the CR2S test statistic and degrees of freedom are both considerably smaller than those from the CR1 test. This reflects the degree of imbalance in allocations across sectors (20,10,10) combined with the cluster-level randomization. Together these result in cases in which using the small sample adjustments greatly impact the findings.

5.3 Effects of minimum legal drinking age on mortality

In this final example, we shift focus from analyses of experiments to panel data, using an example described in Angrist and Pischke (2015). Using data from the Fatal Accident Reporting System maintained by the National Highway Traffic Safety Administration, we examine the effects of changes in the minimum legal drinking age over the time period of 1970-1983 on state-level death rates resulting from motor vehicle crashes. A standard difference-in-differences specification for such a state-by-year panel is

Cite

Note about measure of legal drinking age.

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

In this model, time-point t is nested within state i; the outcome y_{it} is the number of deaths in motor vehicle crashes (per 100,000 residents) in state i at time t; $\mathbf{r}_{it}$ is a vector of covariates; γ_t is a fixed effect for time point t; and μ_i is an effect for state i. The vector $\mathbf{r}_{it}$ consists of a measure of the proportion of the population between the ages of 18 and 20

years who can legally drink alcohol and a measure of the beer taxation rate, both of which vary across states and across time..

In order to estimate the effect of lowering the legal drinking age, we apply both random effects (RE) and fixed effects (FE) estimation. For the RE estimates, we use WLS with weights derived under the assumption that $\mu_1, ..., \mu_m$ are mutually independent, normally distributed, and independent of ϵ_{it}. We also report an artificial Hausman test (Arellano, 1993; Wooldridge, 2002) for correlation between the covariates $\mathbf{r}_{it}$ and the state effects μ_i. Such correlation creates bias in the RE estimator of the policy effect, thus necessitating the use of the FE estimator. The artificial Hausman test amends model (18) to include within-cluster deviations (or cluster aggregates) of the variables of interest, so that the estimating equation is

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

where $\ddot{\mathbf{r}}_{it}$ denotes the within-cluster deviations of the covariate. The parameter $\boldsymbol{\delta}$ captures the difference between the between-panel and within-panel estimates of $\boldsymbol{\beta}$. With this setup, the artificial Hausmann test amounts to testing the null hypothesis that $\boldsymbol{\delta} = \mathbf{0}$, where $\boldsymbol{\delta}$ is estimated using RE.

Hypothesis	Test	F	df	p
Random effects	Standard	8.261	49.00	0.00598
	AHZ	7.785	24.74	0.00999
Fixed effects	Standard	9.660	49.00	0.00313
	AHZ	9.116	22.72	0.00616
Hausman test	Standard	2.930	49.00	0.06283
	AHZ	2.489	8.69	0.13980

Table 3: Tests of effects of minimum legal drink age and Hausman specification test

Table 3 displays the results of the tests for the policy variable and the Hausman tests for each model specification. As in previous examples, our preferred version of the test uses the CR2S adjustment to the variance estimator and the AHZ test. The results of the policy effect tests are quite similar across specifications and versions of the test. Of note is

that the RE and FE estimates have only half the degrees of freedom of the naivë test. For the artificial Hausman test the AHZ test has fewer than 9 degrees of freedom, which leads to a much larger p-value compared to using the CR1 F test.

6 DISCUSSION

Cluster robust standard errors are common practice in economic applications, since in large samples they require few assumptions regarding the error structure of the data. In small and moderate samples, however, the CRVE approach can result in overly liberal tests, with Type I error far above nominal. A promising solution therefore is to incorporate a working model for the error structure into the CRVE approach – bias reduced linearization – as introduced by Bell and McCaffrey (2002), and to use estimated degrees of freedom. In this paper, we have provided solutions to three problems facing the implementation of this BRL approach in economics, making the method useful in a wide range of applications.

While the idea of identifying and implementing a working model for the error structure seems antithetical to the idea of CRVE, simulations studies provided here and elsewhere show that the approach is robust to a high degree of misspecification of this working model. Importantly, the role of this structure while larger when the number of clusters is very small, effectively falls away as the number of clusters increases, converging to the usual CRVE estimator in large samples. This structure greatly improves the performance of CRVE, bringing the Type I error in line with the stated α levels for tests, even in less than ideal situations.

The main driver to these adjustments, as illustrated both in the simulation study and the set of examples, is the fact that in this approach the degrees of freedom are estimated using a Satterthwaite approximation. These degrees of freedom can be much smaller than the number of clusters, particularly when the covariates being tested exhibit a high degree of imbalance or skewness. In economics, these problems are common, particularly in the analysis of experiments and in difference-in-difference models used in panel data. In this paper, we have shown that the small sample adjustments provided here perform well even in situations in which a series of policies are allocated very unequally across groups, and that the approach works by greatly reducing the degrees of freedom. In many cases these

degrees of freedom can be quite small, even when the number of clusters is moderate to large. Most striking, perhaps, is the fact that these differences occur even with as many as 50 clusters, a rule of thumb often given for when CRVE should perform well.

A BRL adjustment matrices

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

A.1 Proof of Theorem 1

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 $\mathbf{\Lambda}$ 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 $\mathbf{B}_i = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$. Then $\mathbf{B}_i^+ = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}'$ and $\mathbf{B}_i^{+1/2} = \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{V}'$.

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,

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

Because $\mathbf{D}_i$, and $\mathbf{\Phi}$ are positive definite and $\mathbf{B}_i$ is symmetric, the eigenvectors $\mathbf{V}$ define an orthonormal basis for the column span of $(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{X}}})_i$. 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 $(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{X}}})_i \mathbf{Z} = \ddot{\mathbf{U}}_i$, 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

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

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

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

It follows that

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

A.2 Proof of Theorem 2

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 $\mathbf{B}_i^+ = (\mathbf{D}_i')^{-1}\,\mathbf{\Omega}_i\mathbf{D}_i^{-1}$. 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 $\mathbf{\Sigma}$. It follows that

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

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

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

Furthermore, the distribution of $\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2$ can be expressed as a weighted sum of χ_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 $\mathbf{p}_{1i}'\mathbf{\Sigma}\mathbf{p}_{2j}$, $i, i = 1, ..., m$.

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. (2015), 'Small sample adjustments for robust variance estimation with meta-regression.', *Psychological Methods* **20**(3), 375–393.

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.