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

September 22, 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

```
> library(knitr)
> library(xtable)
> # set global chunk options
> opts_chunk$set(echo = FALSE, cache = FALSE, fig.path='CR_fig/', fig.align='center', fi
```

The use of cluster-robust variance estimation (CRVE; White, 1984; Liang and Zeger, 1986; Arellan, 1987) is now common in a wide-range of economic analyses. Standard errors are routinely "clustered" to account for correlations arising from the grouping of individuals (e.g., countries, regions, states, villages), as well as measurements across time-periods (as found in panel data). The method is an extension to another economic mainstay -heteroscedasticity-robust standard errors (Huber, 1967, Eicker, 1967, White, 1980) - which are commonly used to account for non-constant variance in regression models. While CRVE has existed for over 30 years, it is in the last 10 years that it has become standard practice, as illustrated by its coverage in major textbooks and review articles (e.g., Woolridge, 2010; Angrist and Pischke, 2009; Cameron and Miller 2011).

Like heteroscedasticity robust standard errors, CRVE allows analysts to estimate causal and structural models using ordinary least squares (OLS), while adjusting standard errors and hypothesis tests thereafter. The method is broadly applicable and can be used to test both individual and multi-parameter hypotheses. For example, an analyst may wish to understand the effect of several state-level policy shifts on employment outcomes, where the policies were implemented at different time-points in each state. The effect of the policies could be estimated using OLS with states and time-periods accounted for using fixed effects. These fixed effects could be included in the model as dummy variables or, more commonly, could be 'absorbed' by first demeaning the data. CRVE would then be used to account for clustering by both state and time, and these clustered standard errors could by used to test hypotheses regarding each policy (i.e., a t-test), as well as across policies (i.e., an F-test). This approach to accounting for clustering via both fixed effects and the use of CRVE is now standard and was popularized by Bertrand, Duflo, and M (2004), who showed via simulation that to do otherwise led to inappropriately small standard errors and incorrect hypothesis tests.

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 results that in small samples, CRVE has a downward bias and that hypothesis tests based on CRVEs can have Type-I error rates that are far from 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 highlight a potentially promising solution to these problems, called the bias-reduced linearization method (BRL), 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, if the fixed effects are accounted for via the inclusion of cluster dummies, the BRL adjustment breaks down and cannot be implemented. This makes the method only useful in cases in which absorbtion or demeaning is preferred. Second, however, as Cameron and Miller (2015) highlight, the standard errors that result from absorption can be substantially different than those produced using dummies. This inconsisency 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 - while Bell and McCaffrey (2002) provide a method for conducting single parameter tests, no such small-sample method has been provided for multiparameter tests. 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.

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 next sub-sections, we begin by reviewing the the small sample CRVE method that is standard in practice in most software applications. In Section 2, we then review the BRL correction to the CRVE estimator, and present advances that address the first two concerns above. This includes the use of generalized inverses, thus allowing BRL to be implemented more broadly, and the definition of a set of conditions under which the absorption and dummy variable approaches result in identical standard error estimates. In Section 3, we then move to hypothesis testing. 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 and performs comparably with current state-of-the-art methods such as the cluster-wild bootstrap procedure described by Cameron, Gelbach and Miller (2008) and Webb and MacKinnon (2013). In Section 4, we illustrate the use of CRVE in small samples, implementing these methods across three examples that span the variety of useages found in economic applications. 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.

1.1 Econometric framework

We begin by considering a generic model of the form,

$$y_{ij} = \mathbf{p}_j\boldsymbol{\tau} + \mathbf{q}_j\boldsymbol{\eta} + \mathbf{s}_j\boldsymbol{\gamma} + \mathbf{t}_j\boldsymbol{\delta} + \epsilon_{ij} \tag{1}$$

where for observation i in cluster j, $\mathbf{p}_j$ is a vecor of p coefficients of primary interest in an analysis (e.g., policy variables), $\mathbf{q}_j$ is a vector of q control variables, $\mathbf{s}_j$ is a vector of s fixed effects that vary across clusters, and $\mathbf{t}_j$ is a vector of t fixed effects that are defined within clusters. In the state-policy example indicated in the introduction, these $\mathbf{p}_j$ would be dummies for each policy under study, $\mathbf{q}_j$ might include demographic controls, $\mathbf{s}_j$ would include state fixed effects, and $\mathbf{t}_j$ would indicate time-period fixed effects. In this example, focus would be on testing hypotheses regarding $\boldsymbol{\tau}$, with estimation of $\boldsymbol{\eta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$ remaining incidental.

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

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

where for cluster j, $\mathbf{R}_j$ is an $n_j \times r$ matrix of covariates that includes both the focal ($\mathbf{p}_j$) and control ($\mathbf{q}_j$) covariates (i.e., $q + p = r$); $\mathbf{S}_j$ is an $n_j \times s$ matrix describing fixed effects that vary across clusters, and $\mathbf{T}_j$ is an $n_j \times t$ matrix describing fixed effects that are identified only within clusters.

We assume that $\mathrm{E}\left(\boldsymbol{\epsilon}_j \,|\mathbf{R}_j, \mathbf{S}_j, \mathbf{T}_j\right) = \mathbf{0}$ and $\mathrm{Var}\left(\boldsymbol{\epsilon}_j \,|\mathbf{R}_j, \mathbf{S}_j, \mathbf{T}_j\right) = \boldsymbol{\Sigma}_j$, for $j = 1, ..., m$, where the form of $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_m$ may be unknown but the errors are independent across clusters. For notational convenience, let $\mathbf{U}_j = [\mathbf{R}_j \ \mathbf{S}_j]$ denote the set of predictors that vary across clusters, $\mathbf{X}_j = [\mathbf{U}_j \ \mathbf{T}_j]$ denote the full set of predictors, $\boldsymbol{\alpha} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\delta}')'$, and $x = r + s + t$. Denote the total number of individual observations by $N = \sum_{j=1}^{m} n_j$. 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 $\boldsymbol{\delta}$ are treated as nuisance parameters. The distinction between the covariates $\mathbf{R}_j$ versus the fixed effects $[\mathbf{S}_j \ \mathbf{T}_j]$ thus depends on context and the analyst's inferential goals. However, the distinction between the two fixed effect matrices $\mathbf{S}_j$ and $\mathbf{T}_j$ is unambiguous, in that the within-cluster fixed effects satisfy $\mathbf{T}_j\mathbf{T}_k' = \mathbf{0}$ for $j \neq k$. We further assume that $\left(\mathbf{U}'\mathbf{U} - \mathbf{U}_j'\mathbf{U}_j\right)$ is of full rank for $j = 1, ..., m$.

We can estimate the $\boldsymbol{\beta}$ using weighted least squares (WLS), where for cluster j we define $\mathbf{W}_j$ to be a symmetric, $n_j \times n_j$ weighting matrix of full rank. Importantly, the WLS

framework includes the unweighted case (where $\mathbf{W}_j = \mathbf{I}_j$, an identity matrix), as well as feasible GLS.[1] In the latter case, it is assumed that $\text{Var}(\mathbf{e}_j | \mathbf{X}_j) = \mathbf{\Phi}_j$, where $\mathbf{\Phi}_j$ 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 $\mathbf{W}_j = \hat{\mathbf{\Phi}}_j^{-1}$, where the $\hat{\mathbf{\Phi}}_j$ 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 $\boldsymbol{\delta}$ 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. Informally, 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 $\mathbf{M_Z} = (\mathbf{Z}'\mathbf{WZ})^{-1}$ and $\mathbf{H_Z} = \mathbf{ZM_Z Z'W}$.

[1] The WLS estimator also encompasses the estimator proposed by Ibragimov and Müller (2010) for clustered data. Assuming that $\mathbf{X}_j$ has rank p for $j = 1, ..., m$, 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 $\mathbf{W}_j = \mathbf{X}_j \left(\mathbf{X}_j'\mathbf{X}_j\right)^{-2} \mathbf{X}_j$.

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 $\mathbf{R}$ 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

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

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

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

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,

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

which depends upon the unknown variance matrices $\boldsymbol{\Sigma}_j$. A model-based approach to estimating this variance would involve assuming a structure to this $\boldsymbol{\Sigma}_j$; 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 $\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)$ empirically. While there are several versions of this approach, all can be written in the form

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

for some n_j by n_j adjustment matrix $\mathbf{A}_j$. Note here that in place of $\mathbf{\Sigma}_j$ now the residuals $\mathbf{e}_j\mathbf{e}_j'$ are used instead.

The form of these adjustments parallels those of the heteroscedastity-consistent (HC) variance estimators proposed by MacKinnon and White (1985). Setting $\mathbf{A}_j = \mathbf{I}_j$, an $n_j \times n_j$ 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 $\mathbf{A}_j = c\mathbf{I}_j$, 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., Stata). Both the CR0 and CR1 estimators rely on asymptotic properties of the residuals in order to consistently estimate $\mathbf{\Sigma}_j$. 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 $\mathbf{X}_j$. 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 $\mathbf{A}_j = \left(\mathbf{I} - \ddot{\mathbf{R}}_j\mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}_j'\mathbf{W}_j\right)^{-1}$. 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

As Cameron and Miller (2015) highlight, the unadjusted CR0 estimator - and to a smaller degree, the CR1 estimator - both tend to under-estimate the true variance of $\hat{\boldsymbol{\beta}}$. 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 $\mathbf{X}_j$. 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 $\mathbf{A}_j \neq c\mathbf{I}_j$ for any constant c. 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 (i.e., $\mathbf{\Sigma}_j = \mathbf{\Phi}_j$) for the correlation structure in the model. The idea is to develop the adjustment matrices $\mathbf{A}_j$ such that when this working model is correct, the CR2 estimator is unbiased. While the idea of specifying a model seems antithetical to the general goal of using CRVE, 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 (CITE these). In these instances, the variance estimator is no longer unbiased, but its bias is greatly reduced (thus the name "bias reduced linearization"). Furthermore, as the number of clusters increases, the reliance on this working model diminishes. In a sense, CR2 provides necessary scaffolding in the small sample case, which falls away when there is sufficient data.

In the original formulation given by Bell and McCaffrey (2002), for a given working model $\mathbf{\Phi}_j$, the BRL approach amounted to defining the matrices $\mathbf{A}_j$ such that,

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

where $\left(\mathbf{I} - \mathbf{H}\right)_j$ denotes the rows of $\mathbf{I} - \mathbf{H}$ corresponding to cluster j. The result was to define $\mathbf{A}_j$ in terms of a matrix square-root inverse. For example, if the working model is set to be the identity matrix $\mathbf{I}_j$ (as suggested by Imbens and Kolesar), then $\mathbf{A}_j = \left(\mathbf{I} - \mathbf{H}\right)_j^{-1}$. This formulation of $\mathbf{A}_j$, however, is problematic when there are fixed effects in a model, as is common in economic applications. In the next two subsections, we address two problems that commonly arise, thus providing a version of BRL that is usable in a wide-range of applications.

2.1 Generalized Inverse

In some instances, the equality defining the $\mathbf{A}_j$ matrices above cannot be solved, because some of the matrices involved cannot be inverted. Angrist and Pischke (2009) highlight that this problem (that the $\mathbf{A}_j$ matrices are undefined) arises in models with balanced state-by-year panels with fixed effects for states and for years because the adjustment matrices

involve inverses of matrices that are not of full rank. Given that the inclusion of clusters via both fixed effects and CRVE is now standard, this problem has made it difficult to implement BRL in a wide variety of economic applications.

In order to address this concern, we provide an alternative formulation of $\mathbf{A}_j$ that meets the requirements for generalized inverses. Begin by defining $\mathbf{\Phi}_j$ as a working model for the covariance of the errors in cluster j, and denote $\mathbf{\Phi} = \text{diag}\left(\mathbf{\Phi}_1, ..., \mathbf{\Phi}_m\right)$. Instead of criterion (EQN ??), now define the adjustment matrices $\mathbf{A}_j$ as those satisfying the following criterion:

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

where $\left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j$ denotes the rows of $\mathbf{I} - \mathbf{H}_{\mathbf{X}}$ corresponding to cluster j. A variance estimator that uses such adjustment matrices will be exactly unbiased when the working model is correctly specified.

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

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

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

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

and $\mathbf{B}_j^{+1/2}$ is the symmetric square root of the Moore-Penrose inverse of $\mathbf{B}_j$. The Moore-Penrose inverse is well-defined and unique even when $\mathbf{B}_j$ is not of full rank (Banerjee and Roy, 2014, Thm. 9.18).

Theorem 1. *Let* $\mathbf{L} = \left(\ddot{\mathbf{U}}'\ddot{\mathbf{U}} - \ddot{\mathbf{U}}_j'\ddot{\mathbf{U}}_j\right)$ *and assume that* $\mathbf{L}$ *has full rank* $r + s$*, so that its inverse exists. Then the adjustment matrices* $\mathbf{A}_j$ *defined in (8) and (9) satisfy criterion (7) and* $\mathbf{V}^{CR2}$ *is exactly unbiased when the working covariance model* $\mathbf{\Phi}$ *is correctly specified.*

Proof: See Appendix A.

The above theorm 3 shows that the adjustment matrices given by (8) and (9) satisfy criterion (7). 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. The

result is that by changing the criterion for defining $\mathbf{A}_j$ and using generalized inverses, we are able to now implement the BRL adjustments in a wide range of applications, including the fixed effects models commonly used in economics.

2.2 Absorption and Dummy Equality

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 $\mathbf{A}_j = c\mathbf{I}_j$, 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-equality problem given above can also arise when implementing the BRL method. Below, we show that by defining the problem in a particular way, the adjustment matrices $\mathbf{A}_j$ 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 $\mathbf{W}_j = \mathbf{\Phi}_j^{-1}$. In this case, the adjustment matrices can be calculated using $\tilde{\mathbf{B}}_j$ in place of $\mathbf{B}_j$, where

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

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

Proof: See Appendix A.

As theorem 4 above demonstrates, using $\tilde{\mathbf{B}}_j$ rather than $\mathbf{B}_j$ leads to algebraically identical adjustment matrices; the form of $\tilde{\mathbf{B}}_j$ 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

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

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. These standard errors, however, are rarely interesting on their own, and are instead used to test hypotheses. In order to do so, Wald-type test statistics are calculated, combining information on the estimated parameters in $\boldsymbol{\beta}$ and the CRVE estimates of the variances of these coefficients. These CRVE based 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 can fail when the number of clusters is small, even when small-sample corrections, including the BRL approach developed in the previous section (CR2), are employed (Bell and McCaffrey, 2002; Bertrand, Duflo and Mullainathan, 2004; Cameron et al., 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 using CRVE in small samples. We begin by considering 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. Equivalently, when $q = 1$, this means that in large samples the test follows a standard normal distribution. 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.

Is this really standard?

By moving to the t- and F-distributions, the CR1 corrections greatly improve the performance of these Wald-type tests in small samples. However, these corrections alone are not adequate, since - like the bias adjustment - the degrees of freedom do not account adequately for features of the covariates that affect performance of the statistics. Cameron and Miller highlight this problem, noting that the BRL approach provides an alternative method for estimating the degrees of freedom using a Satterthwaite approximation. This method has been widely tested and shown to perform well; we review the method and findings in the next sub-section. Importantly, however, no such small-sample approach has been developed for multi-parameter hypothesist tests. In the sub-section that follows, we develop such a test. Our approach is similar to a Satterthwaite approximation, in that it involves approximating the distribution of Q using an F distribution with estimated degrees of freedom.

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, 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 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

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

where $\mathbf{p}_j = (\mathbf{I} - \mathbf{H}_{\mathbf{X}})_j' \mathbf{A}_j \mathbf{W}_j \ddot{\mathbf{R}}_j \mathbf{M}_{\ddot{\mathbf{R}}} \mathbf{c}$. Alternately, for any of the CRVEs one could instead use an "empirical" estimate of the degrees of freedom, constructed by substituting $\mathbf{e}_j\mathbf{e}_j'$ in

Can we use $\mathbf{H}_{\ddot{\mathbf{U}}}$ here instead?

place of $\mathbf{\Sigma}_j$. 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 η degrees of freedom and scale matrix $\mathbf{C}\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{C}'$, 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.

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 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 $\boldsymbol{\Phi}$, 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 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 η_M reduces to ν_M from Equation (13) if $q = 1$.

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.

3.3 Simulation Study

In the previous sub-sections, we have argued that the previously proposed BRL (CR2) t-test and our proposed AHT F-test both perform significantly better than the standard CR1 based tests commonly used in economic applications. This argument is based on extensive simulations that illustrate that the Type I error of the CR1 t-tests can be far above nominal while the CR2 t-test remains close to nominal, and that the differences in these patterns

depends largely on features of the covariates being tested. For example, simulation studies by MacKinnon (2013) indicate that when a covariate has high skew, particularly with a large leverage point, that for a test with $\alpha = 0.05$, the Type I error can be as high as 0.30 (CITE). Similarly, simulations by Tipton (2014) indicate that in these cases, the CR2 test performs vastly better, largely because the degrees of freedom of the CR2 test are much smaller than those of the CR1 test (e.g., df = 5.05 vs df = 38). Imbens and Kolesar .... Cameron and Miller .... While not as widely studied, simulation results for the AHT F-test indicate similar patterns (Tipton and Pustejovsky, in press). In a large set of simulations, these indicate that the CR1 based F-test (with q and $m - 1$ degrees of freedom) can lead to Type I error many times larger than stated, and that like the t-test, these patterns are strongest when the covariates included in the test exhibit large imbalances or skew. Also like the CR2 t-test, the AHT F-test performs close to nominal largely because these covariate features are taken into account in the calculation of the degrees of freedom, which can be much smaller than $m - 1$ in practice.

In this section, we illustrate the performance of the new AHT F-test using a simulation study focused on the data generation models commonly found in economic analyses. Our focus here is on indicating cases in which the results from the standard CR1 F-test may diverge largely from that of this CR2 test. By studying these "extreme" cases, we can have a sense of how these tests perform in wide variety of circumstances, and the factors that affect performance. In particular, we focus here on cases in which the true covariance structure deviates significantly from the 'working model', since we anticipate that economists used to the "model-free" benefits of CRVE may be particularly interested in these findings.

4 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.

4.0.1 Tennessee STAR class-size experiment.

The Tennessee STAR class size experiment is one of the most well studied interventions in education. In the experiment, K âĂŞ 3 students and teachers were randomized within each of 79 schools to one of three conditions: small class-size (targetted to have 13-17 students), regular class-size, or regular class-size with an aide (see Schazenbach, 2006 for a review). Analyses of the original study and follow up waves have found that being in a small class improves a variety of 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. The analytic model is

$$Y_{ijk} = \mathbf{z}'_{jk}\boldsymbol{\alpha}_i + \mathbf{x}'_{jk}\boldsymbol{\beta} + \gamma_k + \epsilon_{ijk} \tag{16}$$

For outcome i, student j is found in school k; $\mathbf{z}_{jk}$ includes dummies for the small-class and regular-plus-aide conditions; and the vector $\mathbf{x}_{jk}$ includes a set of student demographics (i.e., free or reduced lunch status; race; gender; age). Following Krueger (1999), we put the the reading, word recognition, and math scores on comparable scales by converting each outcome to percentile rankings based upon their distributions in the control condition.

We estimated the model in two ways. First, we estimated $\boldsymbol{\alpha}_i$ separately for each outcome i and tested the null hypothesis that $\boldsymbol{\alpha}_i = \mathbf{0}$. Second, we use the seemingly unrelated regression (SUR) framework to test for treatment effects across conditions, using a simultaneous test 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 $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_3 = \mathbf{0}$. In all models, we estimated $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}$ after absorbing the school fixed effects and clustered the errors by school.

4.0.2 Heterogeneous treatment impacts

Angrist and Lavy (2009) reported results from a randomized trial in Israel aimed at increasing matriculation certification for post-secondary education among low achievers. In the Achievement Awards demonstration, 40 non-vocational high schools with the lowest 1999 certification rates nationally were selected (but with a minimum threshold of 3%). This included 10 Arab and 10 Jewish religious schools and 20 Jewish secular schools. The 40 schools were then pair-matched based on the 1999 certification rates, and within each pair one school was randomized to receive a cash-transfer program. In these treatment schools, every student who completed certification was eligible for a payment. The total amount at stake for a student who passed all the milestones was just under $2,400.

Baseline data was collected in January 2001 with follow up data collected in June 2001 and 2002. Following Angrist and Lavy (2009), we focus on the number of certification tests taken as the outcome and report results separately for girls, for boys, and for the combined sample. Given that the program took place in three different types of schools, in this example we focus on determining if there is evidence of variation in treatment impacts across types of schools (i.e., Jewish secular, Jewish religious, and Arab). We use the analytic model:

$$Y_{ij} = \mathbf{z}_j'\boldsymbol{\alpha} + T_j\mathbf{z}_j\boldsymbol{\delta} + \mathbf{x}_{ij}'\boldsymbol{\beta} + \epsilon_{ij} \tag{17}$$

In this model for student i in school j, $\mathbf{z}_j$ is a vector of dummies indicating school type; T_j is a treatment dummy indicating if school j was assigned to the treatment condition; and $\mathbf{x}_{ij}$ contains individual student demographics (i.e., motherâĂŹs and fatherâĂŹs education; immigration status; number of siblings; and an indicator for the quartile of their pre-test achievement from previous years). The components of $\boldsymbol{\delta}$ represent the average treatment impacts in Jewish secular, Jewish religious, and Arab schools. We test the null hypothesis that $\delta_1 = \delta_2 = \delta_3$ to determine if the treatment impact differs across school types. In the second panel of Table 1 we provide the results of this test separately for boys and girls and by year. Importantly, note that the 2000 results are baseline tests, while the 2001 and 2002 results measure the effectiveness of the program.

Add note about program being discontinued in 2002

4.0.3 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 36 time points within 51 states (including the District of Columbia). We focus on the model,

$$Y_{tj} = \mathbf{r}_{tj}'\boldsymbol{\alpha} + \gamma_j + \epsilon_{ij}. \tag{18}$$

In this model, time-point t is nested within state j; the outcome Y_{tj} is log-earnings, which are reported in 1999 dollars; $\mathbf{r}_{tj}$ includes a vector of demographic covariates specific to the time point (i.e., dummy variables for female and white; age and age-squared); and γ_j is a fixed effect for state j.

For sake of example, we focus here on determining whether to use a fixed effects (FE) estimator or a random effects (RE) estimator the four parameters in $\boldsymbol{\alpha}$, based on a Hausmann test. In an OLS model with uncorrelated, the Hausmann test directly compares the vectors of FE and RE estimates using a chi-squared test. However, this specification fails when cluster-robust standard errors are employed, and instead an artificial-Hausman test (Arellano, 1993) is typically used (Wooldridge, 2002, pp. 290-291). This test instead amends the model to additionally include within-cluster deviations (or cluster aggregates) of the variables of interest. In our example, this becomes,

$$Y_{tj} = \mathbf{r}_{tj}'\boldsymbol{\alpha} + \ddot{\mathbf{r}}_{tj}\boldsymbol{\beta} + \gamma_j + \epsilon_{tj}, \tag{19}$$

where $\ddot{\mathbf{r}}_{tj}$ denotes the vector of within-cluster deviations of the covariates (i.e., $\ddot{\mathbf{r}}_{tj} = \mathbf{r}_{tj} - \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_{tj}$). The four parameters in $\boldsymbol{\beta}$ represent the differences between the within-panel and between-panel estimates of $\boldsymbol{\alpha}$. 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 $\gamma_1, ..., \gamma_J$ are mutually independent, normally distributed, and independent of ϵ_{tj}.

5 DISCUSSION

A BRL adjustment matrices

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

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

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

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

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

Because $\mathbf{D}_j$, and $\mathbf{\Phi}$ are positive definite and $\mathbf{B}_j$ is symmetric, the eigenvectors $\mathbf{V}$ define an orthonormal basis for the column span of $\left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right)_j$. We now show that $\ddot{\mathbf{U}}_j$ is in the column space of $\left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j$. Let $\mathbf{Z}_j$ be an $n_j \times (r+s)$ 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 $\mathbf{Z} = \left(\mathbf{Z}_1', ..., \mathbf{Z}_m'\right)'$. Now observe that $\left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right)\mathbf{Z} = \mathbf{Z}$. It follows that

CITATION?

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

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

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

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

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

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

It follows that

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

□

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

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

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

and

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

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

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

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

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

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

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

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

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.

First, note that any of the CR estimatosr can be written in the form $\mathbf{V}^{CR2}=\sum_{j=1}^{M}\mathbf{P}_j\mathbf{e}_j\mathbf{e}_j'\mathbf{P}_j'$ for $r\times n_j$ matrices $\mathbf{P}_j=\mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}_j'\mathbf{W}_j\mathbf{A}_j$. Let $\mathbf{c}_1,\mathbf{c}_2,\mathbf{c}_3,\mathbf{c}_4$ be fixed, $p\times 1$ vectors and consider the linear combination $\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2$. Bell and McCaffrey (2002, Theorem 4) show that the linear combination is a quadratic form in $\mathbf{y}$:

$$\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2=\mathbf{y}'\left(\sum_{j=1}^{m}\mathbf{p}_{2j}\mathbf{p}_{1j}'\right)\mathbf{y},$$

for $N\times 1$ vectors $\mathbf{p}_{sh}=\left(\mathbf{I}-\mathbf{H_X}\right)_h'\mathbf{P}_h'\mathbf{c}_s$, $s=1,...,4$, and $h=1,...,m$.

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

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

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

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

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 $\mathbf{p}_{1i}'\mathbf{\Sigma}\mathbf{p}_{2j}$, $i,j=1,...,m$.

References

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. (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.

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.

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.

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