NCEE 2009-0061 rev.

U. S. DEPARTMENT OF EDUCATION

Technical Methods Report: The Estimation of Average Treatment Effects for Clustered RCTs of Education Interventions

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

Institute of Education Sciences

Technical Methods Report: The Estimation of Average Treatment Effects for Clustered RCTs of Education Interventions

August 2009

Peter Z. Schochet
Mathematica Policy Research

Abstract

This paper examines the estimation of two-stage clustered RCT designs in education research using the Neyman causal inference framework that underlies experiments. The key distinction between the considered causal models is whether potential treatment and control group outcomes are considered to be fixed for the study population (the finite-population model) or randomly selected from a vaguely-defined universe (the super-population model). Appropriate estimators are derived and discussed for each model. Using data from five large-scale clustered RCTs in the education area, the empirical analysis estimates impacts and their standard errors using the considered estimators. For all studies, the estimators yield identical findings concerning statistical significance. However, standard errors sometimes differ, suggesting that policy conclusions from RCTs could be sensitive to the choice of estimator. Thus, a key recommendation is that analysts test the sensitivity of their impact findings using different estimation methods and cluster-level weighting schemes.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

NCEE 2009-0061 rev.
U.S. DEPARTMENT OF EDUCATION

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences under Contract ED-04-CO-0112/0006.

Disclaimer
The Institute of Education Sciences (IES) at the U.S. Department of Education contracted with Mathematica Policy Research to develop methods for estimating average treatment effects in education evaluations. The views expressed in this report are those of the author and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
John Q. Easton
Acting Commissioner

April 2009

Schochet, Peter Z. (2009). *The Estimation of Average Treatment Effects for Clustered RCTs of Education Interventions* (NCEE 2009-0061 rev.). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ncee.ed.gov.

Alternate Formats
Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

The author for this report, Dr. Peter Z. Schochet, is an employee of Mathematica Policy Research with whom IES contracted to develop the methods that are presented in this report. Dr. Schochet and other MPR staff do not have financial interests that could be affected by the content in this report.

Foreword

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

In support of this mission, NCEE promotes methodological advancement in the field of education evaluation through investigations involving analyses using existing data sets and explorations of applications of new technical methods, including cost-effectiveness of alternative evaluation strategies. The results of these methodological investigations are published as commissioned, peer reviewed papers, under the series title, Technical Methods Reports, posted on the NCEE website at http://ies.ed.gov/ncee/pubs/. These reports are specifically designed for use by researchers, methodologists, and evaluation specialists. The reports address current methodological questions and offer guidance to resolving or advancing the application of high-quality evaluation methods in varying educational contexts.

This NCEE Technical Methods paper serves to open up the "black box" of impact estimation for applied education researchers, and highlights both the importance of close attention to the estimation method and the importance of performing sensitivity tests using different estimation methods. Using the Neyman causal inference framework that underlies experiments, the report examines the estimation of impacts in two-stage clustered RCT designs. Several causal models are considered. The key distinction between these models is whether potential treatment and control group outcomes are considered to be fixed for the study population (the finite-population model) or randomly selected from a vaguely-defined universe (the super-population model). Appropriate estimators are derived and discussed for each model, highlighting the differences in underlying assumptions among them. Using data from five large-scale clustered education RCTs, the empirical analysis estimates impacts and their standard errors using the considered estimators to assess whether impact findings are sensitive to the use of different estimation methods and cluster-level weighting schemes each employs.

Contents

List of Tables

Chapter 1: Introduction

In randomized control trials (RCTs) of educational interventions, random assignment is often performed at the school or classroom level rather than at the student level. These group-based designs are common, because RCTs in the education field often test interventions that provide enhanced services to teachers (for example, training in a new reading or math curriculum or mentoring services) or that test interventions that affect the entire school (for example, a school-wide social and character development program or re-structuring initiative). Thus, for these types of interventions, it is infeasible to randomly assign the treatment directly to students.

Under these group-based designs, data are typically collected on students. Thus, using student-level data, the statistical procedures that are used to estimate average treatment effects (ATEs) and their standard errors must account for the potential correlation of the outcomes of students within the same groups. In particular, the standard errors of the ATE estimators must be inflated to account for design effects due to clustering.

Over the past 40 years, a huge statistical literature across multiple disciplines discusses the estimation of treatment effects under two-stage clustered designs (see, for example, Rao 1972, Harville 1977, Laird and Ware 1982, Hsiao 1986, Liang and Zeger 1986, Baltagi and Chang 1994, Murray 1998, Raudenbush and Bryk 2002, Wooldridge 2002, and De Leeuw and Meijer 2008). These models have a number of labels, including random effects models, random coefficient models, one-way models, variance components models, panel models, hierarchical linear models (HLM), and linear mixed models. A number of statistical packages have been developed to estimate these models using analysis of variance (ANOVA), maximum likelihood (ML), restricted ML (REML), generalized estimation equation (GEE), and other methods.

This paper contributes to this literature by discussing the estimation and interpretation of the ATE parameter under clustered RCTs using the non-parametric model of causal inference that underlies experimental designs. This model was introduced for non-clustered designs by Neyman (1923) and later developed in Rubin (1974, 1977) and Holland (1986). This article extends this theory to two-stage clustered RCTs, and develops regression equations that are consistent with this theory. The analysis focuses on continuous outcomes (such as test scores), and discusses relevant ATE parameters assuming that the outcome data are either (1) fixed for the study population (a finite-population model) or (2) random draws from population outcome distributions (the more common super-population model). Appropriate estimation methods and asymptotic moments are discussed for each model, and the methods are linked to the following commonly-used statistical packages: SAS, STATA, R, SUDAAN, and HLM. The paper considers both simple differences-in-means models and those that include baseline covariates.

Finally, ATEs and their standard errors are estimated using the considered methods using data from five recent large-scale clustered RCTs in the education area. The purpose of this analysis is to examine the robustness of study findings to alternative estimation approaches. This is important, because education researchers typically employ statistical packages and estimation routines with which they are most comfortable, and published articles in the evaluation literature rarely report impact results using alternative estimation schemes. Thus, this article can provide information to education researchers about the assumptions underlying commonly-used ATE estimation methods, how these methods work, and the sensitivity of impact findings to alternative estimation strategies. The goal is not to identify the best methods, but to discuss options and interpretation.

The rest of this paper is in six chapters. Chapter 2 discusses the Neyman causal inference model, and Chapters 3 and 4 discuss the estimation of the ATE parameter under the finite- and super-population models, respectively. Chapter 5 discusses methods for estimating variance components for the super-population model, and Chapter 6 presents findings from the empirical analysis. The final chapter presents a summary and conclusions.

Chapter 2: The Neyman Causal Inference Model

This chapter discusses the Neyman finite-population (FP) and super-population (SP) causal inference models under two-stage clustered designs—the most common designs used in education RCTs. The focus is on continuous outcomes. The theory is then used to derive regression equations for estimating the ATE parameters.

The Neyman Finite-Population Model for Two-Stage Clustered Designs

Consider an experimental design where n schools (or classrooms) are randomly assigned to either a single treatment or control condition. The sample contains np treatment and $n(1-p)$ control group schools where p is the sampling rate to the treatment group $(0<p<1)$. It is assumed that the sample contains m_i students from school i and that there are $M=\sum_{i=1}^{n} m_i$ total students in the sample. It is assumed that student outcomes are not affected by the treatment status of other students.

It is assumed for now that the n schools and M students define the population universe—the FP model considered by Neyman for non-clustered designs. Let Y_{Tij} be the "potential" outcome for student j in school i in the treatment condition and Y_{Cij} be the potential outcome for the student in the control condition. The difference between the two fixed potential outcomes, $(Y_{Tij}-Y_{Cij})$, is the student-level treatment effect, and the ATE parameter, β_1, is the average treatment effect over all students:

$$(1) \quad \beta_1 = \overline{Y}_T - \overline{Y}_C = \frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Y_{Tij}-Y_{Cij}).$$

This ATE parameter cannot be calculated directly because potential outcomes for each student cannot be observed in both the treatment and control conditions. Formally, if T_i is a treatment status indicator variable that equals 1 for treatment schools and 0 for control schools, then the *observed* outcome for a student, y_{ij}, can be expressed as follows:

$$(2) \quad y_{ij} = T_i Y_{Tij} + (1-T_i)Y_{Cij}.$$

Importantly, the potential outcomes in (2) are *fixed* and the only source of randomness is T_i. Thus, under the Neyman model, the ATE parameter pertains only to those students and schools at the time the study was conducted. Stated differently, the impact findings have internal validity but do not necessarily generalize beyond the study sample. This approach can be justified on the grounds that schools are usually *purposively selected* for education RCTs, and thus, may be a self-selected sample of schools that are willing to participate, and that are deemed to be suitable for the study based on their student and teacher populations and typical service offerings. Similarly, students in the study sample may not be representative of all students in the study schools, because they could be a potentially nonrandom subset of students whose parents consented to participate in the study, who provided follow-up data, and who did not leave the study schools between baseline and followup.[1]

Under this fixed population scenario, researchers are to be agnostic about whether the study results have external validity. Policymakers and other users of the study results can decide whether the impact evidence is sufficient to adopt the intervention on a broader scale, perhaps by examining the similarity of the observable characteristics of schools and students in the study samples to their own contexts, and using results from subgroup and implementation analyses.

Following the approach for non-clustered designs used by Freedman (2008) and Schochet (2009), a regression model for (2) can be constructed by re-writing (2) as follows:

(3) $y_{ij} = \beta_0 + \beta_1(T_i - p) + \eta_{ij}$, where

- $\beta_0 = p\overline{Y}_T + (1-p)\overline{Y}_C$ and $\beta_1 = \overline{Y}_T - \overline{Y}_C$ are parameters to be estimated

- $\eta_{ij} = \alpha_{ij} + \tau_{ij}(T_i - p)$ is an "error" term, where $\alpha_{ij} = p(Y_{Tij} - \overline{Y}_T) + (1-p)(Y_{Cij} - \overline{Y}_C)$ and $\tau_{ij} = (Y_{Tij} - \overline{Y}_T) - (Y_{Cij} - \overline{Y}_C)$.[2]

The error term η_{ij} is a function of two terms: (1) α_{ij}, the expected observed outcome for the student relative to the expected mean observed outcome; and (2) τ_{ij}, the student-level treatment effect relative to the ATE. Note that α_{ij} and τ_{ij} sum to zero over all students. This model is non-parametric because it does not depend on the distributions of the potential outcomes.

The model in (3) does not satisfy key assumptions of the usual random effects model, because η_{ij} does not have mean zero (over all possible treatment assignment configurations), and, to the extent that τ_{ij} varies across students, η_{ij} is heteroscedastic, $Cov(\eta_{ij}\eta_{ij'})$ is not constant for students in the same schools, $Cov(\eta_{ij}\eta_{i'j'})$ is nonzero for students in different schools (for $i \neq i', j \neq j'$), and η_{ij} is correlated with the regressor $(T_i - p)$:

$$E(\eta_{ij}) = \alpha_{ij}, \quad Var(\eta_{ij}) = \tau_{ij}^2 p(1-p), \quad Cov(\eta_{ij}\eta_{ij'}) = \tau_{ij}\tau_{ij'}p(1-p),$$
$$Cov(\eta_{ij}\eta_{i'j'}) = -\tau_{ij}\tau_{i'j'}p(1-p)/(n-1), \quad E[(T_i - p_i)\eta_{ij}] = \tau_{ij}p(1-p).$$

Note that in this model, the error terms for students within the same schools are correlated only because they have the same treatment status, not because they face similar environments.

[1] For cost reasons, in education RCTs, follow-up data are not usually collected for students in the baseline sample who leave the study districts.

[2] In (3), the term (T_i-p) is used rather than T_i because it simplifies the mathematical proofs presented later in this paper, but this centering has no effect on the findings.

Importantly, the model in (3) should *not* be confused with a *fixed effects* model, where cluster effects are treated as fixed, and cluster-level dummy variables are included in the model. Rather, the model treats cluster-level effects as *random* due to the randomness of treatment status in the model error term.

Finally, (3) implicitly assumes that schools are weighted by their student sample sizes. An alternative specification is to weight schools *equally*. In this case, the ATE parameter is $\beta_1 = \overline{\overline{Y}}_T - \overline{\overline{Y}}_C$, where $\overline{\overline{Y}}_T = (1/n)\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Y_{Tij}/m_i)$ and $\overline{\overline{Y}}_C = (1/n)\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Y_{Cij}/m_i)$ are averages of school-level means. This ATE parameter pertains to the average school effect in the sample rather than to the average student effect. This weighting scheme will result in different impact estimates than the unweighted analysis if student sample sizes vary across schools and impacts vary by school sample size.

The Super-Population Model for Two-Stage Clustered Designs

We now consider a SP version of the Neyman causal inference model where the study schools and students are assumed to be *random* samples from broader populations (see Imbens and Rubin 2007 and Schochet 2008, 2009). This framework is typically used to estimate impacts under clustered RCTs in the education area, and is consistent with popular linear mixed model approaches, such as HLM.

Under this framework, students are nested within schools. Let Z_{Ti} be the potential outcome (mean posttest score) for school *i* in the treatment condition and Z_{Ci} be the potential outcome for school *i* in the control condition. Potential outcomes for the *n* study schools are assumed to be random draws from potential treatment and control outcome distributions in the study super-population. It is assumed that means and variances of these distributions are finite and denoted by μ_T and σ_{uT}^2 for potential treatment outcomes and μ_C and σ_{uC}^2 for potential control outcomes. These two outcome distributions also define the distribution of school-level treatment effects in the super-population, which are assumed to have mean μ_τ and variance σ_τ^2.

Suppose next that m_i students are sampled from the student super-population in study school *i*. The potential student-level outcomes Y_{Tij} and Y_{Cij} are now assumed to be *random* draws from student-level potential outcome distributions (which are conditional on school-level potential outcomes) with respective means Z_{Ti} and Z_{Ci} and respective variances $\sigma_{eT}^2 > 0$ and $\sigma_{eC}^2 > 0$.

Under the SP model, the ATE parameter is $\mu_\tau = E(Z_{Ti} - Z_{Ci}) = \mu_T - \mu_C$. Thus, the impact findings are now assumed to generalize to the super-population of schools that are "similar" to the study schools. How should one interpret this super-population? Does it pertain to the study schools over the "long term" for a broader universe of students and school staff that change over time? Does it pertain to a broader set of schools in the study districts? To similar schools nationwide? The answers to these questions will likely depend on the context (and may not exist), but researchers should be aware that the usual approach for estimating treatment effects in education research makes the implicit assumption of external validity to a school universe that is likely to be vaguely defined. Nonetheless, this approach can be justified on the grounds that policymakers may generalize the findings anyway, especially if the study provides a primary basis for deciding whether to implement the tested interventions more broadly. Furthermore, this approach is more consistent with the Bayesian view that assessing intervention effects is a dynamic process that takes place in a context of continuously increasing knowledge.

As before, we can use (2) to express *observed* student outcomes in terms of potential outcomes, and can rearrange terms to yield the following regression model:

(4) $y_{ij} = \alpha_0 + \alpha_1 T_i + (u_i + e_{ij})$, where

$\alpha_0 = \mu_C$ and $\alpha_1 = \mu_T - \mu_C$ (the ATE parameter) are coefficients to be estimated

$u_i = T_i(Z_{Ti} - \mu_T) + (1 - T_i)(Z_{Ci} - \mu_C)$ is a school-level error term where $E(u_i) = 0$, $E(T_i u_i) = 0$, $Var(u_i \mid T_i = 1) = \sigma^2_{uT}$, and $Var(u_i \mid T_i = 0) = \sigma^2_{uC}$

$e_{ij} = T_i(Y_{Tij} - Z_{Ti}) + (1 - T_i)(Y_{Cij} - Z_{Ci})$ is a student-level error term where $E(e_{ij}) = 0$, $E(T_i e_{ij}) = E(u_i e_{ij}) = 0$, $Var(e_{ij} \mid T_i = 1) = \sigma^2_{eT}$, and $Var(e_{ij} \mid T_i = 0) = \sigma^2_{eC}$.

Furthermore, if we define $\delta_{ij} = u_i + e_{ij}$ as the total error term:

$$Var(\delta_{ij} \mid T_i = 1) = \sigma^2_{uT} + \sigma^2_{eT},\ Var(\delta_{ij} \mid T_i = 0) = \sigma^2_{uC} + \sigma^2_{eC},\ Cov(\delta_{ij}, \delta_{i'j'}) = 0,$$
$$Cov(\delta_{ij}, \delta_{ij'} \mid T_i = 1) = \sigma^2_{uT},\ Cov(\delta_{ij}, \delta_{ij'} \mid T_i = 0) = \sigma^2_{uC}.$$

Thus, this model is the usual random effects model with an exchangeable block diagonal variance-covariance matrix for the error vector except that variances and covariances are allowed to differ for treatments and controls.

Finally, note that (4) can also be derived using the following two-level HLM model (Bryk and Raudenbush, 1992):

$$Level\,1:\quad y_{ij} = z_i + e_{ij}$$
$$Level\,2:\quad z_i = \alpha_0 + \alpha_1 T_i + u_i,$$

where $z_i = T_i Z_{Ti} + (1 - T_i) Z_{Ci}$ is the observed school-level outcome, Level 1 corresponds to students, and Level 2 to units. Inserting the Level 2 equation into the Level 1 equation yields (4). Thus, the HLM approach is consistent with the SP causal inference theory.

Chapter 3: ATE Parameter Estimation for the Finite-Population Model

This chapter discusses ATE parameter and variance estimation for the FP model with and without baseline covariates. Mathematical proofs of asymptotic results are provided in the appendix. It is assumed for the remainder of this article that sample sizes of clusters are large enough so that asymptotic results are approximately valid (see Bingenheimer and Raudenbush, 2004 for a discussion of this issue).

Finite-Population Model Without Covariates

Ordinary least squares (OLS) methods are appropriate for estimating β_1 in (3), because the ATE parameter for the FP model pertains to the study sample only. The following lemma provides the asymptotic moments of the OLS estimator.

Lemma 1. The simple OLS estimator for β_1 under the FP model in (3) is $\hat{\beta}_{1,SR} = (\bar{y}_T - \bar{y}_C)$, where $\bar{y}_T$ and $\bar{y}_C$ are (unweighted) sample means for the treatment and control groups, respectively. As *n* increases to infinity for an increasing sequence of finite populations, $\hat{\beta}_{1,SR}$ is asymptotically unbiased. Furthermore, assume that:

$$(5)\quad \bar{m} = \sum_{i=1}^{n} m_i / n \rightarrow \bar{\bar{m}},\quad \frac{1}{n\bar{m}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} (Y_{Tij} - \bar{Y}_T)(Y_{Tik} - \bar{Y}_T) \rightarrow \bar{S}_T^2,$$

$$\frac{1}{n\bar{m}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} (Y_{Cij} - \bar{Y}_C)(Y_{Cik} - \bar{Y}_C) \rightarrow \bar{S}_C^2, \text{ and } \frac{1}{n\bar{m}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} \tau_{ij}\tau_{ik} \rightarrow \bar{S}_\tau^2,$$

where $\bar{\bar{m}}, \bar{S}_T^2, \bar{S}_C^2,$ and $\bar{S}_\tau^2$ are fixed, nonnegative, real numbers. Then, $\hat{\beta}_{1,SR}$ is asymptotically normal with variance:

$$(6)\quad AsyVar(\hat{\beta}_{1,SR}) = \frac{\bar{S}_T^2}{n\bar{\bar{m}}p} + \frac{\bar{S}_C^2}{n\bar{\bar{m}}(1-p)} - \frac{\bar{S}_\tau^2}{n\bar{\bar{m}}}.$$

The $\bar{S}_T^2$ and $\bar{S}_C^2$ terms pertain to the extent to which *potential outcomes* vary and co-vary across students within the same schools. The $\bar{S}_\tau^2$ term pertains to the extent to which *treatment effects* vary and co-vary across students within schools. Note that if student-level treatment effects are constant, $\bar{S}_\tau^2 = 0$ and $\bar{S}_T^2 = \bar{S}_C^2$.

With heterogeneous treatment effects, it is difficult to find a consistent estimator for $\bar{S}_\tau^2$, because this requires unobserved information on student-level treatment effects. However, because $\bar{S}_\tau^2 \geq 0$, ignoring this term will provide conservative variance estimators. Following this approach, a consistent estimator for the first two terms on the right-hand side in (6) can be obtained using the population averaged generalized estimating equation (GEE) approach developed by Liang and Zeger (1986) for clustered data (see also Hardin and Hilbe 2003).

To describe this method for general applications, it is assumed that $\mathbf{x_{ij}}$ is a row vector of model baseline covariates (including the intercept and $T_i - p$), $\mathbf{y_i}$ is an $m_i x1$ column vector of student outcomes, and $\mathbf{V_i}$ is the assumed ("working") $m_i x m_i$ covariance structure for $\mathbf{y_i}$. The GEE method for estimating the vector of regression parameters $\boldsymbol{\beta}$ solves the following equation for the score function $\mathbf{S(\boldsymbol{\beta})}$:

$$(7) \quad \mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_\mathbf{i}'(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mathbf{V}_\mathbf{i}^{-1}(\mathbf{y_i} - \boldsymbol{\mu}_\mathbf{i}(\boldsymbol{\beta})) = \mathbf{0},$$

where $\boldsymbol{\mu}_\mathbf{i}(\boldsymbol{\beta})$ is the expected value of $\mathbf{y_i}$ that is linked to a linear combination of the covariates through a monotonic differentiable link function g where $g(\mu_{ij}) = \mathbf{x_{ij}}\boldsymbol{\beta}$ and $\mu_{ij} = g^{-1}(\mathbf{x_{ij}}\boldsymbol{\beta})$.

Equation (7) can be solved iteratively using a Taylor series expansion of $\mathbf{S}(\hat{\boldsymbol{\beta}})$ around $\mathbf{S}(\boldsymbol{\beta})$. Under this approach, the estimated parameter vector $\hat{\boldsymbol{\beta}}^{\mathbf{(iter+1)}}$ at iteration $(iter+1)$ can be updated from $\hat{\boldsymbol{\beta}}^{\mathbf{(iter)}}$ as follows:

$$\hat{\boldsymbol{\beta}}^{\mathbf{(iter+1)}} = \hat{\boldsymbol{\beta}}^{\mathbf{(iter)}} + \mathbf{I}_\mathbf{0}^{\mathbf{-1(iter)}} \mathbf{S}(\hat{\boldsymbol{\beta}}^{\mathbf{(iter)}}), \text{ where}$$

$$(8) \quad \mathbf{I_0} = -E(\partial \mathbf{S}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}_\mathbf{i}^{-1} \frac{\partial \boldsymbol{\mu}_\mathbf{i}}{\partial \boldsymbol{\beta}}$$

is the information matrix. The matrix $\mathbf{I_0}$ is sometimes replaced by $\mathbf{J_0} = \partial \mathbf{S}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ (Binder 1983).

The model-based variance estimator of the solution $\hat{\boldsymbol{\beta}}$ is $\mathbf{I}_\mathbf{0}^{-1}$. The empirical or robust "sandwich" variance estimator uses the data to correct for the potential misspecification of $\mathbf{V_i}$ and equals $\mathbf{I}_\mathbf{0}^{-1}\mathbf{I_1}\mathbf{I}_\mathbf{0}^{-1}$ where

$$(9) \quad \mathbf{I_1} = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}_\mathbf{i}^{-1} \mathbf{r_i} \mathbf{r_i'} \mathbf{V}_\mathbf{i}^{-1} \frac{\partial \boldsymbol{\mu}_\mathbf{i}}{\partial \boldsymbol{\beta}},$$

and $\mathbf{r_i} = (\mathbf{y_i} - \boldsymbol{\mu}_\mathbf{i}(\hat{\boldsymbol{\beta}}))$ is an $m_i x1$ vector of regression residuals.

In our application, we assume (1) an *independent* working correlation structure (that is, $\mathbf{V_i}$ is the identity matrix), (2) an identity link function ($\mu_{ij} = \beta_0 + \beta_1(T_i - p)$), and (3) the empirical sandwich variance estimator. The ATE estimator for this linear model is then $\hat{\beta}_{1,GEE} = (\bar{y}_T - \bar{y}_C)$ with the following asymptotic variance estimator:

$$(10) \quad Asy\hat{V}ar(\hat{\beta}_{1,GEE}) = \frac{1}{d^2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} (T_i - p)^2 r_{ij} r_{ik} \approx \frac{1}{(n\bar{m})^2 p(1-p)} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} r_{ij} r_{ik},$$

where $d = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(T_i - p)^2$. This variance estimator is based on the sums of products and cross-products of OLS residuals for students within the same schools. Table 3.1 displays statistical package routines that use this method.

If schools are to be weighted equally under unbalanced designs, the GEE method can be applied by first pre-multiplying the outcome and explanatory variables (including the intercept) by the weights $\sqrt{w_{ij}}$ where $w_{ij} \propto 1/m_i$ (Pfeffermann et al. 1998). Under this approach, it may be reasonable to also weight each school district equally if random assignment is conducted within school districts.

Importantly, as discussed in Murray (1998), the GEE method should be used only if the number of clusters in each research condition is at least 20. For smaller sample sizes, simulations demonstrate that the Type I error rate may not be close to the nominal level.

Finally, for the equal-school weighting scheme, model-free permutation (randomization) tests can also be used to test the strong null hypothesis that all student-level treatment effects are zero (Gail et al., 1996). Under this approach, observed school means are used to construct the distribution of all possible treatment effects under the null hypothesis of no impacts. This is done by (1) allocating schools to all possible combinations of np "pseudo-treatment" schools and $n(1-p)$ "pseudo-control" schools, (2) estimating a treatment effect $\hat{\beta}_1$ for each of the $[n!/\,np!n(1-p)!]$ allocations, (3) sorting these treatment effects from smallest to largest, (4) observing where in the distribution the treatment effect for the *actual* treatment-control allocation lies, and (5) rejecting the null hypothesis if the actual $\hat{\beta}_1$ lies outside the $\alpha/2$ or $1-(\alpha/2)$ quantiles of the permutation distribution (which will have mean 0).[3] The validity of this method does not rely on a model, but only on correct randomization.

Gail et al. (1996) demonstrate through simulations that Type I error rates of these tests are near nominal levels if n is moderate, p is near 0.5, and variances of the outcomes do not differ substantially across the treatment and control conditions. These conditions are likely to hold in practice. Furthermore, Gail et al. (1996) demonstrate that the procedure performs better using school-level residuals from regression models that include baseline covariates (see below).

Finite-Population Model with Covariates

We now examine ATE estimators when the FP models include fixed covariates, $\mathbf{q_{ij}}$, pertaining to the pre-randomization period. The covariates are not indexed by *T* or *C* because their values are independent of treatment status due to randomization. The covariates could include both school-level covariates and student-level covariates that are centered at school-level means. All covariates are assumed to be centered at grand means.

[3] For moderate *n* (say, *n*>30), the number of possible allocations becomes very large. In these cases, the permutation distribution can be estimated from a large random sample of reallocations of school means to the pseudo-treatment and control groups.

Table 3.1: Routines in the Considered Statistical Packages for Estimating ATE Parameters and Their Standard Errors, by Model

Estimation Method	Variance-Related Formulas in Text	Statistical Packages and Routines	Notes on Estimation and Specification
Finite-Population Model			
GEE	(13)	**Sudaan:** *Regress* **SAS:** *Proc Genmod* **Stata:** *xtgee* or *regress vce(cluster)* command **R:** *gee* or *glm* function	An *independent* working correlation structure must be specified to obtain OLS parameter estimates. The empirical sandwich estimator should be specified. The Zeger or Binder optimization method can be specified in most packages.
Permutation	NA	None	Used for hypothesis testing using school-level means or regression residuals
Super-Population Model			
Balanced Design	(22)	**Sudaan:** *Regress* **SAS:** *Proc Reg or GLM* **Stata:** *regress* command **R:** *lm* function	Parameter and standard errors are obtained by applying OLS to the between-school regression model in (21).
ANOVA	(24), (25)	**SAS:** *Proc Panel* **Stata:** *xtreg sa*	Variance component estimates in (24) and (25) are inserted into (16) and (17) to obtain feasible GLS estimates
ML	(28)-(32)	**SAS:** *Proc Mixed* **Stata:** *xtmixed* command **R**: *lme* package; **HLM2, HLM3, HMLM2**	Yields feasible GLS estimates. Statistical packages use different defaults for using ML or REML, and for using Newton-Raphson, Fisher-Scoring or the EM algorithm for optimization.
REML	See (33)	Same as for ML	Same as for ML
GEE	(34), (35)	Same as for GEE above	An *exchangeable* working correlation structure must be specified; yields feasible GLS estimates using the model-based or empirical sandwich variance estimators. The Zeger or Binder optimization method can be specified in most packages.

NA: Not applicable.

In the Neyman model, the covariates are *irrelevant* variables because (3) is the true model. Thus, the ATE parameters considered above without covariates pertain *also* to the models with covariates.

To examine asymptotic moments of the OLS estimator under the FP model with fixed covariates, we assume in addition to (5) that as *n* approaches infinity:

$$(11)\quad \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i}\alpha_{ij}f_{ik}}{n\bar{m}} \rightarrow \overline{S}^2_{\alpha f},\quad \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i}f_{ij}f_{ik}}{n\bar{m}} \rightarrow \overline{S}^2_{ff}, \text{ and } \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i}h_{ij}f_{ik}}{n\bar{m}} \rightarrow \overline{S}^2_{hf},$$

where f_{ij} is the student's predicted value from a full-sample OLS regression of α_{ij} on $\mathbf{q_{ij}}$; h_{ij} is the predicted value from a full-sample OLS regression of τ_{ij} on $\mathbf{q_{ij}}$, and $\overline{S}^2_{\alpha f}$, $\overline{S}^2_{ff}$, and $\overline{S}^2_{hf}$ are fixed, nonnegative real numbers. The following lemma generalizes results in Schochet (2009) and Freedman (2008) to two-stage clustered designs. The proof is provided in the appendix.

Lemma 2. Let $\hat{\beta}_{1,MR}$ be the multiple regression estimator for β_1 under the model in (3) and assume (5) and (11). Then, $\hat{\beta}_{1,MR}$ is asymptotically normal with mean β_1 and variance:

$$(12)\quad AsyVar(\hat{\beta}_{1,MR}) = \left(\frac{\overline{S}^2_T}{n\overline{\overline{m}}p} + \frac{\overline{S}^2_C}{n\overline{\overline{m}}(1-p)} - \frac{\overline{S}^2_\tau}{n\overline{\overline{m}}}\right) - \frac{1}{n\overline{\overline{m}}p(1-p)}\left(2\overline{S}^2_{\alpha f} - \overline{S}^2_{ff} + 2(1-2p)\overline{S}^2_{hf}\right).$$

The first bracketed term in (12) is the variance of the OLS estimator under the FP model without covariates. The $(2\overline{S}^2_{\alpha f} - \overline{S}^2_{ff})$ term is a generalized version of the usual explained sum of squares from a multiple OLS regression, and will typically generate precision gains if the covariates are correlated with potential outcomes. The $2(1-2p)\overline{S}^2_{hf}$ term pertains to regression-adjusted covariances between α_{ij} and τ_{ij} for students within the same school. This term will be zero if $p=0.5$ or if the covariances between potential outcomes are similar in the treatment and control conditions (which would occur, for example, with constant treatment effects); otherwise this term could have any sign.

A variance estimator for (12) can be obtained using the GEE approach discussed above. Let $\mathbf{X_i} = (\mathbf{K}\ \mathbf{\tilde{T}_i}\ \mathbf{Q_i})$, where $\mathbf{K}$ is an $m_i x1$ column of 1s for the intercept, $\mathbf{\tilde{T}_i}$ is an $m_i x1$ vector containing the $T_i - p$ terms, and $\mathbf{Q_i}$ is a matrix of covariates for school i. In this case, a variance estimator is:

$$(13)\quad Asy\hat{V}ar(\hat{\beta}_{1,MR}) = \left[(\sum_{i=1}^{n}\mathbf{X_i'X_i})^{-1}(\sum_{i=1}^{n}\mathbf{X_i'r_ir_i'X_i})(\sum_{i=1}^{n}\mathbf{X_i'X_i})^{-1}\right]_{(2,2)},$$

where the residuals $\mathbf{r_i}$ are calculated from a full-sample OLS regression of $\mathbf{y_i}$ on $\mathbf{X_i}$. The permutation tests discussed above could also be used for significance testing using the school-level residuals $\overline{r}_i$ (for the equal-school weighting scheme).

Chapter 4: ATE Parameter Estimation For The Super-Population Model

This chapter examines ATE parameter estimation for the SP model with and without baseline covariates, where it is assumed that error variances are the same in the treatment and control conditions: $\sigma^2_{uT} = \sigma^2_{uC} = \sigma^2_u$ and $\sigma^2_{eT} = \sigma^2_{eC} = \sigma^2_e$. This assumption is commonly applied and greatly simplifies the presentation.

This chapter focuses on generalized least squares (GLS) methods that are typically used to provide consistent and efficient estimators for α_1 in (4). However, the chapter starts with a discussion of the OLS approach (which produces consistent, but inefficient estimates) so the SP and FP estimators can be compared using a common approach. Methods for estimating variance components to obtain feasible GLS estimates are discussed in Chapter 5.

Super-Population Model Without Covariates

The SP model in (4) for students in school *i* can be expressed in vector notation as follows:

$$(14) \quad \mathbf{y_i} = \alpha_0 + \alpha_1 \mathbf{T_i} + \boldsymbol{\delta}_\mathbf{i},$$

where $\boldsymbol{\Omega}^*_\mathbf{i} = E(\boldsymbol{\delta}_\mathbf{i}\boldsymbol{\delta}'_\mathbf{i})$ is an $m_i x m_i$ positive definite variance-covariance with diagonal terms $\sigma^2_u + \sigma^2_e$ and off-diagonal terms σ^2_u. The estimation of this model using OLS and GLS methods is discussed next.

OLS Methods

Standard methods (see, for example, Schochet 2008) can be used to show that as *n* increases to infinity, the OLS estimator $\hat{\alpha}_{1,SR} = (\overline{y}_T - \overline{y}_C)$ is asymptotically normal with mean α_1 and asymptotic variance that can be estimated as follows:

$$(15) \quad Asy\hat{V}ar(\hat{\alpha}_{1,SR}) = \frac{1}{p(1-p)} \left(\frac{\sum_{i=1}^{n} m_i^2 \hat{\sigma}^2_u}{(\sum_{i=1}^{n} m_i)^2} + \frac{\hat{\sigma}^2_e}{\sum_{i=1}^{n} m_i} \right),$$

where $\hat{\sigma}^2_u$ and $\hat{\sigma}^2_e$ are estimators for σ^2_u and σ^2_e, respectively. Note that this variance is minimized if $p = 0.5$ and $m_i = m$ for all schools (that is, for balanced designs).

The term in parentheses in (15) can be computed by *summing* the elements of $\hat{\boldsymbol{\Omega}}^*_\mathbf{i}$ across schools and dividing by M^2, where $\hat{\boldsymbol{\Omega}}^*_\mathbf{i}$ is an estimator for $\boldsymbol{\Omega}^*_\mathbf{i}$. Thus, (15) is comparable to the $\overline{S}^2_T$ and $\overline{S}^2_C$ terms in (6) for the FP model. Thus, an important difference between the SP and FP models is that unlike the SP model, the FP model contains $\overline{S}^2_\tau$, which reduces variance. Thus, in theory, the variance may be somewhat smaller under the FP model, which is expected, because the SP model assumes external validity, with an associated loss in statistical precision. However, as noted, it is difficult to estimate $\overline{S}_{\tau}^{2}$ for clustered designs; thus, precision gains for the FP model cannot typically be realized in practice.

GLS Methods

Consider a generic regression model where the covariate and variance matrices for school *i* are denoted by $\mathbf{X_i}$ and $\mathbf{\Omega_i}$, respectively. The feasible GLS estimator of the parameter vector $\boldsymbol{\alpha}$ is then:

$$(16) \quad \hat{\boldsymbol{\alpha}}_{\mathbf{GLS}} = (\textstyle\sum_{i=1}^{n} \mathbf{X_i'}\hat{\mathbf{\Omega}}_{\mathbf{i}}^{\mathbf{-1}}\mathbf{X_i})^{-1}(\textstyle\sum_{i=1}^{n} \mathbf{X}_i'\hat{\mathbf{\Omega}}_{\mathbf{i}}^{\mathbf{-1}}\mathbf{y_i}),$$

where $\hat{\mathbf{\Omega}}_{\mathbf{i}}$ is an estimator for $\mathbf{\Omega_i}$.

In our case $\mathbf{X_i} = [\mathbf{K}\ \mathbf{T_i}]$, so (16) reduces to

$$\hat{\alpha}_{1,GLS} = \frac{\sum_{i:T_i=1}^{np} w_i \overline{y}_i}{\sum_{i:T_i=1}^{np} w_i} - \frac{\sum_{i:T_i=0}^{n(1-p)} w_i \overline{y}_i}{\sum_{i:T_i=0}^{n(1-p)} w_i},$$

where $\overline{y}_i$ is the mean outcome in school i and $w_i = [\hat{\sigma}_u^2 + (\hat{\sigma}_e^2 / m_i)]^{-1}$ is the associated school-level weight. This is a weighted differences-in-means estimator, where the weights are inverses of the variances of school-level means.

The weights can also be expressed as $w_i = [ICC + \{(1-ICC)/m_i\}]^{-1}$ where $ICC = \hat{\sigma}_u^2 / (\hat{\sigma}_u^2 + \hat{\sigma}_e^2)$ is the estimated intraclass correlation coefficient. The first ICC term inside the brackets is common to all schools. Thus, the weights differ due to the second term. Schools with smaller variances (more sampled students) receive more weight in the analysis than schools with larger variances (fewer sampled students), because the larger schools provide more information on the super-population parameters μ_T and μ_C. As ICC approaches zero, the SP weights converge to the FP weights where schools are weighted by their sample sizes. Conversely, as ICC approaches one, the SP weights converge to the FP weights where schools are weighted equally. Under the SP approach, it may be reasonable to weight each school district by the size of their school population if random assignment is conducted within school districts.

It is well known that under weak regularity conditions, the feasible GLS estimator is asymptotically normal with mean $\boldsymbol{\alpha}$ and variance $E(\sum_i \mathbf{X_i'}\mathbf{\Omega_i^{-1}}\mathbf{X_i})^{-1}/n$ (see, for example, Wooldridge 2002). This variance can be estimated as follows:

$$(17) \quad Asy\hat{V}ar(\hat{\boldsymbol{\alpha}}_{\mathbf{GLS}}) = (\textstyle\sum_i \mathbf{X_i'}\hat{\mathbf{\Omega}}_{\mathbf{i}}^{\mathbf{-1}}\mathbf{X_i})^{-1},$$

which in our case reduces to

$$(18) \quad Asy\hat{V}ar(\hat{\alpha}_{1,GLS}) = \left[(1-p)^2 \textstyle\sum_{i:T_i=1}^{np} \frac{1}{\hat{\sigma}_u^2 + (\hat{\sigma}_e^2 / m_i)} + p^2 \textstyle\sum_{i:T_i=1}^{n(1-p)} \frac{1}{\hat{\sigma}_u^2 + (\hat{\sigma}_e^2 / m_i)}\right]^{-1}.$$

For known $\mathbf{\Omega_i}$, the GLS estimator is the best linear unbiased estimator (BLUE) (although this may not hold if $\mathbf{\Omega_i}$ is replaced by $\hat{\mathbf{\Omega}}_\mathbf{i}$). The ANOVA, ML, REML, and GEE approaches discussed in Chapter 5 yield feasible GLS estimators where estimators for σ_u^2 and σ_e^2 are inserted into (16) and (17).

For a given sample size, the variance in (18) is minimized when $m_i = m$ and $p = 0.5$. Furthermore, if $m_i = m$, the OLS and GLS estimators of α_1 are identical and yield the following simple variance estimator:

$$(19) \quad Asy\hat{V}ar(\hat{\alpha}_{1,Balanced}) = \frac{1}{p(1-p)}\left(\frac{\hat{\sigma}_u^2}{n} + \frac{\hat{\sigma}_e^2}{nm}\right).$$

Note that replacing m by $\bar{m}$ in (19) is a serviceable variance estimator for designs where sample sizes vary somewhat across schools, which can be seen by setting $m_i = \bar{m}$ in (18).

Super-Population Model With Covariates

Under the SP model with covariates, the covariates $\mathbf{q_{ij}}$ as well as the potential outcomes are considered to be random draws from joint super-population distributions. For the estimation model, the covariate matrix is now $\mathbf{X_i} = [\mathbf{K}\ \mathbf{T_i}\ \mathbf{Q_i}]$ and $\mathbf{\Omega_i}$ is now *conditional* on $\mathbf{Q_i}$. In principle, the covariates should be considered irrelevant variables because (14) is the true model. This procedure, however, considerably complicates the asymptotics for the GLS estimator, because $\mathbf{Q_i}$ will tend to be correlated with the error term, and $\mathbf{\Omega_i}$ will differ from the true $\mathbf{\Omega_i^*}$.[4]

Consequently, the following analysis strays somewhat from the Neyman framework and assumes that the true model contains $\mathbf{Q_i}$. In this case, the GLS formulas in (16) and (17) also apply to the SP model with covariates.

[4] For the OLS estimator, the first problem can be overcome (as it was for the FP model) and the second problem does not occur. The asymptotic variance of the OLS estimator is similar in form to that for the FP model in (12) but does not include terms comparable to $\overline{S}_\tau^2$ (not shown).

Chapter 5: Variance Component Estimation for the Super-Population Model

Feasible GLS estimation requires estimates of the variance components σ_u^2 and σ_e^2. This chapter discusses key features of ANOVA, ML, REML, and GEE estimation methods that can be used to estimate these variance components and that are used in the empirical analysis. To keep the presentation manageable, the discussion does not focus on other methods, such as bootstrap, jackknife, and other resampling methods. De Leeuw and Meijer (2008) provide an excellent, more detailed discussion of GLS estimators for multilevel models.

For simplicity of exposition, in what follows, let the symbol $\mathbf{\Omega_i}$ represent a generic covariance matrix for school *i*, $\mathbf{X_i}$ represent a generic covariate matrix for a school, and $\delta_{ij} = u_i + e_{ij}$ represent a generic normally distributed error for a student.

Balanced Design Estimator

When $m_i = m$ for all schools—that is, for *balanced* designs—a consistent variance estimator for the simple differences-in-means estimator has the following simple form:

$$(20) \quad Asy\hat{V}ar(\hat{\alpha}_{1,Balanced}) = \frac{S_B^2}{np(1-p)},$$

where

$$S_B^2 = \frac{\sum_{i:T_i=1}^{np} (\bar{y}_i - \bar{y}_T)^2 + \sum_{i:T_i=0}^{n(1-p)} (\bar{y}_i - \bar{y}_C)^2}{n-2}$$

is the variance of the mean outcome *between* schools (see, for example, Cochrane 1963). This estimator is consistent because $E(S_B^2) = \sigma_u^2 + (\sigma_e^2 / m)$ (see (19)).

If covariates are included in the model, (20) can be generalized using the following *between-school* regression model:

$$(21) \quad \bar{y}_i = \alpha_0 + \alpha_1 T_i + \mathbf{\bar{q}_i'\alpha_2} + \delta_i,$$

where $\bar{y}_i$ is the school-level mean, $\mathbf{\bar{q}_i}$ is a $k_1 x1$ vector of school-level covariates (that could include *student*-level covariates averaged to the school level), and $\delta_i = (u_i + \bar{e}_i)$ is the school-level error term. Estimating (21) by OLS yields the following variance estimator for α_1:

$$(22) \quad Asy\hat{V}ar(\hat{\alpha}_{1,Balanced}) = (\mathbf{\bar{X}'\bar{X}})_{2,2}^{-1} RSS_B / (n - k_1 - 2),$$

where $\bar{\mathbf{X}} = [\mathbf{K}\,\mathbf{T}\,\bar{\mathbf{Q}}]$ is the covariate matrix and RSS_B is the regression residual sum of squares.

For balanced designs, $\hat{\alpha}_{1,Balanced}$ is an ANOVA or REML estimator and is minimum variance unbiased under normality of the error terms (Searle 1971). This estimator, however, has no optimal properties for unbalanced designs. Nonetheless, it is appealing due to its simplicity, because it is based entirely on the between-school OLS regression, and produces serviceable estimates for designs that are not too highly unbalanced (which is typically the case in practice). As discussed next, estimating variance components to account for unbalanced designs becomes considerably more complex.

ANOVA Estimator

The ANOVA estimator is a method-of-moments estimator that equates regression residual sums of squares to their unobserved expectations and solves these equations to obtain estimators for the variance components. ANOVA methods have the advantage that the variance components can be obtained in one step using easily-understood OLS regression residuals, rather than iteratively, as is the case for the ML, REML, and GEE methods. The disadvantage of the ANOVA methods is that for unbalanced designs, they have no optimal properties beyond asymptotic unbiasedness.

This section discusses the Swamy and Arora (SA; 1972) ANOVA method that was adapted for unbalanced designs by Baltagi and Chang (1994). De Leeuw and Meijer (2008) and Baltagi and Chang (1994) discuss alternative ANOVA estimators that are similar to the SA method.

Under the SA method, an estimator for the student-level variance, σ_e^2, is obtained by first estimating a *within-school* OLS regression:

$$(23) \quad (y_{ij} - \bar{y}_i) = (\mathbf{q_{ij}} - \bar{\mathbf{q}}_\mathbf{i})\gamma + (e_{ij} - \bar{e}_i).$$

This yields the following consistent variance estimator for σ_e^2:

$$(24) \quad \hat{\sigma}^2_{e,ANOVA} = RSS_W / (\textstyle\sum_i^n m_i - n - k_2),$$

where RSS_W is the regression residual sum of squares from (23) and k_2 is the number of student-level (within school) covariates.

To obtain an estimator for σ_u^2, the SA method uses the residual sum of squares RSS_B from the *between-school* regression in (21) where schools are weighted by their sample sizes. In this case, $RSS_B = \hat{\boldsymbol{\delta}}'\mathbf{W}\hat{\boldsymbol{\delta}} = \boldsymbol{\delta}'\mathbf{B}'\mathbf{W}\mathbf{B}\boldsymbol{\delta}$, where $\mathbf{W}$ is an nxn diagonal weight matrix with weights m_i along the diagonal, and $\mathbf{B} = [\mathbf{I_M} - \bar{\mathbf{X}}(\bar{\mathbf{X}}'\mathbf{W}\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}'\mathbf{W}]$. Using matrix algebra, it can be shown that:

$$\begin{aligned} E(RSS_B) &= E(tr[\boldsymbol{\delta}'\mathbf{B}'\mathbf{W}\mathbf{B}\boldsymbol{\delta}]) = \sigma_u^2 tr[\mathbf{B}'\mathbf{W}\mathbf{B}] + \sigma_e^2 tr[\mathbf{B}'\mathbf{W}\mathbf{B}\mathbf{W}^{-1}] \\ &= \sigma_u^2(\textstyle\sum_{i=1}^n m_i - tr[(\bar{\mathbf{X}}'\mathbf{W}\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}'\mathbf{W}\mathbf{W}'\bar{\mathbf{X}}]) + \sigma_e^2(n - k_1 - 2), \end{aligned}$$

where *tr* is the matrix trace operator. Thus, a consistent estimator for σ_u^2 is:

$$(25)\quad \hat{\sigma}^2_{u,ANOVA} = [RSS_B - \hat{\sigma}^2_{e,ANOVA}(n-k_1-2)]/(\textstyle\sum_{i=1}^{n} m_i - tr[(\mathbf{\bar{X}'W\bar{X}})^{\mathbf{-1}}\mathbf{\bar{X}'WW'\bar{X}}]),$$

which could be negative (leading to a negative *ICC* estimate).

The estimators for σ_e^2 and σ_u^2 in (24) and (25) can be inserted into (16) and (17) to obtain feasible GLS estimates. Note that for balanced designs, this approach yields (22). The ANOVA approach can be implemented using SAS (see Table 3.1).

Maximum Likelihood Estimator

ML methods simultaneously estimate ATE parameters and variance components, and are often used to estimate linear mixed models (such as HLM models) that are popular in the education field (see, for example, Raudenbush and Bryk 2002). ML estimators are consistent and asymptotically efficient, but do not take into account the loss in degrees of freedom due to the regression coefficients in estimating the variance components.

To demonstrate the ML method, it is convenient to express $\mathbf{\Omega_i}$ as $\sigma_e^2\mathbf{\Lambda_i}$, where

$$(26)\quad \mathbf{\Lambda_i} = \mathbf{I_{m_i}} + \lambda\mathbf{J_{m_i}},$$

where $\mathbf{I_{m_i}}$ is the identity matrix, $\mathbf{J_{m_i}}$ is an $m_i x m_i$ matrix of 1s, and $\lambda = \sigma_u^2/\sigma_e^2$. Note that $\mathbf{\Lambda_i^{-1}} = \mathbf{I_{m_i}} - (m_i + \lambda^{-1})^{-1}\mathbf{J_{m_i}}$. Because of the normality assumption, the log likelihood is:

$$(27)\quad \log L = -\frac{M}{2}\log(2\pi) - \frac{M}{2}\log\sigma_e^2 - \frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{\Lambda_i}| - \frac{1}{2\sigma_e^2}[\sum_{i=1}^{n}(\mathbf{y_i} - \mathbf{X_i\alpha})'\mathbf{\Lambda_i^{-1}}(\mathbf{y_i} - \mathbf{X_i\alpha})],$$

where $|\mathbf{\Lambda_i}|$ denotes the determinant of $\mathbf{\Lambda_i}$.

Taking derivatives in (27) with respect to the parameters and setting them equal to zero yields the following closed-form solutions for $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}_e^2$ (for a given $\hat{\lambda}$):

$$(28)\quad \hat{\boldsymbol{\alpha}}_{\mathbf{MLE}} = (\textstyle\sum_{i=1}^{n}\mathbf{X_i'\hat{\Lambda}_i^{-1}X_i})^{-1}(\textstyle\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{\hat{\Lambda}_i^{-1}y_i}),\ \text{and}$$

$$(29)\quad \hat{\sigma}^2_{e,MLE} = \frac{1}{M}[\sum_{i=1}^{n}(\mathbf{y_i} - \mathbf{X_i}\hat{\boldsymbol{\alpha}})'\mathbf{\hat{\Lambda}_i^{-1}}(\mathbf{y_i} - \mathbf{X_i}\hat{\boldsymbol{\alpha}})].$$

Equation (28) is the feasible GLS estimator in (16).

The first-order condition for λ is a nonlinear equation that must be solved numerically:

$$(30)\quad 0 = grad = \frac{\partial\log L}{\partial\lambda} = -\frac{1}{2}\sum_{i=1}^{n}tr[\mathbf{\Lambda_i^{-1}J_{m_i}}] + \frac{1}{2\sigma_e^2}[\sum_{i=1}^{n}(\mathbf{y_i} - \mathbf{X_i\alpha})'[\mathbf{\Lambda_i^{-1}J_{m_i}\Lambda_i^{-1}}](\mathbf{y_i} - \mathbf{X_i\alpha})].$$

One common iterative method is the Newton-Raphson method where $\hat{\lambda}^{(iter+1)}$ is updated from $\hat{\lambda}^{(iter)}$ as follows:

$$(31)\quad \hat{\lambda}^{(iter+1)} = \hat{\lambda}^{(iter+1)} - H^{-1(iter)} grad^{(iter)},$$

where

$$(32)\quad H = \frac{\partial^2 \log L}{\partial\lambda\partial\lambda} = \frac{1}{2}\sum_{i=1}^{n} tr[\boldsymbol{\Lambda}_{\mathbf{i}}^{\mathbf{-1}}\mathbf{J}_{\mathbf{m_i}}\boldsymbol{\Lambda}_{\mathbf{i}}^{\mathbf{-1}}\mathbf{J}_{\mathbf{m_i}}] - \frac{1}{\sigma_e^2}[\sum_{i=1}^{n}(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\alpha})'\boldsymbol{\Lambda}_{\mathbf{i}}^{\mathbf{-1}}\mathbf{J}_{\mathbf{m_i}}\boldsymbol{\Lambda}_{\mathbf{i}}^{\mathbf{-1}}\mathbf{J}_{\mathbf{m_i}}\boldsymbol{\Lambda}_{\mathbf{i}}^{\mathbf{-1}}(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\alpha})]$$

is the Hessian matrix.[5] Other iterative methods use $-E(H) = .5\sum_i tr[\boldsymbol{\Lambda}_{\mathbf{i}}^{\mathbf{-1}}\mathbf{J}_{\mathbf{m_i}}\boldsymbol{\Lambda}_{\mathbf{i}}^{\mathbf{-1}}\mathbf{J}_{\mathbf{m_i}}]$ in (31) rather than H (Fisher scoring) or the expectation-maximization (EM) algorithm (see Little and Rubin 2002).

The model parameters can then be estimated using the following steps: (1) obtain an initial value for $\hat{\lambda}$ (for example, using the ANOVA method), (2) calculate $\hat{\boldsymbol{\alpha}}_{\mathbf{MLE}}$ and $\hat{\sigma}^2_{e,MLE}$ using (28) and (29), (3) update $\hat{\lambda}$ using (30)-(32), and (4) return to Step (2) until convergence is achieved. Final feasible GLS estimators can then be obtained using (16) and (17). Table 3.1 displays statistical package routines that use the ML method.

Note that most statistical packages impose a non-negativity constraint for $\hat{\lambda}$ at each iteration. Murray (1998) and Stroup and Littell (2002) demonstrate through simulations, however, that this constraint could deflate the Type I error rate and reduce statistical power. Thus, these authors recommend that options be used in statistical packages that allow for negative variance component estimates. A similar issue applies to the REML estimator discussed next.

Restricted Maximum Likelihood Estimator

Unlike the ML approach, the REML approach for the variance components adjusts for the degrees of freedom loss due to the estimation of the regression parameters (Patterson and Thompson 1971). The REML approach separates the likelihood into two independent parts, one of which depends only on the variance components (the part of interest). The approach profiles out the covariates by finding a linear combination of the outcomes, $\mathbf{y}^* = \mathbf{Ly}$, whose distribution does not depend on $\boldsymbol{\alpha}$, where $\mathbf{L}$ is a $(M-k)xM$ matrix and k is the rank of the covariate matrix $\mathbf{X}$.[6]

To find $\mathbf{L}$, consider first the OLS regression residuals $\mathbf{r} = \mathbf{Py}$, where $\mathbf{P} = (\mathbf{I_M} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{\mathbf{-1}}\mathbf{X}')$ is an MxM idempotent matrix. Note that because $\mathbf{y}$ is assumed to have a multivariate normal distribution,

[5] The formulas in (30) and (32) can be obtained using (26) and the matrix identities: $\partial|\boldsymbol{\Lambda}_{\mathrm{i}}|/\partial\lambda =|\boldsymbol{\Lambda}_{\mathrm{i}}^{-1}|\,tr(\boldsymbol{\Lambda}_{\mathrm{i}}^{-1}\partial\boldsymbol{\Lambda}_{\mathrm{i}}/\partial\lambda)$ and $\partial\boldsymbol{\Lambda}_{\mathbf{i}}^{\mathbf{-1}}/\partial\lambda = -\boldsymbol{\Lambda}_{\mathrm{i}}^{-1}(\partial\boldsymbol{\Lambda}_{\mathrm{i}}/\partial\lambda)\boldsymbol{\Lambda}_{\mathrm{i}}^{-1}$.

[6] The REML approach does not depend on the specific choice of $\mathbf{L}$, so one choice is derived.

$\mathbf{r} \sim \mathbf{N}(\mathbf{0}, \mathbf{P\Omega P'})$, which is independent of $\boldsymbol{\alpha}$. Thus, a solution for $\mathbf{L}$ (which has $(M-k)$ rows) can be obtained from $\mathbf{P}$ by satisfying the relation $\mathbf{P} = \mathbf{L'L}$ subject to the normalizing condition $\mathbf{I_{M\text{-}k}} = \mathbf{LL'}$. Such a solution can be found using the eigenvectors $(\mathbf{E})$ and eigenvalues of $\mathbf{P}$. Note that because $\mathbf{P}$ is idempotent, $(M-k)$ eigenvalues will be one (say, the first ones) and the remaining k will be zero. Thus, $\mathbf{L}$ can be calculated as the first $(M-k)$ rows of $\mathbf{E'}$.

Using this $\mathbf{L}$, it follows that $\mathbf{y}^* = \mathbf{Ly} \sim \mathbf{N}(\mathbf{0}, \mathbf{L\Omega L'})$, whose distribution is independent of $\boldsymbol{\alpha}$. The REML log likelihood can be obtained using this distribution. Harville (1974) shows, however, that an equivalent log likelihood that shows more clearly the way that the regression parameters are profiled out of the likelihood can be expressed as follows:

$$(33) \quad \log L = -\frac{M-k}{2}\log(2\pi) - \frac{M-k}{2}\log(\sigma_e^2) - \frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{\Lambda_i}| - \frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{X_i}'\mathbf{\Lambda_i^{-1}X_i}|$$
$$+\frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{X_i}'\mathbf{X_i}| - \frac{1}{2\sigma_e^2}[\sum_{i=1}^{n}(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\alpha}_{\mathbf{GLS}})'\mathbf{\Lambda_i^{-1}}(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\alpha}_{\mathbf{GLS}})],$$

where $\boldsymbol{\alpha}_{\mathbf{GLS}}$ is given in (16). This likelihood can be maximized with respect to σ_e^2 and λ using the methods discussed above for the ML estimator (not shown). REML estimates of the variance components can then be used in (16) and (17) to obtain feasible GLS estimators. REML estimates are asymptotically equivalent to ML estimates, but the REML approach tends to produce larger standard errors due to the degrees of freedom adjustments. Table 3.1 displays statistical package routines that use the REML method.

GEE Estimator

The GEE estimator (discussed above) is also a feasible GLS estimator for the SP model. The ATE parameter estimate obtained from (7) yields the feasible GLS estimator in (16), and the model-based GEE variance estimator using (8) yields the feasible GLS variance estimator in (17). The GEE empirical sandwich variance estimator is $\mathbf{I_0^{-1}I_1I_0^{-1}}$ where

$$\mathbf{I_1} = \sum_{i=1}^{n}\mathbf{X_i'\Omega_i^{-1}r_ir_i'\Omega_i^{-1}X_i}.$$

Under the GEE method, the variance components are estimated iteratively by updating regression residuals r_{ij}. In Step (1), OLS residuals are used to obtain the following consistent estimates for σ_u^2 and σ_e^2:

$$(34) \quad \hat{\sigma}_{u,GEE}^2 = \hat{\rho}_{GEE}s_{GEE}^2 \text{ and } \hat{\sigma}_{e,GEE}^2 = (1-\hat{\rho}_{GEE})s_{GEE}^2, \text{ where}$$

$$(35) \quad s_{GEE}^2 = \frac{1}{M-k}\sum_{i=1}^{n}\sum_{j=1}^{m_i}r_{ij}^2 \text{ and } \hat{\rho}_{GEE}^2 = \frac{1}{\sum_i m_i(m_i-1)-2k}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k\neq j}^{m_i}(r_{ij}r_{ik}/s_{GEE}^2).$$

These estimates are then used to calculate $\hat{\boldsymbol{\alpha}}_{\mathbf{GLS}}$ in (16). In Steps (2) to (4), new residuals are calculated, (34) and (35) are updated, and new estimates of $\hat{\boldsymbol{\alpha}}_{\mathbf{GLS}}$ are obtained. Steps (2) to (4) are then continued until convergence is achieved. Table 3.1 displays statistical package routines that use this method that require the specification of an *exchangeable* working correlation matrix.

Chapter 6: Empirical Analysis

This chapter presents ATE estimates and their standard errors using five published large-scale RCTs that were funded by the Institute of Education Sciences (IES) at the U.S. Department of Education (ED) and several foundations. These RCTs tested the effects of a wide range of education interventions, including mentoring programs for new teachers (Glazerman et al. 2008), early elementary school math curricula (Agodini et al. 2009), the use of selected computer software in the classroom (Dynarski et al. 2007), selected reading comprehension interventions (James-Burdumy et al. 2009), and Teach for America (Decker et al. 2004). Across the RCTs, random assignment was conducted at either the school or teacher (classroom) level primarily in low-performing school districts, and the key outcome measures were math or reading test scores of elementary school students. Appendix Table B.1 provides information for each study.

All studies (except for the Reading Comprehension study) report impact findings using a SP framework (using HLM models with baseline covariates), although it cannot be determined which specific estimation and optimization methods were used for the analyses. This chapter discusses findings from a re-analysis of the RCT data using the estimation methods considered above for the FP and SP models. The focus is on models that include baseline covariates. Using study documentation, the choice of baseline covariates (including blocking indicators), the construction of the outcome measures, and the treatment of missing data were as similar as possible to those used by the authors of the study reports. For comparable models, the impact results reported below are similar to those presented in the published reports.

SAS was used to estimate the GEE, balanced design, REML, and ML models, because research has shown that the statistical packages considered in this paper yield similar estimates for common model specifications and optimization routines (West et al. 2007, Shah 1998). To keep the presentation manageable, the ML and REML estimates were obtained using the Newton-Raphson algorithm. The SAS code that was used to estimate the models is displayed in the footnotes to Table 6.2 below. The permutation tests were conducted using SAS programs written by the author, where permutation distributions were estimated from 10,000 reallocations of cluster means to the pseudo-treatment and control groups (because the number of possible allocations was too large to delineate for these studies). The ANOVA estimates were also obtained using SAS programs written by the author.[7]

In what follows, information is first presented for each study on cluster-level sample sizes and weights for the FP and SP models. This information is helpful for interpreting the impact findings, which are presented second.

Weights for the Finite-Population and Super-Population Models

As discussed, a key difference between the FP and SP models involves how clusters (schools or classrooms) are weighted in the analysis. In the FP models, clusters are either weighted by their sample sizes or equally, whereas in the SP model, clusters are weighted by the inverses of their variances. The extent to which ATE results differ across the weighting schemes will depend on the variability of cluster sample sizes, ICC values for the outcome variables, and the relationship between cluster-level impacts and cluster sample sizes.

[7] The *Proc Panel* procedure in SAS does not perform the SA ANOVA method that was discussed above, but uses variants of this procedure (which produce results consistent to those presented in this paper).

The top panel of Table 6.1 shows that cluster sample sizes vary for all five studies, but more so for some studies than others. For example, the interquartile range of cluster sizes is about 7 students for the classroom-based Teach for America and Educational Technologies studies, but is 30 students for the school-based Reading Comprehension study. The finding that cluster sample sizes vary within each study suggests that cluster-level weights always differ for the two FP models. There are also differences across the studies in ICC values (Table 6.1). These intraclass correlations range from 0.06 to 0.12 for models that include baseline covariates and from 0.13 to 0.29 for models that exclude covariates.

Finally, the variability of the weights for the SP models lies between the variability of the weights for the two FP models (Table 6.1). For instance, for the Math Curriculum study, the interquartile range for the SP weights for the REML model is 4 (bottom panel of Table 6.1), compared to 14 for the FP model where clusters are weighted by their sample sizes (top panel of Table 6.1), and 0 for the FP model where clusters are weighted equally.

Impact Findings

For all studies, the considered FP and SP estimators yield consistent findings concerning the statistical significance of the ATE estimates (Table 6.2). The estimators show that (1) elementary school students taught by Teach for America teachers performed significantly better on math achievement tests than those taught by traditional teachers, (2) the use of selected software products in the classroom did not improve first graders' math test scores, (3) the offer of teacher induction programs for beginning teachers did not improve math test scores for second to sixth grade students, (4) the Saxon or Math Expressions math curriculum produced significantly higher fifth grade student math test scores than the other tested math curricula, and (5) the Reading for Knowledge reading curriculum produced significantly lower fifth grade student reading scores than the control (status quo) reading curriculum offered in the study schools.

For each study, the ATE impact estimates vary by less than 0.02 or 0.03 standard deviations across the eight estimators (Table 6.2). For example, the impact estimates in effect size units range from 0.261 to 0.273 for the Math Curriculum study, from 0.126 to 0.129 for the Teach for America study, and from -0.147 to -0.159 for the Reading Comprehension study.

The estimated standard errors (and *p*-values), however, range somewhat more across the eight estimators than the ATE point estimates (Table 6.2). For example, standard errors range from 0.038 to 0.075 for the Reading Comprehension study, 0.035 to 0.050 for the Teacher Induction study, and 0.478 to 0.766 for the Educational Technologies study. The finding that the various consistent estimators yield more variable estimates of standard errors than regression coefficients is a pattern that has often been found in the literature for observational studies.

Findings for the SP Estimators. On the basis of the empirical findings and the theory from above, the SP estimators can be divided into two main groups. The first group includes the ANOVA and REML estimators that both account for the loss in degrees of freedom in the variance estimates due to the regression parameters. Across the five studies, these two estimators yield identical ATE impact estimates, and standard errors that differ by at most .003 standard deviations (Table 6.2). The similarity of the ANOVA and REML findings is consistent with Baltagi and Chang (1994), who found using simulations that the ANOVA method performs well for random effects models. Thus, there is reason for education researchers to consider using the ANOVA estimator more often in RCTs.

The second group of SP estimators includes the model-based and empirical sandwich GEE estimators and the ML estimator. Across the five studies, these three estimators yield ATE impact estimates that differ from each other by less than .002 standard deviations, and standard errors that typically differ from each

Table 6.1: Information on Weighting Schemes for the FP and SP Models, by Study

Statistic	Teach for America	Educational Technologies	Teacher Induction	Math Curriculum	Reading Comprehension
Distribution of Cluster Sizes (Percentiles)					
10th	11	10	9	21	28
25th	14	13	13	27	39
50th	17	16	20	31	57
75th	21	19	29	41	69
90th	23	22	47	49	99
ICCs for the SP REML Model					
No covariates	0.29	0.20	0.14	0.19	0.13
Covariates	0.08	0.12	0.07	0.06	0.06
Distribution of Cluster-Level Weights for the SP REML Model With Covariates (Percentiles)[a]					
10th	14	12	17	29	49
25th	16	13	21	32	54
50th	17	14	26	33	60
75th	19	15	29	36	62
90th	20	16	33	38	66
Sample Sizes					
Clusters (*p=% treatment*)	**95 (0.44)**	**137 (0.57)**	**173 (0.52)**	**39 (0.46)**	**39 (0.46)**
Students	**1,630**	**2,176**	**4,381**	**1,309**	**2,256**

Source: Data from studies listed in Appendix Table B.1.

[a]The weights sum to the total student sample size.

Table 6.2: Regression-Adjusted Impact Results, by Study

Model and Estimator	Teach for America	Educational Technologies	Teacher Induction	Math Curriculum	Reading Comprehension
			Finite Population Model		
1. GEE (Empirical)					
a. Clusters Weighted by Sample Sizes	.126 (.048) (.008)*	.032 (.046) (.478)	-.022 (.037) (.548)	.261 (.059) (.000)*	-.147 (.052) (.005)*
b. Clusters Weighted Equally Permutation Tests	.126 (.047) (.007)* (.005)*	.014 (.046) (.766) (.738)	.013 (.045) (.782) (.737)	.273 (.061) (.000)* (.000)*	-.159 (.059) (.007)* (.000)*
			Super-Population Model		
2. Balanced Design	.126 (.055) (.025)*	.014 (.044) (.759)	.013 (.050) (.802)	.273 (.068) (.000)*	-.159 (.075) (.051)
3. ANOVA	.129 (.055) (.023)*	.019 (.045) (.663)	-.001 (.043) (.976)	.269 (.066) (.000)*	-.159 (.069) (.038)*
4. ML	.128 (.048) (.007)*	.020 (.042) (.637)	-.005 (.036) (.888)	.268 (.057) (.000)*	-.153 (.039) (.000)*
5. REML	.129 (.055) (.020)*	.019 (.044) (.661)	-.001 (.043) (.981)	.269 (.067) (.000)*	-.159 (.072) (.027)*
6. GEE					
a. Model-Based	.128 (.049) (.008)*	.020 (.043) (.648)	-.006 (.035) (.859)	.268 (.055) (.000)*	-.151 (.038) (.000)*
b. Empirical	.128 (.047) (.007)*	.020 (.045) (.661)	-.006 (.040) (.874)	.268 (.060) (.000)*	-.151 (.053) (.004)*

Source: Data from studies listed in Appendix Table B.1. See Table 6.1 for sample sizes.

Notes: From left to right, the figures in cells are the ATE impact estimates, estimated standard errors, *p*-values, and *p*-values for the permutation tests for Model 1b. Impact estimates are regression-adjusted using the covariates indicated in Appendix Table 1.

SAS routines were used to estimate the models except for the ANOVA and permutation tests which were performed using SAS programs written by the author. Let CLUS denote the cluster codes, T the treatment dummy, Y the outcome, YC the cluster-level mean outcome, X the list of covariates centered at their cluster-level means, XC the cluster-level mean covariates, and D the input dataset. The following code was then used to estimate the models:

Models 1a and 1b: proc genmod data=D; class CLUS; model Y=T X XC / dist=normal;
repeated subject = CLUS / type = ind; (A weight statement was used for Model 1b to weight clusters equally)
Model 2: proc reg data=D; model YC = T XC;
Model 4: proc mixed data=D method=ml; class CLUS; model Y=T X XC/solution; random CLUS;
Model 5: proc mixed data=D; class CLUS; model Y=T X XC/solution; random CLUS;
Models 6a and 6b: proc genmod data=D; class CLUS; model Y=T X XC/ dist=normal;
repeated subject = CLUS / type = exch models;

*The ATE impact estimate is significantly different from zero at the 0.05 level, two-tailed test.

other by less than .005 standard deviations (Table 6.2). The similarity of estimates for the two GEE estimators suggests that the exchangeable error structure is appropriate for the data. The GEE and ML methods produce smaller standard errors than the REML and ANOVA methods (Table 6.2). This finding is expected for the ML method, which does not adjust for the degrees of freedom loss due to the estimation of the regression parameters.

Finally, the simple balanced design method produces impact and standard error estimates that are consistent with those from the other SP methods, even though this estimator does not account for unbalanced cluster sizes (Table 6.2). Thus, there is good reason to use this simple between-cluster estimator to check the robustness of study findings obtained using the other more complex methods.

Findings for the FP Estimators. Empirical results for the two FP models are displayed in the top panels of Table 6.2 and labeled as "Model 1a" and "Model 1b." Differences in the ATE impact estimates for these two FP models range from 0 to .035 standard deviations across the studies, because of differences in weighting schemes. The differences are most pronounced for the Educational Technologies and Teacher Induction studies where the estimated impacts are not statistically significant.

The ATE point estimates for the FP and SP models typically differ by less .005 standard deviations for the three studies with statistically significant impact estimates (the Teach for America, Math Curriculum, and Reading Comprehension studies; Table 6.2). Furthermore, across all five studies, the standard error estimates for the FP models are similar to each other and to those for the empirical sandwich GEE estimator for the SP model (Table 6.2); the pairwise differences in standard errors are all less than .007 standard deviations. However, as discussed, the standard error estimates for the FP models are *conservative*, because they ignore precision gains from the difficult-to-estimate $\overline{S}_{\tau}^{2}$ terms in (6) and (12).

Finally, for FP Model 1b, the permutation and parametric hypothesis tests yield similar *p*-values (Table 6.2). For example, the respective *p*-values are .005 and .008 for the Teach for America study, .766 and .738 for the Educational Technologies study, and .000 and .007 for the Reading Comprehension study. Thus, the normality assumption underlying the parametric tests appears to be validated using the nonparametric methods, which are much more computationally burdensome.

Chapter 7: Summary and Conclusions

This paper has examined the estimation of two-stage clustered RCT designs in education research using the Neyman causal inference framework that underlies experiments. The key distinction between the considered causal models is whether potential treatment and control group outcomes are considered to be fixed for the study population (the FP model) or randomly selected from a vaguely-defined super-population (the SP model).

In the FP model, the only source of randomness is treatment status, and a clustered design results only because students in the same cluster share the same treatment status. The relevant impact parameter for this model is the average treatment effect for those in the study sample; thus, the impact results are internally valid only. The asymptotic variance for the FP model (that was derived in this paper) can be estimated using a GEE estimator assuming an independent working correlation structure. Two weighting options for this model are (1) to weight each student equally (the OLS approach) or (2) to weight each cluster equally (to estimate ATEs for the average cluster in the sample). The FP variance estimators are likely to be conservative, however, because they ignore precision gains from difficult-to-estimate variance terms that represent the extent to which treatment effects vary and co-vary across students in the same cluster. Thus, in theory, the FP estimators could yield more precise ATE estimates than the SP estimators, but it is difficult to realize these precision gains in practice.

In the SP model, cluster- and student-level potential outcomes are considered to be randomly sampled from respective super-population distributions. In this framework, the relevant ATE parameter is the intervention effect for the average cluster in the super-population. Thus, impact findings are assumed to generalize outside the study sample, although it is often difficult to precisely define the study universe. For estimating the SP model, the paper discussed key features of several feasible GLS estimators (ML, REML, ANOVA, and GEE estimators) assuming an exchangeable random effects error structure. For these estimators, clusters are weighted by the inverses of their variances, and the variability of these weights lies between the variability of the weights under the two FP weighting schemes.

Using data from five recent large-scale clustered RCTs in the education area, the empirical analysis estimated ATEs and their standard errors using the considered estimators. For all five studies, the considered estimators yield consistent findings concerning statistical significance. However, although the estimated impacts are similar across the estimators, the standard errors (and hence, *p*-values) differ more across the estimators. This suggests that in particular studies, policy conclusions could differ using the various estimators.

The choice of the primary estimation method and cluster-level weighting scheme should best fit evaluation research questions and objectives, and should be specified and justified in the analysis protocols. However, there might not always be a scientific basis for making these benchmark choices (that is, there might not be a "true" underlying statistical model for the study). Thus, a key recommendation from this paper is that education researchers consider testing the sensitivity of their benchmark impact findings using alternative estimation methods, rather than relying solely on the methods with which they are most comfortable. These sensitivity analyses could be important for ruling out the possibility that the impact findings are driven by specific distributional assumptions about the data and asymptotic results. Furthermore, it is recommended that findings from sensitivity analyses be reported in study appendixes, that attempts be made to explain discrepancies between sensitivity and benchmark analysis findings, and that the robustness of results be reflected in the study conclusions.

Researchers currently most often report impact findings using the SP framework based on REML or ML methods. Results in this paper suggest that, in the sensitivity analysis, impact estimates could also be estimated using other methods such as the balanced-design, GEE, and FP estimators. The ANOVA method is another approach that could be used more often in education research.

Finally, the choice of whether to adopt the FP or the SP framework is a difficult philosophical issue. In practice, the two methods will tend to blur, however, because standard estimation procedures do not account for precision gains from the FP model, and the empirical results presented in this paper suggest that the FP and SP models yield similar impact findings. Furthermore, the two approaches blur under balanced designs. Nonetheless, researchers should understand the assumptions underlying the SP and FP approaches and their implications for generalizing and interpreting the impact findings.

Appendix A: Proofs

Proof of Lemma 1

Applying standard OLS methods to (3) yields $\hat{\beta}_{1,SR} = \bar{y}_T - \bar{y}_C$. To calculate the asymptotic moments of $\hat{\beta}_1$, we express $\hat{\beta}_1$ as follows:

$$\text{(A.1)} \quad \hat{\beta}_{1,SR} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(T_i - p)y_{ij}}{d} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(T_i - p)[\beta_0 + \beta_1(T_i - p) + \eta_{ij}]}{d} = \beta_1 + \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(T_i - p)\eta_{ij}}{d},$$

where $d = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(T_i - p)^2$. Substituting for η_{ij} using (3) yields:

$$\text{(A.2)} \quad (\hat{\beta}_{1,SR} - \beta_1) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}[\alpha_{ij}(T_i - p) + \tau_{ij}(T_i - p)^2]}{d} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}[\alpha_{ij} + (1-2p)\tau_{ij}]T_i}{d}$$

$$= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} l_{ij}T_i}{d}; \quad l_{ij} = (1-p)(Y_{Tij} - \bar{Y}_T) + p(Y_{Cij} - \bar{Y}_C).$$

Note that $E(d) = n\bar{\bar{m}}p(1-p)$. Thus, $(\hat{\beta}_{1,SR} - \beta_1) \to \sum_{i=1}^{n}\sum_{j=1}^{m_i} l_{ij}p / n\bar{\bar{m}}p(1-p) = 0$ because $\sum_i\sum_j l_{ij} = 0.$ Thus, $\hat{\beta}_{1,SR}$ is asymptotically unbiased.

Using (A.2), the variance of $\hat{\beta}_{1,SR}$ is:

$$Var(\hat{\beta}_{1,SR}) = \frac{Var(\sum_{i=1}^{n}\sum_{j=1}^{m_i} l_{ij}T_i)}{d^2} = \frac{p(1-p)(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i} l_{ij}l_{ik} - \frac{1}{(n-1)}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{j=1}^{m_i}\sum_{j'=1}^{m_i} l_{ij}l_{i'j'})}{d^2},$$

where the last equality holds because $Var(T_i) = p(1-p)$ and $Cov(T_i, T_{i'}) = -p(1-p)/(n-1)$. Because $\sum_i\sum_j l_{ij} = 0$, it follows that $(\sum_i\sum_j l_{ij})^2 = 0$. Hence,

$$\text{(A.3)} \quad Var(\hat{\beta}_{1,SR}) = \frac{n}{n-1}\frac{p(1-p)\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i} l_{ij}l_{ik}}{d^2} \to \frac{1}{n\bar{\bar{m}}p(1-p)}[(1-p)^2\bar{S}_T^2 + p^2\bar{S}_C^2 + 2p(1-p)\bar{S}_{TC}^2],$$

where $\bar{S}_{TC}^2$ is the asymptote of $[1/n\bar{m}]\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Y_{Tij}-\bar{Y}_T)(Y_{Cij}-\bar{Y}_C)$, the covariance between the treatment and control potential outcomes for students within the same schools. A more intuitive variance expression is obtained by writing $\bar{S}_{\tau}^2$ as $\bar{S}_{\tau}^2 = \bar{S}_T^2 + \bar{S}_C^2 - 2\bar{S}_{TC}^2$. Solving for $\bar{S}_{TC}^2$ and substituting into (A.3) yields the variance expression in (7).

The asymptotic normality of $\hat{\beta}_{1,SR}$ follows by expressing (A.1) as $\sqrt{n\bar{m}}p(1-p)(\hat{\beta}_{1,SR}-\beta_1) = \sum_i\sum_j(T_i-p)\eta_{ij}/\sqrt{n\bar{m}}$ and using a central limit theorem for finite populations (see for example, Freedman 2008, Högland 1978, and Hájek 1960).

Proof of Lemma 2

The multiple regression estimator for β_1 is as follows:

$$(A.4)\quad \hat{\beta}_{1,MR} = [\tilde{\mathbf{T}}'(\mathbf{I}-\mathbf{P}_{\mathbf{Q}})\tilde{\mathbf{T}}]^{-1}\tilde{\mathbf{T}}'(\mathbf{I}-\mathbf{P}_{\mathbf{Q}})\mathbf{Y},$$

where $\tilde{\mathbf{T}}$ is an $Mx1$ vector containing $T_i - p$ terms for the full sample, $\mathbf{I}$ is the MxM identity matrix, $\mathbf{P}_{\mathbf{Q}} = \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$ is the projection matrix where $\mathbf{Q}$ is an Mxq matrix of covariates (that are centered around the grand means), and $\mathbf{Y}$ is an $Mx1$ vector of student outcomes.

If we substitute for $\mathbf{Y}$ in (A.4) using the true model in (3), then $\hat{\beta}_{1,MR}$ can be expressed as follows:

$$\text{(A.5)}\quad \hat{\beta}_{1,MR} = \left[\frac{1}{n\bar{m}}\tilde{\mathbf{T}}'(\mathbf{I}-\mathbf{P}_{\mathbf{Q}})\tilde{\mathbf{T}}\right]^{-1}\frac{1}{n\bar{m}}\tilde{\mathbf{T}}'(\mathbf{I}-\mathbf{P}_{\mathbf{Q}})[\mathbf{K}\beta_0 + \tilde{\mathbf{T}}\beta_1 + \boldsymbol{\eta}]$$

$$= \beta_1 + \left[\frac{\tilde{\mathbf{T}}'(\mathbf{I}-\mathbf{P}_{\mathbf{Q}})\tilde{\mathbf{T}}}{n\bar{m}}\right]^{-1}\left[\frac{\tilde{\mathbf{T}}'\boldsymbol{\eta}}{n\bar{m}} - \frac{\tilde{\mathbf{T}}'\mathbf{P}_{\mathbf{Q}}\boldsymbol{\eta}}{n\bar{m}}\right],$$

where $\mathbf{K}$ is a column of 1s and $\boldsymbol{\eta}$ is a vector of error terms in (3). This estimator is biased in finite samples. However, we show that the bias tends to zero as *n* approaches infinity by examining the limiting values of each bracketed term:

$$\left[\frac{\tilde{\mathbf{T}}'(\mathbf{I}\text{-}\mathbf{P}_{\mathbf{Q}})\tilde{\mathbf{T}}}{n\bar{m}}\right]^{-1} \xrightarrow{p} \frac{1}{p(1-p)},$$

$$\frac{\tilde{\mathbf{T}}'\boldsymbol{\eta}}{n\bar{m}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(T_i-p)\alpha_{ij}}{n\bar{m}} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(T_i-p)^2\tau_{ij}}{n\bar{m}} \xrightarrow{p} 0 + p(1-p)(0) = 0,$$

so that $\tilde{\mathbf{T}}$ and $\boldsymbol{\eta}$ are asymptotically uncorrelated, and:

$$\frac{\tilde{\mathbf{T}}'\mathbf{P_Q}\boldsymbol{\eta}}{n\bar{m}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(T_i - p)f_{ij}}{n\bar{m}} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(T_i - p)^2 h_{ij}}{n\bar{m}} \xrightarrow{p} 0,$$

where $\xrightarrow{p}$ denotes convergence in probability. Thus, $\hat{\beta}_{1,MR}$ is a consistent estimator.

To calculate the asymptotic variance of $\hat{\beta}_{1,MR}$, we apply an asymptotic expansion to (A.5):

$$\text{(A.6)} \quad \hat{\beta}_{1,MR} - \beta_1 = \frac{\tilde{\mathbf{T}}'\boldsymbol{\eta}}{n\bar{m}p(1-p)} - \frac{\tilde{\mathbf{T}}'\mathbf{P_Q}\boldsymbol{\alpha}}{n\bar{m}p(1-p)} + o_p(1/n),$$

where $o_p(1/n)$ signifies terms of order $1/n$. Note that the first term on the right-hand side of (A.6) pertains to the regression estimator without covariates. Note also that for the second term, $\tilde{\mathbf{T}}'\mathbf{P_Q}\boldsymbol{\alpha} = \mathbf{T}'\mathbf{P_Q}\boldsymbol{\alpha}$. Thus, (A.6) can be expressed as follows:

$$\text{(A.7)} \quad \hat{\beta}_{1,MR} - \beta_1 = \frac{1}{n\bar{m}p(1-p)}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\alpha_{ij} + (1-2p)\tau_{ij} - \mathbf{q_{ij}}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\boldsymbol{\alpha}\right]T_i + o_p(1/n),$$

where $\mathbf{q_{ij}}$ is a row vector of covariates for student *i*.

The term inside the brackets in (A.7) sums to zero because $\sum_i\sum_j \alpha_{ij} = \sum_i\sum_j \tau_{ij} = 0$, and $\sum_i\sum_j \mathbf{q_{ij}}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\boldsymbol{\alpha} = \sum_i\sum_j \alpha_{ij} = 0$ because it is the sum of fitted values when $\boldsymbol{\alpha}$ is regressed on $\mathbf{Q}$. Thus, if we define l_{ij} as the bracketed term in (A.6), then $\sum_i\sum_j l_{ij} = 0$, and we can use the same methods as for the regression estimator in Lemma 1 to derive the asymptotic variance of $\hat{\beta}_1$ in (10). The asymptotic normality of $\hat{\beta}_1$ follows from (A.6) because both $\tilde{\mathbf{T}}'\boldsymbol{\eta}/\sqrt{n\bar{m}}$ and $\tilde{\mathbf{T}}'\mathbf{P_Q}\boldsymbol{\alpha}/\sqrt{n\bar{m}}$ are asymptotically normal.

Appendix B: Summary of Data Sources

Table B.1: Summary of Data Sources

Study (Authors; Sponsor)[a]	Description of Study	Original and Current Study Populations	Level of Clustering	Outcome for Current Study	Baseline Covariates
Teach for America Evaluation (Decker et al. 2004; SRF; HF, CC)	Study examined the impact of teachers from Teach for America, a highly selective alternative certification program, on the academic achievement of elementary school students. Students were randomly assigned to classrooms taught by Teach for America teachers or traditional teachers in the same grade and school.	1st to 5th graders in the 2001-2002 school year; 17 schools in Baltimore, Chicago, Los Angeles, Mississippi Delta, and New Orleans. Current study focuses on 1st graders.	Teacher	Iowa Test of Basic Skills (ITBS) math score	Baseline test scores in reading and math; grade level indicators; school indicators
Evaluation of Education Technologies (Dynarski et al. 2007; IES)	Study examined the effects of 16 software products on students' academic achievement in 1st grade reading, 4th grade reading, 6th grade math, and algebra in 33 school districts. Within each participating school, teachers were randomly assigned to use a study product or not. For the purposes of our report, outcomes in 1st and 4th grades are used.	Students in 1st grade, 4th grade, 6th grade, and algebra classes in the 2004-05 school year in 33 districts. Current study focuses on 1st graders.	Teacher	1st grade Stanford-9 reading NCE score	Baseline test scores; student's age and gender; teacher's gender, experience, and highest degree; school's racial/ethnic composition; percent of school's students eligible for special education and subsidized lunch

Table B.1: Summary of Data Sources

Study (Authors; Sponsor)[a]	**Description of Study**	**Original and Current Study Populations**	**Level of Clustering**	**Outcome for Current Study**	**Baseline Covariates**
Evaluation of Comprehensive Teacher Induction Programs (Glazerman et al. 2008; IES)	Study examined the effects of comprehensive teacher induction programs on teacher retention, teachers' classroom practices, and student outcomes. The programs provided beginning teachers with an orientation, mentoring sessions, and professional development. Random assignment of elementary schools took place within 17 participating districts.	Beginning teachers in elementary schools within 17 low-income school districts across 13 states in the 2005-06 school year. Current study focuses on 2nd to 6th graders	School	District-specific administered test scores (Z-scores)	Student pretest Z-scores, gender, race/ethnicity, free/reduced price lunch status, special education status, grade level; teacher's age, gender, race/ethnicity, teaching and non-teaching experience, certification status, preparation type, educational attainment
Achievement Effects of Four Early Elementary School Math Curricula: Findings from First Graders in 39 Schools (Agodini et al. 2009; IES)	Study examined the relative impacts of four math curricula on first-grade mathematics achievement. The curricula were selected to represent diverse approaches to teaching elementary school math in the United States. The four curricula are Investigations in Number, Data, and Space; Math Expressions; Saxon Math; and Scott Foresman-Addison Wesley Mathematics.	First graders in 39 Title I schools in four districts in four states for both the original and current study. For the current study, the treatment group was defined as those in schools receiving the Saxon and Math Expressions curricula, and the control group was defined as those receiving the remaining two curricula.	School	ECLS-K total math assessment scale score in five math content areas	ECLS-K pretest score and seven strata (block) indicator variables.

Table B.1: Summary of Data Sources

Study (Authors; Sponsor)[a]	Description of Study	Original and Current Study Populations	Level of Clustering	Outcome for Current Study	Baseline Covariates
Effectiveness of Selected Supplemental Reading Comprehension Interventions (James-Burdumy et al. 2009; IES)	Study examined the impacts of four reading comprehension curricula for a first cohort of fifth graders. The curricula were Project CRISS, ReadAbout, Read for Real, and Reading for Knowledge and were selected based on public submissions and ratings by an expert review panel. Schools were randomly assigned to one of the four intervention groups or to a control group.	5th grade students in the 2006-2007 school year in 89 schools in 10 districts for both the original and current study. For the current study, the treatment group was defined as those in schools offering the Reading for Knowledge curriculum, and the control group includes those in schools that were assigned to the study control group.	School	Composite Z-Score from the Passage Comprehension Subtest of the Group Reading Assessment and Diagnostic Evaluation (GRADE) and the Science and Social Studies (SS) Reading Comprehension Assessments	Indicators of school urban/rural status; teacher race/ethnicity indicators; district indicators; student pretest scores on the GRADE and SS tests; Student race/ethnicity indicators; missing value indicators

[a] Acronyms are defined as follows: IES = Institute of Education Sciences at the U.S. Department of Education; SRF = Smith Richardson Foundation; HF= Hewlett Foundation; CC=Carnegie Corporation.

References

Agodini, R., B. Harris, S. Atkins-Burnett, S. Heavside, T. Novak, R. Murphy (2009). Achievement Effects of Four Early Elementary School Math Curricula: Findings from First Graders in 39 Schools. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Baltagi, B. and Y. Chang (1994). A Comparative Study of Alternative Estimators for the Unbalanced One-Way Error Component Regression Model. *Journal of Econometrics* 62, 67-89.

Binder, D. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review* 51, 279-292.

Bingenheimer, J. and S. Raudenbush. Statistical and Substantive Inferences in Public Health: Issues in the Application of Multilevel Models. *Annual Review of Public Health* 25, 53-77.

Bryk, A. and S. Raudenbush (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Cochran, W. (1963). *Sampling Techniques*. New York: John Wiley and Sons.

De Leeuw, J. and E. Meijer (2008), *Handbook of Multilevel Analysis*, New York: Springer.

Decker, P., D. Mayer, and S. Glazerman (2004). The Effects of Teach For America on Students: Findings from a National Evaluation. Princeton, NJ: Mathematica Policy Research, Inc.

Dynarski, M., R. Agodini, S. Heaviside, T. Novak, N. Carey, L. Campuzano, B. Means, R. Murphy, W. Penuel, H. Javitz, D. Emery, and W. Sussex (2007). Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort. Washington, DC: Institute of Education Sciences.

Freedman, D. (2008). On Regression Adjustments to Experimental Data. *Advances in Applied Mathematics* 40, 180-193.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., and Pee, D. (1996). On Design Considerations and Randomization-Based Inference for Community Intervention Trials, *Statistics in Medicine*, 15, 1069–1092.

Glazerman, S., S. Dolfin, M. Bleeker, A. Johnson, E. Isenberg, J. Lugo-Gil, M. Grider, E. Britton (2008). Impacts of Comprehensive Teacher Induction: Results from the First Year of a Randomized Controlled Study. Washington, DC: U.S. Institute of Education Sciences.

Hájek, J. (1960). Limiting Distributions in Simple Random Sampling from a Finite Population. *Publications of the Mathematics Institute of Hungarian Academy of Science* 5, 361-375.

Hardin , J. and J. Hilbe (2003). *Generalized Estimating Equations.* Boca Raton FL: Chapman and Hall / CRC.

Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association,* 72, 320-340.

Hoglund, T. (1978), Sampling From a Finite Population: A Remainder Term Estimate. *Scandinavian Journal of Statistics* 5 69–71.

Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396), 945-960.

Hsiao, C. (1986). *Analysis of Panel Data,* U.K.: Cambridge University Press.

Imbens, G. and D. Rubin (2007). *Causal Inference: Statistical Methods for Estimating Causal Effects in Biomedical, Social, and Behavioral Sciences,* U.K.: Cambridge University Press.

James-Burdumy, S. et al. (2009). Effectiveness of Selected Supplemental Reading Comprehension Interventions. Washington, DC: U.S. Institute of Education Sciences.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics,* 38, 963–974.

Liang, K. and S. Zeger (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* 73, 13-22.

Murray, D. (1998). *Design and Analysis of Group-Randomized Trials,* New York: Oxford University Press.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments: Essay on Principles. Section 9, Translated in *Statistical Science*, 1990: Vol. 5, No. 4.

Patterson, H. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545-554.

Pfeffermann, D., C. Skinner, D. Holmes, H. Goldstein, J. Rasbash (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. *JRSS B,* 60: 23-40.

Rao, C.R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association,* 69: 112-115.

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Education Psychology*, 66, 688-701.

Rubin, D. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Education Statistics*, 2(1), 1-26.

Schochet, P. (2008). The Late Pretest Problem in Randomized Control Trials of Education Interventions. Technical Methods Paper: U.S. Department of Education: Institute of Education Sciences, National Center for Education Evaluation.

Schochet, P. (2009). Is Regression Adjustment Supported by the Neyman Model for Causal Inference? *Journal of Statistical Planning and Inference*, forthcoming.

Shah, S. V. (1998). Software for GEE: PROC GENMOD and SUDAAN. Paper presented at the SESUG meetings at Norfolk VA.

Swamy, P., and S. Arora (1972). The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models. *Econometrica*, 40:261-275.

West , B., K. Welch, and A. Galecki (2007). *Linear Mixed Models*. Boca Raton FL: Chapman and Hall / CRC.

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