NCEE 2009-4033

U. S. DEPARTMENT OF EDUCATION

The Late Pretest Problem in Randomized Control Trials of Education Interventions

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

The Late Pretest Problem in Randomized Control Trials of Education Interventions

October 2008

Peter Z. Schochet
Mathematica Policy Research, Inc.

Abstract

Pretest-posttest experimental designs are often used in randomized control trials (RCTs) in the education field to improve the precision of the estimated treatment effects. For logistic reasons, however, pretest data are often collected after random assignment, so that including them in the analysis could bias the posttest impact estimates. Thus, the issue of whether to collect and use late pretest data in RCTs involves a variance-bias tradeoff. This paper addresses this issue both theoretically and empirically for several commonly-used impact estimators using a loss function approach that is grounded in the causal inference literature. The key finding is that for RCTs of interventions that aim to improve student test scores, estimators that include late pretests will typically be preferred to estimators that exclude them or that instead include uncontaminated baseline test score data from other sources. This result holds as long as the growth in test score impacts do not grow very quickly early in the school year.

NCEE 2009-4033
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, Inc. to develop methods for examining the late pretest problem 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.

October 2008

Schochet, Peter Z. (2008). *The Late Pretest Problem in Randomized Control Trials of Education Interventions* (NCEE 2009-4033). 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 202-260-98795 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

The author for this report, Dr. Peter Schochet, is an employee of Mathematica Policy Research, Inc. 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.

Contents

Chapter 1: Introduction **1**

Chapter 2: The Late Pretest Problem **3**

Chapter 3: Measuring the Variance-Bias Tradeoff **5**

Chapter 4: Considered Designs **7**

Chapter 5: Theoretical Framework **9**

Chapter 6: The Variance-Bias Tradeoff for Various ATE Estimators **13**
- The Posttest-Only Estimator 13
- The Differences-In-Differences (DID) Estimator 14
- The ANCOVA Estimator 15
- The Unbiased ANCOVA (UANCOVA) Estimator 17
- The Generalized Estimating Equation (GEE) Estimator 18
- HLM Growth Curve Approach 19

Chapter 7: Empirical Analysis **21**
- Structure 21
- Assumptions 23
- Empirical Results for Design I 25
- Empirical Results for Designs II and III 29

Chapter 8: Summary and Conclusions **31**

Appendix A **A-1**

References **R-1**

List of Tables

Table 7.1: Hypothetical Growth Trajectories of Test Score Impacts, by the Number of Months Since the Start of School (Design I) 23
Table 7.2: Maximum Values of β_1 / θ_0 for Which the ANCOVA and DID Estimators Would be Preferred to the Posttest-Only and UANCOVA Estimators (Design I) 26
Table 7.3: Variance, Bias, and *MSE* Estimates for the ANCOVA and UANCOVA Estimators, For Various Values of β_1 / θ_0 (Design I) 28
Table 7.4: School Sample Sizes Needed to Equate MDE Values for the ANCOVA Estimator, by the Size of the Early Treatment Effect (Design I) 28
Table 7.5: Maximum Values of β_1 / θ_0 for Which the ANCOVA Estimator Would be Preferred to the Posttest-Only Estimator (Designs II and III) 29

List of Figures

Figure 7.1: Hypothetical Growth Trajectories of Test Score Impacts (Measured in Standard Deviation Units) 22

Chapter 1: Introduction

Pretest-posttest experimental designs are often used to examine the impacts of educational interventions on student achievement test scores. For these designs, a test is administered to students in the fall of the school year (the pretest) and at a spring follow-up (the posttest). Average treatment effects are then estimated by either examining treatment-control differences on pretest-posttest gain scores or by including pretests as covariates in posttest regression models.

In clustered randomized control trials (RCTs) in the education field, the availability of pretests on individual students is critical for obtaining, at reasonable cost, precise posttest impact estimates (Schochet 2008; Bloom et al. 2005). In these RCTs, groups (such as schools or classrooms) rather than students are typically randomly assigned to the treatment or control conditions. This clustering considerably reduces statistical power due to the dependency of student outcomes within groups. The inclusion of pretests in the analysis, however, can substantially increase precision levels, because group-level pretest-posttest correlations tend to be large. Schochet (2008), for example, demonstrates that for a design in which schools are the unit of random assignment, about 44 total schools are required to detect an impact of 0.25 standard deviations if pretests are used in the analysis, compared to about 86 schools if pretest data are not available. This occurs because pretests tend to explain a large proportion of the variance in posttest scores.

For logistic reasons, however, pretests on individual students are typically collected *after* the start of the school year. In these cases, including late pretests in the analysis could bias the posttest impact estimates in the presence of early treatment effects. Because of variance gains, however, these biased estimators could yield impact estimates that tend to be *closer* to the truth than unbiased estimators that exclude the late pretests. Thus, the issue of whether to collect and use late pretest data in RCTs involves a variance-bias tradeoff.

This paper is the first to systematically examine, both theoretically and empirically, the late pretest problem in education RCTs for several commonly-used impact estimators. The paper addresses three main research questions:

1. ***Under what conditions does the variance-bias tradeoff favor the inclusion rather than exclusion of late pretests in the posttest impact models?*** These conditions are important for assessing whether or not to collect expensive pretest data.

2. ***What are statistical power losses when late pretests are included in the estimation models?*** Large-scale RCTs in the education field are typically powered to detect minimum detectable posttest impacts of about 0.15 to 0.30 standard deviations, ignoring the potential late pretest problem. If pretest data are to be collected, how much larger do school sample sizes need to be in the presence of late pretests to achieve posttest impact estimates with the same level of statistical precision?

3. ***Instead of collecting pretest data, under what conditions is it preferable to collect "true" baseline test score data from alternative sources?*** For example, historic aggregate school-level data could be collected on test scores that are related to the posttest. The correlations between these alternative test scores and the posttests are likely to be smaller than the pretest-posttest correlations, and thus, the alternative test scores will reduce variance less. However, these data are likely to be uncontaminated, and thus, will not bias the posttest impact estimates.

The theory presented in this paper is based on a unified regression approach for group-based RCTs that is anchored in the causal inference and hierarchical linear modeling (HLM) literature. The empirical analysis quantifies the late pretest problem in education RCTs using simulations that are based on key parameter values found in the literature that pertain to achievement test scores of elementary school and preschool students in low-performing school districts. The focus on test scores is consistent with accountability provisions of the No Child Left Behind Act of 2001, and the ensuing federal emphasis on testing interventions to improve reading and mathematics scores of young students.

The rest of this paper is in seven chapters. Chapter 1 discusses the late pretest problem in more detail, and Chapter 2 discusses two measures for quantifying the variance-bias tradeoff when late pretests are included in the impact models. Chapter 3 discusses the considered school-based designs, and Chapter 4 presents the causal inference statistical theory underlying the late pretest problem. Chapter 5 applies this theory to several commonly-used impact estimators, and Chapter 6 presents simulation results. Finally, Chapter 7 presents a summary and conclusions.

Chapter 2: The Late Pretest Problem

Pretests on individual students are typically collected after the start of the school year for several reasons. First, school administrators and teachers typically prefer that baseline testing occur after students and teachers settle into a routine. Second, researchers often want to delay testing until a large percentage of signed study consent forms are returned by parents (many studies in a school setting require active parental consent). Finally, for cost reasons, studies often employ a small number of interviewer teams per site to administer baseline testing in the study schools. Thus, it usually takes time for these teams to set up visiting schedules and to travel to schools that are geographically dispersed. Hence, in many RCTs, baseline testing is not completed until several months after school begins. For example, in the Head Start Impact Study (Puma et al. 2005) baseline testing occurred over a three-month period from October 2002 through December 2002.

The inclusion of late pretests in the posttest impact models could lead to biased impact estimates for several reasons. First, in most evaluations, the tested interventions are implemented in the treatment schools and classrooms prior to the start of the school year. For example, in evaluations testing the effects of a new math or reading curriculum, teacher professional development typically occurs during the summer. Thus, with late pretests, students in the treatment group have already been exposed to the intervention.

A second reason that pretests could be contaminated is if the distribution of baseline testing dates differs across the treatment and control groups. Student test scores tend to increase over time naturally. Thus, pretests could be contaminated if they are administered later for one research group than the other, even if there are no early intervention effects. Well-designed evaluations attempt to evenly disperse testing dates across the treatment and control groups. However, it is sometimes more difficult to schedule testing dates in control schools (who are denied the intervention) than in treatment schools (who are offered intervention services). For example, in the National Evaluation of Early Reading First (Jackson et al. 2007) baseline testing occurred about one month later, on average, in control sites than treatment sites.

Chapter 3: Measuring the Variance-Bias Tradeoff

The main advantage of including late pretests in the posttest impact models is that they can substantially improve the precision of the impact estimates. The main disadvantage of including them is that they could yield biased impact estimates. This paper uses two related loss functions for quantifying this variance-bias tradeoff for a posttest impact estimator $\hat{\gamma}$. The first loss function is the mean square error (*MSE*):

$$(1) \quad MSE(\hat{\gamma}) = E(\hat{\gamma} - \gamma)^2 = Var(\hat{\gamma}) + Bias(\hat{\gamma})^2,$$

where $Var(\hat{\gamma})$ is the variance of the estimator, γ is the true posttest impact, and $Bias(\hat{\gamma}) = [E(\hat{\gamma}) - \gamma]$ is the bias of the estimator. An estimator is preferred to another if it has a lower *MSE* value.

The second loss function, which is typically used in the design stage of impact evaluations to determine appropriate sample sizes, is the minimum detectable impact (*MDI*). The *MDI* represents the smallest program impact that can be detected with a high probability. I follow the usual practice of standardizing minimum detectable impacts into effect size units—that is, as a percentage of the standard deviation of the outcome measures (also known as Cohen's *d*)—to facilitate the comparison of findings across outcomes that are measured on different scales (Cohen 1988). Hereafter, minimum detectable impacts in effect size units are denoted as *MDEs*.

To develop manageable *MDE* formulas for biased estimators, it is assumed that under the null hypothesis of no impacts on posttest scores, there are no impacts on late pretest scores. This assumption rules out early positive or negative intervention effects that disappear by the follow-up test date. This key assumption considerably simplifies the *MDE* calculations because Type I error rates remain the same for all estimators.

Under this assumption, the *MDE* formula for $\hat{\gamma}$ can be obtained by first noting that for significance level α, the critical value for the *t*-statistic under the null hypothesis of no impact on posttest scores is $T^{-1}(1-\{\alpha/2\})$ for a two-tailed test and $T^{-1}(1-\alpha)$ for a one-tailed test, where $T^{-1}(.)$ is the inverse of the student's *t* distribution function with *df* degrees of freedom. For a given *MDI* value, statistical power for a two-tailed test under the alternative hypothesis $H_1: \gamma = MDI$ can then be expressed as follows:

$$(2) \quad P\left(\left|\frac{\hat{\gamma}}{\sqrt{Var(\hat{\gamma})}}\right| > T^{-1}(1-\{\alpha/2\}) \mid \gamma = MDI, Bias(\hat{\gamma})\right) = \beta,$$

where β is the preset statistical power level (for example, 80 percent). The *MDE* formula for $\hat{\gamma}$ can then be obtained by solving for *MDI* in (2) and dividing *MDI* by the standard deviation of the posttest score (θ_1):

$$(3) \quad MDE(\hat{\gamma}) = MDI/\theta_1 = [Factor(\alpha, \beta, df)\sqrt{Var(\hat{\gamma})} - Bias(\hat{\gamma})]/\theta_1,$$

where *Factor*(.) is $[T^{-1}(1-\{\alpha/2\}) + T^{-1}(\beta)]$ for a two-tailed test and $[T^{-1}(1-\alpha) + T^{-1}(\beta)]$ for a one-tailed test. *Factor*(.) becomes larger as α and *df* decrease and as β increases (see Schochet 2008). If $\alpha = .05$ and $\beta = .80$ (typical assumptions) and *df* > 40, *Factor*(.) is about 2.5 for a one-tailed test and 2.8 for a two-tailed test. An estimator is preferred to another if it has a lower *MDE* value.

The *MDE* formula in (3) is appropriate only when the posttest impact estimators are unbiased or biased downwards (that is, when $Bias(\hat{\gamma}) \leq 0$) so that there is a variance-bias tradeoff when comparing estimators. In these cases, relative to the *MSE* criterion, the *MDE* criterion tends to place more weight on the variance component and less weight on the bias component.

Finally, it is important to note that the *MSE* and *MDE* criteria do not include pretest data collection costs. Thus, this paper does not consider these costs when comparing estimators.

Chapter 4: Considered Designs

The focus of this paper is on two-level experimental designs in which students are nested within units (such as schools or classrooms) that are randomly assigned to either a single treatment or control condition. Two-level designs are considered here to keep the presentation manageable and because they are the most common designs used in education research. The two-level considered designs are as follows: **Design I,** where schools are the unit of random assignment; and **Design II**, where classrooms are the unit of random assignment and school effects are treated as fixed (which occurs in the common case where schools are purposively selected for the study and school effects are treated as fixed strata, so that the impact results generalize to the study schools only).

Finally, this paper also considers **Design III**, where students are the unit of random assignment and purposively-selected site (school or district) effects are treated as fixed. This is a nonclustered, stratified RCT design that is a special case (collapsed version) of the two-level designs discussed above. These designs are discussed in more detail in Schochet (2008).

Chapter 5: Theoretical Framework

This chapter discusses the statistical theory underlying the variance-bias tradeoff associated with including pretests in the posttest impact models for two-level clustered RCTs. The theory is discussed in the context of the causal inference theory underlying RCTs (Neyman 1923; Rubin 1974; Holland 1986; Imbens and Rubin 2007; Freedman 2008; Schochet 2007).

It is assumed that students are nested within n units (schools or classrooms) that are randomly assigned to a single treatment or control group. The sample is assumed to contain np treatment units and $n(1-p)$ control units, where p is the sampling rate to the treatment group ($0<p<1$).

This paper considers a "superpopulation" version of the Neyman-Rubin causal inference model (see Imbens and Rubin 2007; Schochet 2007; and Yang and Tsiatis 2001). Let Z_{1Ti} be the "potential" unit-level continuous posttest score for unit i in the treatment condition and Z_{1Ci} be the potential posttest score for unit i in the control condition. Potential posttest scores for the n study units are assumed to be random draws from potential treatment and control posttest distributions in the study population, with means μ_{1T} and μ_{1C}, respectively; a common variance $\sigma_1^2 > 0$ is assumed for each research group to ensure that variance estimates based on standard ordinary least squares (OLS) methods are justified by the Neyman-Rubin causal model (Freedman 2008; Schochet 2007). It is assumed that treatment assignments are independent of potential outcomes (due to random assignment), and that potential outcomes for each unit are unrelated to the treatment status of other units. Finally, let Z_{0Ti}, Z_{0Ci}, μ_{0T}, μ_{0C}, and σ_0^2 denote corresponding quantities for fall *pretest* scores, and let σ_{01} denote the covariance between the potential pretest and posttest scores for both the treatment and control groups (which could depend on how late the pretests are collected).[1]

Suppose next that m students are sampled from the student superpopulation within each study unit. Let Y_{1Tij} be the potential posttest score for student j in unit i in the treatment condition and Y_{1Cij} be the corresponding potential posttest score for the student in the control condition. Y_{1Tij} and Y_{1Cij} are assumed to be random draws from student-level potential treatment and control posttest distributions (which are conditional on school-level potential outcomes) with means Z_{1Ti} and Z_{1Ci}, respectively, and common variance $\tau_1^2 > 0$. Corresponding variables for student-level pretest scores are denoted by replacing subscripts of "1" by subscripts of "0". The covariance between student-level potential pretest and posttest scores within units is denoted by τ_{01}.[2]

Under this causal inference model, the difference between the two potential posttest scores, $(Z_{1Ti} - Z_{1Ci})$, is the unit-level treatment effect for unit i, and the average treatment effect parameter (ATE) is $ATE = E(Z_{1Ti} - Z_{1Ci}) = \mu_{1T} - \mu_{1C}$. The unit-level treatment effects, and hence, the ATE parameter,

[1] Neyman (1923) considered a "finite population" model where potential outcomes are assumed to be fixed for the study population and where the only source of randomness is treatment status.

[2] Equal cluster sample sizes are assumed for simplicity, and because this largely holds in clustered RCT designs in the education area. The results presented in this paper apply approximately for unequal cluster sizes if m is replaced in the formulas by the average cluster size $\bar{m}$ (Kish 1965).

cannot be calculated directly because for each unit and student, the potential outcome is observed in either the treatment or control condition, but not in both. Formally, if T_i is a treatment status indicator variable that equals 1 for treatments and 0 for controls, then the *observed* posttest score for a unit, z_{1i}, can be expressed as follows:

$$(4) \quad z_{1i} = T_i Z_{1Ti} + (1 - T_i) Z_{1Ci}.$$

Similarly, the observed posttest score for a student y_{1ij} is:

$$(5) \quad y_{1ij} = T_i Y_{1Tij} + (1 - T_i) Y_{1Cij}.$$

The simple equations in (4) and (5) form the basis for the causal inference theory presented below.

The terms in (5) can be rearranged to create the following regression model:

$$(6) \quad y_{1ij} = \alpha_0 + \alpha_1 T_i + (u_{1i} + e_{1ij}), \text{ where}$$

1. $\alpha_0 = \mu_{1C}$ and $\alpha_1 = \mu_{1T} - \mu_{1C}$ (the *ATE* parameter) are coefficients to be estimated
2. $u_{1i} = T_i(Z_{1Ti} - \mu_{1T}) + (1 - T_i)(Z_{1Ci} - \mu_{1C})$ is a unit-level error term with mean zero and between-unit variance σ_1^2 that is uncorrelated with T_i
3. $e_{1ij} = T_i(Y_{1Tij} - Z_{1Ti}) + (1 - T_i)(Y_{1Cij} - Z_{1Ci})$ is a student-level error term with mean zero and within-unit variance τ_1^2 that is uncorrelated with u_{1i} and T_i

Importantly, (6) can also be derived using the following two-level HLM model (Bryk and Raudenbush 1992):

$$\begin{aligned} &Level\,1: \quad y_{1ij} = z_{1i} + e_{1ij} \\ &Level\,2: \quad z_{1i} = \alpha_0 + \alpha_1 T_i + u_i, \end{aligned}$$

where Level 1 corresponds to students and Level 2 to units. Inserting the Level 2 equation into the Level 1 equation yields (6). Thus, the HLM approach is consistent with the causal inference theory presented above.[3]

A similar approach can be used to develop a regression model for the observed *pretest* scores:

$$(7) \quad y_{0ij} = \beta_0 + \beta_1 T_i + (u_{0i} + e_{0ij}),$$

[3] It is assumed that there are no biases due to missing posttest data. Davidian et al. (2005) discuss semiparametric estimation of treatment effects in a pretest-posttest study with missing data.

where $\beta_0 = \mu_{0C}$, $\beta_1 = \mu_{0T} - \mu_{0C}$, and u_{0i} and e_{0ij} are between- and within-unit error terms, respectively, with the following properties: $E(u_{0i}) = E(e_{0ij}) = 0$; $E(T_i u_{0i}) = E(T_i e_{0ij}) = 0$; $Var(u_{0i}) = \sigma_0^2$; $Var(e_{0ij}) = \tau_0^2$; $Cov(u_{0i}, e_{0ij}) = 0$; $Cov(u_{0i}, u_{1i}) = \sigma_{01}$; and $Cov(e_{0ij}, e_{1ij}) = \tau_{01}$.

If the pretests are "true" baselines, β_1 will equal zero because of random assignment. Stated differently, with true baselines, $Z_{0Ti} = Z_{0Ci}$ and $Y_{0Tij} = Y_{0Cij}$. With late baselines, the size and sign of β_1 will depend on the growth trajectory of intervention effects, the overall timing of baseline testing, and differences in testing-date distributions across the treatment and control groups. For example, β_1 will tend to be positive if the intervention has early beneficial intervention effects or if pretest testing dates are, on average, later for treatments than for controls.

Chapter 6: The Variance-Bias Tradeoff for Various ATE Estimators

This chapter uses the causal inference regression models in (6) and (7) to mathematically examine the variance-bias tradeoff for several commonly-used ATE impact estimators. All estimators and their asymptotic properties were obtained using standard OLS methods. Appendix A provides a proof of the asymptotic results for the analysis of covariance (ANCOVA) estimator (the most general case); proofs for the other estimators are similar.

The Posttest-Only Estimator

The posttest-only estimator $\hat{\gamma}_{Posttest}$ does not adjust for the pretests and is obtained by applying OLS methods directly to equation (6). The resulting estimator is as follows:

(8) $\hat{\gamma}_{Posttest} = \overline{y}_{1T} - \overline{y}_{1C}$, where

$$\overline{y}_{1T} = \frac{1}{nmp}\sum_{i=1}^{n}\sum_{j=1}^{m} T_i y_{1ij} \text{ and } \overline{y}_{1C} = \frac{1}{nm(1-p)}\sum_{i=1}^{n}\sum_{j=1}^{m}(1-T_i)y_{1ij}.$$

As n approaches infinity (for fixed m), $\hat{\gamma}_{Posttest} \xrightarrow{p} \alpha_1$, where $\xrightarrow{p}$ denotes convergence in probability. Thus, $\hat{\gamma}_{Posttest}$ is an asymptotically unbiased estimator for the ATE parameter. Furthermore, results in Appendix A can be used to show that $\hat{\gamma}_{Posttest}$ converges to a normal distribution with the following asymptotic variance:

$$(9) \quad AsyVar(\hat{\gamma}_{Posttest}) = MSE(\hat{\gamma}_{Posttest}) = \frac{1}{p(1-p)}\left[\frac{\sigma_1^2}{n} + \frac{\tau_1^2}{nm}\right].$$

The within-unit (second) variance term in (9) is the conventional variance expression for an impact estimate for a nonclustered, stratified design (Design III). Design effects in a clustered design arise because of the first variance term, which represents the correlation of student posttest scores within the same units (Murray 1998; Donner and Klar 2000; Raudenbush 1997).

For the empirical work presented below, it is convenient to express the variance expression in (9) in terms of the intraclass correlation (ICC_1) (Cochran 1963; Kish 1965), which is defined as the between-unit variance (σ_1^2) as a proportion of the total variance of the outcome measure ($\theta_1^2 = \sigma_1^2 + \tau_1^2$):

$$(10) \quad AsyVar(\hat{\gamma}_{Posttest}) = \frac{1}{p(1-p)}\left[\frac{\theta_1^2 ICC_1}{n} + \frac{\theta_1^2(1-ICC_1)}{nm}\right].$$

In this formulation, design effects from clustering are small if mean posttest scores do not vary much across units (that is, if ICC_1 is small).

The Differences-In-Differences (DID) Estimator

The DID estimator $\hat{\gamma}_{DID}$ is obtained by applying OLS methods to a gain-score model formed by subtracting the pretest model in (7) from the posttest model in (6). The DID estimator is:

$$\hat{\gamma}_{DID} = (\overline{y}_{1T} - \overline{y}_{1C}) - (\overline{y}_{0T} - \overline{y}_{0C}),$$

where $\overline{y}_{1T}$ and $\overline{y}_{1C}$ are defined as above and

$$\overline{y}_{0T} = \frac{1}{nmp}\sum_{i=1}^{n}\sum_{j=1}^{m} T_i y_{0ij} \text{ and } \overline{y}_{0C} = \frac{1}{nm(1-p)}\sum_{i=1}^{n}\sum_{j=1}^{m}(1-T_i) y_{0ij}\ .$$

As *n* approaches infinity, $\hat{\gamma}_{DID} \xrightarrow{p} (\alpha_1 - \beta_1)$; thus, the asymptotic bias of the DID estimator is $-\beta_1$. This estimator will provide a downwardly biased estimate of the posttest impact if the intervention improves late pretest scores. Conversely, $\hat{\gamma}_{DID}$ will provide an upwardly biased estimate of the posttest impact if the intervention lowers late pretest scores. The DID estimator will be asymptotically unbiased if and only if $\beta_1 = 0$.

The DID estimator converges to a normal distribution with mean $(\alpha_1 - \beta_1)$ and the following asymptotic variance:

$$(11) \quad AsyVar(\hat{\gamma}_{DID}) = \frac{1}{p(1-p)}\left[\frac{\sigma_1^2 + \sigma_0^2 - 2\sigma_{01}}{n} + \frac{\tau_1^2 + \tau_0^2 - 2\tau_{01}}{nm}\right].$$

This expression can also be written as follows:

$$(12) \quad AsyVar(\hat{\gamma}_{DID}) = \frac{1}{p(1-p)}\left[\frac{\sigma_1^2 + \sigma_0^2 - 2\sigma_0\sigma_1\rho_{01}}{n} + \frac{\tau_1^2 + \tau_0^2 - 2\tau_0\tau_1\lambda_{01}}{nm}\right],$$

where $\rho_{01} = (\sigma_{01} / \sigma_1\sigma_0)$ and $\lambda_{01} = (\tau_{01} / \tau_1\tau_0)$ are unit-level and student-level pretest-posttest correlations, respectively. The DID variance does not depend on β_1. The asymptotic variance for the nonclustered Design III can be obtained by setting $\sigma_1^2 = \sigma_0^2 = 0$ in (12).

The comparison of (9) and (12) shows that $\hat{\gamma}_{DID}$ will have smaller variance than $\hat{\gamma}_{Posttest}$ if the pretest-posttest correlations are positive and sufficiently large. For example, if we focus only on the leading unit-level variance term in (12) and assume that $\sigma_1^2 = \sigma_0^2$, the DID estimator will be more efficient if $\rho_{01} \geq 0.5$. This condition is likely to hold in our application, because pretest-posttest correlations of 0.7 to 0.9 are typically found in the education field (Schochet 2008; Bloom et al. 2005). Thus, if $\beta_1 \neq 0$, $\hat{\gamma}_{DID}$ will be asymptotically biased, but *MSE* and *MDE* values could be smaller for $\hat{\gamma}_{DID}$ than for $\hat{\gamma}_{Posttest}$ due to efficiency gains, as discussed in the empirical analysis below.

The ANCOVA Estimator

The ANCOVA estimator $\hat{\gamma}_{ANCOVA}$ is obtained by regressing observed posttest scores on T_i and the pretest scores. The *true* model for the posttest scores is (6), but the observed pretests are included as "irrelevant" covariates to improve the precision of the posttest impact estimates.

It is assumed that two pretest variables are included as covariates in the model: (1) the unit-level mean pretest score, $\overline{y}_{0i}$, and (2) the difference between the student-level pretest score and the unit-level pretest score, $y_{0ij}^{w} = (y_{0ij} - \overline{y}_{0i})$. These two variables are used to allow for separate effects of the pretests in reducing between- and within-unit variance. Thus, the ANCOVA estimation model is:

$$(13) \quad y_{1ij} = \delta_0 + \gamma T_i + \delta_1 \overline{y}_{0i} + \delta_2 y_{0ij}^{w} + (\nu_i + \omega_{ij}),$$

where ν_i and ω_{ij} are mean zero error terms and δ_0, δ_1, and δ_2 are parameters to be estimated.

As shown in Appendix A, as *n* approaches infinity, $\hat{\gamma}_{ANCOVA} \xrightarrow{p} \alpha_1 - \beta_1(\sigma_{01} / \sigma_0^2)$. Thus, the asymptotic bias of the ANCOVA estimator is:

$$(14) \quad AsyBias(\hat{\gamma}_{ANCOVA}) = -\beta_1 \frac{\sigma_{01}}{\sigma_0^2} = -\beta_1 \rho_{01} \frac{\sigma_1}{\sigma_0}.$$

The term $(\sigma_{01} / \sigma_0^2)$ is the OLS parameter estimate from a regression of the unit-level potential posttests on the unit-level potential pretests (that is, when u_{1i} is regressed on u_{0i}). Thus, the asymptotic bias of $\hat{\gamma}_{ANCOVA}$ is the product of this regression coefficient and $-\beta_1$. The ANCOVA estimator will be unbiased only if $\beta_1 = 0$ or $\rho_{01} = 0$.

If $\beta_1 \neq 0$ and $\rho_{01} \geq 0$, the relative bias of $\hat{\gamma}_{ANCOVA}$ compared to $\hat{\gamma}_{DID}$ will depend on the value of $\rho_{01}(\sigma_1 / \sigma_0)$. The bias of $\hat{\gamma}_{ANCOVA}$ will be smaller in absolute value if $\rho_{01}(\sigma_1 / \sigma_0) < 1$. This will occur if σ_0 and σ_1 are similar, which is likely to hold in our application (see below). This condition will also hold if $\sigma_0 > \sigma_1$ or $\rho_{01} = 0$. The bias of the two estimators will be the same if $\rho_{01}(\sigma_1 / \sigma_0) = 1$, which is the assumption underlying the DID model (that is, the regression coefficient on the pretest covariate is 1). The DID estimator will be less biased in absolute value only if $\sigma_1 > (\sigma_0 / \rho_{01})$.

As shown in Appendix A, $\hat{\gamma}_{ANCOVA}$ has an asymptotic normal distribution with mean $\alpha_1 - \beta_1(\sigma_{01} / \sigma_0^2)$ and the following asymptotic variance:

$$(15)\quad AsyVar(\hat{\gamma}_{ANCOVA}) = \frac{1}{p(1-p)}\left[\frac{\sigma_1^2(1-\rho_{01}^2)}{n} + \frac{\tau_1^2(1-\lambda_{01}^2)}{nm}\right]d, \text{ where}$$

$$d = \left[1 + \frac{\beta_1^2 p(1-p)}{\sigma_0^2}\right] = \left[1 + \frac{\beta_1^2}{\theta_0^2}\frac{p(1-p)}{ICC_0}\right],$$

$\theta_0^2 = \sigma_0^2 + \tau_0^2$, and $ICC_0 = \sigma_0^2 / \theta_0^2$.

The term inside the brackets in (15) is similar to the variance expression for the posttest-only estimator in (9) except that σ_1^2 and τ_1^2 are reduced by $(1-\rho_{01}^2)$ and $(1-\lambda_{01}^2)$, respectively, to account for the pretest-posttest correlations (these are regression R^2 adjustments). Countering these precision gains is the design effect $d \geq 1$ which inflates the variance due to the collinearity of T_i and $\overline{y}_{0i}$ in (13). This design effect will typically be small, unless the early treatment effect measured in effect size units (β_1 / θ_0) is unrealistically large compared to the expected size of the posttest impact. For example, assuming p = 0.50 and ICC_0 = 0.15, we find that d = 1.11 if (β_1 / θ_0) = 0.10 and d = 1.07 if (β_1 / θ_0) = 0.20. These values of (β_1 / θ_0) are large relative to the posttest impact that most studies are powered to detect (about 0.15 to 0.30 standard deviations).

By comparing (9) and (15), it can be seen that $\hat{\gamma}_{ANCOVA}$ will typically be more efficient than $\hat{\gamma}_{Posttest}$. This will always be the case if $\beta_1 = 0$ and either $\rho_{01}^2 > 0$ or $\lambda_{01}^2 > 0$. This will also typically be the case if $\beta_1 \neq 0$ except in the unlikely event that R^2 gains from including the pretests are offset by power losses from the design effect d. For example, if we focus on the unit-level variance terms only, $\hat{\gamma}_{ANCOVA}$ will be more efficient than $\hat{\gamma}_{Posttest}$ if $d < 1/(1-\rho_{01}^2)$, which as discussed, will usually be satisfied in practice. Thus, although the ANCOVA estimator could be biased with late pretests, this estimator may produce lower *MSE* and *MDE* values than the posttest-only estimator due to efficiency gains.

The comparison of (11) and (15) shows also that the ANCOVA estimator will typically be more efficient than the DID estimator (see also Oakes and Feldman 2001; Allison 1990; and Reichardt 1979 for a discussion comparing the ANCOVA and DID estimators for nonclustered designs). This will always be the case if $\beta_1 = 0$ and either $\rho_{01}(\sigma_1 / \sigma_0) \neq 1$ or $\lambda_{01}(\tau_1 / \tau_0) \neq 1$. It will also tend to be the case if $\beta_1 \neq 0$. For example, focusing on the unit-level variance terms only, $\hat{\gamma}_{ANCOVA}$ will be more efficient than $\hat{\gamma}_{DID}$ under the following condition:

$$d < 1 + \frac{\left(\rho_{01} - [\sigma_0 / \sigma_1]\right)^2}{(1-\rho_{01}^2)}.$$

Because d values are likely to be small, this inequality is likely to be satisfied for most values of $\rho_{01}(\sigma_1 / \sigma_0) \neq 1$.

These findings suggest then that the ANCOVA estimator will generally be preferred to the DID estimator, because the ANCOVA estimator will typically produce ATE estimates with smaller biases and smaller variances. Thus, in practice, the ANCOVA estimator will tend to produce estimators with smaller *MSE*s and *MDE*s, as quantified in the empirical analysis below.

Finally, for Design III (in which students are randomly assigned within sites), the regression model covariates include the student-level pretests y_{0ij} only (but not the school-level pretests). For this design, the asymptotic bias for $\hat{\gamma}_{ANCOVA}$ is $-\beta_1\lambda_{01}(\tau_1/\tau_0)$, and the asymptotic variance is:

$$AsyVar(\hat{\gamma}_{ANCOVA}) = \frac{1}{p(1-p)}\left[\frac{\tau_1^2(1-\lambda_{01}^2)}{nm}\right]\left[1+\frac{\beta_1^2 p(1-p)}{\tau_0^2}\right].$$

The Unbiased ANCOVA (UANCOVA) Estimator

The UANCOVA estimator is obtained using regression models where the model covariates include "true" baseline variables. This estimator is therefore asymptotically unbiased. I consider two categories of baseline covariates. The first category—which is the focus of the empirical analysis—includes baseline test score data on tests that are related to but not exactly the same as the posttest. These covariates could include school-level standardized test scores for prior cohorts of students in the study schools (who are similar to and in the same grades as the students in the study sample). If available, they could also include school records data from earlier grades for students in the study sample. These alternative baseline data are likely to have lower correlations with the posttests than the student-level pretests that are directly aligned to the posttests. Thus, they may reduce variance less. However, these baseline data are likely to be uncontaminated, and thus, will produce unbiased ATE estimators. They may also be less costly to collect.

The option of collecting alternative baseline data is most pertinent for designs in which schools are the unit of random assignment (Design I). Aggregate school-level data can be obtained from public sources or from school records as part of the evaluation data collection effort. It is usually more difficult to obtain longitudinal school records data for specific teachers and students. Thus, alternative baseline data may not always be a viable substitute for pretest data for designs in which the unit of random assignment is at the classroom level (Design II) or the student level (Design III). Accordingly, the empirical analysis for the UANCOVA estimator presented below focuses on school-based designs.

The second category of covariates includes basic student-level demographic baseline variables that pertain to the period prior to random assignment. Including these variables in posttest-only regression models typically yield R^2 values of about 0.10 to 0.20 (Schochet 2008). These covariates, however, typically yield only small marginal improvements in R^2 values if pretests are also included in the models (Schochet 2008). Thus, the empirical analysis for the ANCOVA and DID estimators ignore these covariates.

The asymptotic properties of the UANCOVA estimator $\hat{\gamma}_{UANCOVA}$ can be obtained by setting $\beta_1 = 0$ in the corresponding formulas for the ANCOVA estimator. Using this approach, we find that $\hat{\gamma}_{UANCOVA}$ has an asymptotic normal distribution with mean α_1 and the following asymptotic variance:

$$(16)\quad AsyVar(\hat{\gamma}_{UANCOVA}) = \frac{1}{p(1-p)}\left[\frac{\sigma_1^2(1-\rho_{U01}^2)}{n}+\frac{\tau_1^2(1-\lambda_{U01}^2)}{nm}\right],$$

where ρ_{U01}^2 and λ_{U01}^2 are, respectively, unit- and student-level correlations between the alternative baselines and the posttests. It is assumed that $\rho_{U01}^2 < \rho_{01}^2$ and $\lambda_{U01}^2 < \lambda_{01}^2$. If the covariates only include school-level aggregate test scores, then $\lambda_{U01}^2 = 0$.

If $\rho_{U01}^2 > 0$ or $\lambda_{U01}^2 > 0$, $\hat{\gamma}_{UANCOVA}$ will be more precise and have lower *MSE* and *MDE* values than $\hat{\gamma}_{Posttest}$. However, $\hat{\gamma}_{UANCOVA}$ will typically be less precise than $\hat{\gamma}_{ANCOVA}$ unless the regression R^2 values are very similar for the two models. For example, focusing only on the leading unit-level variance terms in (15) and (16), $\hat{\gamma}_{ANCOVA}$ will be more precise than $\hat{\gamma}_{UANCOVA}$ if $d < (1-\rho_{U01}^2)/(1-\rho_{01}^2)$, which will typically hold in practice unless ρ_{U01} and ρ_{01} are very similar. The empirical analysis below compares plausible *MSE* and *MDE* values for the two estimators.

The Generalized Estimating Equation (GEE) Estimator

An alternative analytic approach for adjusting for late pretests—that strays somewhat from the causal inference framework discussed above—is to model the growth in impacts as a function of time. The GEE estimator that is considered here involves the simultaneous estimation of the models in (6) and (7) where the unit early treatment effect (β_{1i}) is modeled as a function of the posttest impact (α_1) and the number of months between randomization and the pretest data collection date (t_i). Specifically, it is assumed that $\beta_{1i} = f(\alpha_1, [t_i / l_i])$, where l_i is the length of the follow-up period (for example, 10 months for a spring posttest), and $f(.)$ is a function that specifies how impacts grow over time (the next chapter discusses these functions in more detail).

Using this modeling approach, equation (7) can be rewritten as:

$$(17) \quad y_{0ij} = \beta_0 + f(\alpha_1, c_i)T_i + (u_{0i} + e_{0ij}),$$

where $c_i = (t_i / l_i)$ and $0 \le c_i \le 1$. The parameters in equations (6) and (17) can then be simultaneously estimated using GEE methods (Liang and Zeger 1986; Yang and Tsiatis 2001), which, for tractability, are discussed assuming that the data are aggregated to the unit level.

The GEE estimator for the parameter vector $\boldsymbol{\theta}' = (\alpha_1 \quad \alpha_0 \quad \beta_0)$ can be obtained as the solution to the following equations:

$$(18) \quad \sum_{i=1}^{n} \mathbf{D}_i' \boldsymbol{\Omega}_i^{-1} (\mathbf{z_i} - \hat{\mathbf{z}}_\mathbf{i}) = 0, \quad \text{where}$$

$$\mathbf{z_i} = \begin{pmatrix} z_{1i} \\ z_{0i} \end{pmatrix}, \ \hat{\mathbf{z}}_\mathbf{i} = \begin{pmatrix} \hat{\alpha}_0 + \hat{\gamma}_{GEE} T_i \\ \hat{\beta}_0 + f(\hat{\gamma}_{GEE}, c_i) T_i \end{pmatrix}, \ \boldsymbol{\Omega}_i = \begin{pmatrix} \sigma_1^2 & \sigma_{01} \\ \sigma_{01} & \sigma_0^2 \end{pmatrix}, \text{ and } \mathbf{D_i} = \frac{\partial z_i}{\partial \theta'} = \begin{pmatrix} T_i & 1 & 0 \\ T_i[\partial f(.)/\partial \alpha_1] & 0 & 1 \end{pmatrix}.$$

Because $\hat{\mathbf{z}}_\mathbf{i} = \mathbf{D_i} \hat{\boldsymbol{\theta}}$, the GEE estimator is as follows:

$$(19) \quad \hat{\boldsymbol{\theta}}_{\mathbf{GEE}} = \begin{pmatrix} \hat{\gamma}_{GEE} \\ \hat{\alpha}_0 \\ \hat{\beta}_0 \end{pmatrix} = \left(\sum_{i=1}^{n} \mathbf{D}_i' \hat{\boldsymbol{\Omega}}_\mathbf{i}^{-1} \mathbf{D_i}\right)^{-1} \sum_{i=1}^{n} \mathbf{D}_i' \hat{\boldsymbol{\Omega}}_\mathbf{i}^{-1} \mathbf{z_i},$$

where $\hat{\mathbf{\Omega}}_\mathbf{i}$ is a consistent estimator for the unknown $\mathbf{\Omega}_\mathbf{i}$.

As *n* approaches infinity, Liang and Zeger (1986) show that if $(\sum_{i=1}^{n} \mathbf{D}_i'\hat{\mathbf{\Omega}}_\mathbf{i}^{-1}\mathbf{D}_\mathbf{i})^{-1}$ exists, $\hat{\gamma}_{GEE}$ has an asymptotic normal distribution with mean α_1 and asymptotic variance $[E(\mathbf{D}_i'\mathbf{\Omega}_\mathbf{i}^{-1}\mathbf{D}_\mathbf{i})]^{-1}$. Thus, the GEE estimator is consistent, assuming that *f*(.) is specified correctly.

As an example, suppose that impacts grow linearly over time so that $f(\alpha_1, c_i) = \alpha_1 c_i$. In this case, it can be seen after some algebra that the asymptotic variance of $\hat{\gamma}_{GEE}$ is:

$$(20) \quad AsyVar(\hat{\gamma}_{GEE}) = \frac{1}{np(1-p)}\left[\frac{\sigma_0^2\sigma_1^2(1-\rho_{01}^2)}{\sigma_0^2 + \bar{c}^2\sigma_1^2 - 2\bar{c}\sigma_0\sigma_1\rho_{01}}\right],$$

where $\bar{c}$ is the mean c_i across units. If $c_i = 0$ for each unit (that is, if the pretests are collected before randomization), (20) reduces to the ANCOVA variance expression in (15) with $\beta_1 = 0$ (focusing on unit-level terms only). If $\bar{c} > 0$, (20) will be larger than if $\bar{c} = 0$ as long as $\bar{c} < 2\rho_{01}(\sigma_0 / \sigma_1)$ (which is likely to occur in practice), but the relative efficiency of the ANCOVA and GEE estimators will depend on specific parameter values.

As another example, suppose instead that impacts grow quadratically over time, so that $f(\alpha_1, c_i) = \alpha_1 c_i^2$. In this case, the asymptotic variance of $\hat{\gamma}_{GEE}$ can be obtained from (20) by replacing $\bar{c}$ with $\bar{c}^2$. More generally, (20) applies to impact growth functions that can be expressed as $f(\alpha_1, c_i) = \alpha_1 g(c_i)$, for some function $g(.)$, by replacing $\bar{c}$ with $g(\bar{c})$ in (20).

This GEE approach hinges critically on the correct function form specification for $f(\alpha_1, c_i)$, which is difficult to test. Thus, we do not include the GEE estimator in the empirical analysis, because it is difficult to quantify potential estimator biases. However, this approach is useful for testing the sensitivity of posttest impact findings to alternative estimation procedures.

HLM Growth Curve Approach

Finally, a somewhat related method to the GEE approach is to use an HLM growth curve approach to model student test scores as a function of the time between randomization and data collection. Under this approach, pretests are treated as dependent variables and stacked with the posttests for analysis. This yields a three-level HLM model, where Level 1 corresponds to time, Level 2 to students, and Level 3 to units. This approach is not considered here, because our focus is on designs with a single posttest, so that data on only two time points are available for each student, yielding zero available degrees of freedom for analysis. The growth curve approach would be more appropriate if additional longitudinal test score data were available, so that flexible function forms for the outcome-time relationship could be specified and tested.

Chapter 7: Empirical Analysis

This chapter calculates *MSE* and *MDE* values for the posttest-only, DID, ANCOVA, and UANCOVA estimators using simulations that are based on key parameter values that are found in the literature. The focus is on RCTs for education interventions that aim to improve achievement test scores of elementary school and preschool students in low-performing school districts.

Structure

To help structure and interpret the empirical analysis, it is assumed that the evaluation is designed to detect an intervention effect on spring achievement test scores of 0.15 to 0.30 standard deviation units. These targets are often used in large-scale RCTs in the education field and represent a reasonable compromise between evaluation rigor and evaluation cost (see Schochet 2008 and Hill et al. 2007). These standards are often justified based on what is realistically attainable from meta-analyses of impact findings from previous evaluations in related areas. These effect sizes can also be interpreted by noting that the test performance of young students in math and reading grows by about 0.70 standard deviations per grade (Schochet 2008). Thus, a standardized effect size of 0.25, for example, corresponds with roughly 3.5 months of instruction (assuming a regular 10-month school year).

The empirical results below hinge critically on the growth trajectory of test score impacts, which will partly determine the level of contamination in the late pretest scores. Figure 7.1 displays four hypothetical growth trajectories of test score impacts between the start and end of the school year, where the posttest impact at the end of the school year is expected to be 0.25 standard deviations (Table 7.1 displays monthly impact values for each scenario and formulas). The testing date distributions are assumed to be similar for treatments and controls.

The trajectory of impact growth will likely depend on a number of factors, including the nature of the outcome measure, the nature of the intervention, and the types of students under investigation. For instance, all else equal, initial impact growth is likely to be steeper for an intervention that can be implemented quickly (such as the use of a new textbook) than for an intervention that takes more time to implement (such as a whole-school reform), for more intensive than less intensive interventions, and for students who are more willing and able to grasp intervention components. While there is a large literature on the extent to which test scores grow over time, there is very little evidence on the extent to which *impacts* grow within a school year. Although more research is needed on this issue, it seems plausible that for many interventions, treatment effects on student achievement test scores are likely to grow gradually (linear growth; Panel A in Figure 7.1) or slowly at first but then more quickly after a critical mass of information has been administered and processed (quadratic growth; Panel B or C). Logarithmic growth (Panel D) seems less plausible for test score outcomes but may be more plausible for other educational outcomes (not considered in this paper) such as student behavior, teacher knowledge of specific intervention components, or student assessments that are directly aligned to intervention components.

Figure 7.1 and Table 7.1 suggest then that for many scenarios, intervention effects on late pretest scores will be a relatively small percentage of expected end-of-the-year intervention effects. For example, if the average pretest was collected 2 months after the start of the school year for each research group, the ratio of the pretest-to-posttest impact would be about 20 percent if impacts grew linearly (.05/.25) and 4 percent if impacts grew quadratically (.01/.25). These findings have important implications for the empirical analysis presented below.

Figure 7.1

Hypothetical Growth Trajectories of Test Score Impacts (Measured In Standard Deviation Units)

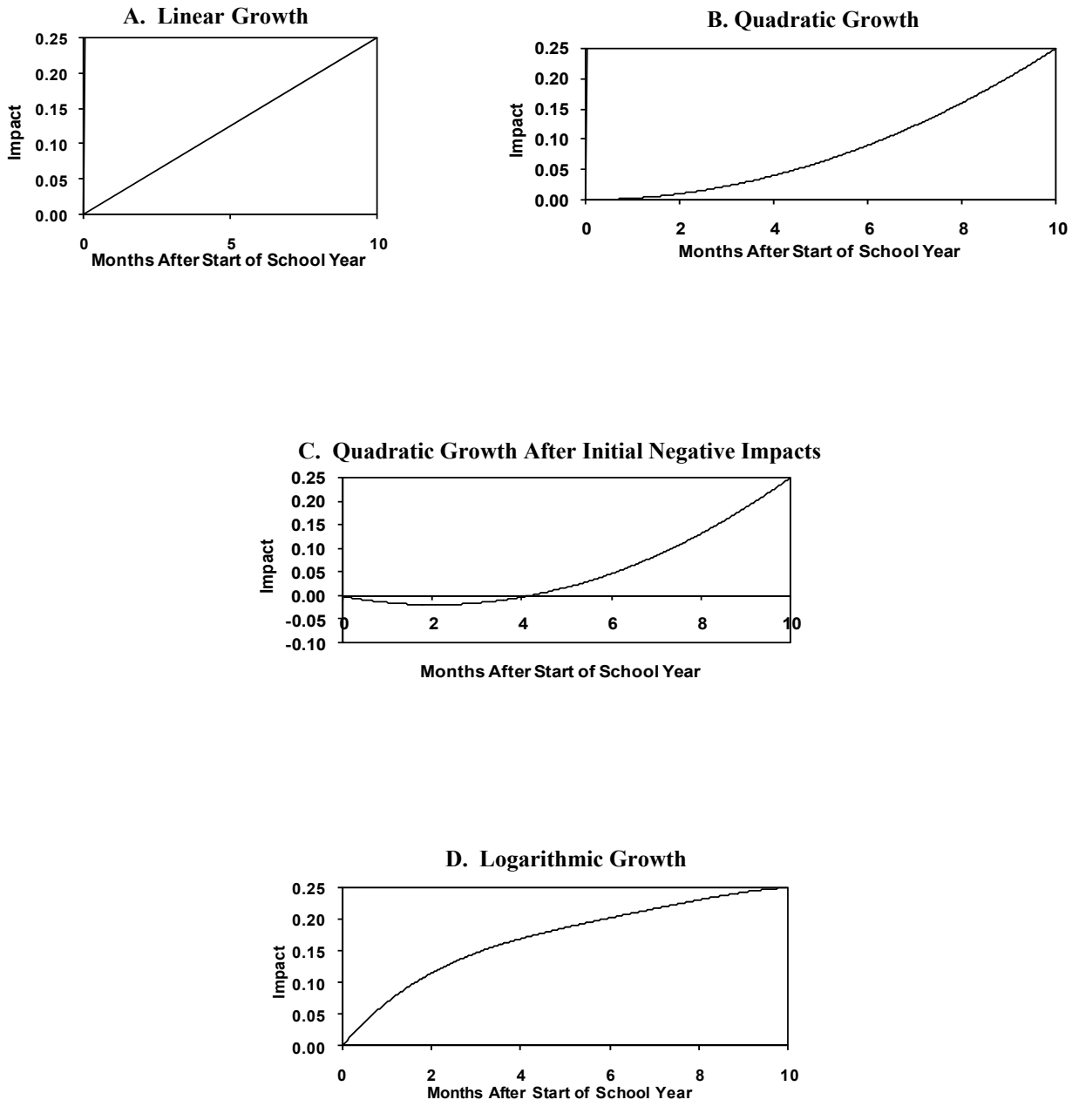

Table 7.1 : Hypothetical Growth Trajectories of Test Score Impacts, by the Number of Months Since the Start of School (Design I)

Months Since Start of School	Growth Trajectory of Test Score Impacts (Measured in Standard Deviation Units)[a]			
	Linear Growth	Quadratic Growth	Quadratic Growth After Initial Negative Impacts	Logarithmic Growth
0	0.00	0.00	0.00	0.00
1	0.03	0.00	-0.01	0.07
2	0.05	0.01	-0.02	0.11
3	0.08	0.02	-0.01	0.14
4	0.10	0.04	0.00	0.17
5	0.13	0.06	0.02	0.19
6	0.15	0.09	0.05	0.20
7	0.18	0.12	0.09	0.22
8	0.20	0.16	0.13	0.23
9	0.23	0.20	0.19	0.24
10	0.25	0.25	0.25	0.25

Note: Testing date distributions are assumed to be similar for the treatment and control groups.

[a] Linear growth assumes that $I(t) = .025t$ where *I(t)* is the impact in month *t*. Quadratic growth assumes that $I(t) = .0025t^2$; Quadratic growth after initial negative impacts assumes that $I(t) = .0069(t-4)^2$ for $t \geq 4$; Logarithmic growth assumes that $I(t) = .104\ln(t+1)$.

Assumptions

To keep the presentation manageable, the *MSE* and *MDE* calculations were performed using the following empirically-based assumptions:

1. *ICC values of 0.15 at the school and classroom levels for both the pretests and posttests (that is, $ICC_1 = ICC_0 = 0.15$)*. Schochet (2008), Hedges and Hedberg (2007), and Bloom et al. (2005) provide empirical evidence for these *ICC* values.

2. *Pretest-posttest squared correlations(R^2 values) equal to 0.50 and 0.70.* These R^2 values are typically found in the literature for achievement test scores of young children (Schochet 2008; Bloom et al. 2005). Thus, results are presented for $\rho_{01}^2 = \lambda_{01}^2 = 0.50$ and $\rho_{01}^2 = \lambda_{01}^2 = 0.70$. These correlations are likely to be larger if the pretests are conducted later rather than earlier, and thus, could be indexed by time. For simplicity, however, the calculations ignore this indexing.

3. *Pretest and posttest variances are equal (that is, $\sigma_1^2 = \sigma_0^2 = \sigma^2$ and $\tau_1^2 = \tau_0^2 = \tau^2$).* This restriction is based on an analysis of test score data from previous RCTs. For example, for the Teach for America (TFA) Evaluation (Decker and Glazerman 2004), the ratio of pretest-to-posttest variances on the Iowa Test of Basic Skills (ITBS) was 1.0 for reading and 0.90 for math for students in grades one to five. Similarly, for the New York City School Voucher

Experiment (Mayer et al. 2002), the corresponding variance ratio for ITBS scores was 1.06 for reading and 0.80 for math, and for the Evaluation of the Effectiveness of Educational Technology Interventions (Dynarski and Agodini 2003), the variance ratio for the Stanford Achievement Test (SAT) was 1.0 for first graders, 1.1 for fourth graders, and 1.2 for sixth graders. Furthermore, the pretest and posttest variances in these studies were very similar for the treatment and control groups (not shown). Finally, based on empirical evidence, it is assumed that $\theta^2 = \sigma^2 + \tau^2 = 15$.

4. *The evaluation includes a total of 40 or 60 schools.* These are typical sample sizes that are included in large-scale education RCTs where schools are the unit of random assignment. These sample sizes typically yield *MDEs* in the 0.15 to 0.30 range (Schochet 2008). Fewer schools (10 to 40), however, are considered for classroom- and student-level designs (Designs II and III), because these designs are less clustered than Design I and yield more precise estimates for a given school sample size.

5. *A 1:1 treatment-control split (that is, p=0.50).* A 1:1 split is a common design used in education RCTs because it yields the most precise impact estimates for a given sample size. Results are very similar for a 2:1 split (not shown).

6. *The intervention is being tested in a single grade with an average of 3 classrooms per school per grade and an average of 23 students per classroom.* It is assumed that 80 percent of students (or 55 students per school) in the baseline sample provide posttest data.

7. *A two-tailed test at 80 percent power and a 5 percent significance level for the MDE calculations.* These are typical assumptions that are used in statistical power calculations for education RCTs and yield a value of about 2.8 for *Factor(.)* in equation (3).

8. *The distributions of testing dates are similar for the treatment and control groups.* Although pretests are sometimes conducted slightly later in control sites than in treatment sites, most well-designed RCTs ensure that testing dates are spread evenly across the two research groups. For simplicity, the same testing date distribution is assumed for treatments and controls. Thus, it is assumed that late pretests could be contaminated by early treatment effects, but not by differences in testing dates across the two research groups.

9. *The covariates for the UANCOVA estimator include only aggregate school-level test scores for prior cohorts of students.* Thus, results are presented for Design I only and it is assumed that $\lambda_{U01}^2 = 0$ in equation (16) and that $\rho_{U01}^2 < \rho_{01}^2$.The UANCOVA results for R^2 values of 0.10 and 0.20 pertain also to an UANCOVA model where the covariates include basic student demographic data rather than baseline test scores.

All calculations were conducted using the asymptotic variance and bias formulas shown above (using an EXCEL spreadsheet). The calculations can easily be revised using alternative assumptions that may pertain to specific evaluations.

Empirical Results for Design I

Table 7.2 displays, for Design I, the *largest* early treatment effect measured in effect size units (that is, β_1 / θ) for which $\hat{\gamma}_{DID}$ and $\hat{\gamma}_{ANCOVA}$ yield smaller *MSE* and *MDE* values than $\hat{\gamma}_{Posttest}$ and $\hat{\gamma}_{UANCOVA}$.[4] The rows with R^2=0 pertain to the posttest-only estimator and the rows with R^2>0 pertain to the UANCOVA estimator. For example, if the sample contains 40 schools and *p*=0.50, $\hat{\gamma}_{ANCOVA}$ will yield a smaller *MSE* value than $\hat{\gamma}_{Posttest}$ if the early treatment effect is less than 0.127 standard deviations. Note that *MDEs* for the posttests for the four ANCOVA specifications are 0.26, 0.20, 0.21, and 0.16, respectively, which were calculated assuming uncontaminated pretest data (the usual power analysis approach for calculating appropriate sample sizes).

The first key finding from Table 7.2 is that under most reasonable assumptions about the growth trajectory of impacts and pretest administration dates, *the DID and ANCOVA estimators will typically be preferred to the posttest-only estimator (see rows with* R^2=0). *MSE* values will be smaller for $\hat{\gamma}_{DID}$ and $\hat{\gamma}_{ANCOVA}$ than $\hat{\gamma}_{Posttest}$ if early treatment effects are less than about 0.10 standard deviations (the figures are somewhat larger using the *MDE* criterion). Assuming an ultimate spring posttest impact of 0.25 standard deviations, this condition will hold if test score impacts grow linearly and the pretests are collected within about 4 months after the start of the school year, or if test score impacts grow quadratically and the pretests are collected within about 7 months (Table 7.1). Even under logarithmic growth, the DID and ANCOVA models will still be preferred if pretests are collected within about 2 months after the start of the school year (Table 7.1). The results are stronger using the *MDE* than *MSE* criterion, because the *MDE* criterion places more weight on the variance component and less weight on the bias component. The results also become stronger as R^2 values increase (and especially so using the *MDE* criterion).

The second main finding from Table 7.2 is that consistent with the theory presented above, *the ANCOVA estimator will typically be preferred to the DID estimator.* The ANCOVA estimator yields lower *MSE* and *MDE* values under our assumptions, because the ANCOVA estimator has both smaller bias in absolute value terms ($\beta_1\rho_{01}$ compared to β_1) and smaller variance (because the design effect *d* is very small and $\rho_{01}(\sigma_1 / \sigma_0) = \rho_{01} \neq 1$). Differences between the two estimators become larger as R^2 values decrease.

The third key finding is that $\hat{\gamma}_{ANCOVA}$ *will typically be preferred to* $\hat{\gamma}_{UANCOVA}$ *as long as test score impacts do not grow very quickly early in the school year.* This result is consistent with the theory presented above, and holds even if R^2 values for the UANCOVA model are a sizeable fraction of R^2 values for the ANCOVA model. For example, for 60 schools and R^2 values of 0.70 for the pretests and 0.50 for the alternative baselines, $\hat{\gamma}_{ANCOVA}$ will yield lower *MSE* and *MDE* values if β_1 / θ_0 is less than about 0.065 standard deviations. This condition will hold if the pretests are collected within 2 months after the start of the school year assuming linear impact growth and within 5 months after the start of the school year assuming quadratic impact growth (Table 7.1).

The somewhat surprising findings for the UANCOVA estimator are due to the importance of R^2 values in reducing variance in clustered RCTs. Loss function improvements due to modest increases in R^2 values tend to offset losses due to estimator biases and collinearity among the covariates.

[4] As discussed above, the *MDE* loss function criterion is pertinent only if it is assumed that $\beta_1 / \theta \geq 0$.

Table 7.2: Maximum Values of β_1 / θ_0 for Which the ANCOVA and DID Estimators Would be Preferred to the Posttest-Only and UANCOVA Estimators (Design I)

			Maximum Values of β_1 / θ_0			
	Model R^2 Values		*MSE* Criterion		*MDE* Criterion	
Number of Schools	ANCOVA and DID	Posttest-Only or UANCOVA[a]	ANCOVA	DID	ANCOVA	DID
40	0.5	0.0[a]	0.127	0.083	0.143	0.085
		0.1	0.115	0.073	0.121	0.068
		0.2	0.101	0.062	0.098	0.050
		0.3	0.086	0.048	0.073	0.032
		0.4	0.066	0.029	0.046	0.012
40	0.7	0.0[a]	0.128	0.106	0.188	0.154
		0.1	0.119	0.098	0.169	0.138
		0.2	0.110	0.090	0.149	0.120
		0.3	0.100	0.081	0.128	0.101
		0.4	0.089	0.072	0.106	0.082
		0.5	0.076	0.060	0.081	0.060
		0.6	0.060	0.046	0.054	0.037
60	0.5	0.0[a]	0.104	0.068	0.118	0.069
		0.1	0.094	0.060	0.100	0.055
		0.2	0.083	0.051	0.080	0.041
		0.3	0.070	0.040	0.060	0.026
		0.4	0.055	0.024	0.038	0.010
60	0.7	0.0[a]	0.105	0.086	0.155	0.126
		0.1	0.098	0.080	0.140	0.112
		0.2	0.090	0.074	0.123	0.098
		0.3	0.082	0.067	0.106	0.083
		0.4	0.073	0.059	0.087	0.067
		0.5	0.062	0.049	0.067	0.049
		0.6	0.049	0.038	0.045	0.030

Note: Testing date distributions are assumed to be similar for the treatment and control groups. See the text for formulas and assumptions underlying the calculations. The calculations assume p=0.50 and that schools are the unit of random assignment.

[a] The posttest-only estimator corresponds to rows with R^2=0 and the UANCOVA estimator corresponds to rows with R^2>0.

To demonstrate this point further, Table 7.3 displays estimated variances, squared biases, and *MSEs* for $\hat{\gamma}_{ANCOVA}$ and $\hat{\gamma}_{UANCOVA}$ assuming 60 schools and R^2 values of 0.70 and 0.50 for the ANCOVA and UANCOVA models, respectively. The variance of $\hat{\gamma}_{ANCOVA}$ (excluding the design effect *d*) is considerably smaller than the variance of $\hat{\gamma}_{UANCOVA}$. This occurs because the ANCOVA R^2 value is larger and affects both the school- and student-level variance terms rather than the school-level term only.[5] Furthermore, for most plausible values of β_1 / θ_0, the design effect *d* for $\hat{\gamma}_{ANCOVA}$ is very small and does not materially inflate the variance estimates. Finally, although the squared bias terms grow quickly, $\hat{\gamma}_{ANCOVA}$ still yields lower *MSE* values than $\hat{\gamma}_{UANCOVA}$ if $\beta_1 / \theta_0 < 0.062$. Thus, from a statistical standpoint, under many realistic scenarios about the growth of impacts and data collection schedules, the pretests will tend to yield estimates that are closer to the truth than the alternative baselines.

These findings provide statistical support for the collection and use of pretest data in education-related RCTs even if the pretests are likely to be collected several months late. It is important to realize, however, that the bias generated by late pretests will erode statistical power, and thus, if pretest data are to be collected, this bias should be taken into account in the statistical power calculations for the study. Table 7.4 demonstrates these power losses for $\hat{\gamma}_{ANCOVA}$ by displaying required school samples to maintain a fixed *MDE* value for various β_1 / θ_0 and R^2 values. For example, if R^2 = 0.50, an *MDE* of 0.228 standard deviations can be achieved with 50 schools if there is no bias ($\beta_1 / \theta_0 = 0$), but 65 schools are required if $\beta_1 / \theta_0 = 0.04$ and more than 100 schools are required if $\beta_1 / \theta_0 = 0.10$. Thus, to achieve desired precision targets, school sample sizes should be increased sufficiently to offset power losses associated with anticipated estimator biases due to contaminated pretest data.

Finally, because it may be difficult to anticipate β_1 / θ_0 values, some studies that collect pretest data may be interested in conducting statistical hypothesis tests to determine whether or not to include late pretests in the analytic models. It is difficult, however, to develop a suitable test for this analysis, because confidence intervals for pretest score impacts are likely to be wide (due to low R^2 values). For example, assuming 60 schools and other assumptions from above, the 95 percent confidence interval around the pretest impact is $\hat{\beta}_1 / \hat{\theta}_0 \pm 0.09$ standard deviations (assuming an R^2 value of 0.20 owing to the use of basic student demographic covariates). Thus, consider a reasonable testing strategy that would include pretests in the analysis only if the following null hypothesis is rejected: $H_0 : (\beta_1 / \theta_0) \geq 0.10$ (where the 0.10 cutoff is selected using results from above). This strategy is similar to the one used by Puma et al. (2005) for the Head Start Impact Study. Because of wide confidence intervals, however, this test would be rejected only if the estimated pretest impacts were very small (or negative). Thus, this approach is too conservative because it would too often exclude the pretests.

Instead, the results from above suggest that in education RCTs, if posttest impacts appear to be nonzero, a reasonable approach in practice would be to include pretests in the analysis unless the estimated impacts on the pretests were quite large (say, greater than 0.15 standard deviations) or if the pretest-posttest correlations were much lower than expected. This strategy is warranted because, as shown, statistical power gains can be achieved even if pretest impacts are a relatively large proportion of expected posttest impacts. It is prudent, however, to assess the robustness of study findings by comparing the impact findings from models with and without the pretests and using the GEE estimator discussed above under various functional form specifications for $f(\alpha_1, c_i)$.

[5] The variance for UANCOVA estimator is 551 if the R^2 is assumed to affect not only the school-level variance term but also the student-level variance term.

Table 7.3: Variance, Bias, and *MSE* Estimates for the ANCOVA and UANCOVA Estimators, for Various Values of β_1 / θ_0 (Design I)

Value of β_1 / θ_0	UANCOVA (R^2=0.50)[a]	ANCOVA (R^2=0.70)[a]			
	Variance = *MSE*[b]	Variance Excluding *d*[b]	Design Effect (*d*)	Bias[2b]	*MSE*[b,c]
0.00	603	331	1.0000	0	331
0.02	603	331	1.0007	28	359
0.04	603	331	1.0027	112	444
0.06	603	331	1.0060	252	585
0.08	603	331	1.0107	448	782
0.10	603	331	1.0167	700	1,036
0.12	603	331	1.0240	1,008	1,347

Note: Testing date distributions are assumed to be similar for the treatment and control groups. See the text for formulas and assumptions underlying the calculations.

[a] The figures assume a sample size of 60 schools and that schools are the unit of random assignment.

[b] The figures are multiplied by 10^5 and divided by θ_0^2.

[c] The *MSE* calculations were obtained using the following formula: (Variance Excluding *d**Design Effect)+Bias^2.

Table 7.4. School Sample Sizes Needed to Equate MDE Values for the ANCOVA Estimator, by the Size of the Early Treatment Effect (Design I)

Early Treatment Effect (β_1 / θ_0)	School Sample Sizes Needed to Achieve an *MDE* of:					
	0.255 (R^2=0.50)	0.197 (R^2=0.70)	0.228 (R^2=0.50)	0.177 (R^2=0.70)	0.208 (R^2=0.50)	0.161 (R^2=0.70)
No Bias: 0	**40**	**40**	**50**	**50**	**60**	**60**
0.02	45	48	57	61	69	75
0.04	51	59	65	76	81	96
0.06	58	73	76	97	95	127
0.08	67	93	89	129	114	177
0.10	78	123	107	180	140	262

Note: Testing date distributions are assumed to be similar for the treatment and control groups. See the text for formulas and assumptions underlying the calculations. The calculations assume p=0.50 and that schools are the unit of random assignment.

Empirical Results for Designs II and III

Table 7.5 displays figures, comparable to those in Table 7.2, that compare the posttest-only and ANCOVA estimators for Designs II and III.[6] The calculations assume R^2=0.50 for the ANCOVA model and that the sample includes 10 to 40 schools rather than 40 or 60 because Designs II and III are less clustered than Design I, and thus, can achieve similar power levels with fewer study schools.

The results for Designs II and III are very similar to those for Design I (Table 7.5). The ANCOVA estimator yields lower *MDE* and *MSE* values than the posttest-only estimator for most plausible assumptions about test score impact growth and pretest data collection schedules. The results are robust to the number of schools that are included in the evaluation.

Table 7.5. Maximum Values of β_1 / θ_0 for Which the ANCOVA Estimator Would be Preferred to the Posttest-Only Estimator (Designs II and III)

Number of Schools	*MSE* Criterion	*MDE* Criterion	*MDEs for the ANCOVA Estimator Assuming Uncontaminated Pretests*
Design II: Classrooms Are the Unit of Random Assignment			
10	0.158	0.175	*0.320*
20	0.113	0.128	*0.227*
30	0.093	0.106	*0.185*
40	0.080	0.092	*0.160*
Design III: Students Are the Unit of Random Assignment			
10	0.078	0.091	*0.155*
20	0.055	0.064	*0.110*
30	0.045	0.053	*0.090*
40	0.039	0.046	*0.078*

Note: Testing date distributions are assumed to be similar for the treatment and control groups. See the text for formulas and assumptions underlying the calculations. The unit of random assignment is at the classroom-level for Design II and at the student-level for Design III. The calculations assume R^2=.50 for the ANCOVA model.

[6] This chapter considers only the ANCOVA estimator because, as discussed, it is preferred to the DID estimator.

Chapter 8: Summary and Conclusions

This paper has examined theoretical and empirical issues related to the inclusion of late pretests in posttest impact models for clustered RCT designs in a school setting. The inclusion of late pretests will increase the precision of the estimated posttest impacts but could also introduce bias. Accordingly, the theoretical work examined, using a loss function approach, the conditions under which these biased estimators will produce impact estimates that are likely to be closer to the truth than unbiased estimators that either exclude the pretests or use uncontaminated test score data from other sources. The empirical work quantified the variance-bias tradeoffs for several commonly-used impact estimators.

The first research question that the paper addressed is: Under what conditions should late pretest data be collected and included in the posttest impact models? The answer to this question is clear: From a loss function perspective, estimators that include late pretests will typically be preferred to estimators that exclude them. This finding is supported by both the theoretical and empirical work, and will hold under most reasonable assumptions about the growth trajectory of impacts and pretest collection dates. In particular, the two most common pretest-posttest estimators—the DID and ANCOVA estimators—will typically yield smaller loss function values than the posttest-only estimator. This remains true even if the early treatment effect is a relatively large fraction of the expected posttest impact, and for designs in which schools, classrooms, or students are the unit of random assignment.

Another analysis finding is that the ANCOVA estimator will typically have smaller biases and smaller variances than the more restrictive DID estimator. Thus, the ANCOVA approach will often be preferred to the DID approach, because it will generate estimators with smaller loss function values.

The second research question that this paper addressed is: If pretest data are to be collected in education RCTs, what are statistical power losses when late pretests are included in the estimation models? The answer is that relative to a design with uncontaminated pretests, power losses with late pretests can be large, even if pretest contamination is modest. Thus, school sample sizes for RCTs in the education field should be increased to offset power losses if pretest data are expected to be collected several months after the start of the school year.

The final research question that this paper addressed is: Instead of collecting pretest data, is it preferable to collect uncontaminated baseline test score data from alternative sources? The answer is generally "no." Under the assumption that R^2 values for these alternative test scores are somewhat smaller than those for the pretests, the ANCOVA estimator will tend to dominate the UANCOVA estimator as long as the growth in test score impacts do not grow very quickly early in the school year. These somewhat surprising results hold because even relatively small increases in R^2 values will likely offset estimator biases and variance increases due to the collinearity of the model covariates.

The results comparing the ANCOVA and UANCOVA estimators, however, will not hold if R^2 values using school records and pretest data are similar. Bloom et al. (2005) and Cook et al. (2008) provide preliminary evidence that aggregate school-level R^2 values using school records data can be large, but this issue has not been systematically explored in the literature. Thus, comparing R^2 values using pretest and school records data is an important area for future research. Another important future research topic is to examine the relative costs of obtaining the two types of data. To the extent that school records data are cheaper to collect than pretest data, the UANCOVA estimator could be preferred to the ANCOVA estimator if the loss functions account not only for variance and bias, but also for data collection costs.

Another important issue that affects the findings is the growth trajectory of test score impacts over the school year. Although it is reasonable to assume that impacts grow linearly (the most agnostic assumption) or quadratically, there may be contexts where test score impacts grow very quickly and then

level off. In these instances, the biased estimators may perform worse than the unbiased ones. To obtain a base of knowledge about actual patterns of impact growth, future studies could be designed to administer tests at several points throughout the school year.

Finally, the methods developed in this paper could also be applied to examine the late pretest problem for RCTs in fields other than education. The main conclusions presented here, however, could differ in other contexts due to differences in the growth trajectory of treatment effects, the timing of pretest data collection, pretest-posttest correlations, and other key parameter values.

Appendix A: Proof of Asymptotic Results for the ANCOVA Estimator

Lemma 1. Let $\hat{\gamma}_{ANCOVA}$ be the OLS estimator for γ in the two-level model in (13). Then, as the number of units, n, increases to infinity and for fixed m, $\hat{\gamma}_{ANCOVA}$ converges to a normal distribution with mean $\alpha_1 - \beta_1(\sigma_{01}/\sigma_0^2)$ and the following asymptotic variance:

$$(A.1) \quad AsyVar(\hat{\gamma}_{ANCOVA}) = \frac{1}{p(1-p)}\left[\frac{\sigma_1^2(1-\rho_{01}^2)}{n} + \frac{\tau_1^2(1-\lambda_{01}^2)}{nm}\right]\left[1 + \frac{\beta_1^2 p(1-p)}{\sigma_0^2}\right]$$

Proof: It is convenient to express (6) and (7) in terms of centered random variables:

$$(A.2) \quad y_{1ij}^* = \alpha_1 T_i^* + (u_{1i}^* + e_{1ij}^*)$$
$$(A.3) \quad y_{0ij}^* = \beta_1 T_i^* + (u_{0i}^* + e_{0ij}^*),$$

where $T_i^* = T_i - p$, $y_{kij}^* = y_{kij} - E(y_{kij})$, $u_{ki}^* = u_{ki} - E(u_{ki})$ and $e_{kij}^* = e_{kij} - E(e_{kij})$ for $k = 0,1$. Let $\tilde{y}_{kij}$, $\tilde{T}_i$, $\tilde{u}_{ki}$ and $\tilde{e}_{kij}$ be respective *empirically* centered variables. Furthermore, let $\mathbf{X_{ij}} = [T_i \ \bar{y}_{0i} \ y_{0ij}^w]$, and $\mathbf{X_{ij}^*}$ and $\tilde{\mathbf{X}}_{\mathbf{ij}}$ be associated centered row vectors of model covariates. Finally, let $\mathbf{y_1^*}$ denote the vector of y_{1ij}^* values, $\mathbf{X}^*$ denote the matrix of X_{ij}^* values, and $\boldsymbol{\delta}' = (\gamma \ \delta_1 \ \delta_2)$ denote the parameter vector in (13).

As n approaches infinity (for fixed m) the OLS estimator $\hat{\boldsymbol{\delta}}$ in (13) converges to the following vector:

$$(A.4) \quad \hat{\boldsymbol{\delta}} = \left[\sum_{i=1}^{n}\sum_{j=1}^{m}\tilde{\mathbf{X}}_{\mathbf{ij}}'\tilde{\mathbf{X}}_{\mathbf{ij}}/nm\right]^{-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\tilde{\mathbf{X}}_{\mathbf{ij}}'\tilde{y}_{1ij}/nm\right] \xrightarrow{p} \left[E(\mathbf{X}_{\mathbf{ij}}^{*\prime}\mathbf{X}_{\mathbf{ij}}^*)\right]^{-1}\left[E(\mathbf{X}_{\mathbf{ij}}^{*\prime}y_{1ij}^*)\right].$$

By inserting (*A*.2) and (*A*.3) into (*A*.4), it can be shown that:

$$(A.5) \quad \left[E(\mathbf{X}_{\mathbf{ij}}^{*\prime}\mathbf{X}_{\mathbf{ij}}^*)\right]^{-1} = \frac{1}{p(1-p)\sigma_0^2\tau_0^2}\begin{bmatrix} [\sigma_0^2 + \beta_1^2 p(1-p)]\tau_0^2 & -\beta_1 p(1-p)\tau_0^2 & 0 \\ -\beta_1 p(1-p)\tau_0^2 & p(1-p)\tau_0^2 & 0 \\ 0 & 0 & p(1-p)\sigma_0^2 \end{bmatrix},$$

and

$$(A.6) \quad E(\mathbf{X}_{\mathbf{ij}}^{*\prime}y_{1ij}^*) = \begin{bmatrix} \alpha_1 p(1-p) \\ \alpha_1\beta_1 p(1-p) + \sigma_{01} \\ \lambda_{01} \end{bmatrix}.$$

It follows then that:

$$(A.7)\quad \hat{\gamma}_{ANCOVA} \xrightarrow{p} \gamma = \alpha_1 - \beta_1(\sigma_{01}/\sigma_0^2),$$
$$\hat{\delta}_1 \xrightarrow{p} \delta_1 = (\sigma_{01}/\sigma_0^2), \text{ and}$$
$$\hat{\delta}_2 \xrightarrow{p} \delta_2 = (\lambda_{01}/\tau_0^2).$$

To obtain the asymptotic distribution of the two-level OLS estimator, we rewrite the right-hand-side of (*A*.4) as follows:

$$(A.8)\quad \sqrt{nm}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) = \sqrt{nm}\left[\left[E(\mathbf{X}_{\mathbf{ij}}^{*/}\mathbf{X}_{\mathbf{ij}}^{*})\right]^{-1}\frac{\mathbf{X}^{*/}(\mathbf{y}_{\mathbf{1}}^{*} - \mathbf{X}^{*/}\boldsymbol{\delta})}{nm}\right] + o_p(1),$$

where $o_p(1)$ denotes a vector that converges in probability to zero. Because $E\left[\mathbf{X}^{*/}(\mathbf{y}_{\mathbf{1}}^{*} - \mathbf{X}^{*/}\boldsymbol{\delta})\right] = \mathbf{0}$, a simple application of the central limit theorem (see, for example, Rao 1973) can be used to show that $\hat{\boldsymbol{\delta}}$ is asymptotically normally distributed with the following variance:

$$(A.9)\quad AsyVar(\hat{\boldsymbol{\delta}}) = \left[E(\mathbf{X}_{\mathbf{ij}}^{*/}\mathbf{X}_{\mathbf{ij}}^{*})\right]^{-1} E[\mathbf{X}^{*/}(\mathbf{y}_{\mathbf{1}}^{*} - \mathbf{X}^{*/}\boldsymbol{\delta})(\mathbf{y}_{\mathbf{1}}^{*} - \mathbf{X}^{*/}\boldsymbol{\delta})'\mathbf{X}^{*}]\left[E(\mathbf{X}_{\mathbf{ij}}^{*/}\mathbf{X}_{\mathbf{ij}}^{*})\right]^{-1}.$$

The variance in (*A*.1) then follows using formulas from above and additional algebra to calculate the expectations in (*A*.9).

References

Allison, P. (1990). Change Scores as Dependent Variables in Regression Analysis. In *Sociological Methodology* (20). Edited by C. Clogg. Oxford, UK: Blackwell, 93-114;

Bloom, H., L. Hayes, and A. Black (2005). Using Covariates to Improve Precision. New York, NY: MDRC.

Byrk, A. and S. Raudenbush (1992). *Hierarchical Linear Models for Social and Behavioral Research. Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

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

Cohen, J. (1988). *Statistical Power Analysis for Behavioral Sciences.* Hillside, NJ: Lawrence Erlbaum.

Cook, T. et al. (2008). Impacts of School Improvement Status on Students with Disabilities. Technical Work Group Materials. Washington, DC: American Institutes for Research.

Davidian, M., A. Tsiatis, and S. Leon (2005). Semiparametric Estimation of Treatment Effect in a Pretest-Posttest Study with Missing Data. *Statistical Science*, 20(3), 261-301.

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

Donner, A. and N. Klar (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Dynarski, M. and R. Agodini (2003). The Effectiveness of Educational Technology: Issues and Recommendations for the National Study. Princeton, NJ: Mathematica Policy Research.

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

Gleason, P. and R. Olsen (2004). Impact Evaluation of Charter School Strategies. Design Documents. Princeton, NJ: Mathematica Policy Research, Inc.

Heckman, J. and E. Vytlacil (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3), 669-738.

Hedges, L. (2004). Correcting Significance Tests for Clustering. Chicago, IL: University of Chicago Working Paper.

Hedges, L. and E. Hedberg (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis,* 29(1), 60-87.

Hill, C., H. Bloom, A. Black, and M. Lipsey. Empirical Benchmarks for Interpreting Effect Sizes in Research. New York, NY: MDRC.

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

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

Jackson, R. et al. (2007). National Evaluation of Early Reading First. Final Report to Congress. U.S. Department of Education, Institute of Education Sciences: Washington DC.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

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

Mayer, D., P. Peterson, D. Myers, C. Tuttle, and W. Howell. (2002). School Choice in New York City: An Evaluation of the School Choice Scholarships Program. Washington, DC: Mathematica Policy Research, Inc.

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

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments: Essay on Principles. Chapter 9, Translated in *Statistical Science*, 1990: 5(4), 465-472.

Oakes, J and H. Feldman (2001). Statistical Power for Nonequivalent Pretest-Posttest Designs: The Impact of Change-Score Versus ANCOVA Models. *Evaluation Review*, 25(3), 3-28.

Puma, M. et al. (2005). Head Start Impact Findings: First Year Findings. Final Report to the U.S. Department of Health and Human Services, Administration for Children and Families: Washington DC.

Rao, C. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley and Sons.

Raudenbush, S. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. *Psychological Methods*, 2(2), 173-185.

Reichardt, C. (1979). *Quasi-Experimentation: Design and Analysis for Field Settings*. Boston: Houghton Mifflin.

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. (2007). Is Regression Adjustment Supported by the Neyman Model for Causal Inference?. Working Paper: Mathematica Policy Research, Inc.: Princeton NJ.

Schochet, P. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62-87.

Yang, L. and Tsiatis, A. (2001). Efficiency Study of Estimators for a Treatment Effects in a Pretest-Posttest Trial. *American Statistician* 55(4), 314-321.