NCEE 2010-4003

U.S. DEPARTMENT OF EDUCATION

Precision Gains from Publically Available School Proficiency Measures Compared to Study-Collected Test Scores in Education Cluster-Randomized Trials

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

Precision Gains from Publically Available School Proficiency Measures Compared to Study-Collected Test Scores in Education Cluster-Randomized Trials

October 2010

John Deke
Lisa Dragoset
Ravaris Moore
Mathematica Policy Research

Abstract

In randomized controlled trials (RCTs) where the outcome is a student-level, study-collected test score, a particularly valuable piece of information is a study-collected baseline score from the same or similar test (a pre-test). Pre-test scores can be used to increase the precision of impact estimates, conduct subgroup analysis, and reduce bias from missing data at follow up. Although administering baseline tests provides analytic benefits, there may be less expensive ways to achieve some of the same benefits, such as using publically available school-level proficiency data. This paper compares the precision gains from adjusting impact estimates for student-level pre-test scores (which can be costly to collect) with the gains associated with using publically available school-level proficiency data (available at low cost), using data from five large-scale RCTs conducted for the Institute of Education Sciences. The study finds that, on average, adjusting for school-level proficiency does not increase statistical precision as well as student-level baseline test scores. Across the cases we examined, the number of schools included in studies would have to nearly double in order to compensate for the loss in precision of using school-level proficiency data instead of student-level baseline test data.

NCEE 2010-4003
U.S. DEPARTMENT OF EDUCATION

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

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 assess the precision gains from school proficiency measures in randomized controlled trials. The views expressed in this report are those of the authors and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

October 2010

This report is in the public domain. While permission to reprint this publication is not necessary, the citation should be:
Deke, John, Dragoset, Lisa, and Moore, Ravaris (2010). *Precision Gains from Publically Available School Proficiency Measures Compared to Study-Collected Test Scores in Education Cluster-Randomized Trials* (NCEE 2010-4003). 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

There are three authors for this report with whom IES contracted to develop the discussion of issues presented. Drs. John Deke and Lisa Dragoset and Mr. Ravaris Moore are employees of Mathematica Policy Research, Inc. (Mathematica). The authors and other 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 compares the precision gains from adjusting for a study-collected pre-test score (which can be costly to collect) with the gains associated with publically available school proficiency data (available at low cost from extant sources), using data from five large-scale RCTs conducted by Mathematica Policy Research for the Institute of Education Sciences (IES). The study finds that, on average, adjusting for school-level proficiency does not increase statistical precision as well as study-collected baseline test scores. Across the cases examined, the number of schools included in studies would have to nearly double in order to compensate for the loss in precision of using proficiency data instead of study-collected baseline test data.

Acknowledgements

The authors would like to thank the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences (IES), U.S. Department of Education for supporting this work. We also gratefully acknowledge the review and comments provided by several IES staff and the members of the IES Methods Working Group.

The authors would like to thank Mathematica Policy Research, Inc. staff members, including Dr. Peter Schochet for his careful review of the draft document. We also thank Dr. Thomas Cook of Northwestern University for his thoughtful review and helpful input.

The views expressed herein are those of the authors and do not reflect the policies or opinions of the U.S. Department of Education. Any errors or omissions are the responsibility of the authors.

We appreciate the willingness of reading developers to engage in a large-scale, rigorous evaluation and to contribute their perspectives and insights during interviews. We could not have conducted this study without the districts, schools, and teachers who agreed to participate in the study, use the reading curricula, permit observation of their classroom instruction, and share their views.

Contents

List of Tables and Figures

Chapter 1: Introduction

Over the past 10 years, evaluators of educational programs have increasingly used randomized controlled trials (RCTs) to estimate causal relationships between interventions in education and student outcomes, particularly in the federal studies funded by the Institute of Education Sciences (IES). In education studies where the outcome is a student-level, study-collected test score, a particularly valuable piece of information is a study-collected baseline score from the same test (a pre-test). Pre-test scores can be used to increase the precision of impact estimates, which in turn reduces the cost of a study by enabling researchers to detect effects of a similar size with smaller sample sizes. Pre-tests can also be used to conduct subgroup analysis to understand how impacts vary by student achievement level, and to reduce bias from missing data at follow up, either through regression adjustment or imputation of missing data. See James-Burdumy et al. (2009) for an example of how study-collected baseline test scores can be used in these ways. In addition, the presence of study staff in schools at the beginning of the year provides opportunities to learn about school-specific issues that might affect the study in unanticipated ways.

Although administering baseline tests reduces study costs and provides analytic benefits, there may be less expensive ways to achieve some of the same benefits. One way might be to rely on district-administered tests for all the students in the study. In districts that provide easy access to student-level test data, this could be a cost-effective alternative. However, experience on large nationwide studies conducted for IES has shown that there is considerable variation in the accessibility of school-district data and the willingness of school districts to cooperate with data-collection efforts. Since school districts are included in evaluations based on whether they meet study eligibility requirements (for example, a willingness to implement an intervention and allow random assignment) and not on the accessibility of their data, researchers conducting these studies often find that collecting test score data from school districts can be challenging.

An even less expensive alternative to baseline test administration that does not require researchers to request data from local districts is to use publically available school-level proficiency data. The State Education Data Center (SEDC), by way of the schooldatadirect.org website, provides a central data source for school-level proficiency data covering all school districts in the United States for which data are available. If the school-level proficiency data from this single, centralized source were nearly as effective at increasing the statistical precision of experimental impacts it could lead to significant cost reduction in studies conducted for IES.

This paper compares how publically available school-level proficiency data from SEDC perform relative to study-collected pre-test data in terms of improving the precision of impacts in RCTs in which schools (rather than students) are randomly assigned to treatment and control groups and a study-collected post-test score is the outcome of interest.[1] Prior research has demonstrated the large precision gains that are possible in these RCTs from covariate adjustment for pre-tests, particularly at the school level (Raudenbusch 1997; Hedges and Hedberg 2007; Bloom et al. 2007; Schochet 2008a). In fact, Bloom et al. (2007) find that "the precision-enhancing power of pretests declines only slightly as the number of years between the pretest and posttests increases; improves only slightly with pretests for more than one baseline year; and is substantial even when the pretest differs from the posttest," all of which suggest that proficiency data (which is from a different test and often an earlier year) might perform well. The unique contribution of this paper is to compare the precision gains from adjusting for a study-collected pre-test score (which can be costly to collect) with the gains associated with publically available school proficiency data (available at low cost from extant sources). This study is strictly empirical and does not

[1] For the purposes of this paper, we assume that the outcome of interest is a student-level study-collected post-test score, as opposed to a school-level proficiency measure. However, this may not always be the case. See Perez-Johnson et al. (2009) for a discussion of considerations for choosing appropriate outcomes in an educational RCT.

attempt to provide a conceptual model for the precision gains associated with different types of pre-test variables. This paper focuses primarily on elementary schools. This paper does not address the approach of collecting student-level data from school districts.

We compare the precision gains from using school proficiency data to the gains from using study-collected pre-test data based on data from five large-scale RCTs conducted by Mathematica Policy Research for the Institute of Education Sciences (IES), many of which included multiple post-tests and pre-tests. Specifically, we examine the average precision loss from adjusting for school proficiency data instead of using pre-test data and the risks for an individual study of experiencing a much larger precision loss when using proficiency data. We also express this loss in terms of the number of schools that a study would need to add in order to make up for the loss in precision associated with using school proficiency data instead of study-collected pre-test data.

If school-level proficiency data are an acceptable alternative to study-collected data for the purpose of improving statistical precision, it might still be necessary for studies to collect student-level data in order to adjust for non-response bias. To investigate whether non-response bias is a serious concern for IES-funded studies, we assess the extent to which different types of attrition bias appear to exist in past education RCTs.

The paper is organized as follows. In Chapter 2 we discuss the parameters of interest in this study. In Chapter 3 we describe our data sources. In Chapter 4 we compare the precision gains associated with study-collected student test score data to the gains associated with school-level proficiency data. In Chapter 5 we discuss additional analyses intended to better understand the findings of Chapter 4 and explore alternative approaches to reducing the costs of collecting baseline test scores. In Chapter 6 we examine the potential implications for attrition bias of not collecting pre-test data (since pre-test data are often used in imputation or regression adjustment strategies intended to reduce attrition bias). Chapter 7 concludes.

Chapter 2: Minimum Detectable Effects in Clustered Randomized Controlled Trials

In this chapter we define the key parameters that we estimate in this paper and describe how we estimate them. The key parameters are the (1) intraclass correlation coefficient (ICC), (2) student-level and school-level R^2 values, and (3) minimum detectable effect (MDE).

We use a super-population framework for thinking about the distribution of impacts. We assume that treatment schools are sampled from a normally distributed population with mean m_T and variance σ_B^2 and that control schools are sampled from a normally distributed population with mean m_C and variance σ_B^2. Within schools, we assume that students are sampled from a normally distributed population with a school-specific mean of $\overline{y_s}$ and a variance of σ_W^2. The impact is the difference between the mean outcome in the treatment group and the mean outcome in the control group. We assume that each school has the same number of students and that equally sized samples are drawn from each school. The number of schools in the treatment and control groups is allowed to differ.

For a variable that is observed for a population of students clustered within schools, the ICC is the proportion of the total variance that is due to variation between schools. Following the notation in Hedges and Hedberg (2007), if the total variance of the post-test score can be decomposed into within (student-level) and between (school-level) terms $\sigma_T^2 = \sigma_B^2 + \sigma_W^2$, then the ICC is simply $\rho = \sigma_B^2 / \sigma_T^2$. In this paper we estimate the terms σ_B^2 (school-level variance) and σ_W^2 (student-level variance) using the linear mixed effects (lme4) package (Bates 2008) in R without covariate adjustment.

Reducing both σ_B^2 and σ_W^2 through covariate adjustment can increase statistical precision of impacts in RCTs where schools are the unit of random assignment. Again following the notation in Hedges and Hedberg (2007), we define the population school-level R^2 as $R_B^2 = 1 - \sigma_{AB}^2 / \sigma_B^2$ and the population student-level R^2 as $R_W^2 = 1 - \sigma_{AW}^2 / \sigma_W^2$, where σ_{AB}^2 and σ_{AW}^2 are the residual between and within variance terms after covariate adjustment. We estimate σ_{AB}^2 and σ_{AW}^2 using the lme4 package in R with covariate adjustment.

The ICC, student-level R^2, and school-level R^2 values can be used to calculate the minimum detectable effect, which is the smallest program impact that can be detected with a high probability. We define the program impact as the difference between the mean outcome in the treatment group and the mean outcome in the control group. The MDE formula is:

$$(1)\ MDE = \left[T^{-1}\left(1-\alpha/2, df\right) + T^{-1}\left(\beta, df\right) \right] * \sqrt{Var(impact)}$$

where T is the cumulative density function of the t-distribution, α is the probability of a type 1 error, β is the probability of a type 2 error, and df is the number of degrees of freedom. The formula for the variance of the impact is:

$$(2)\ Var(impact) = \frac{\sigma_{AB}^2}{N_{sch} * p * (1-p)} + \frac{\sigma_{AW}^2}{N_{sch} * p * (1-p) * N_{stu}}$$

where N_{Sch} is the total number of schools in the study, p is the proportion of schools in the treatment group, and N_{stu} is the number of students in each school. The residual between and within variance terms

after covariate adjustment, σ^2_{AB} and σ^2_{AW}, can be expressed in terms of the ICC, student-level R^2, and school-level R^2 values as follows:

$$(3)\ \sigma^2_{AB} = \sigma^2_T \rho(1-R^2_B)$$
$$(4)\ \sigma^2_{AW} = \sigma^2_T (1-\rho)(1-R^2_W)$$

Plugging (2)-(4) into (1) gives the formula for the covariate-adjusted MDE:

$$(5)\ MDE = \left[T^{-1}(1-\alpha/2, df) + T^{-1}(\beta, df)\right] * \sigma_T \sqrt{\frac{\rho(1-R^2_B)}{N_{sch} * p * (1-p)} + \frac{(1-\rho)(1-R^2_W)}{N_{sch} * p * (1-p) * N_{stu}}}.$$

See Murray (1998) and Bloom (2004) for derivations of these formulas.

Dividing the MDE by σ_T yields the minimum detectable effect size (MDES), which will be our primary focus (instead of the MDE). A smaller MDES is more desirable than a larger MDES; studies often target an MDES of 0.20 or 0.25.

In practice, the school-level R^2 has a much greater effect on the MDE than the student-level R^2. In this study, we chose to focus strictly on the school-level R^2 because the precision gains from a student-level R^2 are generally quite small.[2] For example, if we have a sample size of 40 schools and 1,800 students (equally divided among the schools), an ICC of 0.15, a school-level R^2 of 0.5, and a student-level R^2 of 0.5, then the MDE would be 0.26. Reducing the school-level R^2 to zero would increase the MDE to 0.36 but reducing the student-level R^2 to zero would only increase the MDE to 0.28.

[2] For this reason, Spybrook et al. (2009) do not take the student-level R^2 into account at all in their optimal design software.

Chapter 3: Data

Student Test Score Data

This study uses baseline and follow-up student level test score data from five large-scale experimental studies previously conducted for IES by Mathematica.[3] See Table 3.1 for a brief description of each study, including the grades covered. Together, these studies yield 25 separate test score outcomes covering a total of 30,000 students drawn from kindergarten to grade 9 in 27 states and 500 schools. All but one of the studies included only elementary schools. Baseline and follow-up tests were created by a diverse set of test developers to measure a range of skills related to reading and mathematics proficiency for students of different ages. In the case of the Evaluation of Teacher Induction Programs, student pre- and post-test scores were collected from school districts' administrative records. For the other five studies, the pre- and post-tests were selected and administered by the study.

For each study, we have the following data. First, we have student-level baseline and follow-up test scores (some studies have more than one of each, because students were tested in multiple subject areas). For all of the studies, we analyzed one year of data, meaning that the baseline test was conducted at the beginning of the school year and the follow-up test was conducted at the end of that same school year. None of the studies suffer from attrition at the school level. However, most studies experience attrition at the student level. We use the study-collected baseline and follow-up tests in their original form, and we also aggregate up to school-level means for some analyses. We also have each school's National Center for Education Statistics School identification number. This unique national identifier was used to link schools to external data sources. For several studies, we have a classroom identification variable. Finally, since all of our studies are randomized controlled trials, we have treatment assignment variables.

School-Level Proficiency Data and Constructed Variables

School-level proficiency data were retrieved from the State Education Data Center (SEDC) by way of the schooldatadirect.org website. We chose this data source because it is a single source for school-level proficiency data from school districts across the United States that could be easily used by future evaluations with no cost beyond downloading and merging the data. SEDC is funded by a nonprofit organization with the aim of (1) advocating for quality education data collection, standards, and use, and (2) serving the U.S. as a free provider of state education data and analytical tools. The site offers data on most U.S. schools including student demographics, economic characteristics of the surrounding community, and proficiency rates for the school overall as well as by grade level. SEDC proficiency data are obtained from individual states, and proficiency rates are based on state-specific proficiency standards. Specifically, the proficiency rate is reported at the school-level and is the proportion of students in the school that are deemed "proficient" using the state's definition of proficiency. In some cases, multiple categories of proficiency are available in SEDC, but for consistency in variable construction across states we always used the binary categorizations "proficient" and "not proficient".

Researchers considering the use of SEDC school-level proficiency data as an alternative to collecting student-level baseline test data should be aware of the following potential limitations. First, there is limited middle- and high-school data in the SEDC database. Second, proficiency rates by grade level are less commonly available than school-wide proficiency rates. Similarly, proficiency rates for student

[3] These data were collected by Mathematica on behalf of IES and are available as restricted use files. The five studies were: (1) the Evaluation of Reading Comprehension Interventions (James-Burdumy et al. 2010), (2) the Evaluation of the Impact of Teacher Induction Programs (Isenberg et al. 2009), (3) the Impact Evaluation of Teacher Preparation Models (Constantine et al. 2009), (4) the Evaluation of Mathematics Curricula (Agodini et al. 2009), and (5) the Evaluation of Educational Technology Interventions (Campuzano et al. 2009).

subgroups (such as gender, race, ELL status, and disability status) are not commonly available. Third, this study focuses on large multi-district evaluations, in which it is often difficult to obtain a good measure of school-level baseline achievement for many of the schools/districts in the study. In contrast, a researcher conducting a small-scale evaluation (with only one or a few districts) might do better analytically by focusing his/her efforts on obtaining school-level mean scores on state achievement tests for the schools of interest, rather than relying on SEDC data.

Proficiency rates from the SEDC are available for multiple years and a range of student subgroups. The most commonly available proficiency variables are school-wide math and reading proficiency rates, defined as the proportion of students in the school (regardless of grade level) that are deemed "proficient". These school-wide proficiency rates are available as far back as 2002. On average across the RCTs examined in this study, 88 percent of schools had school-wide proficiency rates available for at least one year. Proficiency rates are also available for various student subgroups, but we did not use subgroup proficiency variables because rates of missing data were too high. When proficiency rates were available for multiple years we created composite variables. We used two approaches to constructing composites. The first simply takes the mean of the yearly variables, which could be the best measure (in terms of precision gains) if school-level proficiency varies across time only because of measurement error; we call this construction the "average" composite The second takes the most recent non-missing value, which could be the best measure if the true achievement of students at the school is changing over time; we call this construction the "most recent" composite. For example, if a study's baseline testing year

Table 3.1: Descriptive Statistics for Previously Completed RCT Studies

Study	Purpose	Student Grade	Student Outcome Measures	Unit of Random Assignment	Number of States	Number of Districts	Number of Schools	Number of Students	Response Rate Pretest	Response Rate Posttest
Evaluation of Reading Comprehension Interventions	This study evaluates the impact of four interventions on fifth-grade reading achievement.	5	Group Reading Assessment and Diagnostic Evaluation (GRADE), Educational Testing Service (ETS) Science Reading Comprehension Assessment, ETS Social Studies Reading Comprehension Assessment	School	8	10	89	6,350	0.99	0.88
Evaluation of Early Elementary School Mathematics Curricula	This study compares the effects of four different elementary math curricula on improving student math achievement.	1, 2	Early Childhood Longitudinal Study Mathematics Assessment	School	4	4	39	1,583	0.96	0.87
Evaluation of Teacher Induction Programs	The study examines whether comprehensive teacher induction programs lead to higher teacher retention rates and other positive teacher and student outcomes as compared to prevailing, generally less comprehensive approaches to supporting new teachers.	2-6	District-Administered Standardized Achievement Tests	School	12	15	235	8,292	NA[a]	NA[a]
Evaluation of Teacher Preparation Models	This study examines the effect of different approaches to teacher preparation on teacher practice and student performance.	K-5	Reading Comprehension, Vocabulary, Math Concepts and Applications, and Math Computation subtests of the California Achievement Tests, 5th Edition	Student	7	20	63	2,491	0.97	0.90
Evaluation of the Effectiveness of Reading and Mathematics Software Products (EERMSP)	This study randomly assigned teachers to a treatment group that uses a specified educational technology, or a control group that used conventional teaching approaches. The study consisted of four sub-studies of different interventions at different grade levels (see four rows below).	-	-	-	-	-	-	-		
EERMSP Grade 1	-	1	Stanford Achievement Test (version 10) Reading , and Test of Word Reading Efficiency	Teacher	12	15	53	4,424	0.97	0.95

Study	Purpose	Student Grade	Student Outcome Measures	Unit of Random Assignment	Number of States	Number of Districts	Number of Schools	Number of Students	Response Rate Pretest	Response Rate Posttest
EERMSP Grade 4	-	4	Stanford Achievement Test (version 10), Reading	Teacher	9	12	44	3,109	0.93	0.93
EERMSP Grade 6	-	6	Stanford Achievement Test (version 10), Math	Teacher	7	10	28	4,261	0.96	0.89
EERMSP Algebra	-	8, 9	Educational Testing Service's (ETS) End-of-Course Algebra Assessment	Teacher	8	11	24	3,009	0.82	0.81

Source: Previously completed RCT studies.

[a]Response rates are not applicable for the Teacher Induction study because this study used all available test score data from school districts covering the grades included in the study.

RCT = randomized controlled trial.

was 2006, and a 2006 proficiency rate is available for a given school, then the "most recent" composite variable equals the 2006 value. Otherwise, we revert to the next most recent value (2005, 2004, etc.). Under both constructions, the final composite is missing only if data are missing in all years. When a composite has a missing value, we impute to the mean of that composite across all schools in a given RCT and include a missing value dummy in regression analyses.

Finally, we also included additional school-level information available from the SEDC in our analyses. In particular, we included the racial composition of the student body, the proportion of students who are eligible for free or reduced-price lunch (FRPL), and the percentage of students classified as English language learners (ELL). For each RCT study, we used these variables only if they were non-missing for at least half the schools in the study. As a result, the regression analysis conducted below includes the race and FRPL variables for most of the studies, but includes the ELL variable for only one study. As with the proficiency data, when race, FRPL, or ELL data was available for multiple years we created two composite variables. The first simply takes the mean of the yearly variables; we call this construct the "average" race, ELL, and FRPL. The second takes the most recent non-missing value; we call this construct the "most recent" race, ELL, and FRPL.

State-Level Proficiency Rates, NAEP Scores, and Related Constructs

School-level proficiency rates may vary across states because of true differences in student academic achievement or because of differences in the stringency of the performance standards adopted by the states. Unfortunately, there is no way to directly compare school-level proficiency rates across states because each state establishes its own performance standards. However, state-level National Assessment of Educational Progress (NAEP) scores can be used as an anchor to partially disentangle true variation in student achievement across states from variation in how proficiency is defined across states.

We collected data on state-level proficiency rates and state-level NAEP scores (mean and standard deviation) in order to control for variation across states in how proficiency is defined. State-level proficiency rates were retrieved from educational data sites of the specific state, and the rates indicate the proportion of students who met the state's proficiency requirements. Some state proficiency measures used multiple proficiency ratings (such as, "not proficient," "limited proficiency," "advanced proficiency"). In these cases, we collapsed the ratings into two categories, "proficient" and "not proficient." NAEP scores were retrieved from the National Center for Educational Statistics website, and they indicate how well a representative subset of students from each state performed against the same proficiency measure.

The method that we use to anchor state-level proficiency measures using the NAEP scores is as follows. Assuming that percentiles of the NAEP distribution are equal to percentiles of the state tests used to determine proficiency, and assuming that the state test scores are distributed normally, we infer the proficiency cutoff used in each state and express that cutoff on the NAEP scale. We then calculate an imputed NAEP score for every school based on the school-level proficiency rate. Clearly, the assumptions of normality and equal percentiles between the NAEP and state test scores are unlikely to hold; however, this is the best available approach to deal with the challenging issue of incomparable proficiency metrics across states. Several studies have clarified the challenges of linking state tests at the student level (Linn 1993; Feuer et al. 1999; Koretz et al. 1999). However, for our purposes here, a failure of this method would simply reduce the precision gains from covariate adjustment—it would not introduce bias into the impact estimates. For other examples of equipercentile linking approaches, see McLaughlin and Bandeira de Mello (2002, 2003) and Braun and Qian (2007).

Though NAEP scores are attractive because they can serve as a benchmark for all states, they also have some limitations. NAEP scores are only available for certain years, in specific subjects, and for a limited number of grade levels. On several occasions, the NAEP score that was most relevant for our sample may

have been based on students that were a few grades above and a few years behind our sample. Also, the correlation between NAEP scores and state tests may vary across states.

Table 3.2 provides a list of the various models used in the MDES analyses. Each model includes a constant term, district dummies, demographic variables, and different combinations of the proficiency measures described above. These various models are examined in order to uncover any general patterns in the results that indicate whether certain proficiency variables perform better than others. Specifically, we examine whether math or reading proficiency measures perform better, whether average or most recent scores perform better, and whether the NAEP adjustment makes a difference in terms of the explanatory power of each proficiency measure. In Chapter 4, we examine the precision gains associated with these models.

Table 3.3 provides a key for all of the study outcomes examined throughout the report. A capital letter is used to reference each study outcome. The outcomes are sorted in descending order by the school-level R^2 using the study-collected baseline test.

Table 3.2: Regression Models Used in the MDES Analysis

Model Number	Model Description
Model 1	Constant term, district dummies, and most recent race, English language learners (ELL), and free or reduced-price lunch (FRPL).
Model 2	Constant term, district dummies, and average race, ELL, and FRPL.
Model 3	Constant term, district dummies, school-level ***math and reading*** proficiency rates for the ***most recent*** year available that was on or before the year of study-level baseline testing, and most recent race, ELL, and FRPL.
Model 4	Constant term, district dummies, school-level ***reading*** proficiency rates for the ***most recent*** year available that was on or before the year of study-level baseline testing, and most recent race, ELL, and FRPL.
Model 5	Constant term, district dummies, school-level ***math*** proficiency rates for the ***most recent*** year available that was on or before the year of study-level baseline testing, and most recent race, ELL, and FRPL.
Model 6	Constant term, district dummies, the ***average*** school-level ***math and reading*** proficiency rates across all available years on or before the year of study-level baseline testing, and average race, ELL, and FRPL.
Model 7	Constant term, district dummies, the ***average*** school-level ***reading*** proficiency rates across all available years on or before the year of study-level baseline testing, and average race, ELL, and FRPL.
Model 8	Constant term, district dummies, the ***average*** school-level ***math*** proficiency rates across all available years on or before the year of study-level baseline testing, and average race, ELL, and FRPL.
Model 9	Constant term, district dummies, school-level ***math and reading*** proficiency rates for the ***most recent*** year available, transformed into ***NAEP*** scores, and most recent race, ELL, and FRPL.
Model 10	Constant term, district dummies, school-level ***reading*** proficiency rates for the ***most recent*** year available, transformed into ***NAEP*** scores, and most recent race, ELL, and FRPL.
Model 11	Constant term, district dummies, school-level ***math*** proficiency rates for the ***most recent*** year available, transformed into ***NAEP*** scores, and most recent race, ELL, and FRPL.
Model 12	Constant term, district dummies, ***average*** school-level ***math and reading*** proficiency rates across all available years, transformed into ***NAEP*** scores, and average race, ELL, and FRPL.
Model 13	Constant term, district dummies, ***average*** school-level ***reading*** proficiency rates across all available years, transformed into ***NAEP*** scores, and average race, ELL, and FRPL.
Model 14	Constant term, district dummies, ***average*** school-level ***math*** proficiency rates across all available years, transformed into ***NAEP*** scores, and average race, ELL, and FRPL.

Source: State Education Data Center (SEDC), National Center for Educational Statistics, and educational data sites for individual states.

Table 3.3: Study/Outcome Codes

Letter	Study	Outcome
A	Reading and Mathematics Software Products: Grade 4	Stanford Achievement Test Version 10: Total Reading Score
B	Reading and Mathematics Software Products: Grade 4	Stanford Achievement Test Version 10: Work Study Skills Score
C	Reading and Mathematics Software Products: Grade 6	Stanford Achievement Test Version 10: Problem Solving Score
D	Reading and Mathematics Software Products: Grade 6	Stanford Achievement Test Version 10: Total Math Score
E	Teacher Preparation Models	California Achievement Test, 5th Edition: Reading Comprehension
F	Reading Comprehension Interventions	Group Reading Assessment and Diagnostic Evaluation
G	Reading and Mathematics Software Products: Grade 6	Stanford Achievement Test Version 10: Procedures Score
H	Reading and Mathematics Software Products: Grade 4	Stanford Achievement Test Version 10: Reading Vocabulary Score
I	Reading and Mathematics Software Products: Grade 4	Stanford Achievement Test Version 10: Reading Comprehension Score
J	Reading Comprehension Interventions	Educational Testing Service Social Studies Reading Comprehension Assessment
K	Teacher Preparation Models	California Achievement Test, 5th Edition: Math Concepts and Applications
L	Teacher Preparation Models	California Achievement Test, 5th Edition: Vocabulary
M	Reading and Mathematics Software Products: Algebra	Educational Testing Service End-of-Course Algebra Assessment: Processes Score
N	Reading and Mathematics Software Products: Algebra	Educational Testing Service End-of-Course Algebra Assessment: Skills Score
O	Reading Comprehension Interventions	Educational Testing Service Science Reading Comprehension Assessment
P	Teacher Induction Programs	District-Administered Standardized Achievement Test: Math
Q	Teacher Induction Programs	District-Administered Standardized Achievement Test: Reading
R	Reading and Mathematics Software Products: Algebra	Educational Testing Service End-of-Course Algebra Assessment: Total Score
S	Early Elementary School Mathematics Curricula	Early Childhood Longitudinal Study Mathematics Assessment
T	Reading and Mathematics Software Products: Grade 1	Stanford Achievement Test Version 10: Word Reading Score
U	Reading and Mathematics Software Products: Grade 1	Stanford Achievement Test Version 10: Total Reading Score
V	Reading and Mathematics Software Products: Grade 1	Stanford Achievement Test Version 10: Sentence Reading Score
W	Teacher Preparation Models	California Achievement Test, 5th Edition: Math Computation
X	Reading and Mathematics Software Products: Algebra	Educational Testing Service End-of-Course Algebra Assessment: Concepts Score
Y	Reading and Mathematics Software Products: Grade 1	Stanford Achievement Test Version 10: Sounds and Letters Score

Source: Previously completed RCT studies.

Chapter 4: Precision Gains from School-Level Proficiency Data

We compare the precision gains from study-collected baseline tests to models that incorporate different combinations of proficiency measures in Tables 4.1 and 4.2 and in Figure 4.1. Table 4.1 shows the ICC; the school-level R^2 achieved using the base model (that is, district dummies only); the school-level mean of the study-collected baseline test score; the school-level mean of the study-collected baseline test and the most recent school-level FRPL, ELL, and racial composition data[4]; the "most recent" proficiency measure and the most recent school-level FRPL, ELL, and racial composition data; and the "best" proficiency, FRPL, ELL, and racial composition measures, which are the measures that yield the highest R^2. We included results from the model using the "most recent" proficiency measure for every study/outcome to show what precision gains are possible using a consistent approach across all studies. We included results from the model using the "best" proficiency measure to illustrate what additional gains might be possible in cases where a researcher believes that gains which might be specific to their data set are still valuable.[5] The last column of the table shows the difference between the school-level R^2 achieved using the baseline test and the school-level R^2 achieved using the "best" proficiency and other school-level measures. Figure 4.1 plots this difference for all 25 outcomes. Table 4.2 shows the MDES values achieved using these same four models; the second-to-last column shows the difference between the MDES achieved using the baseline test and the MDES achieved using the "best" school-level measures. In all cases we calculated the MDES assuming a sample size of 40 schools and 1,800 students, evenly divided between the treatment and control groups (we did not use the study's actual sample size so that variation in the MDES is due entirely to differences in the ICC and R^2). The last column shows the number of additional schools that would need to be added to a study using the "best" proficiency model in order to achieve the same MDES that was achieved with the study-collected test scores.

We find that, for 22 of 25 outcomes, including the "best" school-level proficiency and other school-level measures as covariates will reduce the MDE relative to a model that only includes school district dummies (see table 4.2). For example, the school-level R^2 value for outcome E rises from 0.37 to 0.63 when moving from the base model to the "best" proficiency and other school-level measures (see Table 4.1), and the MDES falls from 0.35 to 0.28 (see table 4.2), Across all 25 outcomes the average ratio of the MDES using the best proficiency and demographic covariates to the MDES using only district dummies is 0.82.

Among the school-level proficiency variables examined, we find that the NAEP-standardized versions tend to perform approximately as well as the variables that are not NAEP-standardized and that the "most-recent" variable construction performs the same as the "average" construction. We calculated the ratio of school-level R^2 and MDES values using NAEP-standardized variables to those values using variables that are not NAEP-standardized. Those medians of those ratios are 0.998 and 1.001 (both slightly favoring the non-NAEP-standardized variables). We interpret these findings to suggest that there is no practical advantage to NAEP-standardization. We also constructed similar ratios of school-level R^2

[4] We included the race, ELL, and FRPL variables in the models that contain SEDC proficiency rates because they are easy to download simultaneously with the proficiency rates and may provide additional explanatory power. In order to make valid comparisons between the models that contain study-collected baseline test scores and the models that contain SEDC proficiency data, the race, ELL, and FRPL variables are included in both types of models.

[5] The higher school-level R^2 achieved using the "best" proficiency measure (relative to the "most recent" measure) might be due to a spurious correlation between that proficiency measure and the outcome measure (the follow-up test score), and thus might not be observed in a sample of different schools. In other words, for each completed RCT, for each outcome, we have identified the school-level proficiency measure that results in the highest school-level R^2 *for the particular sample of schools included in the study*. Had the study included a different set of schools, a different proficiency measure might have been chosen as the "best," and the school-level R^2 achieved using that measure might not be as high as the school-level R^2 value displayed in Table 4.1.

Table 4.1: School-Level R^2 Values (R_B^2) Achieved Using Various Measures of Baseline Student Achievement

		School-Level R^2 (R_B^2)						
Outcome	ICC	Base Model	School-Level Mean of Study-Collected Baseline Test	School-Level Mean of Study-Collected Baseline Test + Most Recent Race, ELL, FRPL	Most Recent Proficiency Measure + Most Recent Race, ELL, FRPL (Model 3)	"Best" Proficiency Measure + Race, ELL, FRPL	Model Number of "Best" Proficiency Measure	Difference Between (R_B^2) Using Baseline Test and "Best" Proficiency Measure
A	0.27	0.69	0.97	0.97	0.70	0.73	Model 9	0.24
B	0.24	0.82	0.98	1.00	0.83	0.86	Model 9	0.12
C	0.16	0.24	0.93	0.95	0.22	0.28	Model 6	0.65
D	0.16	0.22	0.93	0.97	0.22	0.23	Model 6	0.70
E	0.21	0.37	0.92	0.91	0.63	0.63	Model 3	0.29
F	0.14	0.69	0.92	0.92	0.84	0.85	Model 10	0.07
G	0.12	0.28	0.91	0.96	0.30	0.30	Model 3	0.62
H	0.22	0.70	0.89	0.88	0.73	0.74	Model 9	0.14
I	0.22	0.57	0.89	0.88	0.57	0.60	Model 10	0.29
J	0.12	0.68	0.91	0.90	0.90	0.93	Model 9	-0.02
K	0.24	0.19	0.86	0.85	0.32	0.36	Model 1	0.50
L	0.30	0.29	0.93	0.94	0.58	0.59	Model 11	0.33
M	0.10	0.68	0.86	0.74	0.89	1.00	Model 9	-0.14
N	0.16	0.75	0.86	0.75	0.93	1.00	Model 9	-0.14
O	0.12	0.71	0.85	0.83	0.81	0.84	Model 11	0.02
P	0.15	-0.08	0.67	0.74	0.34	0.38	Model 14	0.29
Q	0.14	-0.08	0.66	0.72	0.35	0.37	Model 14	0.29
R	0.20	0.68	0.71	0.40	0.85	0.99	Model 9	-0.28
S	0.19	0.34	0.82	0.88	0.53	0.59	Model 2	0.23
T	0.07	0.23	0.65	0.76	0.60	0.61	Model 5	0.04
U	0.10	0.26	0.66	0.80	0.63	0.64	Model 7	0.02
V	0.11	0.30	0.55	0.66	0.62	0.63	Model 5	-0.08
W	0.12	0.11	0.53	0.62	0.10	0.15	Model 5	0.38
X	0.08	0.73	0.71	0.66	0.83	0.88	Model 9	-0.17
Y	0.10	0.32	0.57	0.69	0.55	0.59	Model 7	-0.03

Source: Previously completed RCT studies.

Note: District dummies are included in all models. The base model includes only district dummies. The "best" proficiency measure is defined as the school-level proficiency measure that achieves the highest school-level R^2. Detailed descriptions of each model are listed in Table III.2.

ELL = English language learner; FRPL = free or reduced-price lunch; ICC = intraclass correlation coefficient; RCT = randomized controlled trial.

Table 4.2: Minimum Detectable Effect Size (MDES) Achieved Using Various Measures of Baseline Student Achievement

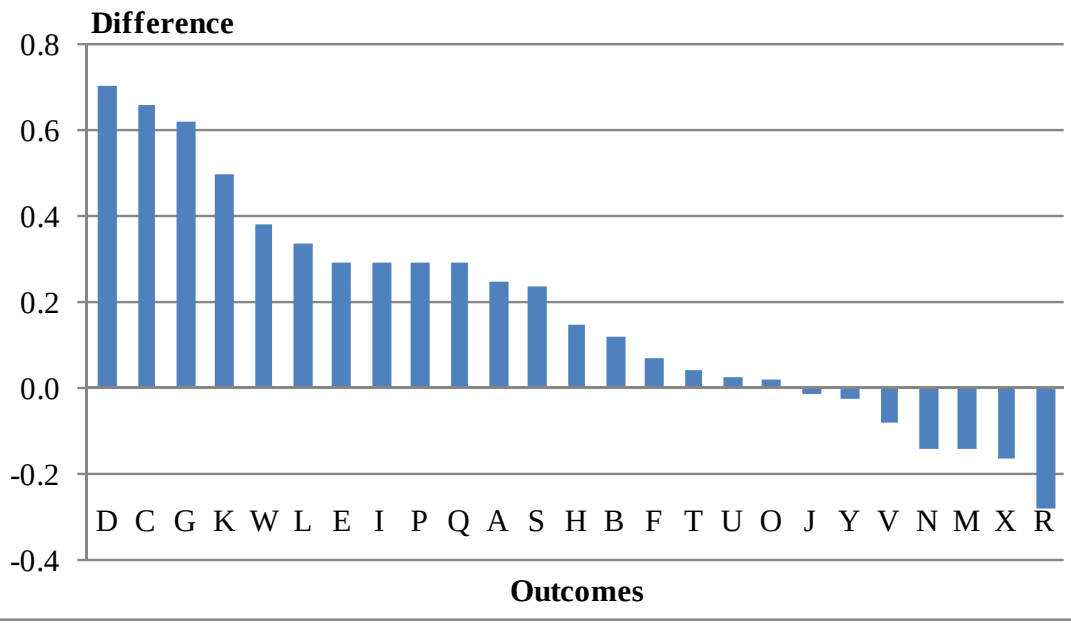
		MDES							
Outcome	ICC	Base Model	School-Level Mean of Study-Collected Baseline Test	School-Level Mean of Study-Collected Baseline Test + Most Recent Race, ELL, FRPL	Most Recent Proficiency Measure + Most Recent Race, ELL, FRPL (Model 3)	"Best" Proficiency Measure + Race, ELL, FRPL	Model Number of "Best" Proficiency Measure	Difference Between MDES Using Baseline Test and "Best" Proficiency Measure	Number of Additional Schools Needed to Achieve Original MDES
A	0.27	0.29	0.14	0.15	0.29	0.28	Model 9	-0.13	101
B	0.24	0.22	0.13	0.12	0.22	0.20	Model 9	-0.07	49
C	0.16	0.34	0.16	0.15	0.35	0.34	Model 6	-0.18	133
D	0.16	0.34	0.16	0.14	0.35	0.34	Model 6	-0.19	141
E	0.21	0.35	0.17	0.17	0.28	0.28	Model 3	-0.11	66
F	0.14	0.23	0.16	0.16	0.19	0.18	Model 10	-0.02	11
G	0.12	0.30	0.16	0.14	0.30	0.30	Model 3	-0.14	96
H	0.22	0.27	0.19	0.19	0.25	0.25	Model 9	-0.06	29
I	0.22	0.31	0.19	0.19	0.30	0.30	Model 10	-0.11	58
J	0.12	0.22	0.16	0.16	0.16	0.15	Model 9	0.01	-3
K	0.24	0.42	0.21	0.21	0.39	0.38	Model 1	-0.17	88
L	0.30	0.44	0.18	0.17	0.34	0.34	Model 11	-0.16	99
M	0.10	0.21	0.17	0.20	0.16	0.13	Model 9	0.04	-15
N	0.16	0.23	0.19	0.22	0.16	0.13	Model 9	0.06	-21
O	0.12	0.21	0.18	0.18	0.19	0.18	Model 11	-0.01	2
P	0.15	0.39	0.24	0.22	0.32	0.31	Model 14	-0.07	24
Q	0.14	0.38	0.24	0.22	0.31	0.30	Model 14	-0.06	23
R	0.20	0.26	0.25	0.34	0.20	0.13	Model 9	0.13	-28
S	0.19	0.35	0.21	0.18	0.30	0.28	Model 2	-0.07	32
T	0.07	0.26	0.20	0.18	0.21	0.20	Model 5	-0.01	2
U	0.10	0.29	0.22	0.18	0.22	0.22	Model 7	0.00	2
V	0.11	0.28	0.24	0.22	0.23	0.22	Model 5	0.02	-5
W	0.12	0.33	0.25	0.23	0.33	0.32	Model 5	-0.07	23
X	0.08	0.19	0.19	0.20	0.17	0.16	Model 9	0.03	-12
Y	0.10	0.27	0.23	0.21	0.23	0.23	Model 7	0.00	-2

Source: Previously completed RCT studies.

Note: We calculated all MDES using a sample size of 40 schools and 1,800 students, evenly divided between the treatment and control groups (we did not use each study's actual sample size so that variation in the MDES is due only to variation in the ICC and R^2). District dummies are included in all models. The base model includes only district dummies. The "best" proficiency measure is defined as the school-level proficiency measure that achieves the highest school-level R^2. The last column shows the number of additional schools that would need to be added to a study in order to achieve the same MDES using the "best" proficiency measure that was obtained using the study-collected baseline test. Detailed descriptions of each model are listed in Table III.2.

ELL = English language learner; FRPL = free or reduced-price lunch; ICC = intraclass correlation coefficient; RCT = randomized controlled trial.

Figure 4.1: Difference Between School-Level R^2 Values Achieved Using the Baseline Test and the "Best" Proficiency Measure for 25 Outcomes



and MDES values comparing the most recent proficiency variables to the average (averaging across multiple years of data) proficiency variables. Those ratios were 1.007 and 0.995. Finally, we constructed ratios of school-level R^2 and MDES values comparing the math-and-reading-combined proficiency variables to the reading-only and math-only proficiency variables. Those ratios were 1.02 and 0.99 for reading-only and 1.005 and 0.997 for math-only, suggesting that the combined reading and math proficiency variables performed no better than the reading-only and math-only variables.

While including school-level proficiency variables definitely improves statistical precision compared to not including any prior achievement measures, we find that, for the majority of outcomes, some precision is lost when school-level proficiency data are used instead of the study-collected baseline test data, and that the size of this loss can be quite large. This is demonstrated in Tables 4.1 and 4.2 by smaller school-level R^2s and higher MDES values when school-level proficiency data are used instead of study-collected baseline test data[6]. For example, the school-level R^2 value for outcome E falls from 0.92 to 0.63 when moving from the baseline test to the "best" proficiency and other school-level measures, and the MDES rises from 0.17 to 0.28. Across all 25 outcomes, the average ratio of the R^2 using the "best" proficiency and other school-level measures to the R^2 using just district dummies and the study-collected baseline test data was 0.79. Across all 25 outcomes, the average ratio of the MDES using the "best" proficiency and other school-level measures to the MDES using just district dummies and the study-collected baseline test data was 1.31. Adding the FRPL, ELL, and racial composition variables to a regression that includes the study-collected baseline test has a negligible benefit. Adding the best school-level proficiency variables to a regression that includes the study-collected baseline test (not shown in table) increases the average

[6] Note when comparing MDE and R^2 values across studies holding the covariate choice fixed that a low school-level R^2 does not necessarily mean a high MDE, because a low school-level R^2 is more likely to be accompanied by a low ICC: across all 25 outcomes, the correlation between the ICC and the school-level R^2 achieved using the study-collected baseline test is 0.59 (which is statistically significant, with a p-value of 0.002).

school-level R^2 by 0.05, which reduces the average MDE by 0.02. For the average study, this gain in precision is equivalent to adding nine schools to the study sample.

Figure 4.1 shows the difference between the school-level R^2 values achieved using the baseline test and the school-level R^2 values achieved using the "best" proficiency measure for all 25 outcomes. Positive values in this figure correspond to cases where the study-collected baseline test performs better than the school-level proficiency variables. Negative values correspond to cases where the school-level proficiency variables performed better than the study-collected variables. We see that for the majority of outcomes (18 out of 25), the school-level proficiency data do not perform as well as the study-collected baseline test data in terms of explaining the school-level variation in the outcome variable.

In only 7 out of 25 cases do we find that school-level proficiency data perform as well as or better than study-collected baseline test data in terms of the school-level R^2. For example, for outcome J, the "best" school-level proficiency measure produced a school-level R^2 that was slightly greater than the school-level R^2 achieved using the study-collected baseline test data (0.93 versus 0.91), and produced an MDES of 0.15 instead of the MDES of 0.16 that was achieved using the baseline test data.

Another way to interpret the rise in the MDES when moving from a study-collected baseline test to a school-level proficiency measure is in terms of the number of additional schools that would be needed in the study in order to achieve the MDES using the original baseline test score. The last column of Table 4.2 provides this information. We see that, for most outcomes, additional schools would be needed in the study sample in order to achieve the MDES achieved using baseline test data. Across all 25 outcomes, the number of additional schools needed ranges from -28 to 141, with an average of 36 (that is, the average sample size would have to nearly double to compensate for the loss in precision) and a median of 23.

For each of the 25 outcomes, we also calculated the number of schools needed to achieve an MDES of 0.20, using the school-level mean of the baseline test and the "best" school-level proficiency measure, and using both measures. These results are displayed in Table 4.3. On average, the number of schools needed to achieve an MDES of 0.20 when using study-collected baseline test data is 37; the average number of schools needed when using school-level proficiency data is 64. Thus, on average, the number of schools needed in a study would need to increase by nearly three-quarters in order to achieve an MDES of 0.20 when using proficiency data instead of a study-collected baseline test.

Performance Correlates: When Do Proficiency Data Perform Well?

We conducted some additional analyses to test whether proficiency data perform better (that is, achieve higher school-level R^2s and lower MDES values) in certain contexts. Specifically, we explored whether certain aspects of studies or the school-level proficiency data were correlated with the difference in the school-level R^2 achieved using the "best" proficiency measure and the school-level R^2 achieved using the study-collected baseline test. In particular, we hypothesized that school-level proficiency data might perform better when: (1) the study has a larger number of schools, (2) the study has a smaller number of states, (3) the students in the study are older, and (4) the amount of missing proficiency data is lower. Because small sample sizes might lead to spurious underestimates of the school-level R^2, proficiency data might perform better in studies with larger numbers of schools. In addition, states often have very different definitions of "proficient," resulting in school-level proficiency measures that are not necessarily comparable across states. Therefore, one might expect that proficiency data will perform better in studies where schools are concentrated in fewer states, because the proficiency measure will be comparable across all schools within the same state. Among younger students (kindergarten, first grade, and second grade), school-level proficiency measures may be less effective in explaining differences in follow-up test scores across schools because the school-level measures of proficiency are less likely to be based on these students (since these younger students were less likely to be in the schools at the time of earlier testing). Thus, one might expect that school-level proficiency data will perform better when the students in a study are older. Finally, recall that when proficiency data were missing for a particular school, we imputed the

Table 4.3: Number of Schools Needed to Achieve an MDES of 0.2

Outcome	ICC	School-Level R^2 (R_B^2)		Number of Schools Needed to Achieve an MDES of 0.2	
		Using School-Level Mean of Study-Collected Baseline Test	Using "Best" Proficiency Measure	Using School-Level Mean of Study-Collected Baseline Test	Using "Best" Proficiency Measure
A	0.27	0.97	0.73	21	72
B	0.24	0.98	0.86	19	42
C	0.16	0.93	0.28	26	107
D	0.16	0.93	0.23	26	113
E	0.21	0.92	0.63	29	77
F	0.14	0.92	0.85	26	34
G	0.12	0.91	0.30	26	83
H	0.22	0.89	0.74	35	61
I	0.22	0.89	0.60	35	85
J	0.12	0.91	0.93	26	24
K	0.24	0.86	0.36	42	136
L	0.30	0.93	0.59	31	111
M	0.10	0.86	1.00	29	18
N	0.16	0.86	1.00	34	17
O	0.12	0.85	0.84	32	32
P	0.15	0.67	0.38	56	90
Q	0.14	0.66	0.37	54	86
R	0.20	0.71	0.99	61	18
S	0.19	0.82	0.59	43	77
T	0.07	0.65	0.61	37	40
U	0.10	0.66	0.64	44	46
V	0.11	0.55	0.63	56	49
W	0.12	0.53	0.15	62	97
X	0.08	0.71	0.88	36	26
Y	0.10	0.57	0.59	51	50

Source: Previously completed RCT studies.

Note: District dummies, race, ELL, and FRPL are included as covariates in all models. The "best" proficiency measure is defined as the school-level proficiency measure that achieves the highest school-level R^2.

ELL = English language learner; FRPL = free or reduced-price lunch; ICC = intraclass correlation coefficient; MDES = m2inimum detectable effect size; RCT = randomized controlled trial.

missing value with the mean of the variable and included a missing value dummy. If proficiency data are missing for a large portion of schools in the study, then we would expect that the explanatory power of the proficiency measure would decrease, resulting in lower school-level R^2 values and higher MDES values. We also examined whether proficiency data perform better when the outcome is a math or reading test, but we do not have a clear hypothesis as to why proficiency data would perform better for one outcome than the other.

Table 4.4 shows the correlations (across all 25 outcomes) of the school-level R^2 and the MDES associated with using school-level proficiency (and FRPL, ELL, and racial composition), with the number of schools in the study, the number of states in the study, the average grade level of students in the study, and the proficiency data missing rate.[7] We also examined the correlations between various study aspects and the log ratio of the R^2 and MDES values achieved using the "best" school-level proficiency data (and FRPL, ELL, and racial composition) to the R^2 and MDES values achieved using the study-collected baseline test score. We include the log ratio in order to "difference out" all common inputs into the R^2 and MDES (such as the ICC and the contribution of school district dummies to the school-level R^2). Thus, variation in the log ratios across studies is due to the difference between adjusting for the study-collected test versus adjusting for school-level proficiency rates.

We found that neither the school-level R^2 nor the MDES achieved using the "best" proficiency measure is correlated with the number of schools in the study or the number of states in the study. However, both the school-level R^2 and the MDES are correlated with the average grade level of students in the study (correlation coefficients of 0.43 and -0.48), and these correlations are statistically significant at the 5 percent significance level. Thus, it appears that school-level proficiency data perform better when the students in a study are older. However, the *difference* in R^2 and MDES values between proficiency and study-collected baseline tests is not significantly correlated with grade level. Thus, it would seem that, while proficiency measures yield a higher R^2 for older students, so do study-collected baseline test scores.

We find that the MDE and school-level R^2 are not correlated with the rate of missing proficiency data.

Finally, we compared the average school-level R^2 for math outcomes to that of reading outcomes (not shown in table). The average school-level R^2 for reading outcomes was 0.69 and the average for math outcomes was 0.56. The difference was not statistically significant (p-value: 0.28).

[7] The p-values do not take into account dependencies among the 25 outcomes resulting from the fact that there are multiple outcomes within each RCT study.

Table 4.4: Correlations of School-Level R^2 (R_B^2) and MDES Values with Various Study Aspects

Study Aspect		Correlation with School-Level R^2 (R_B^2)	Correlation with MDES	Correlation with the Log Ratio of the School-Level R^2 Using the "Best" Proficiency Measure to the School-Level R^2 Using the Baseline Test	Correlation with the Log Ratio of the MDES Using the "Best" Proficiency Measure to the MDES Using the Baseline Test
Number of Schools in the Study	Correlation coefficient	-0.27	0.22	-0.11	0.04
	P-value[a]	0.20	0.29	0.61	0.84
Number of States in the Study	Correlation coefficient	0.02	-0.16	0.29	-0.25
	P-value[a]	0.91	0.46	0.16	0.22
Average Grade of Students	Correlation coefficient	0.43	-0.48	0.17	-0.36
	P-value[a]	0.03	0.02	0.42	0.08
Proficiency Data Missing Rate	Correlation coefficient	0.24	-0.36	0.30	-0.29
	P-value[a]	0.25	0.07	0.14	0.16

Source: Previously completed RCT studies.

Note: This table shows the correlations of the school-level R^2 and the MDES achieved using the "best" proficiency measure with the number of schools in the study, the number of states in the study, the average grade level of students in the study, and the amount of missing proficiency data. The table also shows the correlations of the log ratio of the school-level R^2 (or MDES) using the "best" proficiency measure to the school-level R^2 (or MDES) using the study-administered baseline test with the number of schools in the study, the number of states in the study, the average grade level of students in the study, and the amount of missing proficiency data. The "best" proficiency measure is defined as the school-level proficiency measure that achieves the highest school-level R^2.

[a]P-values do not take into account dependencies among the 25 outcomes resulting from the fact that there are multiple outcomes within each RCT study.

MDES = minimum detectable effect size; RCT = randomized controlled trial.

Chapter 5: Additional Analyses

We conducted two additional analyses in order to better understand the findings of Chapter 4 and to explore an alternative approach to cost savings in collecting baseline test data. First, we partially decomposed the source of the precision loss when using proficiency measures instead of a study-collected test score. Second, we examined an alternative strategy for cost savings in which subsamples of students are tested at baseline instead of testing all students.

Decomposing Precision Loss

In Chapter 4, we found that, on average, the ratio of the school-level R^2 using the best proficiency measure to the school-level R^2 using the study-collected pre-test was 0.73. There are several reasons why school-level achievement measures might not perform as well as study-collected student-level achievement measures in terms of increasing precision and decreasing the MDES. First, publically available school-level achievement measures might represent different students than those participating in the study. This could be because the school-level measures pertain to a different cohort of students, a different grade, or even a different set of students within the same grade and cohort (because students transfer in and out of schools). Second, school-level achievement measures could be for a different test than the study-collected follow-up test (for example, reading versus math) or a different aspect of the same subject (such as, vocabulary versus comprehension in reading).

While we cannot tell which of these reasons may be driving the results in Chapter 4, we can partially decompose the source of the precision loss when using proficiency measures instead of study-collected test score data. Specifically, we can examine how school-level R^2 values change when the test, grade-level, and student cohort are held constant, but a different set of students is used to calculate the school-level baseline achievement measure.

We explore this issue using the study-collected test scores by comparing two different school-level R^2 values: one in which we calculate a school-level R^2 using aggregate pre-test data for the same set of students for whom we have post-test scores and another in which we calculate the school-level R^2 using a disjoint subsample of students. Specifically, we can divide students into two equally sized groups, A and B, within each school. We can then calculate a school-level R^2 by regressing the post-test for students in group A on the school-level aggregate of their pre-test scores. We can also calculate a school-level R^2 by regressing the post-test for students in group A on the school-level aggregate of the pre-tests of students in group B. Comparing these two R^2s shows the precision loss associated with using a pre-test for a different set of students in the same school, holding all other factors constant.

In Table 5.1 we present these two R^2 values for all 25 cases. The values reported in this table are the average R^2 values across 1,000 simulation replications. For each replication, we re-randomized students into the groups A and B described above. The average ratio of the school-level R^2 for disjoint samples to the school-level R^2 for the same samples is 0.72. Recall that the ratio of the school-level R^2 using the best proficiency measure to the school-level R^2 using the study-collected pre-test was 0.73. Despite the similarity in these numbers, we cannot claim that this issue completely explains the difference in school-level R^2 between the best proficiency measure and the study-collected pre-test because the proficiency scores are not always for completely disjoint samples and because the proficiency scores are based on different tests than the study-collected pre-test. Furthermore, we had to cut our student sample in half to conduct this analysis, and smaller samples of disjoint students that are randomly sampled from the same school are more likely to differ from one another by chance. Nevertheless, this analysis shows that the difference in

Table 5.1: School-Level R^2 (R_B^2) When Disjoint Student Samples Are Used at Baseline and Follow-Up

Outcome	School-level R^2 (R_B^2) Using the Same Students at Baseline and Follow-up	School-level R^2 (R_B^2) Using Disjoint Sets of Students at Baseline and Follow-up	Difference
A	0.87	0.68	0.19
B	0.78	0.62	0.16
C	0.83	0.72	0.11
D	0.82	0.70	0.12
E	0.74	0.56	0.18
F	0.75	0.49	0.26
G	0.73	0.61	0.12
H	0.79	0.61	0.18
I	0.77	0.58	0.19
J	0.56	0.25	0.31
K	0.67	0.50	0.17
L	0.71	0.58	0.13
M	0.54	0.49	0.05
N	0.63	0.56	0.07
O	0.49	0.21	0.28
P	0.62	0.30	0.32
Q	0.61	0.34	0.27
R	0.65	0.60	0.05
S	0.55	0.35	0.20
T	0.55	0.35	0.20
U	0.55	0.35	0.20
V	0.50	0.34	0.16
W	0.41	0.29	0.12
X	0.29	0.28	0.01
Y	0.27	0.16	0.11

Source: Previously completed RCT studies.

Note: For each outcome, this table shows the school-level R^2 achieved using the same set of students at baseline and follow-up and the school-level R^2 achieved using disjoint samples of students at baseline and follow-up. The R^2 values reported here are averages across 1,000 simulation replications. Each replication randomly sampled 50% of the students in each school, calculated the school-level baseline test score using only those students, calculated the school-level follow-up test score using the other 50% of students, and then calculated the school-level R^2. District dummies are not included as covariates.

school-level R^2 between the proficiency measure and the study-collected pre-test is (1) of a plausible magnitude and (2) possibly due in large part to disjoint samples of students.

In summary, we have shown that if the sample of baseline students is totally disjoint from the sample at follow-up, even if these students come from the same grade and cohort, and even if the baseline test is the same as the follow-up test, the result may be large reductions in the school-level R^2. School-level proficiency measures obtained from public sources might represent different cohorts, different students, and different tests. Therefore, it is not surprising that using school-level proficiency measures as covariates rather than student-level achievement measures resulted in lower precision levels (that is, higher MDES values) for most of the outcomes we examined in Chapter 4.

Alternative Cost Savings Strategy

In Chapter 4, we found that, on average, the number of schools in a study would need to double in order to compensate for the loss in precision from using school-level proficiency measures instead of study-collected baseline test data. An alternative strategy, at least for studies where the marginal cost of collecting another student's pre-test is high, could be to only test a random subsample of students at baseline. To explore the potential benefits of this approach, we randomly drew subsamples of 20 percent, 40 percent, 60 percent, and 80 percent of the full sample in each study and calculated the school-level R^2 when adjusting for the school-level mean pre-test based on these subsamples. We repeated this 1,000 times and calculated the average school-level R^2 across simulation replications. Table 5.2 shows results for all 25 cases.

On average, the ratio of the school-level R^2 associated with using a 20 percent subsample at baseline to the school-level R^2 associated with the full sample is 0.80. The average ratios for the 40, 60, and 80 percent subsamples are 0.91, 0.95, and 0.98, respectively. On average, the number of schools that a study would need to add (assuming that it starts with 40 schools) to compensate for the lower precision is 26, 11, 5, and 2, respectively (not shown in table). Thus, when testing 20, 40, 60, or 80 percent subsamples of students at baseline, the number of schools in the study would need to increase by 65, 27.5, 12.5, or 5 percent, respectively, in order to maintain the same MDES.

The answer to whether collecting baseline test scores for a subsample of students is a cost-effective alternative to using publically available school proficiency data or collecting baseline test scores for all students will depend upon the various costs for each individual study. This option may be cost-effective for some studies, depending on the tradeoff between the marginal cost of testing another student at baseline and follow-up and the marginal cost of including another school in the study (which includes the cost of implementing the intervention). For example, in a study where the marginal cost of testing another student is high (perhaps because the testing is done individually by pulling each student out of class), but the marginal cost of including another school in the study is relatively low, it may be cost-effective to collect baseline test scores for an 80 percent subsample and increase the number of schools slightly.

Table 5.2 Sampling Distribution of the School-Level R^2 R_B^2 When Taking 20%, 40%, 60%, and 80% Student Subsamples At Baseline

Outcome	R_B^2 Using Full Sample	R_B^2 Using: 20 Percent Subsample	40 Percent Subsample	60 Percent Subsample	80 Percent Subsample
A	0.94	0.80	0.88	0.91	0.93
B	0.91	0.74	0.84	0.88	0.90
C	0.92	0.85	0.89	0.91	0.92
D	0.91	0.83	0.88	0.90	0.91
E	0.89	0.73	0.83	0.86	0.88
F	0.89	0.69	0.81	0.85	0.88
G	0.87	0.77	0.83	0.85	0.86
H	0.90	0.75	0.84	0.88	0.89
I	0.88	0.73	0.82	0.85	0.87
J	0.81	0.52	0.67	0.74	0.78
K	0.84	0.68	0.77	0.81	0.83
L	0.86	0.74	0.82	0.84	0.85
M	0.80	0.62	0.72	0.77	0.79
N	0.84	0.71	0.78	0.81	0.83
O	0.74	0.49	0.61	0.68	0.72
P	0.77	0.55	0.66	0.71	0.75
Q	0.74	0.52	0.64	0.69	0.72
R	0.83	0.74	0.79	0.81	0.83
S	0.75	0.60	0.69	0.72	0.74
T	0.74	0.63	0.70	0.72	0.73
U	0.74	0.64	0.70	0.72	0.73
V	0.72	0.61	0.67	0.70	0.71
W	0.69	0.55	0.63	0.66	0.68
X	0.63	0.44	0.54	0.58	0.61
Y	0.52	0.44	0.48	0.50	0.52
Average	**0.81**	**0.65**	**0.74**	**0.77**	**0.79**

Source: Previously completed RCT studies.

Note: For each outcome, this table shows the school-level R^2 achieved using the full sample of students in each school, and the quantiles of the school-level R^2 achieved using only a subsample of students from each school to calculate the school-level baseline test score. The full sample of students is always used to calculate the school-level follow-up test score. The last row shows the mean of the school-level R^2 achieved using student subsamples, averaged across all 25 outcomes. Based on 1,000 replications. Each replication randomly sampled X% of the students in each school (where X was 20, 40, 60 or 80), calculated the school-level baseline test score using only those students, and then calculated the school-level R^2. District dummies are not included as covariates.

RCT = randomized controlled trial.

Chapter 6: Attrition Bias

In this chapter, we address the topic of attrition bias, which is a related but separate issue from this study's main research topic. The main topic of concern is whether publically available school-level proficiency data can provide the same precision gains as a student-level baseline test score, when used as covariates in a regression of a follow-up test score. Even if school-level proficiency performed as well as student-level study-collected baseline tests, it might still be desirable to collect student-level baseline tests if they could be used to reduce non-response bias in the follow-up test data. Thus, we conducted the analyses in this chapter to assess whether non-response bias has been a problem in the five RCT studies examined in this report. If it has, then a student-level test may be useful in addressing this problem in future studies; if it hasn't, then a student-level test may not be needed for this purpose. In Chapter 4, we present evidence that school-level proficiency variables do not perform as well as student-level study-collected tests, for the purpose of improving the precision of impact estimates, which renders the question of whether student-level test scores should be collected in order to reduce attrition bias somewhat immaterial. However, empirical data about whether non-response bias exists (and to what extent it exists) in previous IES studies may be potentially useful for planning future studies and are therefore presented in this chapter.

Attrition bias (or nonresponse bias) is a bias that arises when individuals or schools with missing data differ systematically from those without missing data, and when individuals or schools in the treatment group without missing data differ from those in the control group without missing data in terms of their pre-intervention characteristics. Attrition may happen at either the student or school level; attrition bias can come from either source. Student-level baseline test scores and school-level proficiency data can be used to diagnose and adjust for missing data at follow-up. A variety of techniques are available that can use either student-level baseline test scores or school-level proficiency data to partially mitigate attrition bias. Corrections for this bias include the maximum likelihood method of Griliches et al. (1978) and Heckman's (1979) two-stage procedure. An alternative method is to impute missing data. Under this approach, multiple imputation as pioneered by Rubin (1987, 1996) is considered to be the "gold standard." Puma et al. (2008) examine the performance of various methods for handling missing data in RCTs.

When attrition occurs at the school level, either the school-level aggregate study-collected pre-test or school-level proficiency data could be used in analyses that diagnose or attempt to mitigate attrition bias. However, when attrition only occurs at the student level, school-level measures are of little value in adjusting for the within-school differences between students who are missing post-test scores and those who are not. Because there is no school-level attrition in the studies examined here, we cannot compare the performance of proficiency data to that of a study-collected pre-test in terms of diagnosing and reducing bias. Instead, the focus of our analyses in this chapter is to assess the extent to which different types of attrition bias exist in education RCTs. For these studies, we can examine the extent to which student-level baseline test scores have been needed to address bias due to missing follow-up test score data. If attrition bias is small, then the loss of student-level baseline covariates might be inconsequential. However, if attrition bias is a significant issue, then the loss of baseline covariates that could be used to mitigate that bias could be a serious issue.

Types of Attrition Bias

Missing outcome data can cause two types of bias: (1) biased impacts for the subgroup of individuals for whom outcome data are available, and (2) biased impacts for the original sample that was randomly assigned to treatment and control groups. The first type of bias arises when we are unable to establish a causal link between the intervention and the outcome. That is, the treatment impact cannot be solely attributed to the intervention if treatment and control students differ in other systematic ways. Therefore, if students with follow-up tests in the treatment group differ systematically from students with follow-up tests in the control group, treatment impacts could be biased for the subgroup of individuals for whom

outcome data are available. In this type of situation, including the baseline test score as a covariate in the regression will control for one way in which treatment group students may differ from control group students, thus reducing attrition bias (using the baseline test score to impute the missing values of the follow-up test, so long as that imputation is conducted separately for the treatment and control groups, is another approach).

The second type of bias arises when students without follow-up data differ systematically from students with follow-up data and impacts vary between these two groups. In this type of situation, treatment impacts could be biased for the original sample that was randomly assigned to treatment and control groups. For example, if follow-up test scores are more likely to be missing for low achievers than for high achievers, and if the intervention has a different impact for low achievers than for high achievers, the estimated impacts of the intervention might be unbiased for the high achievers, but will not be unbiased for the full sample that was randomly assigned and included both low and high achievers. In this type of situation, the baseline test score can be used to impute the missing values of the follow-up test score, thus reducing attrition bias (so long as the imputation is conducted separately for the treatment and control groups).

Data Analyses

Although it is impossible to measure the true extent of attrition bias (since by definition we do not observe outcomes for students with missing outcome data), the high correlation between pre- and post-tests in education creates a unique opportunity to develop a plausible estimate of the extent to which attrition bias may exist. We assessed the level of attrition bias in past education RCTs using study-collected baseline test scores in several ways. To address the issue of estimating unbiased impacts for those individuals for whom outcome data are available, we first calculated the extent to which rates of missing data differed among treatment and control groups. We then examined whether students with follow-up tests in the treatment group differed systematically from students with follow-up tests in the control group in terms of their baseline test scores. To address the issue of estimating unbiased impacts for the original sample that was randomly assigned to treatment and control groups, we examined whether students without follow-up tests were systematically different from students with follow-up tests. Specifically, we calculated the extent to which baseline test scores of students without follow-up data differed from the baseline test scores of students with follow-up data.

Results

Tables 6.1 and 6.2 examine treatment-control differences in rates of missing data and baseline test scores. Table 6.1 shows the differences in attrition rates between treatment and control groups. For 22 out of 24 outcomes, this difference is not statistically significant. Table 6.2 shows the differences in regression-adjusted[8] mean baseline test scores between treatment and control groups for students with follow-up tests. Because baseline test scores are highly correlated with follow-up test scores, these differences are a measure of bias resulting from differential response patterns between the treatment and control groups at follow-up. The absolute value of the difference between the treatment and control groups ranges from 0.004 standard deviations to 0.12 standard deviations, with an average absolute deviation of 0.05 standard deviations (for comparison, many experimental evaluations in education are designed to detect effects of 0.20 standard deviations). For only three outcomes the difference in means is statistically significant at the 5 percent significance level.

[8] The numbers in this table are based on a regression of the baseline test score on treatment status, a dummy variable indicating nonresponse follow-up, and an interaction between those two variables. Adjusting baseline test scores in this manner enables us to compare treatment-control differences in baseline test scores for follow-up respondents that are due to nonresponse bias only, not differences that may occur by chance as part of randomization.

Together, Tables 6.1 and 6.2 provide evidence that, in the education RCTs examined in this study, the proportion of students lacking a follow-up test score typically does not differ between the treatment and control groups and the baseline test scores of students with follow-up test scores typically do not differ between the treatment and control groups, meaning that there is little need for a student-level baseline test score in order to adjust for non-response bias. An important caveat is that the studies examined here focused on students in grades K-9 and either on curriculum changes (math, reading, and technology) or teacher training/induction programs. These findings may not generalize to other ages or types of interventions. For example, they might not apply in an evaluation of charter high schools.

Table 6.3 shows the differences in mean baseline test scores between students with and without follow-up tests. For all but three outcomes, we find that students without a follow-up test scored lower on the baseline test than students with a follow-up test by an average of 0.30 standard deviations. This difference is statistically significant (at the 5 percent significance level) for 18 outcomes. The results in Table 6.3 provide evidence that, in education RCTs, the ability to estimate unbiased impacts for the original sample that was randomly assigned to treatment and control groups will be compromised by missing outcome data if impacts differ between high and low achievers.

Table 6.1: Differences in Attrition Rates Between Treatment and Control Groups

Outcome	Treatment Group Attrition Rate	Control Group Attrition Rate	Difference	P-Value
A	0.146	0.127	0.018	0.824
B	0.121	0.100	0.021	0.908
C	0.130	0.085	0.046	0.984
D	0.136	0.099	0.037	0.931
E	0.099	0.086	0.013	0.232
F	0.108	0.124	-0.016	0.059
G	0.137	0.103	0.034	0.905
H	0.144	0.121	0.023	0.886
I	0.117	0.096	0.021	0.916
J	0.116	0.144	-0.028	0.033
K	0.102	0.087	0.015	0.181
L	0.099	0.087	0.012	0.284
M	0.196	0.224	-0.028	0.242
N	0.196	0.224	-0.028	0.242
O	0.120	0.140	-0.019	0.096
P	0.012	0.010	0.002	0.738
Q	0.000	0.002	-0.002	0.007
R	0.196	0.224	-0.028	0.242
S	0.000	0.000	0.000	N/A
T	0.162	0.157	0.005	0.555
U	0.193	0.205	-0.011	0.389
V	0.181	0.182	0.000	0.495
W	0.081	0.085	-0.004	0.779
X	0.196	0.224	-0.028	0.242
Y	0.174	0.186	-0.013	0.370

Source: Previously completed RCT studies.

RCT = randomized controlled trial.

Table 6.2: Differences in Baseline Test Scores Between Treatment and Control Groups for Students with Follow-Up Tests

	Adjusted[a] Mean Baseline Test Score				
Outcome	Control Group	Treatment Group	Difference in Means	Difference in Effect Size Units	P-Value
A	45.308	44.889	-0.419	-0.031	0.773
B	13.001	12.797	-0.205	-0.047	0.563
C	18.538	18.084	-0.454	-0.068	0.665
D	29.885	29.128	-0.758	-0.072	0.648
E	38.212	39.757	1.546	0.073	0.055
F	100.491	100.532	0.042	0.003	0.918
G	11.405	11.038	-0.367	-0.078	0.577
H	13.124	13.129	0.005	0.001	0.992
I	19.106	18.765	-0.341	-0.054	0.605
J	100.595	100.561	-0.034	-0.001	0.932
K	42.689	43.270	0.581	0.026	0.488
L	40.108	41.546	1.439	0.071	0.076
M	31.057	29.927	-1.130	-0.064	0.482
N	34.779	34.191	-0.588	-0.029	0.824
O	100.469	100.552	0.083	0.003	0.844
P	0.034	-0.039	-0.073	-0.073	0.284
Q	0.022	-0.027	-0.049	-0.049	0.438
R	33.445	32.671	-0.774	-0.056	0.676
S	31.767	31.430	-0.337	-0.039	0.581
T	31.275	31.046	-0.229	-0.056	0.719
U	84.866	83.886	-0.980	-0.080	0.555
V	20.297	19.902	-0.396	-0.064	0.520
W	59.095	59.571	0.475	0.019	0.685
X	34.589	34.145	-0.444	-0.025	0.792
Y	32.805	32.563	-0.242	-0.062	0.627

Source: Previously completed RCT studies.

[a]The numbers in this table are based on a regression of the baseline test score on treatment status, a dummy variable indicating nonresponse follow-up, and an interaction between those two variables. Adjusting baseline test scores in this manner enables us to compare treatment-control differences in baseline test scores for follow-up respondents that are due to nonresponse bias only, not differences that may occur by chance as part of randomization.

RCT = randomized controlled trial.

Table 6.3: Differences in Baseline Test Scores Between Students With and Without Follow-Up Tests

Outcome	Mean of Baseline Test Score: Students with Follow-up Scores	Mean of Baseline Test Score: Students without Follow-up Scores	Difference in Means	Difference in Effect Size Units	P-value
A	45.448	41.131	4.317	0.346	0.000
B	13.020	11.953	1.067	0.317	0.000
C	18.550	14.809	3.741	0.581	0.000
D	29.901	24.810	5.091	0.497	0.000
E	38.269	39.460	-1.190	-0.057	0.543
F	100.350	96.153	4.196	0.319	0.000
G	11.409	8.994	2.415	0.524	0.000
H	13.168	12.092	1.077	0.252	0.001
I	19.159	15.697	3.463	0.534	0.000
J	100.473	98.075	2.399	0.187	0.091
K	42.728	43.581	-0.853	-0.039	0.678
L	40.118	43.157	-3.040	-0.144	0.119
M	30.905	27.502	3.403	0.210	0.011
N	34.176	30.043	4.132	0.228	0.005
O	100.343	94.653	5.690	0.442	0.000
P	0.057	-0.638	0.695	0.699	0.000
Q	0.042	-0.329	0.371	0.374	0.231
R	33.090	29.399	3.691	0.313	0.002
S	31.411	N/A	N/A	N/A	N/A
T	31.298	28.961	2.337	0.280	0.004
U	84.920	77.238	7.682	0.372	0.000
V	20.302	17.456	2.845	0.366	0.000
W	59.156	55.309	3.847	0.165	0.265
X	34.221	31.419	2.802	0.168	0.041
Y	32.820	30.494	2.326	0.342	0.000

Source: Previously completed RCT studies.

RCT = randomized controlled trial.

Chapter 7: Conclusion

The purpose of this study has been to assess whether publically available school proficiency data can be an acceptable alternative to collecting baseline test scores in order to increase the statistical precision of impacts in RCTs. We also examined the extent to which attrition bias is a problem that needs to be addressed by statistical adjustments that use baseline test scores.

The results and conclusions presented here are based on school-level proficiency data from the State Education Data Center (SEDC) and student test score data from large, multi-district evaluations. With only one exception, all of these large-scale evaluations were conducted in elementary schools. Thus, the results may not be generalizable to studies involving middle- and high-schools. Furthermore, the results may not be generalizable to other publically-available data sources, such as school-level mean scores on state achievement tests available from school, district, or state websites. This study focused specifically on the SEDC data because it is a convenient and low-cost source for proficiency data on most US schools. This makes it a likely choice for researchers conducting large-scale, multi-district evaluations.

With respect to precision, we find that, on average, adjusting for school-level proficiency does not increase statistical precision as well as study-collected baseline test scores. Across the cases we examined, the number of schools included in studies would have to nearly double in order to compensate for the loss in precision of using proficiency data instead of study-collected baseline test data. This finding is consistent with the finding in Schochet (2008b) that small differences in the school-level R^2 can have substantial power implications, which in the context of that paper meant that it is better to accept the small bias associated with adjusting for a late pre-test than to accept the considerable power loss associated with a smaller R^2. Nevertheless, in cases where it is not feasible to collect a baseline test, using school-level proficiency data is a low-cost way to improve precision. Also, adding school-level proficiency variables as covariates in addition to the study-collected baseline test score increased the school-level R^2 by 0.05 in the data we examined, which was equivalent to increasing the number of schools in a study from 40 to 49.

With respect to attrition bias, we found little evidence that student attrition biases the impacts calculated for the subgroup of students who have follow-up test scores. However, we do find that students with follow-up test scores tend to have higher baseline test scores. This means that the impact for the subgroup of students with follow-up test scores might be different from the impact for the full sample if impacts vary by prior achievement. If student-level baseline test scores are available, they can be used to impute missing outcomes, thereby mitigating this difference.

References

Agodini, Roberto, Barbara Harris, Sally Atkins-Burnett, Sheila Heaviside, Timothy Novak, and Robert Murphy. "Achievement Effects of Four Early Elementary School Math Curricula: Findings from First Graders in 39 Schools (NCEE 2009-4052)." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Bates, Douglas. "Computational Methods for Mixed Models." Working paper, Department of Statistics, University of Wisconsin–Madison, 2009. Retrieved from http://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf.

Bloom, Howard. "Randomizing Groups to Evaluate Place-Based Programs." New York: MDRC, 2004.

Bloom, Howard, Lashawn Richburg-Hayes, and Alison Black. "Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions." *Educational Evaluation and Policy Analysis*, vol. 29, no. 1, 2007, pp. 30-59.

Braun, Henry I., and Jiahe Qian. "An Enhanced Method for Mapping State Standards onto the NAEP Scale." In *Linking and Aligning Scores and Scales*, edited by Neil J. Dorans, Mary Pommerich, and Paul Holland. New York: Springer Science + Business Media LLC, 2007, pp. 313-338.

Campuzano, Larissa, Mark Dynarski, Roberto Agodini, and Kristina Rall. "Effectiveness of Reading and Mathematics Software Products: Findings from Two Student Cohorts." Report prepared for the U.S. Department of Education, Institute of Education Sciences. Princeton, NJ: Mathematica Policy Research, February 2009.

Constantine, J., D. Player, T. Silva, K. Hallgren, M. Grider, and J. Deke. "An Evaluation of Teachers Trained Through Different Routes to Certification, Final Report (NCEE 2009-4043)." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Feuer, M. J., P. W. Holland, B. F. Green, M. W. Bertenthal, and F. C. Hemphill. "Uncommon Measures: Equivalence and Linkage Among Educational Tests." Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council. Washington, DC: National Academy Press, 1999.

Griliches, Z., B. H. Hall, and J. A. Hausman. "Missing Data and Self-Selection in Large Panels. *Annales de L'INSEE*, vols. 30-31, 1978, pp. 137-176.

Heckman, J. J. "Sample Selection Bias as a Specification Error." *Econometrica*, vol. 47, 1979, pp. 153–161.

Hedges, L., and E. Hedberg. "Intraclass Correlation Values for Planning Group-Randomized Trials in Education." *Educational Evaluation and Policy Analysis,* vol. 29, no. 1, 2007, pp. 60-87.

Isenberg, E., S. Glazerman, M. Bleeker, A. Johnson, J. Lugo-Gil, M. Grider, and S. Dolfin. *Impacts of Comprehensive Teacher Induction: Results From the Second Year of a Randomized Controlled Study* (NCEE 2009-4072). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

James-Burdumy, S., J. Deke, J. Lugo-Gil, N. Carey, A. Hershey, R. Gersten, R. Newman-Gonchar, J. Dimino, K. Haymond, and B. Faddis. *Effectiveness of Selected Supplemental Reading*

Comprehension Interventions: Findings from Two Student Cohorts (NCEE 2010-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2010.

James-Burdumy, Susanne, Wendy Mansfield, John Deke, Nancy Carey, Julieta Lugo-Gil, Alan Hershey, Aaron Douglas, Russell Gersten, Rebecca Newman-Gonchar, Joseph Dimino, Bonnie Faddis, and Janice Dole. "Effectiveness of Selected Supplemental Reading Comprehension Interventions: Impacts on a First Cohort of Fifth-Grade Students." Princeton, NJ: Mathematica Policy Research, May 2009.

Koretz, D. M., M. W. Bertenthal, and B. F. Green. "Embedding Questions: The Pursuit of a Common Measure in Uncommon Tests." (Report of the Committee on Embedding Common Test Items in State and District Assessments, National Research Council.) Washington, DC: National Academy Press, 1999.

Linn, R. L. "Linking Results of Distinct Assessments." *Applied Measurement in Education*, vol. 6, no. 1, 1993, pp. 83-102.

McLaughlin, D., and V. Bandeira de Mello. "Comparing State Reading and Math Performance Standards Using NAEP." Paper presented at the National Conference on Large-Scale Assessment, San Antonio, June 2003.

McLaughlin, D., and V. Bandeira de Mello. "Comparison of State Elementary School Mathematics Achievement Standards, Using NAEP 2000." Paper presented at the American Educational Research Association Annual Meeting, New Orleans, April 2002.

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

Perez-Johnson, Irma, Joshua Haimson, Samina Sattar, Phil Gleason, and Henry May. "Using State Tests in Education Experiments: A Discussion of the Issues." Princeton, NJ: Mathematica Policy Research, July 2009.

Puma, Michael J., Robert Olsen, Stephen Bell, and Christopher Price. "Missing Data Issues in Randomized Control Trials: What to Do When Data Are Missing?" Annapolis, MD: Chesapeake Research Associates, LLC, August 2008.

Raudenbush, S.W. "Statistical Analysis and Optimal Design for Cluster Randomized Trials." *Psychological Methods*, vol. 2, no. 2, 1997, pp. 173-185.

Rubin, Donald B. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association*, vol. 91, 1996, pp. 473-489.

Rubin, Donald B. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley, 1987.

Schochet, Peter. "Statistical Power for Random Assignment Evaluations of Education Programs." *Journal of Educational and Behavioral Statistics*, vol. 33, no. 1, 2008a, pp. 62-87.

Schochet, Peter. "The Late Pretest Problem in Randomized Control Trials of Education Interventions." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, October 2008b.

Spybrook, Jessaca, Stephen W. Raudenbush, Richard Congdon, and Andrés Martínez. “Optimal Design for Longitudinal and Multilevel Research: Documentation for the “Optimal Design” Software.” 2009. Retrieved from http://sitemaker.umich.edu/group-based/files/od-manual-v200-20090722.pdf.