NCEE 2012-4016

U. S. DEPARTMENT OF EDUCATION

Estimating the Impacts of Educational Interventions Using State Tests or Study-Administered Tests

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

Estimating the Impacts of Educational Interventions Using State Tests or Study-Administered Tests

October 2011

Authors

Robert B. Olsen
Fatih Unlu
Cristofer Price
Abt Associates

Andrew P. Jaciw
Empirical Education

Project Officer

Meredith Bachman
Institute of Education Sciences

NCEE 2012-4016
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 (IES), under contract ED-04-C0-0012/0006.

Disclaimer
The Institute of Education Sciences at the U.S. Department of Education contracted with Abt Associates to develop a report that produces empirical evidence on the differences in impact estimates and standard errors resulting from statistical models that use state assessments to measure student achievement, statistical models that use study-administered tests to measure student achievement, and statistical models that use a combination of these two types of tests. 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 2011

Olson, R.B., Unlu, F., Jaciw, A.P., and Price, C. (2011). *Estimating the Impacts of Educational Interventions Using States Tests or Study-Administered Tests.* (NCEE 2012-4016). 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 Institute of Education Sciences 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 Conflict of Interest

Three of the authors for this report, Robert B. Olsen, Fatih Unlu, and Cristofer Price, are employees of Abt Associates Inc., with whom IES contracted to develop the methods that are presented in this report. Andrew P. Jaciw is an employee of Empirical Education Inc., with whom Abt subcontracted to contribute to the study. None of the authors or other staff involved in this study have financial interests that could be affected by the content in this report.

Foreword

The National Center for Education Evaluation and Regional Assistance (NCEE) within the Institute of Education Sciences (IES) is responsible for (1) conducting evaluations of federal education programs and other programs of national significance to determine their impacts, particularly on student achievement; (2) encouraging the use of scientifically valid education research and evaluation throughout the United States; (3) providing technical assistance in research and evaluation methods; and (4) supporting the synthesis and wide dissemination of the results of evaluation, research, and products developed.

In line with its mission, NCEE supports the expert appraisal of methodological and related education evaluation issues and publishes the results through two report series: the *NCEE Technical Methods Report* series that offers solutions and/or contributes to the development of specific guidance on state of the art practice in conducting rigorous education research, and the *NCEE Reference Report* series that is designed to advance the practice of rigorous education research by making available to education researchers and users of education research focused resources to facilitate the design of future studies and to help users of completed studies better understand their strengths and limitations.

This *NCEE Reference Report* examines the differences in impact estimates and standard errors that arise when these are derived using state achievement tests only (as pre-tests and post-tests), study-administered tests only, or some combination of state- and study-administered tests. State tests may yield different evaluation results relative to a test that is selected, and administered, by the research team for several reasons. For instance, (1) because state tests vary in content and emphasis, they also can vary in their coverage of the types of knowledge and skills targeted by any given intervention. In contrast, a study-administered test will correspond to the intervention being evaluated. In addition to differences in alignment with treatment, state tests may yield divergent evaluation results due to differences in (2) the stakes associated with the test, (3) missing data, (4) the timing of the tests, (5) reliability or measurement error, and (6) alignment between pre-test and post-test. Olsen, Unlu, Jaciw, and Price (2011) discuss how these six factors may differ between state- and study-administered tests to influence the findings from an impact evaluation.

Specifically, Olsen et al. use data from three single-state, small-scale evaluations of reading interventions that collected outcomes data using both study-administered and state achievement tests to examine this and other issues. The authors found that (1) impact estimates based on study-administered tests had smaller standard errors than impact estimates based on state tests, (2) impacts estimates from models with "mismatched" pre-tests (e.g., a state pre-test used in combination with a study-administered post-test) had larger standard errors than impact estimates from models with matched pre-tests, and (3) impact estimates from models that included a second pre-test covariate had smaller standard errors than impact estimates from models that included a single pre-test covariate. Study authors caution that their results may not generalize to evaluations conducted in other states, with different study-administered tests, or with other student samples.

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 IES staff and the members of the IES Methods Working Group. The authors would like to thank Ricky Takai and Jacob Klerman of Abt Associates and Irma Perez-Johnson, John Burghardt, and John Deke of Mathematica Policy Research for their helpful suggestions on earlier drafts of this report, Qingfeng Zhao from Empirical Education for estimating the analysis models, and David Cook from Abt Associates for conducting hypothesis tests and reviewing earlier drafts of the report. Finally, the authors are grateful to Jan Nicholson and Katheleen Linton for their expert assistance in producing this document.

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.

Contents

List of Tables

Page

List of Figures

Page

A. Introduction

1. Background

Many evaluations of educational interventions estimate the impacts of the intervention on student achievement. A key design question for these evaluations is how to measure student achievement. While some evaluations aim to estimate an intervention's impacts on some specific subdomain, many evaluations are focused on achievement in reading or mathematics more generally. Therefore, many evaluations face the challenge of deciding how to obtain general measures of reading and/or mathematics achievement for the students in the study sample.

Since the passage of the No Child Left Behind Act (NCLB) in 2001, states have been required to test students in both reading and mathematics in every grade between grades 3 and 8, and in at least one grade in high school. Furthermore, some states and districts have been willing to provide data on student scores on these tests for federally funded evaluations. Several evaluations sponsored by the Institute of Education Sciences (IES) rely on state tests, including an ongoing evaluation of charter schools, an evaluation of teacher induction programs (Glazerman et al. 2010), an evaluation of teacher professional development in early reading (Garet et al. 2008), and a recent evaluation of the Student Mentoring Program (Bernstein et al. 2009). Therefore NCLB has created a new option for measuring student achievement in evaluations of educational effectiveness—to rely on the scores from state-required tests.

In addition, many states have received federal funding to improve their state data systems. Improved state data systems will help to make student test scores from state assessments more easily accessible for research purposes. Under the Educational Technical Assistance Act of 2002, IES received funding for a grant program to support states in their efforts to improve their longitudinal data systems (Institute of Education Sciences 2008). To date, 41 states and the District of Columbia have received at least one grant under this program.[1]

There are at least four reasons for evaluators to consider relying less on study-administered tests and more on state tests to measure student achievement. First, it would reduce the burden on students who already face substantial testing. Second, it could substantially reduce the costs of conducting evaluations: testing students is expensive, while collecting state test scores is relatively inexpensive. Third, scores on state tests can have consequences for the students who take them, so these tests may elicit a higher level of effort than the "no-stakes" tests that studies administer. Fourth, state assessments are the primary tool that policymakers use to assess student achievement and hold schools accountable for it.

At the same time, state tests can present challenges to researchers conducting educational evaluations. The most notable example is that in multi-state evaluations, researchers who collect state test scores

[1] See the data on the grant program's website for more details: http://nces.ed.gov/Programs/SLDS/stateinfo.asp.

must develop a defensible approach to addressing the fact that each state has its own test. A common approach is to standardize student test scores so that the impact estimates reflect effect sizes relative to a well-defined reference population of students in the same state. However, whether pooling data from different tests can be justified, the conditions under which it is defensible, and whether pooled impact estimates are sensitive to the approach selected are open questions. For a discussion of the issues associated with using state tests in educational evaluations, see May et al. (2009).

In choosing between tests for any particular evaluation, we should first ask ourselves if, from a substantive perspective, one test is clearly preferable to the other. To answer this question, we must first choose the achievement domain (e.g., math or reading) and in some cases the subdomain (e.g., vocabulary or fluency) we hope to measure. Then we face a choice between alternative tests that we believe measure the same underlying domain or subdomain. In some studies, there may be strong substantive reasons to prefer one test to another test within the same domain.

In many evaluations, there may be a group of tests in the same domain where none of the tests is clearly "better" for the evaluation from a substantive perspective than the other tests. In these instances, the optimal choice will be heavily influenced by the relative costs of different options. The current trend toward using state tests in educational evaluations is presumably driven largely by costs: the marginal cost of administering one additional achievement test is much larger than the marginal cost of obtaining state test scores for one additional student.

However, there are two reasons why it is difficult to assess the cost implications of choosing state tests over study-administered tests. First, data on the relative costs are not easily accessible. While the marginal costs of administering study tests are surely many times larger than the marginal costs of collecting state test scores, data on these costs are not generally available in the public domain. In addition, while our experience suggests that the costs of negotiating access to state assessment data from a state or district are non-trivial, systematic data on these costs are also not publicly available.

Second, there is no systematic evidence on whether the parameters that determine sample size requirements in impact studies differ between state tests and study-administered tests. The sample size requirements for Randomized Controlled Trials (RCTs) depend on key parameters such as the intra-class correlations and R-squares of the regression (e.g., Hedges & Hedberg 2007, Schochet 2008b). *However, to the best of our knowledge, there is no published research on the values of these parameters for state assessments.* Furthermore, there is no reason to be confident that the results from the literature, which are based on data from study-administered tests (e.g., Jacob & Zhu 2009) and from pre-NCLB district tests (e.g., Bloom, Richburg-Hayes, & Black 2007), would apply to state tests.

Therefore, the possibility of using state tests in education evaluations provides both opportunities for substantial cost savings and some challenges and uncertainties about the cost implications.

2. Research Questions

This report takes an important first step in assessing the consequences of relying on state tests versus study-administered tests for general, student-level measures of reading and math achievement in evaluations of educational effectiveness. In this study, we address four research questions.

Question 1: Will impact evaluations in education yield different impact estimates and statistical precision of the impact estimates if they use state tests to measure student achievement at both baseline and follow-up instead of administering standardized tests at both points in time as part of the evaluation? In the previous section, we identified several reasons why the two types of tests could yield systematically different impact estimates. Whatever the reason, evidence that state tests tend to yield larger or smaller impact estimates than study-administered tests would justify a reassessment of how large the impacts of educational interventions must be to be considered educationally meaningful. For example, if the impacts of different middle school math curricula tend to be twice as large for study-administered tests as for state tests, it would make no sense to set standards for judging the magnitude of the impact estimates that are the same for both types of tests. In contrast, if there were systematic empirical evidence indicating that differences in impacts between the two types of tests are small and random, then the choice between the two types of tests in any evaluation would be less important, and we would ignore the distinction when interpreting the results from educational evaluations that vary in the type of test they chose to use.

In addition, evidence suggesting that state and study-administered tests yield systematically different impact estimates or standard errors would have implications for the sample size requirements of individual studies and the overall cost of a research portfolio that relies on state test scores. If one type of test yields smaller impacts than the other type of test, evaluations will need larger samples to detect impacts for the test that yields smaller impacts than for the test that yields larger impacts. For example, suppose that the impact estimates from state tests tend to be smaller than the impact estimates from study-administered tests (e.g., if state tests are less "well-aligned" to the intervention). Under this scenario, we would need larger samples to detect impacts using state test scores than to detect impacts using study-administered test scores. In addition, if one type of test yields less precise impact estimates than the other (holding sample size constant), evaluations will need larger samples to detect impacts of a given size for the test that yields less precise estimates. For example, suppose that holding the sample size constant, state tests yield impact estimates with larger standard errors than do study-administered tests. This would suggest we need larger samples in evaluations that rely on state tests than in evaluations that rely on study-administered tests.

Question 2: Does measuring student achievement using one type of test at baseline and another type of test at follow-up reduce the statistical precision of the impact estimates? The precision of the impact estimates depends on the R-square of the regression and thus on the correlation between pre-test scores and post-test scores. Furthermore, we would expect a lower pre-post correlation for a "mismatched" pre-test (e.g., a pre-test that differs from the post-test, like a state pre-test for a study-administered post-test) than for a "matched" pre-test. Therefore, other things held constant, we would expect less precise impact estimates or larger sample size requirements when studies choose a mismatched pre-test than when they choose a matched pre-test.

Differences in precision may have sample size and thus cost implications for educational evaluations. Studies that use a mismatched pre-test may require a larger sample to aim to detect impacts of a given size than if they had instead used a matched pre-test. However, the magnitude of this difference is an empirical question.

Question 3: Does controlling for both types of student achievement measures at baseline (i.e., pre-test scores) increase the statistical precision of the impact estimates? A richer set of control variables will increase the R-square of the impact regression, increase the precision of the impact estimates for a fixed sample size, and reduce the study's sample size requirements. Therefore, an obvious question is whether controlling for baseline measures of achievement from *both* study-administered tests and state tests increases the R-square of the regression and yields more precise impact estimates than controlling for only one of the two achievement measures. If so, studies could be conducted with smaller samples if they would collect baseline achievement scores from both types of tests.

Question 4: Can using both types of student achievement measures at follow-up (i.e., post-test scores) increase the statistical precision of the impact estimates? Under some conditions, collecting multiple outcome measures in the same domain may lead to more precise impact estimates. For multiple outcomes in the same domain, Schochet (2008b) proposes constructing composite outcomes by averaging the scores on the individual tests, and basing the "confirmatory" impact analysis on the composite measures. Under certain conditions, composite outcome measures will produce more reliable outcome measures, reduce the variance of the measurement error, produce more precise impact estimates, and reduce the sample size requirements of the study.

To see this, suppose the two tests were as similar as two different forms of the same test. If the two forms had equal reliability, we would expect the average of the two scores to provide a more reliable measure of student achievement than a student's score from either test individually.

However, study-administered reading tests and state reading tests are *not* different forms of the same test, and they may not have equal reliability. Therefore, it is not clear whether we should expect more precise impact estimates from simple averages of scores from the two tests than from either test alone.

The remainder of this report will proceed as follows. In Section B, we discuss the possible reasons why the choice between state and study-administered tests may affect the impact estimates or their standard errors. In Section C, we describe our data and methods. In Section D, we report the results from the analysis. In Section E, we offer some concluding thoughts and suggestions for future research.

B. Why the Choice between Tests May Matter in Educational Evaluations

This section provides a conceptual or theoretical assessment of how the choice between tests could affect the impacts we estimate, the precision with which they are estimated, or both. In this section, we first specify the types of models that researchers typically estimate in educational impact evaluations. We then use these models to provide a framework for the discussion that follows. This discussion introduces six factors that may influence an evaluation's sample size requirements, the magnitude of the impact estimates, or both: (1) reliability or measurement error, (2) missing data, (3) alignment between pre-test and post-test, (4) the timing of the tests, (5) alignment between treatment and post-test, and (6) the stakes associated with the test.

Models for estimating the impacts of educational interventions. To assess the potential consequences of selecting different measures of student achievement, it is helpful to specify one or more formal models of the type we estimate in educational impact evaluations. In an *unclustered design*, for which the unit of random assignment is the same as the unit of analysis (e.g., random assignment of students within schools), a standard model of student achievement can be expressed as follows:

$$(1) \qquad Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \varepsilon_i$$

where Y_i is the outcome of interest (i.e., post-test) for student *i*, T_i is the treatment indicator of student *i* (equals 1 for students assigned to treatment and 0 for those assigned to control), X_i is the value of the pre-test variable for student *i*, and ε_i is the usual student-level error term.[2] Schochet (2008a) and others show that the variance of the impact estimate ($\hat{\beta}_1$) from this regression can be represented as:

$$(2) \qquad Var(\hat{\beta}_1) = \frac{\sigma^2(1-R^2)}{N\bar{T}(1-\bar{T})}$$

where σ^2 is the variation in the outcome of interest across students, R^2 is the R-square of the regression model (i.e., the proportion of the total variation that is explained by the pre-test variable), N is the total number of students, and $\bar{T}$ is the proportion of students assigned to the treatment group. Equation 2 shows that the variance of the impact estimate is simply a function of the unexplained variance in the outcome variable ($\sigma^2(1-R^2)$).

[2] For simplicity and without loss of generality, we consider an impact model that includes only the pre-test as a covariate.

In a two-level *clustered design,* where the unit of random assignment is at a higher level (e.g., school or classroom) than the unit of analysis (e.g., student), the impact regression model can be expressed as follows:

$$(3) \qquad Y_{ij} = \gamma_0 + \gamma_1 T_j + \gamma_2 X_{ij} + \mu_j + \varepsilon_{ij}$$

where Y_{ij} is the outcome for student *i* in cluster *j*, T_j is the treatment indicator for cluster *j and* X_{ij} is the pre-test variable for student *i* in cluster *j*. The cluster-level and student-level terms, μ_j and ε_{ij} respectively, are assumed to be independent of each other and of the covariates in the model.

In a clustered design, the variance of the impact estimate ($\hat{\gamma}_1$) can be calculated using the following formula (Murray 1998, Schochet 2008a, Spybrook et al. 2009):

$$(4) \qquad Var(\hat{\gamma}_1) = \frac{\sigma_c^2(1-R_c^2)}{J\overline{T}(1-\overline{T})} + \frac{\sigma_s^2(1-R_s^2)}{Jn\overline{T}(1-\overline{T})}$$

where J is the number of clusters, n is the number of students per cluster, and $\overline{T}$ is the proportion of clusters assigned to the treatment group. The variance of the treatment effect can be decomposed into the contribution due to variation across clusters and the contribution due to variation within clusters. At the cluster level, σ_c^2 denotes the cluster-level variance (i.e., the variance of μ_j), and R_c^2 is the proportion of the cluster-level variance explained by the pre-test variable. At the student level, σ_s^2 denotes the student-level variance (i.e., the variance of ε_{ij}), and R_s^2 is the proportion of the student-level variance explained by the pre-test variable.

Reliability or measurement error. Reliability is defined as the proportion of a measure's variability that is free of measurement error. Measurement error in either the outcome variable (post-test) or in the pre-intervention measure of the outcome variable (pre-test) will increase the sample size requirements of the study to detect impacts of a given size. Whether state or standardized tests tend to be more reliable is an open question, and we would conjecture that there is far more variability within each of these broad categories of tests than between categories. Nonetheless, given the feasible test options for a particular evaluation, differences in the amount of measurement error could have important implications for an evaluation's sample size requirements.

Research on the role of measurement error in estimation has a long history in both the psychometric literature (e.g., Sutcliffe 1958, Williams & Zimmerman 1995) and the econometrics literature (e.g., Griliches & Hausman 1986, Bound & Krueger 1991). The seemingly most benign form of measurement error is referred to by econometricians as "classical measurement error," in which the error is normally distributed with mean zero and constant variance—and is uncorrelated with any of the independent variables in the model. However, even classical measurement error can have unfortunate consequences for studies of the impacts of educational interventions.

Based on a careful examination of standard variance formulas in Equations 2 and 4, we reach two conclusions about the effects of classical measurement error on the precision of the impact estimate:

1. ***Classical measurement error in the post-test variable will reduce the precision of the impact estimate, holding the sample size constant, and increase the sample size required to detect impacts of a given size.*** This is consistent with the results from the econometrics literature (e.g., Angrist & Krueger 1999) and the psychometric literature (e.g., Williams & Zimmerman 1995) and it can be demonstrated formally using the variance formulas in Equations 2 and 4. For example, consider an unclustered design, and suppose we normalize the outcome variable to have a variance equal to 1. Under this scenario, an increase in the classical measurement error of the outcome variable, holding its total variance constant, decreases the R^2 value since it decreases correlation between pre-test scores and post-test scores. Holding other factors constant, this increases the variance of the impact estimate (see Equation 2).[3] Similarly, in a clustered design, classical measurement error in the outcome at the cluster level decreases R_c^2, and measurement error in the outcome at the student level decreases R_s^2, thereby decreasing the precision of the impact estimates.

2. ***Classical measurement error in the pre-test variable will also reduce the precision of the impact estimate and increase the sample size requirements of the study.*** As with classical measurement error in the post-test variable, classical measurement error in the pre-test variable will reduce the correlation between pre-test scores and post-test scores, reduce the R-square of the regression, and decrease the precision of the impact estimate. More specifically, classical measurement error will reduce the R-square of the regression model in unclustered designs (R^2 in Equation 2) and the "within-cluster" R-square in clustered designs (R_s^2 in Equation 4), which reduces the precision of the impact estimates (holding the total outcome variance constant).

We show below that *classical measurement error in pre-test scores or post-test scores will not bias the estimated impact of the treatment in studies that randomize units to treatment or control conditions*. Classical measurement error in an independent variable will bias the estimated coefficient *on the mismeasured variable* toward zero (Angrist & Krueger 1999). However, the key question for impact evaluations is whether measurement error in one independent variable (e.g., a pre-test covariate) yields bias in the estimated coefficient on a different independent variable (e.g., the treatment indicator).

We show that the answer to this question depends on the study's research design. In general, since the covariates may be correlated with the treatment, it can be shown that measurement error leads to bias in the estimated treatment effect. However, in experimental studies, randomization ensures that

---

[3] In studies where the variance of the outcome variable is *not* normalized to 1, classical measurement error also increases the variance of the impact estimate through an increase in the variance of the outcome variable (σ^2).

in expectation, the treatment will be uncorrelated with the covariates, and we show below that this ensures that the estimated treatment effect will be unbiased *even if the covariates suffer from measurement error*.

To see this, consider the following formula for the treatment effect (see Equation 1):

$$(5) \qquad \beta_1 = \frac{\beta_{Y|T} - \beta_{Y|X}\beta_{X|T}}{1 - R^2_{X|T}}$$

where $\beta_{Y|T}$ is the bivariate regression coefficient for the regression of the outcome variable (i.e., the post-test variable) on the treatment indicator, $\beta_{Y|X}$ is the bivariate regression coefficient for the regression of the outcome variable on the covariate in the model (i.e., the pre-test variable), $\beta_{X|T}$ is the bivariate regression coefficient for the regression of the covariate on the treatment indicator, and $R^2_{X|T}$ is the R-square of that regression model.

Because randomization ensures that the pre-test and treatment variables are uncorrelated with each other, $\beta_{X|T}$ and $R^2_{X|T}$ are both zero in experimental studies. Therefore, with random assignment, Equation 5 simplifies to Equation 6:

$$(6) \qquad \beta_1 = \beta_{Y|T}$$

Equation 6 indicates that the underlying treatment effect parameter that we estimate from a regression with the pre-test as a covariate (β_1) is identical to the treatment effect parameter that we estimate from a regression that excludes the pre-test ($\beta_{Y|T}$)—or put differently, is equal to the expected mean difference between the treatment and control groups. Furthermore, this conclusion holds regardless of the amount of measurement error in either the pre-test or the post-test. Therefore, we conclude that classical measurement error in the pre-test does not create any bias in the estimated treatment effect.

At the same time, it is important to note that measurement error in test scores may not be "classical" or random. Since there are an infinite number of types of non-random measurement error, we cannot make any general statements about how non-random measurement error would affect the magnitude of the impact estimates or the precision of those estimates. However, some forms of non-random measurement error lead to biased impact estimates. For example, suppose that an intervention has a positive effect on student achievement and reduces the fraction of students who are at risk of "failing" a particular achievement test. Furthermore, suppose teachers "cheat" and provide extra assistance during the testing time to students at risk of failing the test. Under this scenario, teacher cheating would lead to more systematic (and positive) measurement error in the control group than in the treatment group, and the estimated impact would be biased toward zero.

Finally, May et al. (2009) raise particular concerns about the reliability of state tests that would lead to non-random measurement error, such as their reliability for very low-scoring students whose level of achievement falls far below the threshold for proficiency set by the state. While these concerns are

valid, it is an open question whether, in practice, state tests tend to produce more or less reliable measures of achievement than the study-administered standardized tests that researchers often choose for their studies.

Missing data. Missing data in either the outcome variable (post-test) or in the pre-intervention measure of the outcome variable (pre-test) can introduce bias in the impact estimates, alter the sample so that the impact estimates characterize a different population than initially intended, and increase the sample size requirements of the study due to the loss of sample. Therefore, if the missing data rates or patterns are likely to differ substantially between state and study-administered tests, the choice between these two options could have important implications for the internal validity of the study design and the costs of the evaluation.

For state test scores, the rate and pattern of missing data will depend on state testing policies, the amount of student mobility across states, and even the amount of student mobility within states if state test scores are collected from schools or districts instead of from states.[4] For study-administered tests, missing data rates may also depend on student mobility, since study tests are usually administered in or near participating schools, and the missing data problem may be more serious than for state tests if any movement out of a participating school leads to study attrition. However, missing data for study-administered tests may also depend on the requirements associated with informed consent (e.g., passive versus active consent) and on factors that affect a student's willingness to take the test and parental willingness to allow it (e.g., financial incentives, the convenience of the testing site, or the level of encouragement from school officials to participate in the evaluation).

If data on either pre-test scores or post-test scores are Missing Completely at Random (Rubin 1976), and researchers delete cases with missing values, the size of the analysis sample and the precision of the impact estimates will depend on the missing data rate.[5] Loss of sample for purely random reasons has the same effect as simple random sampling: the sample size is reduced and statistical power is reduced. In unclustered designs, missing data increases the variance of the impact estimate by reducing the number of students in the analysis sample (see Equation 2). In clustered designs, missing data can increase the variance of the impact estimate by reducing either the number of clusters, if data are missing for entire classes or schools, or the number of students per cluster, if data are missing for individual students within clusters. To compensate, evaluations need to select a larger initial sample to achieve any pre-specified level of precision (e.g., Minimum Detectable Effect Size).

If data on either pre-test scores or post-test scores are not Missing Completely at Random, the story is more complicated. While a summary of the methods to address more problematic forms of missing data is beyond the scope of this report, readers may want to review selected scholarly papers (e.g.,

---

[4] While districts will typically be able to provide state test score data for students who change schools within the district, only states will be able to provide state test score data for students who change districts within the same state.

[5] While imputation would seem to be a more appealing option than case deletion to avoid the loss of sample, simulations in Puma et al. (2009) suggest that imputation methods do not always yield more precise impact estimates than case deletion.

Rubin 1976, 1987, 1996), papers that summarize missing data methods for empirical researchers (e.g., Graham 2009, Schafer 1999, Allison 2002), and a recent technical methods report focused on missing data in random assignment studies in education (Puma et al. 2009).

However, it is clear that when pre-test scores or post-test scores are missing for some cases on a non-random basis, biased impact estimates can result (e.g., Puma et al. 2009). Bias can result from a missing data mechanism that leaves us with an analysis sample of students, those with non-missing test scores, for which there are systematic differences between the treatment and control groups. Bias can also result from a missing data mechanism that leaves us with an analysis sample of students that is not representative of the population of interest.

To the best of our knowledge, there is no empirical evidence on whether state tests tend to yield more or less bias from missing data than commonly used study-administered tests (using conventional procedures for administering these tests in educational evaluations). However, if one approach to collecting achievement data does systematically yield greater missing data bias than the other approach, this should be an important factor in choosing between collecting state test scores and administering a test as part of the evaluation.

Alignment between pre-test and post-test. It is well known that pre-test scores can make an important contribution to RCTs by increasing the precision of the impact estimates and reducing the evaluation's sample size requirements (e.g., Bloom et al. 2007). The gain in precision from including pre-test scores in the model depends on the R-square of the regression—or better put, on the increase in the R-square from including the pre-test variable as a covariate in the model. For example, Bloom et al. (2007) provides estimates of the precision gains from including pre-test covariates in the model for studies that randomize schools to experimental conditions.

The choice between pre-tests with different levels of alignment with the intervention may have important implications for the sample size requirements of the study. Consider two different tests, one that is well-aligned with the post-test, perhaps because it is a different form of the same test, and another which is not so well-aligned with the post-test. In this scenario, we would expect the well-aligned pre-test variable to be more highly correlated with the post-test, yield a larger R-square in the regression model, and produce impact estimates with smaller standard errors than the pre-test variable that is less well-aligned.

This analysis has potentially important implications for the choice between study-administered pre-tests and pre-test scores from state assessments. In general, other things being equal, we would expect an evaluation to have larger sample size requirements if the evaluation chooses to collect pre-test scores from state assessments when post-test scores are obtained from a study-administered test, or vice versa, than if both pre-test and post-test scores come from the state test.

To the best of our knowledge, there is no empirical evidence on the consequences of choosing a state pre-test for a study-administered post-test, or vice versa. However, Bloom et al. (2007) present related evidence. The authors found that pre-tests that are mismatched with respect to subject area (e.g., a math pre-test for a reading post-test) can sometimes result in significantly less statistical power and significantly higher sample size requirements than tests that are matched on subject area

(e.g., a reading pre-test for a reading post-test). However, whether mismatched pre-tests from the same subject area, or domain, yield substantially less precise impact estimates is an open question.

Timing of the test. The timing of the pre-tests and post-tests may affect the estimated impacts and their standard errors. Since state testing is an important activity, few schools would agree to participate in a study that requires study-related testing during the time in which state tests are scheduled to be administered. Therefore, study-administered tests need to be administered either before or after state testing.

For post-tests, conducting study testing before the state tests leaves a shorter follow-up window, while conducting it after the state tests produces a longer follow-up window. If shorter-term impacts tend to be smaller than longer-term impacts, we might expect larger impacts from later tests than from earlier tests. However, the implications of the timing of the test for the magnitude of the impacts could depend on the nature of the intervention and the pattern of impacts over time.

For pre-tests, the story may be clearer. While state pre-test scores always come from tests taken in the previous spring, study pre-tests are often administered in the fall after students return to school. If later pre-tests reduce the length of the follow-up period, this may increase the correlation between pre-test scores and post-test scores, which could increase the R-square of the impact regression, increase the precision of the impact estimates for fixed sample, and reduce the sample size requirements of the study.

Shorter follow-up windows may have implications for the precision of the impact estimates as well. For example, consider two design options, both of which involve a post-test administered in May. Under Option A, the study will rely on a pre-test from the previous May; under Option B, the study will rely on a pre-test score from September after students return to school. Other things being equal, we would expect the correlation between pre-test and post-test to be larger for Option B than Option A because the amount of time between tests is shorter.[6] Furthermore, if the pre-post correlation is larger for Option B than for Option A, we would expect the R-square of the impact regression to be higher and sample size requirements to be lower under Option B than under Option A.

As shown in Schochet (2008c), fall pre-testing can generate biased impact estimates if pre-testing occurs after treatment group members are first exposed to the intervention. If the intervention has positive effects on achievement very shortly after the intervention is introduced, and before the pre-test is administered, the intervention will affect the pre-test scores of students in the treatment group. Under this scenario, the impact estimates will be biased toward zero (see Schochet 2008c). When study pre-tests are administered in the fall, and after the start of the intervention, we would expect

[6] This phenomenon could be exacerbated by summer learning loss. If summer learning loss were constant across the sample, it would have no affect on correlation between pre-test scores and post-test scores. However, if summer learning loss varied across individuals, we would expect this to reduce the pre-post correlation in test scores (relative to measuring post-test scores just prior to the summer). In practice, if summer learning loss reduces the pre-post correlation in test scores, it will reduce the R-square of the regression and reduce the precision of the resulting impact estimates.

study-administered pre-tests to yield smaller impact estimates than state pre-tests administered in the spring (other things being equal).

Alignment between treatment and post-test. Education researchers often worry about selecting tests that are insufficiently aligned to the intervention under the fear that the study will fail to capture the effects of the intervention. On the other hand, education researchers sometimes worry about selecting tests that are overly aligned to the intervention so that we would expect a positive impact almost by construction. In general, we would expect larger impacts on post-tests that are better aligned to the intervention than on post-tests that are less well-aligned to the intervention. For example, consider an intervention that is designed to boost reading comprehension, but that does not focus on vocabulary. In this instance, we might expect to find larger impacts on tests of reading comprehension than on general reading achievement tests that measure reading comprehension, vocabulary, and perhaps other subdomains.

However, even if we focus on achievement measures in the same domain or subdomain, different tests may weigh the subdomains differently. Returning to our example from the previous paragraph, if an intervention has a positive effect on reading comprehension but no effect on vocabulary, its effect on general measures of reading achievement will depend on how the two subdomains are weighted in constructing an overall score. We would expect this intervention to have larger effects on general reading achievement tests that give more weight to the subdomains affected by the intervention and less weight to the subdomains not affected by the intervention.

We have no theoretical or empirical basis for expecting state tests to be consistently better aligned, or consistently worse aligned, with the interventions we study than study-administered tests. Of course, we would expect study-administered tests that are narrow in their scope and chosen to be well-aligned to the intervention (e.g., a nationally normed vocabulary test for a vocabulary-focused intervention) to produce larger impact estimates on average than broader tests of achievement of any type. However, among broad tests of reading or mathematics achievement, we have no priors on whether state tests or study-administered tests in the same domain will tend to be better aligned with the interventions we evaluate. On the one hand, we might expect researchers to select broad study-administered tests that give substantial weight to the subdomains on which the intervention focuses. On the other hand, if the intervention was developed to boost student performance relative to particular state standards, and the intervention is being tested and evaluated in the same state, the state test may be more closely aligned to the intervention than any of the possible study-administered tests.

Stakes associated with the test. Different tests vary in the stakes associated with student performance on the tests. State tests are linked to state accountability systems, and these tests have high stakes for schools, especially those that are close to the threshold for making Adequate Yearly Progress. In some states (e.g., Texas), state tests also have high stakes for students and can affect whether students advance to the next grade level. In contrast, students have little incentive to perform well on study-administered tests. Student scores on these tests have no effect on the students who take them. Given the differences in incentives, we would expect that students in general exert more effort on state tests than on study-administered tests.

It is not clear how the stakes associated with the post-test would affect the estimated impacts or the precision with which they are measured. We can imagine simple models under which the stakes

associated with the test affect neither the impact estimates nor the precision of the impact estimates. Consider two different tests that are identical except for the stakes associated with the outcomes (or consider two different administrations of the same test, one with high stakes attached and the other with low stakes attached). Furthermore, suppose that (1) test stakes have a constant positive effect on effort exerted in taking a test, and (2) the effort exerted on the test has a constant positive effect on student test scores. Under this model, we would expect to find higher scores on high-stakes tests than low-stakes tests, no difference in the reliability of individual scores between the two tests, no difference in impacts between the two tests, and no difference in the precision of the two impact estimates between the two types of tests.

However, under other models, the stakes associated with the test *could* differentially affect student test scores in the treatment and control groups, which would lead to different impacts. For example, suppose the effect of additional effort varies with student achievement *at the time the test is administered*.[7] Furthermore, suppose that at the time the post-test is administered, scores are higher for the treatment group than for the control group—that is, the treatment had a positive effect on student achievement. Under this scenario, raising the stakes associated with a test could affect the effort exerted by both groups equally, but affect their post-test scores unequally, because the effect of additional effort on student test scores may not be the same for the higher-scoring treatment group than for the lower-scoring control group.

In addition, under some scenarios, the stakes associated with the test may affect the precision of the impact estimates through the reliability of the test scores. Suppose that when the stakes associated with a test's outcome are low, students are inclined to guess randomly on some questions which they could answer correctly, without guessing, simply to conserve effort. Furthermore, suppose that as the stakes associated with the test increases, student effort increases, and the fraction of questions for which the student will guess randomly decreases. Since random guessing introduces random measurement error into students' test scores, we would expect a positive relationship between the stakes of the test and reliability of the test scores. If increasing the stakes associated with the post-test increases the reliability of the post-test scores, we would expect it to increase the precision of the impact estimates (see the subsection earlier in this section titled "Reliability or measurement error").

We are not aware of any empirical evidence on the effect of the stakes associated with achievement tests on student effort, or on the effects of test-taking effort on student test scores. This means that we have no prior expectations on whether to expect educational interventions to have larger impacts on higher-stakes tests or on lower-stakes tests. In addition, we are not aware of any empirical evidence on the relationship between the stakes associated with the test and the reliability of the test score measures, which affects the precision of the impact estimates. Therefore, the effect of testing stakes on the precision of the impact estimates in educational evaluations is unknown at this time.

[7] For example, additional effort exerted in taking a test might have a smaller effect on test scores for higher achieving students due to ceiling effects.

C. Research Questions, Data and Analysis Methods

This section describes the research design for the study, the data used in the analysis, and the methods that we used to address the four research questions specified in Section A.

1. Research Questions and Hypotheses

This study addresses the four questions presented in Section A:

Question 1: Will impact evaluations in education yield different impact estimates and statistical precision of the impact estimates if they use state tests to measure student achievement at both baseline and follow-up instead of administering standardized tests at both points in time as part of the evaluation? To address this question, we conduct formal statistical tests of whether relying on study tests yields different impact estimates and standard errors than relying on state tests.

Question 2: Does measuring student achievement using one type of test at baseline and another type of test at follow-up reduce the statistical precision of the impact estimates? To address this question, we conduct a formal statistical test of whether "mismatched pre-tests" (i.e., a state pre-test for a study-administered post-test, or vice versa) yield less precise impact estimates than matched pre-tests.

Question 3: Does controlling for both types of student achievement measures at baseline (i.e., pre-test scores) increase the statistical precision of the impact estimates? To address this question, we conduct a formal statistical test of whether a second pre-test in the same domain increases the precision of the impact estimates.

Question 4: Can using both types of student achievement measures at follow-up (i.e., post-test scores) increase the statistical precision of the impact estimates? To address this question, we conduct a formal statistical test of whether models that specify the average score between the two post-tests as the outcome variable yield more precise impact estimates than models that specify either post-test individually as the outcome variable.

The four research questions, along with the hypotheses we test, are listed in Table 1.

2. Data

For this study, we needed data from one or more evaluations with four variables: (1) post-test scores from a study-administered achievement test, (2) post-test scores from state tests in the same domain as the study-administered post-test, (3) pre-test scores from the same study-administered test used to measure outcomes, and (4) pre-test scores from the same state tests used to measure outcomes. To obtain the necessary data, we selected three previously completed random assignment studies in education. See Table 2 for a summary of these studies.

Each of these three studies randomized classrooms to treatment or control conditions. In addition, each study drew its sample from a single district. However, from this point forward, we will refer to each study by the state in which the district is located—Arizona, California, or Missouri.[8] All three of these studies were conducted by Empirical Education, Inc.[9] For descriptions of the three studies and their data, see Appendix A.[10]

Table 1: Research Questions and Hypotheses to be Tested

Research Question	Hypothesis
1. Will impact evaluations in education yield different impact estimates and statistical precision if they use state tests to measure student achievement at both baseline and follow-up instead of administering standardized tests at both points in time as part of the evaluation?	1a. Relying on study tests yields different impacts than relying on state tests.
	1b. Relying on study tests yields different standard errors than relying on state tests.
2. Does measuring student achievement using one type of test at baseline and another type of test at follow-up reduce the statistical precision of the impact estimates?	2. Mismatched pre-tests yield larger standard errors than matched pre-tests.
3. Does controlling for both types of student achievement measures at baseline (i.e., pre-test scores) increase the statistical precision of the impact estimates?	3. A second pre-test in the same domain reduces the standard error of the impact estimates.
4. Can using both types of student achievement measures at follow-up (i.e., post-test scores) increase the statistical precision of the impact estimates?	4. Models that specify the average score between the two post-tests as the outcome variable yield smaller standard errors than models that specify either post-test individually as the outcome variable.

[8] The California study included a sample that spanned several states. However, in this study, state test score data were collected only in California. In Missouri, state pre-test scores were unavailable, so the study used scores from a district-required test instead.

[9] The results from the California study were published in Miller et al. (2007). The results from the other two studies are proprietary and were not publically released.

[10] The data collected for these studies are owned by the school systems involved and are not generally available for use by researchers. However, Empirical Education has an agreement with each school district to use these data for research purposes.

Table 2: Summary of the Data from the Three Evaluations Selected for This Study

State	Study Test	State Test	Unit of Randomization	Grade Levels	Number of Classrooms	Number of Students
AZ	NWEA-MAP	AIMS	Classes	3-5	15	98
CA	NWEA-ALT	CST	Classes	3-5	20	564
MO	NWEA-MAP	Missouri Assessment Program	Classes	7-8	28	567

Note: For the full names of the tests, see the accompanying text. For more details on these three studies, see Appendix A.

An important limitation of these studies is that they are based on relatively small samples. In fact, our analyses suggest that the impact estimates in all three studies were statistically insignificant. However, this does not mean that the impacts were zero. We address the limitation of small sample sizes in each of the three studies by designing our confirmatory analysis to benefit from the combined sample from the three studies, as described later in this section. The significant differences reported in Section D suggest that the data and methods used for this study had adequate power for the analyses that we conducted. This means that while the sample size for each of the three studies was small, the combined sample was adequate for our purposes.

The three studies are, in effect, a convenience sample of studies that met the requirements for our methodological study. Each study collected scores from the state test used to measure student achievement in reading: Arizona's Instrument to Measure Standards (AIMS), the California Standards Test (CST), and the Missouri Assessment Program.[11]

With one exception, the studies collected both pre-test scores and post-test scores from the state tests. In Missouri, the study did not obtain state pre-test scores and instead collected pre-test scores from a district-required reading test. This limitation led us to exclude Missouri from some of the analyses—in particular, the analysis to address Question 2—as we describe and justify later in the report.

In addition, each of the three studies collected an additional baseline and follow-up measure of reading achievement by administering a reading achievement test offered by the Northwest Evaluation Association (NWEA). The studies in Arizona and Missouri administered NWEA's Measures of Academic Progress (MAP) reading test, a computer adaptive test that is often used for formative assessment (Northwest Evaluation Association 2003). The study in California administered NWEA's Achievement Level Test Series (ALT), which is a paper and pencil adaptive test (NWEA 2003). For the sake of simplicity, we will refer to both tests as the MAP test in the remainder of this report.[12]

[11] Each of the three studies collected student-level scale scores (not just their proficiency levels as defined by state accountability standards).

[12] For more information about the MAP, see the NWEA website: www.nwea.org/products-services/computer-based-adaptive-assessments/map. For more information about the ALT, again see the NWEA website: www.nwea.org/support/article/711. NWEA claims the MAP and the ALT can be used interchangeably (Kingsbury 2001).

To describe the test scores data used in our analysis, we conducted a descriptive analysis. For the estimated correlations between MAP scores and scores on the state and district assessments, see Table 3. These correlations show the strength of the linear relationship between the scores on the two tests. For scatterplots of the test score data, see Appendix B. For cross-tabulations or contingency tables between quartiles in the distribution of MAP reading scores and quartiles in the distribution of state or district reading scores, see Appendix C.

One important question about the MAP is whether it was highly aligned, or even "overaligned," with the interventions being tested. If so, the results from this study might not be generalizable to evaluations that choose more general achievement measures, like many large-scale evaluations conducted by the Institute of Education Sciences (IES). In fact, the MAP was selected for these three studies to a large extent for ease of implementation, and not because the researchers expected the MAP to be especially closely aligned with the interventions being studied.[13] In addition, none of the interventions tested in these three studies were developed by the NWEA, which developed the MAP. Finally, if the study test were overaligned with the interventions, we would expect to see larger impacts on MAP post-test scores than on state post-test scores. However, as we show later, the difference in impacts was statistical insignificant. Therefore, there is no reason to believe that the MAP is overaligned with the interventions tested in these three studies.

Table 3: Correlations between MAP Reading Test Scores and State Reading Test Scores

	Pre-test Scores			Post-test Scores		
State	Control Group	Treatment Group	Pooled	Control Group	Treatment Group	Pooled
Arizona	.60 (n=41)	.57 (n=32)	.57 (n=73)	.53 (n=41)	.69 (n=32)	.60 (n=73)
California	.70 (n=127)	.76 (n=151)	.76 (n=278)	.84 (n=127)	.82 (n=151)	.83 (n=278)
Missouri	.52 (n=186)	.55 (n=175)	.55 (n=361)	.73 (n=186)	.69 (n=175)	.71 (n=361)

Notes: In Missouri, the study relied on pre-test scores from the district test instead of the state test. The *p*-value for each of these correlations is less than .001.

Another important question about the MAP is whether it is more highly aligned with state tests than are the tests selected for most large-scale impact evaluations. If this were true, our study results would be biased toward finding no differences in impact estimates and standard errors between study and state tests. One possible concern arises from the fact that the MAP is often used as a formative assessment to predict student performance on state assessments. Therefore, we might worry that the MAP was designed to be more closely aligned with state tests than other study-administered tests that are not used as formative assessments.

[13] This claim is based on correspondence with researchers at Empirical Education, which conducted all three studies.

To assess this possibility, we assembled evidence on the correlations between scores on different tests in different states. If the MAP were more closely aligned with state tests than other study tests, we would expect the correlation between MAP and state test scores to be higher than the correlation between state test scores and scores from other study-administered tests. The correlations we assembled from the literature are reported in Appendix J. In summary, we find no evidence that the alignment with state assessments is substantially greater for the MAP than for other study-administered tests.

Another possible concern arises from the fact that, in Arizona and Missouri, the "generic" form of the MAP test was customized with the goal of making it more closely aligned to state standards. If this effort had much effect on the MAP's alignment with state standards in those states, then we would expect the correlation with state test scores to be higher in the Arizona and Missouri studies than in the California study. However, the correlations presented in Table 3 provide no evidence that the MAP's alignment with state assessments was greater in Arizona and Missouri than in California, where the generic form of the MAP was used.

Finally, to understand the data used in this study, it is important to note that all three evaluations tested interventions that were designed to improve reading achievement. However, the three studies used in this analysis tested the effectiveness of three *different* interventions. Two of the three interventions were clearly interventions focused on reading instruction. While the third intervention was focused on science, it had a reading component. The interventions tested in the three evaluations are summarized below (see Appendix A for more details):

- **Arizona.** The treatment was a reading intervention system that provides explicit, systematic instruction with ongoing progress monitoring. This intervention was designed for struggling readers in elementary schools.
- **California.** The treatment was Pearson Education's Scott Foresman Science, a year-long science curriculum for daily instruction that is based on inquiry-rich content with a sequence of structured and supportive inquiry activities. A key feature of the curriculum is the Leveled Reader, which helps teachers differentiate instruction by reading level. Although the main purpose of the intervention is to improve science skills, the program provides reading supports to make the science content accessible.
- **Missouri.** The treatment was a supplemental reading program that provides explicit, systematic instruction with ongoing progress monitoring. This program was developed for struggling adolescent readers.

3. Analysis Methods

This section describes the analysis methods used to test the confirmatory hypotheses specified for the study. In summary, we:

- Estimated impacts and standard errors using seven different models for each of the three studies;
- Computed differences in impacts and standard errors between models for each study;

- Estimated standard errors of these differences using bootstrapping;
- Conducted formal statistical tests in each study for differences in impact estimates and standard errors between models (exploratory analysis);
- Computed pooled estimates of the differences in impacts and standard errors by averaging the state-level differences (using inverse variance weights);
- Tested the five hypotheses by conducting formal statistical tests for pooled differences in impact estimates and standard errors between models.

The seven models are summarized in Table 4. These models are used to address the four research questions, as summarized in Table 5. For example to address the first research question, we estimated Model A and Model B, and we compared the impact estimates and standard errors to test Hypothesis 1a and 1b, respectively.

Table 4: Summary of the Analysis Models

	Post-test		Pre-test	
Model	Study Test	State Test	Study Test	State Test
A	✓		✓	
B		✓		✓
C	✓			✓
D		✓	✓	
E	✓		✓	✓
F		✓	✓	✓
G	Simple average of two *z*-scores		✓	✓

Table 5: Using Models A-G to Address the Four Research Questions

Research Question	Hypothesis Tests	Comparison between Models	Summary of this Comparison
1	1a and 1b	Model A vs. Model B	Comparison of a model based on study tests to a model based on state tests
2	2	Model A vs. Model C	Comparison of models based on matched pre-tests to models based on mismatched pre-tests
		Model B vs. Model D	
3	3	Model A vs. Model E	Comparison of models that include one pre-test covariate to models that include two pre-test covariates
		Model B vs. Model F	
4	4	Model E vs. Model G	Comparison of models in which the dependent variable is the average of the two post-test scores to models in which the dependent variable is based on a single post-test score
		Model F vs. Model G	

It is important to note that Missouri was excluded from one of these comparisons. Because the data in Missouri included pre-test scores from the district test instead of the state test, we excluded Missouri from comparisons between Models B and D to address Question 2 because both models included a mismatched pre-test in Missouri (i.e., Model B was based on a state post-test and a mismatched district pre-test). Therefore, in Missouri, the comparison between Models B and D does not offer a test of whether mismatched pre-tests yield less precise impact estimates than matched pre-tests.

However, we included Missouri in all of the other comparisons because we saw no compelling reason to exclude it. For example, we included Missouri in the comparison between Models A and E because the Missouri data, with pre-test scores from both the MAP and district tests, allowed us to test whether two pre-test covariates in the same domain yielded more precise impact estimates than a single pre-test covariate.

Estimate impacts and standard errors. Separately for each of the three data sets, we estimated seven different models of the impacts of the treatment on student reading achievement. Each of these models regresses a measure of reading achievement on a treatment indicator and one or more pre-test measures of reading achievement.

More details on how we specified and estimated the seven models are provided below. Our goal was to specify models to be as similar as possible to the models estimated in impact evaluations for IES.

- **Dependent variables.** The dependent variable is a post-intervention measure of reading achievement. To compute impacts in effect size units, we transformed the scale scores into *z*-scores. More precisely, we re-scaled the post-test scores such that the control group had a mean of zero and standard deviation of one.[14]
- **Independent variables.** The independent variables include one or more pre-intervention measures of achievement (i.e., pre-test scores), the blocking factor (pairs of classrooms within a school), and student demographic characteristics (gender, race and ethnicity, and eligibility for free or reduced-price lunches).[15] Pre-test scores were re-scaled in the same manner as the post-test scores. These independent variables were included in the regression models to improve statistical power by reducing the unexplained variation in the dependent variable.
- **Model specification.** Because the outcome variables are continuous, all seven analysis models were specified as linear models. Furthermore, in all three experiments, students are nested within classrooms, and classrooms were randomly assigned within matched pairs of classrooms. Therefore, we estimated two-level models that account for the clustering of students within classrooms, and we included dummy variables for each pair of classrooms.[16]

---

[14] In principal, the magnitude of the impact estimates would be more easily interpretable if we normalized the test scores using the mean and standard deviation from some population of policy interest—perhaps the scores for all students in the same grade level and state (e.g., specifying a norming population of all students in grades 3-5 in the state of Arizona for the Arizona study). However, state-level means and standard deviations are not available for the MAP test. To ensure that we could make valid comparisons between state and study-administered tests, we selected a common sample so we could compute the mean and standard deviation for both tests: the control group in each respective study. For example, in Arizona, we normalized students' MAP post-test scores using the mean and standard deviation of MAP post-test scores in the study's control group (i.e., subtracting the mean and dividing by the standard deviation), and we normalized students' state post-test scores using the mean and standard deviation of state post-test scores in the study's control group.

[15] Data on eligibility for free or reduced-price lunches were not available in California.

[16] This model allowed schools to vary in their average achievement levels, but it specified a constant treatment effect, as in many educational evaluations.

The estimated variance at the student and classroom levels is reported for each model in Appendix K.

- **Missing covariates.** To address missing covariates, we used the dummy variable method. In short, the dummy variable method involves three steps: (1) create a dummy variable that equals one if the value of the variable is missing and zero otherwise, (2) add the dummy variable to the impact model as a covariate, and (3) replace the missing value from the original variable with any constant, such as zero or the mean for non-missing cases (see Puma et al. 2009 for more details).

The analysis sample for the confirmatory analysis was restricted to the students for which we had non-missing values of all four test scores: (1) MAP post-test, (2) state post-test, (3) MAP pre-test and (4) state or district pre-test. We refer to this sample as "the common sample," to contrast it "the full sample" which includes all students in the sample.[17] Restricting the confirmatory analysis to the common sample was a difficult decision because missing data are common in educational evaluations, and analyses of whether the choice of tests affects the impact estimates or the precision of the estimates would ideally account for the influence of missing data. However, the missing data rate and pattern are a function of the study-specific data collection strategy that was implemented by the study team. For example, the missing data rate for study-administered tests will depend on whether tests were administered in the students' regular classrooms or off-site, whether active consent was required, and whether the study offered incentives and make-up testing dates to boost response rates. And given student mobility across schools, the missing data rate for state-administered tests may depend on whether the study collected these scores from participating schools, districts, or states. Therefore, to reduce the likelihood that our study findings are driven by the particular strategies that the three studies used in collecting the data, we decided to focus the confirmatory analysis on the common sample. As an exploratory analysis, we estimated the seven models on the full sample and present the results in Appendix E.

The remainder of this section describes the steps we took to use the impact estimates and standard errors in addressing the four research questions posed for this study.

Compute differences in impact estimates and standard errors. To test the specified hypotheses, we computed the difference in impacts and standard errors between models specified in Table 5 for each of the three studies. More specifically, we computed the difference in impact estimates between Models A and B to test Hypothesis 1a, and we computed the difference in standard error estimates between each pair of models specified in Table 5 to test Hypothesis 1b and Hypotheses 2-4.

Estimate standard errors of these differences. Testing for differences between models was complicated by the fact that the estimates from two different models computed from the same sample will be correlated. To address this challenge, we developed a parametric bootstrapping approach for this study, as described in Appendix F. In summary, this procedure involved estimating the seven models in each study and using them to generate 1,000 bootstrap samples with similar distributional properties as the original data. In each of these 1,000 bootstrap samples, we estimated each pair of

[17] In the analysis based on the full sample, we used the dummy variable method to address missing pre-test scores.

models (e.g., Models A and C) and computed the difference in the impact estimates and standard errors between models. To obtain the standard error of each difference, we computed the standard deviation across the 1,000 bootstrap samples. For more details, see Appendix F.

Conduct formal statistical tests in each study. While our confirmatory analysis was based on a pooled analysis (as described later), we first conducted exploratory hypothesis tests of whether there were non-zero differences in impacts and standard errors between models for each of the three studies. For each hypothesis, and separately in each study, we conducted a *t*-test. The *t*-statistic was computed as the difference in the impact or standard error estimates between the two models divided by the bootstrap estimate of the standard error of this difference.

Finally, we computed the *p*-value associated with the formal hypothesis of no difference between models. Using this approach, we conducted a formal test for each of the differences in estimates between models. These *p*-values were not adjusted for multiple comparisons (i.e., for the fact that we conducted separate tests for each of the three studies) because the state-level analyses were treated as exploratory.

Compute pooled estimates of the differences. To increase statistical power, we pooled the estimates across studies, *and we based our confirmatory tests on the pooled results.* In pooling across studies, we used standard meta-analytic techniques to "average" the estimates across the three data sets (Cooper, Hedges, & Valentine 2002, Lipsey & Wilson 2001). Pooling both increases the power of the test and reduces the likelihood that study findings are driven by a single idiosyncratic study (e.g., a study with an unusual state test).

To create a pooled estimate of the difference in impact estimates or standard error estimates between models, we construct a weighted average of the three state-level differences. For this analysis, we set the state-level weights proportional to the inverse of the variance of the state-level difference (e.g., the variance of the difference in standard errors between Models A and C in Arizona). If the value of the underlying parameter (e.g., the difference in standard errors across models) is the same for the three studies, inverse variance weights yield the most precise estimates of these parameters.

More precisely, let Δ_s be the difference between two estimates (impact estimates or standard error estimates) for state *s*, and let $\text{var}(\Delta_s)$ be the bootstrap estimate of the variance of the Δ_s. To construct pooled estimates of the difference between the two impact estimates or standard error estimates, we constructed a weighted average of the state-level estimates. Let Wgt_s be the weight constructed for state *s*. Note that Wgt_s is inversely proportional to the variance of the impact estimate for state *s*, as shown below, and that the weights sum to one:

$$Wgt_{AZ} = \left\{ \text{var}(\Delta_{AZ}) \left[\frac{1}{\text{var}(\Delta_{AZ})} + \frac{1}{\text{var}(\Delta_{CA})} + \frac{1}{\text{var}(\Delta_{MO})} \right] \right\}^{-1}$$

$$Wgt_{CA} = \left\{ \text{var}(\Delta_{CA}) \left[\frac{1}{\text{var}(\Delta_{AZ})} + \frac{1}{\text{var}(\Delta_{CA})} + \frac{1}{\text{var}(\Delta_{MO})} \right] \right\}^{-1}$$

$$Wgt_{MO} = \left\{ \text{var}(\Delta_{MO}) \left[\frac{1}{\text{var}(\Delta_{AZ})} + \frac{1}{\text{var}(\Delta_{CA})} + \frac{1}{\text{var}(\Delta_{MO})} \right] \right\}^{-1}$$

Using these weights, we created the average difference between the corresponding estimates from two models, pooling across the three studies:

$$\Delta_{pooled} = Wgt_{AZ}(\Delta_{AZ}) + Wgt_{CA}(\Delta_{CA}) + Wgt_{MO}(\Delta_{MO})$$

Test the five hypotheses by conducting formal statistical tests. The most important hypothesis tests conducted for this study are the five confirmatory tests of the study's five hypotheses. Each of the confirmatory tests involves one or more formal hypothesis tests of whether a pooled difference in impacts or standard errors is different from zero.

For these tests, we needed an estimate of the variance for each pooled difference. We took the expression for Δ_{pooled} given above and derived the following formula for the variance of the pooled difference:[18]

$$\text{var}(\Delta_{pooled}) = \left(\frac{1}{\text{var}(\Delta_{AZ})} + \frac{1}{\text{var}(\Delta_{CA})} + \frac{1}{\text{var}(\Delta_{MO})} \right)^{-1}$$

For each pooled difference between models, we conducted a *t*-test for whether the difference was statistically significant. The *t*-statistic was computed by dividing the pooled difference by an estimate of its standard error:

$$t_{pooled} = \frac{\Delta_{pooled}}{\sqrt{\text{var}(\Delta_{pooled})}}$$

From Table 5, it is clear that three of the hypotheses (2-4) involve multiple (in particular, two) comparisons.[19] For each of these hypotheses, a significant difference in *either* of the two comparisons was treated as evidence supporting the hypothesis. Therefore, a multiple comparisons correction is appropriate.

---

[18] This formula was derived using the standard formula for the variance of a linear combination of random variables under the assumption that the three state-level samples were independent of each other.

[19] The multiple comparisons problem has received substantial attention in evaluation circles in education, as evidenced by the convening of a working group at the Institute of Education Sciences to consider the challenges associated with multiple comparisons, and the completion of a report based on the results of this effort (Schochet 2008b). The most important problem with multiple comparisons is the risk that researchers will overstate their level of confidence in estimates that, by themselves, would be classified as statistically significant.

To adjust the *p*-values for multiple comparisons in testing Hypotheses 2-4, we applied a Bonferroni correction.[20] For example, to address Hypothesis 2 (mismatched pre-tests yield less precise impact estimates than matched pre-tests), we multiplied the *p*-values for each of the two comparisons by a factor of two to create Bonferroni-adjusted *p*-values.[21] While the Bonferroni correction is known to be conservative, the evidence suggests that the difference in power between the Bonferroni correction and other more sophisticated methods is small when the number of comparisons is small (see Schochet 2008b, Table B.4).

For the confirmatory analysis, we characterize the strength of the evidence based on the pooled results. We set an alpha level of .05 as our standard for evidence because it is conventional and consistent with the standards for hypothesis testing established by the National Center for Education Statistics.[22] We set a more liberal alpha level of .10 as our standard for "suggestive" evidence because many evaluation reports identify estimates that are significant at the 10 percent level. For example, to address Question 2 on the mismatch hypothesis, if either of the two tests for the pooled differences yields a Bonferroni-adjusted *p*-value of less than .05, then we would conclude that there is evidence in favor of the mismatch hypothesis. However, if neither of the two tests yields a *p*-value of less than .05, but at least one of them yields a *p*-value of less than .10, then we would conclude that there is suggestive evidence in favor of the mismatch hypothesis.

---

[20] In implementing this approach, we combine the *p*-values reported in Appendix E with the number of comparisons required to address each of the four questions.

[21] This is algebraically equivalent to the standard approach, which involves dividing the alpha level of the test (e.g., .05 or 5 percent) by a factor of two.

[22] See these standards at http://nces.ed.gov/statprog/2002/stdtoc.asp, as downloaded on February 20, 2010.

D. Empirical Results

This section presents the results of the empirical analysis that we conducted to address the study's four research questions.

1. Summary of Results

Our analyses provide some evidence that studies with mismatched pre-tests need larger samples than studies with matched pre-tests, that controlling for multiple measures in the same domain can reduce a study's sample size requirements (by a small amount), and that averaging two post-test scores in the same domain can also reduce a study's sample size requirements (again by a small amount). These results are summarized below.

Question 1: Will impact evaluations in education yield different impact estimates and statistical precision of the impact estimates if they use state tests to measure student achievement at both baseline and follow-up instead of administering standardized tests at both points in time as part of the evaluation? On average, across the three studies,[23] we found no evidence of differences in impacts, but suggestive evidence of differences in their standard errors—in particular, that evaluations based on state test scores may produce larger standard errors than evaluations based on study-administered tests.

Question 2: Does measuring student achievement using one type of test at baseline and the other type of test at follow-up reduce the statistical precision of the impact estimates? On average, across the three studies, we found evidence that the answer to this question is yes. Furthermore, on average in these three studies, our estimates suggest that the required sample size would be 45 percent larger for a mismatched state pre-test than for a matched study pre-test, if the post-test is measured using the study test, and 100 percent larger for a mismatched study pre-test than for a matched state pre-test, if the post-test is measured using the state test.

Question 3: Does controlling for both types of student achievement measures at baseline (i.e., pre-test scores) increase the statistical precision of the impact estimates? On average, across the three studies, we found evidence that the answer to this question is yes. In addition, on average in these three studies, our estimates suggest that the required sample size would be 10 percent smaller with both pre-tests than with only a matched study pre-test, if the post-test is measured using the study test, and 25 percent smaller with both pre-tests than with only a matched state pre-test, if the post-test is measured using the state test. However, the results seem to be attributable to a potentially anomalous result in one of the three studies. Therefore, we recommend caution in interpreting the results from this test (see the next subsection for more details).

---

[23] Throughout the discussion of the study findings, we refer to the confirmatory test results based on the pooled estimates using language like "on average, across the three studies" or "on average in these three studies."

Question 4: Can using both types of student achievement measures at follow-up (i.e., post-test scores) increase the statistical precision of the impact estimates? On average, across the three studies, we found evidence that the answer to this question is yes. When we created a composite measure of achievement that averages the student's score from the state test with his or her score from the study-administered test, the estimated impact on the composite measure was more precisely estimated than the estimated impact on either of the two post-test scores individually. In addition, on average in these three studies, our estimates suggest if we specify the dependent variable as the average of the two post-test scores, the required sample size would be 15 percent smaller than if we specified the study post-test as the dependent variable, and 25 percent smaller than if we specified the state post-test as the dependent variable.

It is important to conclude this summary of results with a note of caution. The analysis was limited by the data from three relatively small random assignment studies. These data met our requirements for this study because they offered pre-test scores and post-test scores from both state tests and study-administered tests. However, the small size of the samples limited the power of the analysis and the precision of the estimates. While the point estimates suggest that the sample size implications of choosing different tests are quite large (e.g., estimated differences in sample size requirements as large as 100 percent), these estimates include sampling error, so the true sample size implications may be substantially smaller (or larger). *Therefore, we should base any conclusions about the sample size implications of choosing different tests on the growing body of research in this area, and not on the results from any single study.*

In addition, it is not clear whether the results from these three small studies generalize to the larger evaluations that IES typically funds. Additional studies are necessary to build the body of evidence researchers need to make informed decisions when designing evaluations.

At the same time, this report provides the first empirical evidence that directly addresses these four questions. Despite limited statistical power, we found statistically significant evidence that the choice between state and study-administered tests, and the specification of the models that use them in estimating the impacts of educational interventions, can have important implications for the sample sizes required by educational impact evaluations.

2. Confirmatory Test Results

The remainder of this section presents the results from the confirmatory analysis designed to address each of the four research questions. This analysis involves estimating the models described in Table 4 and conducting the analysis described in Section C. As indicated in Section C, the confirmatory analysis was conducted using the common sample, which excludes students with missing values for any of the four test scores—MAP post-test, MAP pre-test, state post-test, or state or district pre-test. Exploratory results for the full sample are presented in Appendix E.

Question 1: Will impact evaluations in education yield different impact estimates and statistical precision of the impact estimates if they use state tests to measure student achievement at both baseline and follow-up instead of administering standardized tests at both points in time as part of the evaluation? To address this question, we compared the impact estimates and standard errors from Model A (MAP post-test and MAP pre-test) to the impact estimates and standard errors from

Model B (state post-test and state pre-test). For this analysis, we had no *a priori* expectations about what the results would be.

To address Question 1, we tested two separate hypotheses:

- **Hypothesis 1a: Relying on study tests (Model A) yields different *impacts* than relying on state tests (Model B).**
- **Hypothesis 1b: Relying on study tests (Model A) yields different *standard errors* than relying on state tests (Model B).**

To test these two hypotheses, we estimated Models A and B separately for each state to produce impact estimates and estimates of their standard errors, and we computed pooled estimates of the impact and standard error for each model across studies using inverse variance weights. Then we tested Hypothesis 1a by assessing whether there was a significant difference between the pooled impact estimate for Model A and the pooled impact estimate for Model B, and we tested Hypothesis 1b by assessing whether there was a significant difference between the pooled standard error estimate for Model A and the pooled standard error estimate for Model B. For more details on the analysis, see Section C.

The confirmatory test results are presented in Table 6. The left panel of Table 6 shows the results for Hypothesis 1a (impacts); the right panel of Table 6 shows the results for Hypothesis 1b (standard errors).

Table 6 does not provide evidence supporting Hypothesis 1a. The pooled difference in impacts between Models A and B was statistically insignificant. An insignificant difference does not mean that the actual difference was zero, and it does not rule out the possibility that the difference was simply too small to detect.[24] However, the results in this report provide no evidence that evaluations which rely on state tests will yield systematically different impact estimates than evaluations that rely on study-administered tests.

[24] For the Minimum Detectable Differences for all of the analyses reported in this chapter, see Appendix H, Exhibit H.3 and Exhibit H.5.

Table 6: Comparing Effect Sizes and Standard Errors between State and Study Tests, Estimates to Address Question 1

Analysis	Effect Sizes			Standard Errors		
	Model A (MAP post, MAP pre)	Model B (State post, state pre)	*p*-value† (Model A – Model B)	Model A (MAP post, MAP pre)	Model B (State post, state pre)	*p*-value† (Model A – Model B)
Pooled	**-0.050**	**-0.071**	**.340**	**0.101**	**0.126**	**.060***
By state						
Arizona	-0.126	0.001	.537	0.202	0.200	.956
California	-0.053	-0.154	.103	0.077	0.106	.054*
Missouri	-0.028	0.004	.719	0.103	0.116	.406

† For the pooled analyses, which are the basis for our confirmatory tests, we report *p*-values that are not adjusted for multiple comparisons because we used a single statistical test for each of the two hypotheses (Hypothesis 1a for impacts or effect sizes and Hypothesis 1b for standard errors). For the state-level analyses, we report unadjusted *p*-values because the analyses should be classified as exploratory.

Notes: * indicates $p < .10$, ** indicates $p < .05$, and *** indicates $p < .01$. For a description of the models, see Table 4. The estimates in this table were computed from the common sample, after excluding students with missing values in either of the pre-test scores or post-test scores. For pooled estimates of the average effect size or standard error across studies for each model, we computed a weighted average of the state-level estimates, where the weight for each state was proportional to the inverse of the variance of the estimate for that state. For the pooled hypothesis test of no difference between models, we computed a weighted average of the state-level differences in the estimates between models, where the weight for each state was proportional to the inverse of the variance of the estimated difference for that state (see Section C for more details).

Table 6 provides suggestive evidence that supports Hypothesis 1b. More specifically, we found that Model B (state post-test and state pre-test) produced larger standard errors than Model A (MAP post-test and MAP pre-test). The pooled difference in standard errors between Models A and B was statistically significant at the 10 percent level.[25] Therefore, based on the decision rule described in

Section C, we classify this evidence as "suggestive." We conclude that on average in these three studies, the analysis provides suggestive evidence that using state tests for both pre-test and post-test measures yields a less precise impact estimate than using the study-administered test for both measures.

Question 2: Does measuring student achievement using one type of test at baseline and the other type of test at follow-up reduce the statistical precision of the impact estimates? To address this question, we compared the standard errors from models that included a mismatched pre-test variable to the standard errors from otherwise equivalent models that included a matched pre-test. We would expect that relative to a matched pre-test covariate, using a mismatched pre-test covariate would reduce the R-square of the impact regression, increase the estimated standard error of the impact estimate, and increase the sample size requirements of the evaluation, as explained more formally in

[25] The *p*-value of the difference is equal to .060, as shown in Exhibit 7.

Section B. However, whether this effect is large or small, and whether it is large enough to be detected in our analysis, is an empirical question.

To address Question 2, we tested the hypothesis listed below:

- **Hypothesis 2: Mismatched pre-tests yield larger standard errors than matched pre-tests.**

To test Hypothesis 2, we estimated impacts and the standard errors of the impact estimates using Models A, B, C, and D separately for each state, and we pooled the standard errors for each model across studies using inverse variance weights. Then we tested the hypothesis with two statistical tests—a test for whether the standard errors differ between Model A (MAP post-test and *matched* MAP pre-test) and Model C (MAP post-test and *mismatched* state or district pre-test), and a test for whether the standard errors differ between Model B (state post-test and *matched* state pre-test) and Model D (state post-test and *mismatched* MAP pre-test)—after correcting for multiple comparisons. The multiple comparisons correction allows us to conclude that we have found evidence supporting the hypothesis if either difference (i.e., the difference between Models A and C or the difference between Models B and D) is statistically significant. For more details on the analysis, see Section C.

The confirmatory test results for Question 2 are presented in Table 7. The left panel of Table 7 shows the results for the MAP post-test, comparing Models A and C; the right panel of Table 7 shows the results for the state post-test, comparing Models B and D.

Table 7 provides evidence that supports Hypothesis 2. In the pooled analysis, the estimated standard error was larger for Model C (MAP post-test and *mismatched* state or district pre-test) than for Model A (MAP post-test and *matched* MAP pre-test), and the difference was statistically significant at the 5 percent level, even after accounting for multiple comparisons.[26] We conclude that on average in these three studies, the analysis provides evidence that mismatched pre-tests yield less precise impact estimates than matched pre-tests.

Question 3: Does controlling for both types of student achievement measures at baseline (i.e., pre-test scores) increase the statistical precision of the impact estimates? To see if additional pre-test covariates reduce the standard error of the impact estimate, we compared the standard errors from models with one pre-test covariate to the standard errors from models with two pre-test covariates in the same domain. We would expect the additional pre-test covariate to increase the R-square of the impact regression, reduce the standard error of the estimated impact, and reduce the sample size requirements of the evaluation, as explained more formally in Section B. However, whether this effect is large or small, and whether it is large enough to be detected in our analysis, is an empirical question.

[26] The Bonferroni-adjusted *p*-value of this difference equals .0102, as shown in Exhibit 7 (after rounding).

Table 7: Estimating the Increase in Standard Errors from a Mismatched Pre-test, Question 2

Analysis	Standard Errors			Standard Errors		
	Model A (MAP post, MAP pre)	Model C (MAP post, state pre)	*p*-value[†] (Model A – Model C)	Model B (State post, state pre)	Model D (State post, MAP pre)	*p*-value[†] (Model B – Model D)
Pooled	**0.101**	**0.152**	**.010****	**0.139**	**0.170**	**.515**
By state						
Arizona	0.202	0.230	.282	0.200	0.230	.210
California	0.077	0.153	.059*	0.106	0.113	.830
Missouri	0.103	0.129	.033**	NA[††]	NA[††]	NA[††]

[†] For the pooled analysis, which is the basis for our confirmatory test, we report *p*-values that are adjusted for multiple comparisons because we used two statistical comparisons to test a single hypothesis (Hypothesis 2): the adjusted *p*-value equals two times the unadjusted *p*-value. As a result, some adjusted *p*-values may be greater than one. For the state-level analyses, we report unadjusted *p*-values because the analyses should be classified as exploratory.

[††] Missouri was excluded from the analysis for comparisons of Models B and D because district pre-test scores were collected instead of state pre-test scores. This means that Model B is based on a mismatched pre-test (state post-test and district pre-test), and the data from Missouri cannot be used to test Hypothesis 2 when post-test scores come from state assessments.

Notes: * indicates $p < .10$, ** indicates $p < .05$, and *** indicates $p < .01$. For a description of the models, see Table 4. The estimates in this table were computed from the common sample, after excluding students with missing values in either of the pre-test scores or post-test scores. For pooled estimates of the average standard error across studies for each model, we computed a weighted average of the state-level estimates, where the weight for each state was proportional to the inverse of the variance of the estimate for that state. For the pooled hypothesis test of no difference in the standard error estimate between models, we computed a weighted average of the state-level differences in the estimates between models, where the weight for each state was proportional to the inverse of the variance of the estimated difference for that state (see Section C for more details).

To address Question 3, we tested the hypothesis listed below:

- **Hypothesis 3: A second pre-test in the same domain reduces the standard error of the impact estimates.**

To test Hypothesis 3, we estimated impacts and the standard errors of the impact estimates using Models A, B, E, and F separately for each state, pooled the standard errors for each model across studies using inverse variance weights, and tested the hypothesis with two statistical tests—a test for whether the standard errors differ between Model A (MAP post-test and MAP pre-test) and Model E (MAP post-test and *both* pre-tests), and a test for whether the standard errors differ between Model B (state post-test and state pre-test) and Model F (state post-test and *both* pre-tests)—after correcting for multiple comparisons. The multiple comparisons correction allows us to conclude that we have found evidence supporting the hypothesis if either difference (i.e., the difference between Models A and E or the difference between Models B and F) is statistically significant. For more details on the analysis, see Section C.

The confirmatory test results for Question 3 are presented in Table 8. The left panel of Table 8 shows the results for comparisons between Models A and E (for the MAP post-test); the right panel of Table 8 shows the results for comparisons between Models B and D (for the state post-test).

Table 8 provides evidence that supports Hypothesis 3. In the pooled analysis, the estimated standard error was smaller for Model E (MAP post-test and *both* pre-tests) than for Model A (MAP post-test and MAP pre-test), and the difference was statistically significant at the 1 percent level, even after accounting for multiple comparisons.[27] Therefore, we conclude that Table 8 provides evidence that in these three studies, a second pre-test in the same domain increases the precision of the impact estimates.

However, this result seems to be driven heavily by a potentially anomalous result in Missouri, where the difference in standard errors was small but highly significant. Our investigations suggest that this result may be attributable to the bootstrap estimate of the correlation between the two standard error estimates, which is very close to one.[28] Therefore, caution is warranted in interpreting the results of this analysis.

Question 4: Can using both types of student achievement measures at follow-up (i.e., post-test scores) increase the statistical precision of the impact estimates? In particular, we tested whether averaging the two post-test scores can produce more precise impact estimates than either post-test alone.[29] As explained earlier, if the two tests provide "noisy" measures of the same underlying construct, then averaging the two tests could reduce the measurement error in the dependent variable, yield impact estimates with smaller standard errors, and reduce the sample size requirements of the evaluation, as shown more formally in Section B. However, whether the simple average of the two tests produces impact estimates with smaller standard errors than either test individually is an empirical question.[30]

---

27 The Bonferroni-adjusted *p*-value for this difference is less than .001, as shown in Exhibit 9.

28 As a result, the standard error of the difference was close to zero. Because this standard error enters the *t*-test in the denominator of the *t*-statistic, a near zero standard error estimate can produce a very large *t*-statistic and a very small *p*-value, as shown in Appendix H, Exhibit H.5.

29 More precisely, we took the simple mean between the student's *z*-score on each of the two tests, where scores from each test of have been rescaled to have a mean of zero and a standard deviation of one (among control students).

30 More sophisticated weighting approaches are possible and could be tested in future work (see Schochet 2008b).

Table 8: Estimating the Decrease in Standard Errors from Using Both Pre-tests, Question 3

Analysis	Standard Errors			Standard Errors		
	Model A (MAP post, MAP pre)	Model E (MAP post, both pre)	*p*-value† (Model A – Model E)	Model B (State post, state pre)	Model F (State post, both pre)	*p*-value† (Model B – Model F)
Pooled	**0.101**	**0.096**	**<.001*** **	**0.126**	**0.105**	**.614**
By state						
Arizona	0.202	0.206	.228	0.200	0.214	.251
California	0.077	0.073	.135	0.106	0.082	.362
Missouri	0.103	0.098	<.001***	0.116	0.101	.080*

† For the pooled analysis, which is the basis for our confirmatory test, we report *p*-values that are adjusted for multiple comparisons because we used two statistical comparisons to test a single hypothesis (Hypothesis 3): the adjusted *p*-value equals two times the unadjusted *p*-value. As a result, some adjusted *p*-values may be greater than one. For the state-level analyses, we report unadjusted *p*-values because the analyses should be classified as exploratory.

Notes: * indicates $p < .10$, ** indicates $p < .05$, and *** indicates $p < .01$. For a description of the models, see Table 4. For more details on the computations presented in this table, see the notes below Table 7.

To address Question 4, we tested the hypothesis listed below:

- **Hypothesis 4: Models that specify the average score between the two post-tests—one from state tests and the other from a study-administered test—as the outcome variable yield smaller standard errors than models that specify either post-test individually as the outcome variable.**

To test Hypothesis 4, we estimated impacts and the standard errors of the impact estimates using Models E, F, and G separately for each state, and we pooled the standard errors for each model across studies using inverse variance weights. Then we tested the hypothesis with two statistical tests—a test for whether the standard errors differ between Model G (average or composite post-test and both pre-tests) and Model E (MAP post-test and both pre-tests), and a test for whether the standard errors differ between Model G and Model F (state post-test and both pre-tests)—after correcting for multiple comparisons. The multiple comparisons correction allows us to conclude that we have found evidence supporting the hypothesis if either difference (i.e., the difference between Models G and E or the difference between Models G and F) is statistically significant. For more details on the analysis, see Section C.

Table 9 provides evidence that supports Hypothesis 4. In the pooled analysis, the estimated standard error was smaller for Model G (average or composite post-test and both pre-tests) than for Model F (state post-test and both pre-tests), and the difference was statistically significant at the 5

Table 9: Estimating the Decrease in Standard Errors from Averaging the Two Post-tests, Question 4

Analysis	Standard Errors			Standard Errors		
	Model G (mean post, both pre)	Model E (MAP post, both pre)	*p*-value† (Model G – Model E)	Model G (mean post, both pre)	Model F (State post, both pre)	*p*-value† (Model G – Model F)
Pooled	**0.085**	**0.096**	**.243**	**0.085**	**0.105**	**.035****
By state						
Arizona	0.182	0.206	.360	0.182	0.214	.187
California	0.064	0.073	.436	0.064	0.082	.132
Missouri	0.086	0.098	.265	0.086	0.101	.146

† For the pooled analysis, which is the basis for our confirmatory test, we report *p*-values that are adjusted for multiple comparisons because we used two statistical comparisons to test a single hypothesis (Hypothesis 4): the adjusted *p*-value equals two times the unadjusted *p*-value. As a result, some adjusted *p*-values may be greater than one. For the state-level analyses, we report unadjusted *p*-values because the analyses should be classified as exploratory.

Notes: * indicates $p < .10$, ** indicates $p < .05$, and *** indicates $p < .01$. For a description of the models, see Table 4. For more details on the computations presented in this table, see the notes below Table 7.

percent level, even after accounting for multiple comparisons.[31] Therefore, we conclude that on average in these three studies, Table 9 provides evidence that specifying the outcome variable as the simple average between the *z*-scores of the two tests yields more precise impact estimates than either post-test individually.

3. Differences in Sample Size Requirements

From a practical perspective, it is helpful to translate the differences in standard errors between models into differences in sample size requirements. The analysis presented in this section thus far holds constant other factors that influence the standard errors—most notably, the size of the sample. However, researchers typically choose a sample size target to achieve a particular Minimum Detectable Effect Size (MDES) (e.g., 0.15 standard deviations). If researchers had evidence on the key power parameters (e.g., ICCs and R-squares) separately for state tests and study-administered tests, this would allow them to set sample size targets that account for the type of test chosen for the evaluation.

How much would we expect the sample size targets to be affected by the choice between different tests? This is a critically important question because the cost of an evaluation depends on both the per-student cost of obtaining achievement data, which is lower for state tests than study tests, *and* on

[31] The Bonferroni-corrected *p*-value of this difference equals .034. Therefore, the difference is significant at the .05 or 5 percent level.

the size of the sample. To address this question, we apply standard formulas to the standard error estimates presented earlier in this section. For more details on these calculations, see Appendix I.[32] Using these formulas shown in Appendix I, we estimated the consequences of choosing different tests for a hypothetical evaluation with 20 classrooms. This hypothetical evaluation shares three features with the evaluations in Arizona, California, and Missouri: (1) the number of classrooms is comparable to the sample sizes in these three studies, (2) the research design is the same as in the three studies (i.e., the unit of randomization is the classroom), and (3) the statistical power parameters are comparable by construction (because we used the pooled standard error estimates across the three studies in our sample size calculations for the hypothetical evaluation).[33]
For each pair of models, we:

- Classified one model as the "primary analysis model" and the other model as the "alternative analysis model";
- Set the sample size for the primary analysis model to 20 classrooms;
- Estimated the Minimum Detectable Effect Size (MDES) under the primary analysis model with 20 classrooms; and
- Computed the number of classrooms required to achieve the same MDES with the alternative analysis model.

The resulting estimates are presented in Table 10.

Table 10 suggest that the hypothetical evaluation would need a larger sample size if it relied on state tests to measures student achievement at baseline and follow-up than if it relied on the study-administered test. In particular, on average in these three studies, our best estimate suggests that choosing Model B over Model A would increase the required sample size by 45 percent. Recall that we found suggestive evidence supporting Hypothesis 1b—that relying on state tests (Model B) yields less precise impact estimates than relying on study-administered tests (Model A). For the hypothetical evaluation, our estimates suggest that the difference in precision translates into an increase in the required sample size from 20 classrooms to 29 classrooms.

[32] In designing studies, researchers typically rely on published estimates of the intra-class correlation and the R-square of the regression model both within and between clusters. For estimates from the three studies used in our analysis, see Appendix I.

[33] For these calculations, we relied on the estimates from the common sample for the same reasons that we relied on these estimates for the confirmatory analysis (see Section C for a discussion of this choice). Therefore, it is important to recognize that our estimates do not account for the role of missing data in the determining sample size requirements. If the two tests have very different missing data rates, this could produce differences in the sample size requirements that are not captured by our analysis.

Table 10: Sample Size Implications of Choosing Different Design Options: A Hypothetical RCT with 20 Classrooms

	Primary Design Option			Alternative Design Option			Estimated Change in Sample Required by Alternative Option (95% CI)
Question	Data to Collect	Primary Analysis Model	Number of Classrooms	Data to Collect	Alternative Analysis Model	Number of Classrooms (95% CI)	
1	Study post-test Study pre-test	Model A	20	State post-test State pre-test	Model B	29 (17,46)	45% increase (-15%, +130%)
2	Study post-test Study pre-test	Model A	20	Study post-test State pre-test	Model C	40 (22,64)	100% increase (+10%, +220%)
2	State post-test State pre-test	Model B	20	State post-test Study pre-test	Model D	28 (14,47)	40% increase (-30%, +135%)
3	Study post-test Study pre-test	Model A	20	Study post-test Study pre-test State pre-test	Model E	18 (17,20)	10% decrease (-15%, 0%)
3	State post-test State pre-test	Model B	20	State post-test Study pre-test State pre-test	Model F	15 (13,20)	25% decrease (-35%, 0%)
4	Study post-test Study pre-test State pre-test	Model E	20	Study post-test Study pre-test State post-test State pre-test	Model G	17 (13,24)	15% decrease (-35%, +20%)
4	State post-test Study pre-test State pre-test	Model F	20	Study post-test Study pre-test State post-test State pre-test	Model G	15 (13,20)	25% decrease (-35%, 0%)

Note: For the formulas used to conduct the calculations for the alternative analysis model, see Appendix I. The last two columns include approximate 95% confidence intervals, as described in Appendix I.

Estimates from Table 10 also suggest that the hypothetical evaluation would need a larger sample if it selected a mismatched pre-test than if it selected a matched pre-test. In particular, we found that:

- **Choosing Model C over Model A would increase the required sample size by 100 percent.** Recall that we found evidence supporting Hypothesis 2—that mismatched pre-tests yield less precise impact estimates than matched pre-tests. For the hypothetical evaluation, when post-test scores come from the study-administered test, our estimates suggest that the decrease in precision from choosing Model C translates into an increase in the required sample size from 20 classrooms to 40 classrooms.
- **Choosing Model D over Model B would increase the required sample size by 40 percent.** For the hypothetical evaluation, when post-test scores come from the state test, our estimates suggest that the decrease in precision from choosing Model D translates into an increase in the required sample size from 20 classrooms to 28 classrooms.

Estimates from Table 10 suggest that the hypothetical evaluation would need a somewhat *smaller* sample if both pre-test scores were collected and included as covariates in the model, relative to collecting scores from a single pre-test. In particular, we found that:

- **Choosing Model E over Model A would reduce the required sample size by 10 percent.** Recall that we found evidence supporting Hypothesis 3—that a second pre-test covariate increases the precision of the impact estimates. For the hypothetical evaluation, when post-test scores come from the study-administered test, our estimates suggest that the increase in precision translates into a reduction in the required sample size from 20 classrooms to 18 classrooms.
- **Choosing Model F over Model B would reduce the required sample by 25 percent.** For the hypothetical evaluation, when post-test scores come from the state test, our estimates suggest that the increase in precision translates into a reduction in the required sample size from 20 classrooms to 15 classrooms.

Finally, estimates from Table 10 suggest that the hypothetical evaluation would need a somewhat smaller sample if both post-test scores were collected and averaged together to create a composite post-test score, relative to relying on scores from either post-test individually. In particular, we found that:

- **Choosing Model G over Model E would reduce the required sample size by 15 percent.** Recall that we found evidence supporting Hypothesis 4—that a composite post-test increases the precision of the impact estimates. For the hypothetical evaluation, when post-test scores come from the study-administered test, our estimates suggest that the increase in precision translates into a reduction in the required sample size from 20 classrooms to 17 classrooms.
- **Choosing Model G over Model F would reduce the required sample size by 25 percent.** For the hypothetical evaluation, when post-test scores come from the state test, our estimates suggest that the increase in precision translates into a reduction in the required sample size from 20 classrooms to 15 classrooms.

Finally, we conclude with two cautions. First, the sample size effects of choosing the alternative designs over the base designs, as shown in Table 10, are measured with sampling error. For example,

consider the comparison between Models A and B. While the difference in estimated standard errors was significant at the 10 percent level, it was insignificant at the 5 percent level. This means that while our best estimate suggests that choosing Model B over Model A would increase the required number of classrooms by 45 percent, the 95 percent confidence interval for this effect includes zero, which means we cannot be 95 percent confident that the effect is not zero.

To account for the statistical uncertainty associated with the estimated sample size requirements reported for the alternative models in Table 10, the last two columns of the table report approximate 95-percent confidence intervals. The width of the confidence interval around the estimated sample size requirement is a measure of the uncertainty associated with the estimate. In addition, confidence intervals that include 20 classrooms suggest that we cannot be 95 percent confident about whether the true sample size requirement is greater than or less than the 20 classrooms required under the primary analysis model.

Second, the generalizability of the estimates from Table 10 is limited by the scope of the data and research design used in the three evaluations. The generalizability of study findings is addressed in the next and final section of the report.

E. Generalizability of Study Findings and Call for Additional Research

Most evaluations of educational interventions and programs sponsored by the U.S. Department of Education include one or more measures of student achievement as key outcomes. Because it is much less expensive to collect state test scores than to administer standardized tests, these evaluations are increasingly relying on state tests to provide at least some of the key achievement measures for the study.

In this report, we consider the possible reasons why state tests may yield different impact estimates from study-administered tests. In addition, we consider the possible reasons why the precision of the estimates, and thus the study sample size requirements, may depend on the choice between the two types of tests. Even if these factors do not drive decisions regarding which type of test to use for an evaluation, they may have important consequences for how large the sample needs to be, and even how we interpret the magnitude of the estimated achievement impacts. Sample size implications are important because they have real resource implications for an individual study and for a portfolio of funded projects.

In this study, we found evidence that the choice between state and study-administered tests can "matter" in terms of the study's sample size requirements. While the analysis had limited statistical power to detect differences, some significant differences were detected nonetheless. In particular, for the three studies we used for this analysis, we found some level of support, as defined earlier, for four hypotheses:

- **Hypothesis 1b:** Relying on study tests (Model A) yields different *standard errors* than relying on state tests (Model B). In particular we found suggestive evidence that the standard errors were larger when we relied on state tests than when we relied on study-administered tests.
- **Hypothesis 2:** Mismatched pre-tests yield larger standard errors than matched pre-tests.
- **Hypothesis 3:** A second pre-test in the same domain reduces the standard error of the impact estimate.
- **Hypothesis 4:** Models that specify the average score between the two post-tests—one from state tests and the other from a study-administered test—as the outcome variable yield smaller standard errors than models that specify either post-test individually as the outcome variable.

However, it is important to recognize that these hypotheses are not mathematical axioms that can be proven or disproven to hold in all cases. We tested these hypotheses because there are good reasons to believe that they will hold *when other factors are held constant* (as described earlier in the report). However, in real evaluations, other factors may vary across tests, and they may differ in important ways between the state tests and study-administered tests. Therefore, we would not expect the four hypotheses to hold in all studies.

Two important ways in which tests may vary are their reliability and the amount of missing data. Both factors influence the precision of the impact estimates. Other things held constant, we would expect: (1) more reliable tests to yield more precise impact estimates, and (2) tests with less missing data to yield more precise impact estimates. The relative reliability of state and study-administered tests in this study may not be generalizable to other evaluations conducted in other states and with other study-administered tests. Furthermore, the amount of missing data will vary depending on the tests used and the data collection strategy employed by the research team. For state tests, the amount of missing data will depend on several factors, including the fraction of students exempted from state testing, the amount of student mobility, and whether the evaluation team collects data from schools, districts, or state agencies. For study-administered tests, the amount of missing data may also depend on whether the tests are administered in school or outside of school, as well as the incentives offered to students and parents to participate. *More generally, there is no reason to expect that in other evaluations, the combined effects of reliability, missing data, and other factors on the precision of the impact estimates will be the same in magnitude—or even in direction—as we found for the three evaluations selected for this study.*

To see that the hypotheses will not hold in all cases, consider the following hypothetical examples:

- **Example 1.** Suppose that for the states included in the evaluation, the reliability of the state tests is equal to the reliability of the study-administered tests that the researchers are considering. In this example, we would not expect the choice between state and study-administered tests to yield different levels of precision, and Hypothesis 1b would be false.

- **Example 2.** Suppose that the researchers prefer to measure student outcomes with a study-administered test, but that missing data rates for the pre-test will be much higher for the study test than for the state tests (e.g., if it will be difficult to provide students with adequate incentives to show up for testing outside of school hours). If the difference in missing data rates is sufficiently large, the mismatched state pre-test could yield more precise impact estimates than the matched study pre-test because the state pre-test would yield a much larger analysis sample than the study pre-test. If this were true, Hypothesis 2 would be false.

Example 1 reinforces that our study has not proved that study-administered tests always yield more precise impact estimates than state tests. Example 2 reinforces that our study has not proved that matched pre-test will always yield more precise impact estimates than study tests. Multiple factors influence the precision of the impact estimates in an educational study, and the net effect of these factors is theoretically ambiguous.

Finally, one additional caution about generalizing the findings from this study to other evaluations is warranted: the effects of choosing different tests for the impact analysis may depend in part on the study's research design. It is well known that the precision of impact estimates depends heavily on the amount of unexplained variation at the level of random assignment. Furthermore, there is some evidence that suggests that the ability of covariates to explain the variation in outcomes may differ between the student, classroom, and school levels (see Bloom et al. 2007, Xu & Nichols 2010). This raises the question of whether the magnitude of the differences in precision measured in this paper would be different in other studies that randomize either students or schools. At the same time, we

would expect the direction of the effects, holding other factors constant, to be the same across study designs (e.g., mismatched pre-tests will reduce the precision of the impact estimates).

While our study addresses four specific questions, the more general question of interest is the following: *For educational impact studies, what are the consequences of using state tests instead of study-administered tests on the magnitude of the impact estimates and the size of the samples required to detect the impacts?* To build a stronger evidence base for addressing this question, it would be useful for researchers to conduct additional studies like the one we conducted, but with different samples. Additional studies, especially those based on data in other states and with different study-administered tests, would produce evidence on whether the four hypotheses for which we found support in this study usually hold, or only hold in rare cases.

Since education studies are increasingly relying on state tests, more evidence on the sample size requirements for these studies would be useful—even when comparisons with study-administered tests are not possible. Most of the papers that have contributed to our understanding of the sample size requirements in random assignment evaluations were based on data from either study-administered tests or pre-NCLB district tests (e.g., Bloom et al. 2007). Fortunately, some researchers have begun to compute and publish estimates based on state tests; these estimates are more clearly applicable when computing sample size requirements for studies based on state tests. For example, Xu and Nichols (2010) estimate key power parameters for randomized and cluster randomized designs based on state tests using data from Florida and North Carolina; additional studies like this one, but in other states, would be useful.

References

Allison, P.D. (2002). *Missing Data*. Sage University Paper 136.

Angrist, J.D., & A.B. Krueger (1999). Empirical Strategies in Labor Economics. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (pp. 1277-1366). Elsevier.

Bernstein, L., C. Dun Rappaport, L. Olsho, D. Hunt, & M. Levin (2009). *Impact Evaluation of the U.S. Department of Education's Student Mentoring Program* (NCEE 2009-4047). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Black, A.R., Doolittle, F., Zhu, P., Unterman, R., & Grossman, J. B. (2008). *The evaluation of enhanced academic instruction in after-school programs: findings after the first year of implementation* (NCEE 2008-4021). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Bloom, H. S., L. Richburg-Hayes, & A.R. Black (2007). Using Covariates to Improve Precision For Studies that Randomize Schools to evaluate Educational Interventions. *Educational Evaluation and Policy Analysis*, 29(1), 30-39.

Bound, J., & A.B. Krueger (1991). The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right. *Journal of Labor Economics*, 9(1), 1-24.

Cooper, H., L.V. Hedges, & J.C. Valentine (2002). *The Handbook of Research Synthesis and Meta-Analysis*, 2nd edition. New York: Russell Sage Foundation.

Cronin, J. (2004). *Aligning the NWEA RIT Scale with the South Carolina High School Assessment Program*. Lake Oswego, OR: Northwest Evaluation Association.

CTB/McGraw-Hill (2008a). *Teacher's Guide to TerraNova*, 3rd edition. Monterey, CA: CTB/McGraw-Hill Companies, Inc.

CTB/McGraw-Hill (2008b). *TerraNova*, 3rd edition. Technical Bulletin 1. Monterey, CA: CTB/McGraw-Hill Companies, Inc.

Dynarski, M., R. Agodini, S. Heaviside, T. Novak, N. Carey, L. Campuzano, M. Means, R. Murphy, W. Penuel, H. Javitz, D. Emery, & W. Sussex (2007). *Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort* (NCEE 2007-4005). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Educational Testing Service (2009). *California Standards Tests Technical Report, Spring 2008 Administration*. Report for California Department of Education, Standards and Assessment Division.

Garet, M.S., S. Cronen, S. Eaton, A. Kurki, M. Ludwig, W. Jones, K. Uekawa, A. Falk, H. Bloom, F. Doolittle, P. Zhu, & L. Sztejnberg (2008). *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement* (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Glazerman, S., E. Isenberg, S. Dolfin, M. Bleeker, A. Johnson, M. Grider, & M. Jacobus (2010). *Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study* (NCEE 2010-4027). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Graham, J.W. (2009). Missing Data Analysis: Making it Work in the Real World. *Annual Review of Psychology*, 60, 549-576.

Harcourt Educational Measurement (1996). *Stanford Achievement Test Series*, 9th edition. San Antonio, TX: Harcourt Educational Measurement.

Harcourt Assessment, Inc. (2004). *Stanford Achievement Test Series*, 10th edition. Technical data report. San Antonio, TX: Harcourt Assessment, Inc.

Griliches, Z., & J.A. Hausman (1986). Error in Variables in Panel Data. *Journal of Econometrics*, 31, 93-118.

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

Institute of Education Sciences, U.S. Department of Education (2008). *Rigor and Relevance Redux: Director's Biennial Report to Congress* (IES 2009-6010). Washington DC: Institute of Education Sciences, U.S. Department of Education.

Jacob, R. & P. Zhu (2009). *New Empirical Evidence for the Design of Group Randomized Trials in Education* (MDRC Working Paper on Research Methodology). New York: MDRC.

Kendall, M.G., A. Stuart, & J.K. Ord (1997). *Kendall's Advanced Theory of Statistics*, 6th edition. New York: Oxford University Press.

Kingsbury, G.G. (2001). *A Comparison of MAP and ALT Scores*. Lake Oswego, OR: Northwest Evaluation Association.

Kovacevic, M., R. Huang, & Y. You (2006). Bootstrapping for Variance Estimation in Multi-Level Models Fitted to Survey Data. *ASA Proceedings of the Survey Research Methods Section*, 3260-3269.

Lipsey, M.W., & D.B. Wilson (2001). *Practical Meta-Analysis*. Thousand Oaks, CA: Sage Publications.

May, H., I. Perez-Johnson, J. Haimson, S. Sattar, & P. Gleason (2009). *Using State Tests in Education Experiments: A Discussion of the Issues* (NCEE 2009-013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Miller, G.I., A. Jaciw, B. Ma, & X. Wei (2007). *Comparative Effectiveness of Scott Foresman Science: A Report of a Randomized Experiment in Visalia Unified School District.* (Empirical Education Rep. No. EEI_PEdSFSci-05-FR-Y1-S5.1). Palo Alto, CA: Empirical Education Inc.

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

Northwest Evaluation Association (2004). *A Few Notes about Reliability and Validity as They Are Reported In "NWEA Reliability and Validity Estimates: Achievement Level Tests and Measures of Academic Progress*." Lake Oswego, OR: Northwest Evaluation Association.

Northwest Evaluation Association (2003). *Technical Manual for the NWEA Measures of Academic Progress and Achievement Level Tests.* Portland, OR: Northwest Evaluation Association.

Puma, M.J., R.B. Olsen, S.H. Bell & C. Price (2009). *What to Do when Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

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

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8, 3-15.

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

Schochet, P.Z. (2008b). *Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schochet, P.Z. (2008c). *The Late Pretest Problem in Randomized Control Trials of Education Interventions* (NCEE 2009-4033). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shields, J. (2008). *A Comparison of the NWEA Measures of Academic Progress and the Missouri Assessment Program*. Doctoral Dissertation, University of Missouri.

Spybrook, J., S. Raudenbush, R. Congdon, & A. Martinez (2009). *Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software*. New York: W.T. Grant Foundation.

Sutcliffe, J.P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika*, 23, 9-17.

Thacker, A.A., E.R. Dickinson, & M.E. Koger (2004). *Relationships Among the Pennsylvania System of School Assessment (PSSA) and Other Commonly Administered Assessments.* Report by the Human Resources Research Organization for Central Susquehanna Intermediate Unit.

Torgesen, J., A. Schirm, L. Castner, S. Vartivarian, W. Mansfield, D. Myers, F. Stancavage, D. Durno, R. Javorsky, & C. Haan (2007). *National Assessment of Title I, Final Report: Volume II: Closing the Reading Gap, Findings from a Randomized Trial of Four Reading Interventions for Striving Readers* (NCEE 2008-4013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Torgesen, J.K., R.K. Wagner, & C.A. Rashotte. (1999). *TOWRE: Test of Word Reading Efficiency.* Austin, TX: PRO-ED.

van der Leeden, R., E. Meijer, & F.M.T.A Busing (2005). Resampling Multilevel Models. In J. de Leeuw (Ed.), *Handbook of Multilevel Analysis*. New York: Springer.

Williams, K.T. (2001a). *Group Reading Assessment and Diagnostic Evaluation (GRADE) Teacher's Scoring & Interpretive Manual.* Circle Pines, MN: American Guidance Service.

Williams, K.T. (2001b). *Group Reading Assessment and Diagnostic Evaluation (GRADE) Technical Manual.* Circle Pines, MN: American Guidance Service, Inc.

Williams, R., & D. Zimmerman (1995). Impact of Measurement Error on Statistical Power: Review of an Old Paradox. *Journal of Experimental Education*, 63(4), 363.

Woodcock, R.W. (1998). *Woodcock Reading Mastery Tests-Revised NU* (WRMT-R/NU). Circle Pines, MN: American Guidance Service.

Xu, Z., & A. Nichols (2010). *New Estimates of Design Parameters for Clustered Randomization Studies: Findings from North Carolina and Florida* (CALDER Working Paper 43). Washington, DC: Urban Institute.

Appendix A: Description of the Three Experiments

This appendix describes the three experiments that we have reanalyzed to address the four research questions specified for this project. Each study description presents information on the type of intervention tested, the design of the evaluation, including random assignment, the sample sizes, and the specific reading tests used to measure achievement in the study. Table A.1 provides a summary of the three studies. All three of these experiments were conducted by Empirical Education Inc. (EEI). The data collected for these studies are owned by the school systems involved and are not generally available for use by researchers. However, Empirical Education has an agreement with each school district to use these data for research purposes. The remainder of this appendix presents more details on the Arizona, California, and Missouri experiments.

1. The Arizona Experiment

Type of Intervention Tested

The treatment program was a reading intervention system created for struggling elementary readers. It was intended to supplement core reading instruction and designed to accelerate students toward grade-level reading performance. The program provides explicit, systematic instruction with ongoing progress monitoring. Treatment classrooms received the first half of the intervention only. (Arizona started the program in the middle of the year and could only use half of the year-long program.) Teachers received a half day in-service training led by a representative from the vendor company. Teachers also received monthly implementation support through on-site observations and meetings with vendor company staff.

Control classes received the "business as usual" program. Teachers used a kit supplied by the district's core reading program designed to meet the needs of struggling readers.

The Design of the Evaluation, Including Site Selection and Random Assignment

The evaluation was a randomized controlled trial (RCT). The participating district was identified by the vendor as a district interested in the product and willing to conduct a structured research study with a subset of their classrooms. The district identified interested schools whose principals invited teachers to an after-school meeting. Teachers volunteered to participate. Twenty-two classroom teachers and two reading specialists volunteered for the study.

The classroom was the unit of assignment. Matched pairs were formed from the 22 regular classroom teachers (i.e., those having one class of students each). Teachers who came from the same school and taught the same grade were paired and a coin toss was used to randomly assign one member of each pair to the treatment group and the other to the control group.[34] The two reading specialists were responsible for multiple classes: one was responsible for six classes, and the other was responsible

[34] Students were assigned to classrooms as they normally were (non-randomly) prior to random assignment.

for two classes. For each reading specialist, half of the classes were assigned to the treatment group and half were assigned to the control group.

Table A.1: Summary of the Three Randomized Controlled Trials

State	Intervention	Unit of Randomization	Grade Levels	Number of Classrooms	Number of Students
AZ	Supplemental reading program	Classes	3-5	15	98
CA	Scott Foresman Science	Classes	3-5	20	564
MO	Supplemental reading program	Classes	7-8	28	567

The Sample Sizes

This study involved one district, six schools, 24 teachers (and reading specialists) and 959 students from grades 3 – 5. For the purpose of this project, we have excluded grade 3 because state pre-tests were not available for that grade. In addition, the study defined a group of "focal children" on whom the intervention was focused. The remaining sample for the analysis included five schools, 12 teachers, 15 classrooms and 98 students (see Table A.2).

The Key Outcomes of Interest and Specific Measures Selected

Study-administered standardized test: Reading test scores were obtained from a study-administered standardized test. The pre-test consists of Northwest Evaluation Association's Measures of Academic Progress (NWEA MAP) Reading Survey 2 – 5 AZ V2. The post-test consists of MAP Reading Goals Survey 2 – 5 AZ V2. The version of the NWEA MAP used for the study was a state-aligned computerized adaptive assessment.

State-mandated test: Reading test scores were also obtained from Arizona's Instrument to Measure Standards (AIMS). AIMS is a vertically scaled test that is administered as part of the state assessment system in grades 3 – 8. For the study, the district provided 2005 AIMS reading scores as pre-test measures and 2006 AIMS reading scores as post-test measures for all students in grades 3 – 5. Since current grade 3 students were in grade 2 during the 2005 school year, they were not administered the AIMS and did not receive an AIMS score.

Summary Statistics and Missing Data

Table A.3 provides summary statistics for each of the variables used in the analysis. The final column presents the results of a statistical test for whether there are significant differences between the treatment and control samples at baseline, which could result from sampling error.[35] Table A.4 presents information on the amount of missing data for each variable used in the analysis.

[35] This statistical test accounted for the clustering of students within classrooms.

Table A.2: Sample Sizes for the Arizona Experiment

Condition	Schools	Teachers	Classrooms	# of Students
Treatment	N.A.	N.A.	7	44
Control	N.A.	N.A.	8	54
Total	**5**[a]	**12**[a]	**15**	**98**

[a] The experiment included five schools in total, and classrooms were randomly assigned. Therefore, we do not report the number of schools or teachers separately for the treatment group and the control group.

Table A.3: Summary Statistics for the Arizona Experiment

	Treatment Group		Control Group		
Variable	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**	**Significant Difference?**
State pre-test	401	27	401	29	No
State post-test	431	32	437	38	No
NWEA pre-test	194	17	192	14	No
NWEA post-test	191	16	191	14	No
Male	.59		.72		No
Eligible for free/reduced price lunch	.68		.74		No
Eligible for free lunch	.57		.65		No
Eligible for reduced price lunch	.11		.09		No
Asian	.00		.02		No
Black	.02		.04		No
Hispanic	.75		.61		No
White	.14		.19		No
Number of Students	*44*		*54*		

Table A.4: Missing Data in the Arizona Experiment

	Treatment Group		Control Group	
Variable	**Non-missing values**	**Missing values**	**Non-missing values**	**Missing values**
State pre-test	35	9	44	10
State post-test	44	0	52	2
NWEA pre-test	44	0	54	0
NWEA post-test	40	4	51	3
Male	44	0	54	0
Eligible for free or reduced price lunch	44	0	54	0
Eligible for free lunch	44	0	54	0
Eligible for reduced price lunch	44	0	54	0
Asian	44	0	54	0
Black	44	0	54	0
Hispanic	44	0	54	0
White	44	0	54	0

2. The California Experiment

Type of Intervention Tested

The treatment program was Pearson Education's Scott Foresman Science, a year-long science curriculum intended to be used as daily instruction. Based on inquiry-rich content with a sequence of structured and supportive inquiry activities, the science curriculum provides materials for both students and teachers in print, video, and online. A main feature of the curriculum is the Leveled Reader, which is designed to provide the teacher with an easy way to differentiate instruction and provide reading support at different reading levels. Although the main purpose of the intervention is to improve science skills, the program provides reading supports to make the science content accessible. The experiment treated reading achievement as an outcome under the premise that improved reading scores could be an important impact of the program.

Control classes received the "business as usual" science program offered by the district.

The Design of the Evaluation, Including Site Selection and Random Assignment

The evaluation was a randomized controlled trial (RCT). Pearson Education, the parent company of Scott Foresman, worked with a separate marketing company to identify districts interested in participating in research involving science curriculum. The district in the study was identified and contact information was forwarded to the study team. After contacting the district and identifying schools, the study team met with district staff members and principals to explain the details and procedures of the study. Principals identified eligible teachers, and 21 teachers volunteered to participate.

The unit of randomization was the teacher. Matched pairs were formed based on school assignment and grade level taught. Ten pairs were formed and members of each pair were randomly assigned to either treatment or control. One teacher was unpaired and was randomly assigned to treatment. After random assignment, one control teacher moved out of the area before the beginning of the school year and was excluded from the study, leaving 20 teachers in the experiment.

The Sample Sizes

This study involved one district, two schools, 20 teachers and 616 students from grades 3 – 5. However, the study team identified a group of focal children prior to the experiment to whom the intervention was targeted. After excluding other students, the remaining sample included two schools, 20 teachers, 20 classrooms and 564 students (see Table A.5).

The Key Outcomes of Interest and Specific Measures Selected

Study-administered standardized test: Reading test scores were obtained from a study-administered standardized test. The pre-test consisted of the Northwest Evaluation Association (NWEA) Achievement Level Test (ALT) of reading achievement. The post-test also consisted of the ALT test of Reading. This test is an adaptive and comprehensive paper and pencil test designed to measure growth over time. The difficulty level of the test given was determined using a short locator test.

Table A.5: Sample Sizes for the California Experiment

	No. of Schools	No. of Teachers	No. of Classes	Total Students
Treatment	N.A.	11	11	320
Control	N.A.	9	9	244
Total	**2**[a]	**20**[b]	**20**	**564**

[a] In this experiment, classrooms/teachers were randomly assigned. Therefore, we do not report the number of schools separately for the treatment group and the control group.

[b] One of the teachers randomly assigned to the control group moved out of the area before the start of the study and before student rosters were available. This teacher was not included in the study (and is not included in the table above).

State-mandated test: Pre-test and post-test reading scores were also obtained from the California Standards Test (CST), which is administered as part of the state assessment system in grades 2 – 11.

Summary Statistics and Missing Data

Table A.6 provides summary statistics for each of the variables used in the analysis. The final column presents the results of a statistical test for whether there are significant differences between the treatment and control samples at baseline, which could result from sampling error.[36] Table A.7 presents information on the amount of missing data for each variable used in the analysis.

[36] This statistical test accounted for the clustering of students within classrooms.

Table A.6: Summary Statistics for the California Experiment

Variable	Treatment Group		Control Group		Significant Difference?
	Mean	Standard Deviation	Mean	Standard Deviation	
State pre-test	342	57	349	51	No
State post-test	347	55	350	54	No
NWEA pre-test	194	14	196	14	No
NWEA post-test	203	14	203	14	No
Male	.48		.50		No
Eligible for free/reduced price lunch	NA		NA		No
Eligible for free lunch	NA		NA		No
Eligible for reduced price lunch	NA		NA		No
Asian	.10		.10		No
Black	.03		.04		No
Hispanic	.54		.58		No
White	.31		.27		No
Number of Students	*320*		*244*		

Table A.7: Missing Data in the California Experiment

Variable	Treatment Group		Control Group	
	Non-missing values	Missing values	Non-missing values	Missing values
State pre-test	289	31	210	34
State post-test	301	19	235	9
NWEA pre-test	204	116	172	72
NWEA post-test	266	54	210	34
Fraction male	320	0	244	0
Eligible for free/reduced price lunch	NA		NA	
Fraction receiving free lunch	NA		NA	
Fraction receiving reduced price lunch	NA		NA	
Fraction Asian	320	0	244	0
Fraction Black	320	0	244	0
Fraction Hispanic	320	0	244	0
Fraction White	320	0	244	0

3. The Missouri Experiment

Type of Intervention Tested

The treatment program was a middle school reading curriculum created for struggling adolescent readers. The program is content aligned with the National Standards for Reading, and it is intended to supplement core reading instruction. The complete program requires 30 weeks of instruction, and it provides explicit, systematic instruction with ongoing progress monitoring.

The Design of the Evaluation, Including Site Selection and Random Assignment

The evaluation was a randomized controlled trial (RCT). The participating district was identified by the vendor as a district interested in the product and willing to conduct a structured research study with a subset of their classrooms. Researchers corresponded with district staff to explain the procedures. The district identified interested schools whose principals invited teachers to participate. Seven teachers representing two middle-schools volunteered to participate.

Classes were the unit of assignment. For each teacher, similar classes were paired, and from each pair, one class was randomly assigned to the treatment group and the other class was assigned to the control group.

The Sample Sizes

This study involved one district, two schools, seven teachers, 28 classes, and 610 students from grades 7 and 8. However, as in Arizona, the Missouri experiment defined a group of focal children to whom the intervention was targeted prior to the experiment. After excluding other students, the remaining sample includes two schools, seven teachers, 28 classrooms and 567 students (see Table A.8).

The Key Outcomes of Interest and Specific Measures Selected

Study-administered standardized test: Reading test scores were obtained from a study-administered standardized test. The pre-test consists of Northwest Evaluation Association's Measures of Academic Progress (NWEA MAP) Reading Survey 6+ MO V3. The post-test was the MAP Reading Goals Survey 6+ MO V3. The version of NWEA MAP used for this study was a state-aligned computerized adaptive assessment.

State-mandated test: The district also provided results from their own testing and from the Missouri Assessment Program.[37] The district provided pre-test scores from the district test because pre-test scores were not available from the Missouri Assessment Program; it provided post-test scores from the Missouri Assessment Program because the district test was only administered in grade 7.

[37] Students in the sample took the state test in the previous year. However, the district was unable to provide state pre-test scores to the study team due to computer-related problems.

Table A.8: Sample Sizes for the Missouri Experiment

Condition	Schools	Teachers	Classrooms	# of Students
Treatment	N.A.	N.A.	14	274
Control	N.A.	N.A.	14	293
Total	**2**[a]	**7**[a]	**28**	**567**

[a] For each teacher, one classroom was randomly assigned to the treatment group and one to the control group. Therefore, we do not report the number of schools or teachers separately for each group.

Summary Statistics and Missing Data

Table A.9 provides summary statistics for each of the variables used in the analysis. The final column presents the results of a statistical test for whether there are significant differences between the treatment and control samples at baseline, which could result from sampling error.[38] Table A.10 presents information on the amount of missing data for each variable used in the analysis.

[38] This statistical test accounted for the clustering of students within classrooms.

Table A.9: Summary Statistics for the Missouri Experiment

Variable	Treatment Group		Control Group		Significant Difference?
	Mean	Standard Deviation	Mean	Standard Deviation	
District pre-test	40	14	38	15	No
State post-test	663	34	658	29	No
NWEA pre-test	216	11	214	11	No
NWEA post-test	211	13	211	13	No
Male	.45		.51		No
Eligible for free/reduced price lunch	.86		.85		No
Eligible for free lunch	.77		.74		No
Eligible for reduced price lunch	.09		.11		No
Asian	.00		.00		No
Black	.98		.97		No
Hispanic	.00		.00		No
White	.02		.03		No
Number of Students	*274*		*293*		

Table A.10: Missing Data in the Missouri Experiment

Variable	Treatment Group		Control Group	
	Non-Missing Values	Missing Values	Non-Missing Values	Missing Values
District pre-test	253	21	247	46
State post-test	246	28	253	40
NWEA pre-test	256	18	260	33
NWEA post-test	198	76	223	70
Fraction male	274	0	293	0
Eligible for free/reduced price lunch	257	17	253	40
Fraction receiving free lunch	257	17	253	40
Fraction receiving reduced price lunch	257	17	253	40
Fraction Asian	274	0	293	0
Fraction Black	274	0	293	0
Fraction Hispanic	274	0	293	0
Fraction White	274	0	293	0

Appendix B: Scatter Plots of Student Test Scores

The scatter plots presented in this appendix show the relationship between student reading scores from two different types of tests: (1) the NWEA study-administered test, and (2) state or district tests. Each exhibit presents the scatter plots for one of the three experiments. In addition, each exhibit includes four figures: (1) pre-test scores for the treatment group, (2) pre-test scores for the control group, (3) post-test scores for the treatment group, and (4) post-test scores for the control group.

Each point in a scatter plot represents a single student in the sample. In most figures, the points display student scale scores. The exception to this rule are the pre-test scores from the district test in Missouri, where the scores reported to the evaluation team and used in the analysis for this report are the raw scores. The scores have not been transformed into *z*-scores, as they are for the impact analysis presented in the text of this report. The line drawn through the points represents the best fitting regression line.

Figure B.1: Reading Scores in the Arizona Experiment, NWEA (MAP) vs. State Test

Figure 1: Treatment group pre-tests (N=35)

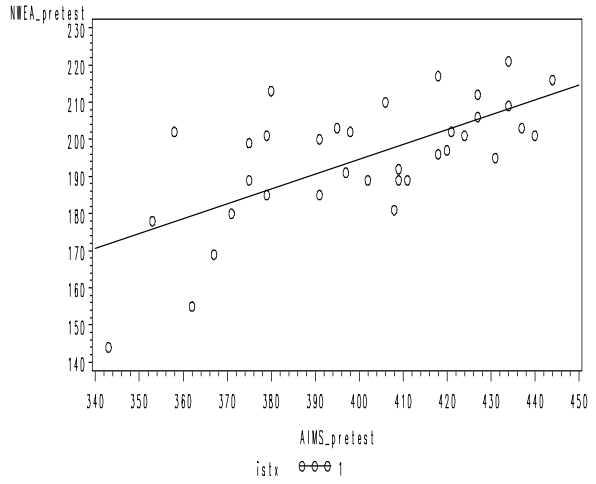


Figure 2: Control group pre-tests (N=44)

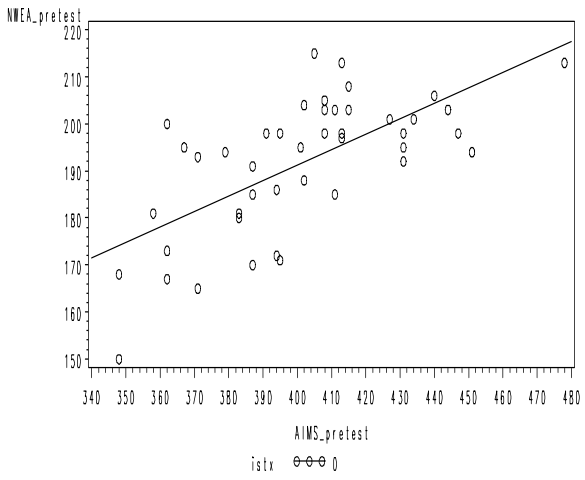


Figure 3: Treatment group post-tests (N=40)

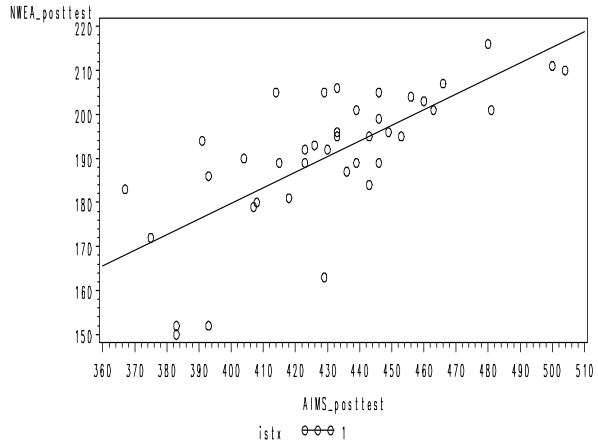


Figure 4: Control group post-tests (N=50)

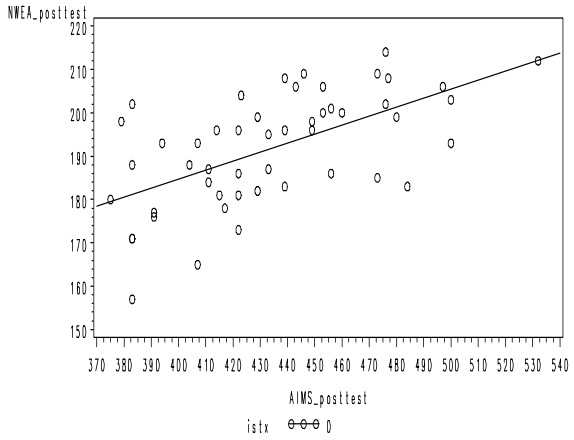


Figure B.2: Reading Scores in the California Experiment, NWEA (ALT) vs. State Test

Figure 1: Treatment group pre-tests (N=179)

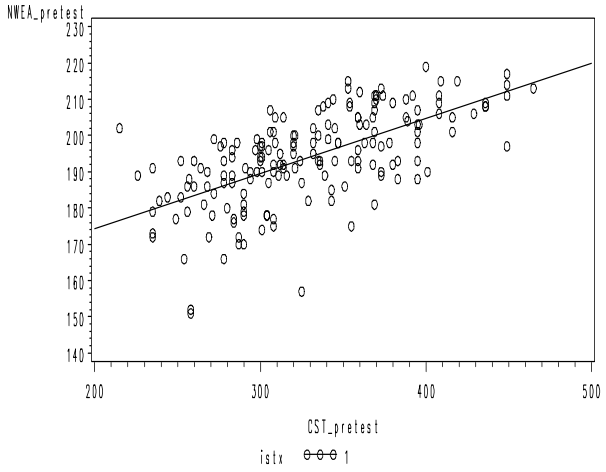


Figure 2: Control group pre-tests (N=140)

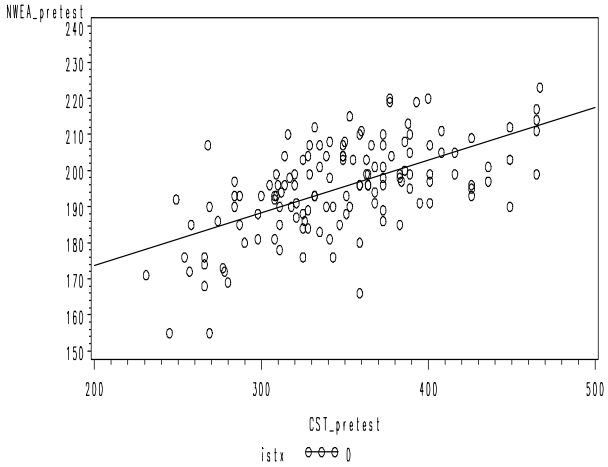


Figure 3: Treatment group post-tests (N=266)

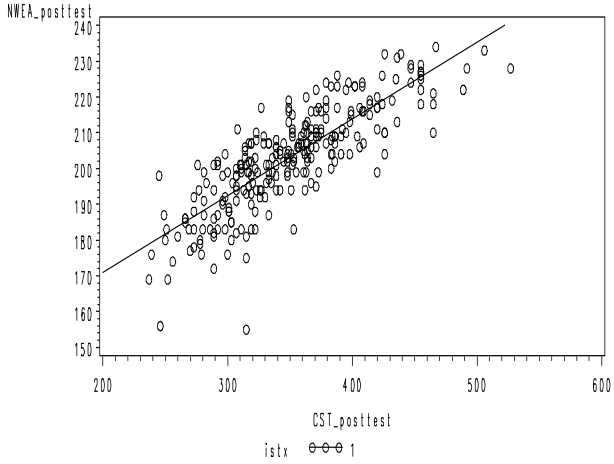


Figure 4: Control group post-tests (N=210)

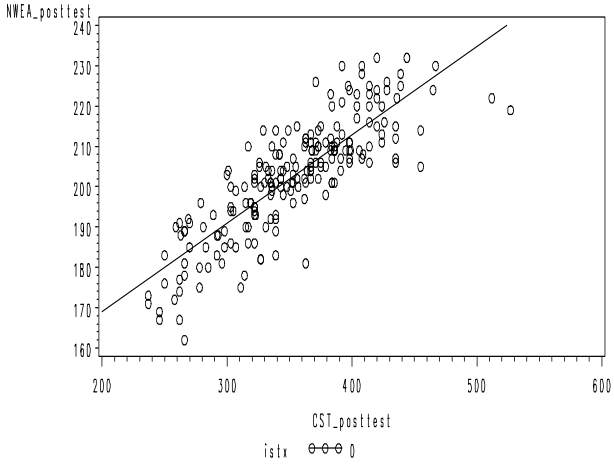


Figure B.3: Reading Scores in the Missouri Experiment, NWEA (MAP) vs. State Test

Figure 1: Treatment group pre-tests (N=241)

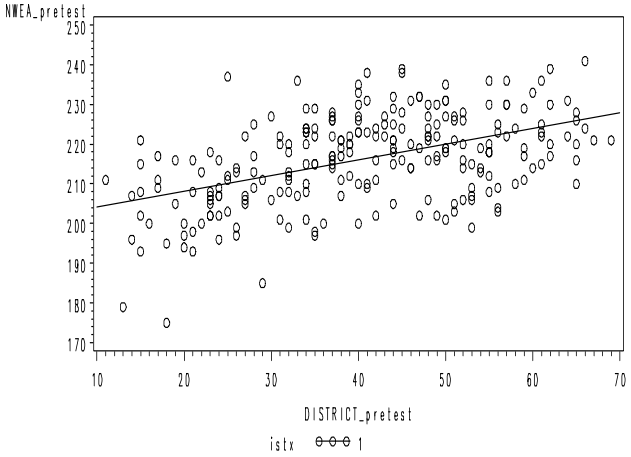


Figure 2: Control group pre-tests (N=228)

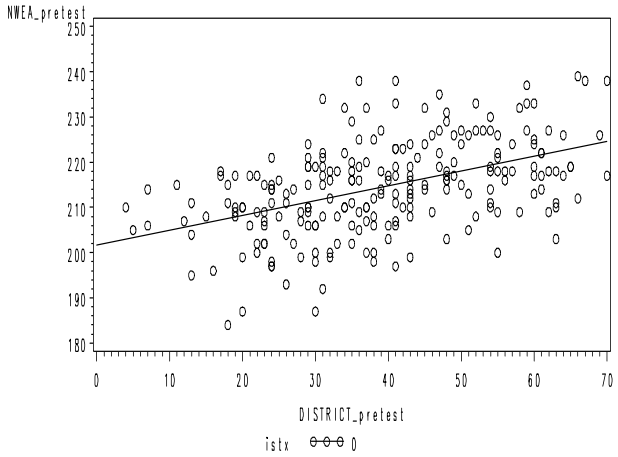


Figure 3: Treatment group post-tests (N=191)

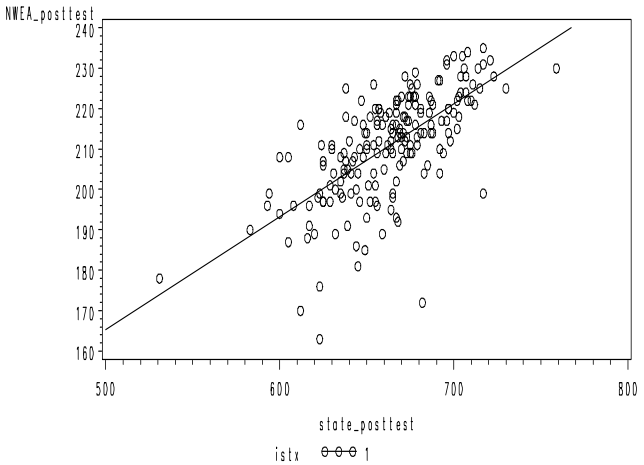
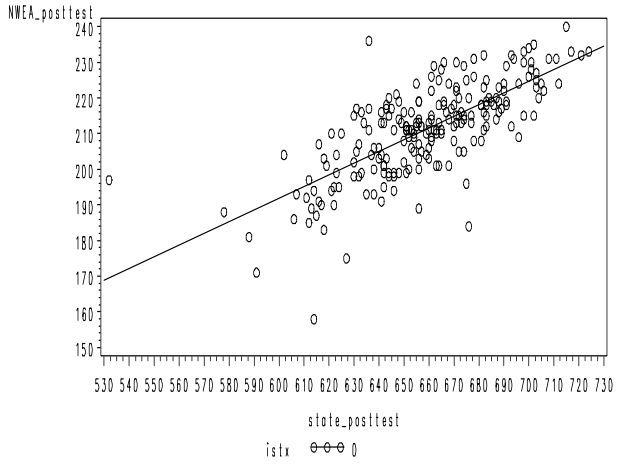


Figure 4: Control group post-tests (N=217)



Appendix C: Quartiles of the Test Score Distribution

Table C.1: Test Score Quartiles in Arizona Experiment, Control Group

State Test Scores	Study-Administered Test (NWEA, MAP)			
Quartiles	***1***	***2***	***3***	***4***
	Pre-test Scores			
1 (Bottom 25 percent)	16%	7%	2%	0%
2	9%	9%	5%	2%
3	0%	5%	7%	14%
4 (Top 25 percent)	0%	5%	11%	9%
	Post-test Scores			
1	14%	6%	2%	2%
2	10%	8%	6%	2%
3	0%	4%	12%	8%
4	0%	8%	4%	14%

Table C.2: Test Score Quartiles in Arizona Experiment, Treatment Group

State Test Scores	Study-Administered Test (NWEA, MAP)			
Quartiles	***1***	***2***	***3***	***4***
	Pre-test Scores			
1 (Bottom 25 percent)	17%	3%	3%	0%
2	6%	6%	11%	3%
3	3%	14%	3%	6%
4 (Top 25 percent)	0%	3%	9%	14%
	Post-test Scores			
1	18%	5%	3%	0%
2	5%	13%	3%	5%
3	3%	8%	10%	5%
4	0%	0%	10%	15%

Table C.3: Test Score Quartiles in California Experiment, Control Group

State Test Scores	Study-Administered Test (NWEA, ALT)			
Quartiles	***1***	***2***	***3***	***4***
	Pre-test Scores			
1 (Bottom 25 percent)	14%	9%	1%	1%
2	7%	7%	6%	4%
3	3%	5%	9%	9%
4 (Top 25 percent)	1%	4%	9%	11%
	Post-test Scores			
1	19%	5%	0%	0%
2	5%	14%	4%	1%
3	0%	6%	13%	6%
4	0%	0%	7%	18%

Table C.4: Test Score Quartiles in California Experiment, Treatment Group

State Test Scores	Study-Administered Test (NWEA, ALT)			
Quartiles	***1***	***2***	***3***	***4***
	Pre-test Scores			
1 (Bottom 25 percent)	15%	6%	3%	0%
2	6%	8%	9%	2%
3	3%	6%	8%	8%
4 (Top 25 percent)	1%	4%	5%	15%
	Post-test Scores			
1	18%	7%	0%	0%
2	7%	12%	6%	0%
3	0%	6%	12%	7%
4	0%	1%	6%	18%

Table C.5: Test Score Quartiles in Missouri Experiment, Control Group

District/State Test Scores	Study-Administered Test (NWEA, MAP)			
Quartiles	***1***	***2***	***3***	***4***
District Test Scores	***Pre-test Scores***			
1 (Bottom 25 percent)	12%	9%	4%	0%
2	8%	4%	7%	5%
3	4%	6%	7%	8%
4 (Top 25 percent)	1%	6%	7%	11%
State Test Scores	***Post-test Scores***			
1	16%	6%	3%	0%
2	7%	10%	7%	1%
3	2%	6%	9%	7%
4	0%	3%	6%	17%

Table C.6: Test Score Quartiles in Missouri Experiment, Treatment Group

District/State Test Scores	Study-Administered Test (NWEA, MAP)			
Quartiles	***1***	***2***	***3***	***4***
District Test Scores	***Pre-test Scores***			
1 (Bottom 25 percent)	15%	7%	1%	1%
2	5%	7%	9%	5%
3	3%	5%	7%	10%
4 (Top 25 percent)	3%	6%	7%	9%
State Test Scores	***Post-test Scores***			
1	14%	9%	1%	1%
2	7%	8%	8%	1%
3	3%	5%	9%	8%
4	1%	3%	6%	16%

Note: In the Missouri experiment, district test scores were used as pre-test measures while state test scores were used as post-test measures.

Appendix D: Estimates from Other Evaluations

A comprehensive review of other evidence that addresses the four research questions is beyond the scope of this study. However, a small number of IES-funded studies have both collected reading scores from state or district tests *and* administered one or more general reading tests as part of the evaluation. In particular, we are aware of three such evaluations:

1. **Closing the Reading Gap.** This evaluation estimated the effects of four different reading pull-out programs for struggling readers in grades 3 and 5. This study administered two tests of reading comprehension at baseline and follow-up: the Passage Comprehension test from the Woodcock Reading Mastery Test–Revised (Woodcock 1998), and the Group Reading Assessment and Diagnostic Evaluation, or GRADE (Williams 2001a, Williams 2001b). In addition, follow-up reading scores from the state assessments were collected. Using these data, the authors estimated regressions that correspond to Models A and D. For more information about the study, see Torgesen et al. (2007).

2. **Evaluation of the Effectiveness of Educational Technology Interventions.** This evaluation is estimating the impacts of education technology interventions, including some interventions focused on reading and other interventions focused on math. The evaluation administered two reading tests, the Stanford 9 (Harcourt Educational Measurement 1996) and the Test of Word Reading Efficiency (Torgesen, Wagner, & Rashotte 1999), and it also collected district or state reading scores. While the primary impact estimates in both reports were based on study-administered tests, the first study report presents the results of a sensitivity analysis in which they estimated impacts on reading scores from district tests. However, since the report does not specify which pre-test variable (i.e., district or study-administered) was used to estimate impacts on scores from the district or state reading test, it is not clear whether the sensitivity analysis reflect a comparison of Models A and B or a comparison of Models A and D. For more information about the study, see Dynarski et al. (2007).

3. **Impact Evaluation of Academic Instruction for After-School Programs.** This evaluation is estimating the impacts of Harcourt's Mathletics and Success For All's Adventure Island, two academic programs designed for after-school programs. The evaluation administered the abbreviated Stanford 10 reading test (Harcourt Assessment, Inc. 2004) at baseline and follow-up, but it also collected follow-up test scores from district or state reading tests. Using these data, the authors estimated regressions that correspond to Models A and D. For more information about the study, see Black et al. (2008).

Unfortunately, none of these studies provide evidence that directly addresses any of the four research questions for this study. The first and third evaluations present evidence that allows us to compare Model A (study-administered pre-test, study-administered post-test) to Model D (study-administered pre-test, state or district post-test). However, this comparison does not directly address any of our four research questions. In summary, we are not aware of other evaluations that have produced evidence that directly addresses the hypotheses that we test in this study.

Appendix E: Estimates from the Full Sample

Section D of this report presents estimates based on the common sample, which includes all of the students for which scores from all four tests were available (i.e., non-missing). The decision to focus the confirmatory analysis on the common sample was described and justified in Section C of the report.

In this appendix, we present the results from the analysis for the full sample. The full sample includes all of the observations for which the model's dependent variable was non-missing. Note that the dependent variable is the MAP test in Models A, C, and E, the state test in Models B, D, and F, and the average of the two post-tests in Model G (see Table 4 in Section C). Therefore, the full sample analysis includes all students with non-missing MAP post-test scores for Models A, C, and E, all students with non-missing state post-test scores for Models B, D, and F, and all students with non-missing scores on both post-tests for Model G.

The results from the full sample analysis are presented in Tables E.1 – E.4.

Table E.1: Comparing Effect Sizes and Standard Errors between State and Study Tests, Estimates to Address Question 1, Full Sample

	Standard Errors			Standard Errors		
Analysis	Model A (MAP post, MAP pre)	Model B (State post, state pre)	*p*-value† (Model A – Model B)	Model A (MAP post, MAP pre)	Model B (State post, state pre)	*p*-value† (Model A – Model B)
Pooled	**0.005**	**-0.012**	**.348**	**0.106**	**0.109**	**.494**
By State						
Arizona	-0.041	-0.094	.808	0.171	0.238	.091*
California	0.070	-0.089	.017**	0.115	0.093	.147
Missouri	-0.019	0.060	.243	0.086	0.083	.848

† For the pooled analyses, we report *p*-values that are not adjusted for multiple comparisons because we used a single statistical test for each of the two hypotheses (Hypothesis 1a for impacts or effect sizes and Hypothesis 1b for standard errors). For the state-level analyses, we report unadjusted *p*-values because the analyses should be classified as exploratory.

Notes: * indicates $p < .10$, ** indicates $p < .05$, and *** indicates $p < .01$. For a description of the models, see Table 4. The estimates in this table were computed from the full sample. For each model, we deleted cases with missing values in the dependent variable, and we used the dummy variable adjustment method to address missing values in the independent variables. For pooled estimates of the average effect size or standard error across studies for each model, we computed a weighted average of the state-level estimates, where the weight for each state was proportional to the inverse of the variance of the estimate for that state. For the pooled hypothesis test of no difference between models, we computed a weighted average of the state-level differences in the estimates between models, where the weight for each state was proportional to the inverse of the variance of the estimated difference for that state (see Section C for more details).

Table E.2: Estimating the Increase in Standard Errors from a Mismatched Pre-test, Question 2, Full Sample

Analysis	Standard Errors			Standard Errors		
	Model A (MAP post, MAP pre)	Model C (MAP post, state pre)	*p*-value† (Model A – Model C)	Model B (State post, state pre)	Model D (State post, MAP pre)	*p*-value† (Model B – Model D)
Pooled	**0.106**	**0.128**	**.026****	**0.118**	**0.211**	**.071***
By State						
Arizona	0.171	0.195	.361	0.238	0.263	.306
California	0.115	0.081	.399	0.093	0.163	.032**
Missouri	0.086	0.119	.008***	NA	NA	NA

† For the pooled analysis, we report *p*-values that are adjusted for multiple comparisons because we used two statistical comparisons to test a single hypothesis (Hypothesis 2): the adjusted *p*-value equals two times the unadjusted *p*-value. As a result, some adjusted *p*-values may be greater than one. For the state-level analyses, we report unadjusted *p*-values because the analyses should be classified as exploratory.

†† Missouri was excluded from the analysis for comparisons of Models B and D because district pre-test scores were collected instead of state pre-test scores. This means that Model B is based on a mismatched pre-test (state post-test and district pre-test), and the data from Missouri cannot be used to test Hypothesis 2 when post-test scores come from state assessments.

Notes: * indicates $p < .10$, ** indicates $p < .05$, and *** indicates $p < .01$. For a description of the models, see Table 4. The estimates in this table were computed from the full sample. For each model, we deleted cases with missing values in the dependent variable, and we used the dummy variable adjustment method to address missing values in the independent variables. For pooled estimates of the average standard error across studies for each model, we computed a weighted average of the state-level estimates, where the weight for each state was proportional to the inverse of the variance of the estimate for that state. For the pooled hypothesis test of no difference in the standard error estimate between models, we computed a weighted average of the state-level differences in the estimates between models, where the weight for each state was proportional to the inverse of the variance of the estimated difference for that state (see Section C for more details).

Table E.3: Estimating the Decrease in Standard Errors from Using Both Pre-tests, Question 3, Full Sample

Analysis	Standard Errors			Standard Errors		
	Model A (MAP post, MAP pre)	Model E (MAP post, both pre)	*p*-value† (Model A – Model E)	Model B (State post, state pre)	Model F (State post, both pre)	*p*-value† (Model B – Model F)
Pooled	**0.106**	**0.082**	**<.001*****	**0.109**	**0.100**	**.001*****
By State						
Arizona	0.171	0.174	.431	0.238	0.253	.242
California	0.115	0.059	<.001***	0.093	0.102	.736
Missouri	0.086	0.088	.096*	0.083	0.072	.202

† For the pooled analysis, we report *p*-values that are adjusted for multiple comparisons because we used two statistical comparisons to test a single hypothesis (Hypothesis 3): the adjusted *p*-value equals two times the unadjusted *p*-value. As a result, some adjusted *p*-values may be greater than one. For the state-level analyses, we report unadjusted *p*-values because the analyses should be classified as exploratory.

Notes: * indicates $p < .10$, ** indicates $p < .05$, and *** indicates $p < .01$. For a description of the models, see Table 4. For more details on the computations presented in this table, see the notes below Table E.2.

Table E.4: Estimating the Decrease in Standard Errors from Averaging the Two Post-tests, Question 4, Full Sample

Analysis	Standard Errors			Standard Errors		
	Model G (mean post, both pre)	Model E (MAP post, both pre)	*p*-value† (Model G – Model E)	Model G (mean post, both pre)	Model F (State post, both pre)	*p*-value† (Model G – Model F)
Pooled	**0.083**	**0.082**	**1.000††**	**0.083**	**0.100**	**.052***
By State						
Arizona	0.190	0.174	.555	0.190	0.253	.011**
California	0.072	0.059	.257	0.072	0.102	.012**
Missouri	0.073	0.088	.152	0.073	0.072	.920

† For the pooled analysis, we report *p*-values that are adjusted for multiple comparisons because we used two statistical comparisons to test a single hypothesis (Hypothesis 4): the adjusted *p*-value equals two times the unadjusted *p*-value. As a result, some adjusted *p*-values may be greater than one. For the state-level analyses, we report unadjusted *p*-values because the analyses should be classified as exploratory.

†† While the unadjusted *p*-value is less than one, the *p*-value after adjusting for multiple comparisons is greater than or equal to one.

Notes: * indicates $p < .10$, ** indicates $p < .05$, and *** indicates $p < .01$. For a description of the models, see Table 4. For more details on the computations presented in this table, see the notes below Table E.2.

Appendix F: Hypothesis Tests and Minimum Detectable Differences

In this study, we compared the results from pairs of different models to estimate the impacts of an educational intervention. For a description of all of the models, see Section C.

To determine whether the differences in impact estimates and standard errors between different models could easily be attributed to sampling error, we conducted hypothesis tests. For example, in comparing Models A and B to address Question 1, we conducted two hypothesis tests for each state—one to test the null hypothesis that the two impacts were equal, and the other to test the null hypothesis that the two standard errors were equal.

However, conducting the tests was complicated by two factors. First, the impact estimates and standard error estimates for any two models estimated from the same sample (e.g., the sample from the Arizona experiment) are positively correlated. Therefore, hypothesis tests that are based on the independence assumption would be too conservative—and potentially by a large margin if the correlation between the estimates is high. Second, standard estimation techniques do not yield estimates of the variance of the standard error estimates, much less correlations between the standard error estimates, which are necessary to test the hypothesis that two standard errors are equal.

To address these challenges, we implemented, tested, and utilized a bootstrapping algorithm designed specifically for cluster randomized trials, including the three studies we reanalyzed for this report. In most applications, researchers conduct non-parametric versions of bootstrapping to relax the assumptions of the mixed model. However, in this application, our goal was not to relax these assumptions; instead, we simply wanted to compute variance and covariance estimates that would allow us to conduct these hypothesis tests. In addition, given the research design implemented in the three RCTs, the best approach to non-parametric bootstrapping was unclear. In particular, it was not clear whether the bootstrapping algorithm should resample blocks or classes within blocks. Given these uncertainties, we decided to develop a parametric bootstrapping algorithm. For formal discussions of parametric bootstrapping, as well as specific approaches to implementing it, see Kovacevic, Huang, & You (2006) and van der Leeden, Meijer, & Busing (2005).

In broad terms, to implement parametric bootstrapping to facilitate hypothesis tests that compare two models (e.g., Model A and Model B), we:

- Estimated both models in the data;
- Used the estimated model to produce a predicted value of the dependent variable or variables conditional on the fixed covariates in the model;
- Added class-level and student-level noise to generate a random bootstrap value for the outcome for each individual;
- Repeated the process 1,000 times to generate 1,000 bootstrap samples for each pair of models.

The particular bootstrapping procedure implemented for this project was developed by the authors of this report. Because, to the best of our knowledge, this bootstrapping algorithm has not been used in prior studies, we subjected the algorithm to a series of tests. For these tests, we used data from an unrelated evaluation. Once the algorithm passed the tests using the data from that evaluation, we implemented the algorithm in the three RCTs used in this project, and re-tested the algorithm.

The tests we developed were designed to ensure that the bootstrapping algorithm yielded bootstrap sample data with distributional properties that were consistent with the observed data. In particular, we wanted the bootstrap sample data to match the actual data on the following dimensions:

- **Means and standard deviations.** The means and standard deviations of simulated outcomes should converge (over many replications) to the observed means and standard deviations of the observed data.
- **Correlations between the scores from the state and study-administered tests.** The correlation between the *relevant* test scores (e.g., between the study-administered post-test in Model A and the state post-test in Model B) in the bootstrap samples should converge (over many replications) to the correlation in the true (observed) data set.

In addition, we wanted to make sure that the bootstrap algorithm generated impact estimates and standard errors that were close to the impact estimates and standard error estimates from the two-level hierarchical linear model (HLM) which was used to produce the estimates reported in the main body of this report. In most applications, differences between the two sets of estimates would not necessarily suggest a problem with the bootstrapping algorithm; instead, it could indicate violations of the assumptions imposed by the two-level HLM. However, since our bootstrapping algorithm was designed to impose the same assumptions as the two-level HLM, we would have interpreted differences in the estimates as evidence that the bootstrapping algorithm was not performing properly.

More specifically, in our diagnostic testing, we checked to make sure that the bootstrapping samples generated estimates that were consistent with the two-level HLM estimates from the actual data with respect to:

- **Variance components in the unconditional impact model.** For each type of outcome (study-administered and state), the variance components estimated from a two-level unconditional model would converge to the observed estimates of the variance components from the unconditional two-level model of the observed data. By "unconditional model" we mean a model with no right-hand side fixed effects terms other than the intercept.
- **Variance components in the conditional or full impact model.** For each type of outcome (study-administered and state), the variance components estimated from a two-level conditional (fully specified) HLM would converge to the observed estimates of the variance components from the full two-level model of the observed data. By "full model" we mean the models with pre-test and all other covariates used in the final impact models.
- **The treatment effect.** For each type of outcome (study-administered and state), the treatment impact estimate and all other fixed effects parameter estimates from the full two-level model would converge to the observed fixed effect parameter estimates from the full two-level model in the observed data.

- **The standard error of the treatment effect.** For each type of outcome (evaluator and state), the standard deviation of the impact estimates over the 1,000 replicates would be equal to the standard errors of the impact estimates from the true (observed data).

The remainder of this appendix describes how we constructed the bootstrap samples. Sections 1.1 – 1.6 describe how bootstrap samples were constructed to compare Model B to Model A. Section 1.7 summarizes how the same approach was used to compare other pairs of models. Appendix G provides a more general description of our approach to generating correlated residuals. Appendix H reports the results from this bootstrapping process.

1. Fit Models to the Observed Data

The impact models that were fit to the observed data correspond to a cluster randomized design where classes were randomized to treatment and control conditions, and students were nested in classes. For each model pair, we fit an unconditional model and a full model. To describe our approach, we focus on the comparison between Model A and Model B to address the study's first research question. This pair of models can be written as follows:

Model A.1 - Unconditional:

$$Yevaluator_{post.ij} = \beta_0^u + \alpha_{0.j}^u + \varepsilon_{ij}^u$$

Model A.2 - Full:

$$Yevaluator_{post.ij} = \beta_0 + \alpha_{0j} + \beta_1(Yevaluator_{pre.ij}) + \beta_2(Trt_j) + \sum_{m=1}^{M} \beta_{m+2}(X_{m.ij}) + \varepsilon_{ij}$$

where

The "u" superscript for parameters in the unconditional model emphasizes that they are different than the parameters in the full model, and

$Yevaluator_{post.ij}$	is the evaluator-administered (study-administered) test score, measured post intervention, of the i^{th} student in the j^{th} class.
$Yevaluator_{pre.ij}$	is the evaluator-administered test score, measured pre intervention, of the i^{th} student in the j^{th} class.
Trt_j	is a treatment indicator, =1 for treatment and 0 for control.
$X_{m.ij}$	is the m^{th} (m=1…M) additional covariates that are included in the model to account for the sampling design (e.g. dummies for class pairs) or to increase statistical power.[39]
α_{0j}	is a random intercept term for the j^{th} class and is assumed to be distributed normally with mean 0 and variance τ^2, and is assumed to be independent of ε_{ij}.

[39] Note that although we have subscripted X with both "i" and "j," some of these covariates were "level-2" covariates (i.e., all students within a class share the same value on the variable), while others varied among the students within a class.

ε_{ij} is the random level-1 residual term, and is assumed to be distributed normally with mean 0 and variance σ^2, and is assumed to be independent of α_{0j}.

Model B.1 - Unconditional:

$$Ystate_{post.ij} = \beta_0^{u*} + \alpha_{0.j}^{u*} + \varepsilon_{ij}^{u*}$$

Model B.2 - Full:

$$Ystate_{post.ij} = \beta_0^{*} + \alpha_{0.j}^{*} + \beta_1^{*}(Ystate_{pre.ij}) + \beta_2^{*}(Trt_j) + \sum_{m=1}^{M} \beta_{m+2}^{*}(X_{m.ij}) + \varepsilon_{ij}^{*}$$

where all of the terms are as described for Model A.2, but where the test scores come from the state test. The "u" superscript for parameters in the unconditional model emphasize that they are different than the parameters in the full model, and the "*" superscript on parameters in Models B.1 and B.2 emphasize that they are different than those in Models A.1 and A.2.

Each of the four models described above produces estimates of Level-2 (class) and Level-1 (student) variance components. We denote these estimates as:

L2Var.ModA.uc	Level 2 (class) variance from Model A.1
L1Var.ModA.uc	Level 1 (student) variance from Model A.1
L2Var.ModA.full	Level 2 (class) variance from Model A.2
L1Var.ModA.full	Level 1 (student) variance from Model A.2
L2Var.ModB.uc	Level 2 (class) variance from Model B.1
L1Var.ModB.uc	Level 1 (student) variance from Model B.1
L2Var.ModB.full	Level 2 (class) variance from Model B.2
L1Var.ModB.full	Level 1 (student) variance from Model B.2

2. Obtain Predicted Values from Full Models

We used SAS Proc Mixed (version 9.2) to fit models to data. There are two types of predicted values that can be output as options to the model statement in SAS Proc Mixed. One type generates predicted values using only the fixed effect parameters from the model. The second type uses both the fixed effect parameters and the random intercept estimates for each class to calculate the predicted values. We used the second type. These predicted values, denoted as $\tilde{Y}evaluator_{post.ij}$ and $\tilde{Y}state_{post.ij}$ were calculated as:

$$\tilde{Y}evaluator_{post.ij} = \hat{\beta}_0 + \hat{\alpha}_{0j} + \hat{\beta}_1(Yevaluator_{pre.j}) + \hat{\beta}_2(Trt_j) + \sum_{m=1}^{M} \hat{\beta}_{m+2}(X_{m.ij})$$

and

$$\tilde{Y}state_{post.ij} = \hat{\beta}_0^* + \hat{\alpha}_{0j}^* + \hat{\beta}_1^*(Ystate_{pre.j}) + \hat{\beta}_2^*(Trt_j) + \sum_{m=1}^{M} \hat{\beta}_{m+2}^*(X_{m.ij})$$

where the "hats" above the parameters indicate that they are parameter estimates.

3. Obtain Estimates of Variance Components when Predicted Values are Outcomes

We next obtain the variance components when the predicted values are fit to unconditional models. The models are of the form

$$\tilde{Y}evaluator_{post.ij} = \beta_0^u + \alpha_{0.j}^u + \varepsilon_{ij}^u$$

$$\tilde{Y}state_{post.ij} = \beta_0^{u*} + \alpha_{0.j}^{u*} + \varepsilon_{ij}^{u*}$$

Each of the two models described above produces estimates of Level-2 (class) and Level-1 (student) variance components. We denote these estimates as:

L2Var.PredA.uc	Level 2 (class) variance from predicted values from Model A – unconditional model
L1Var.PredA.uc	Level 1 (student) variance from predicted values from Model A – unconditional model
L2Var.PredB.uc	Level 2 (class) variance from predicted values from Model B – unconditional model
L1Var.PredB.uc	Level 1 (student) variance from predicted values from Model B – unconditional model

4. Random Assignment of Classes to Treatment or Control Conditions

Next, we randomly assigned classes in the simulated data set to treatment or control conditions. In particular, to replicate the design of the three studies, we randomly assign classes *within the blocks* (usually teachers) in which they were randomly assigned in the actual studies.

First, for the actual (observed) treatment classes, we subtracted off the treatment effect from the predicted values for the students assigned to the treatment group. *PredA1* and *PredB2* are the new predicted values where the treatment effect is removed (subtracted out). In particular, we implemented this step as follows:

if Trt=0 then *PredA1* = $\tilde{Y}evaluator_{post.ij}$

if Trt=1 then *PredA1* = $\tilde{Y}evaluator_{post.ij}$ - $\hat{\beta}_2$

where $\hat{\beta}_2$ is the estimated treatment effect from Model A.2.

if Trt=0 then *PredB1* = $\tilde{Y}state_{post.ij}$

if Trt=0 then *PredB1* = $\tilde{Y}state_{post.ij}$ - $\hat{\beta}_2^*$

where $\hat{\beta}_2^*$ is the estimated treatment effect from Model B.2.

Next, within each randomization block, we randomly assigned the classes to treatment or control conditions. After this "simulated random assignment of classes" we named the simulated treatment indicator as "*RanTrt*."

For the classes that were randomly assigned to the simulated treatment group (i.e. where *RanTrt*=1) we added back in the treatment effect to the predicted value, where the new predicted values were named *PredA2* and *PredB2*. In particular, we implemented these steps as follows:

if *RanTrt*=0 then *PredA2* = *PredA1*
if *RanTrt*=1 then *PredA2* = *PredA1*+ $\hat{\beta}_2$

if *RanTrt*=0 then *PredB2* = *PredB1*
if *RanTrt*=1 then *PredB2* = *PredB1*+ $\hat{\beta}_2^*$

5. Generating the Simulated Values

The simulated values were obtained by adding normally distributed random deviates to the predicted values. There are three kinds of random deviates added to each predicted value. For the simulated values corresponding to Model A.2, they are:

- A level-2 (class level) random deviate (named "*L2NoiseA*")
- A level-1 (student level) random deviate that was common to both A and B models (named "*L1CommonNoise*")
- A level-1 (student level) random deviate that was unique model A (named "*L1UniqueNoiseA*")

The simulated value corresponding to model A (named "*SimA*") was obtained as:

SimA = *PredA2* + *L2NoiseA* + *L1CommonNoise* + *L1UniqueNoiseA*

Similarly, the simulated value corresponding to Model B (named "*SimB*") was obtained as:

SimB = *PredB2* + *L2NoiseB* + *L1CommonNoise* + *L1UniqueNoiseB*

Details regarding the generation of the random deviates are provided in the sections that follow.

Generate Level-2 Random Deviates

Often, the level-2 variance of the predicted values was less than the level-2 variance from unconditional models of the observed data. To increase the level-2 variance in the simulated values, we add a level-2 random deviate to the predicted values. In order to accomplish this, we made a class-level data set that had one record per class. For each of the two test scores, we generated a random deviate, using the SAS "normal" function with mean zero and standard deviation equal to the square root of the difference between the level-2 variance in the observed data and the level-2 variance of predicted value.

For the evaluator test we named the deviate "L2NoiseA," and it was generated as a random normal deviate with mean zero and variance equal to the square root of *L2Var.ModA.uc -L2Var.PredA.uc*. In particular, the SAS code used to generate this random deviate is:

L2NoiseA=normal(-1)[40]*sqrt(*L2Var.ModA.uc -L2Var.PredA.uc*);

Similarly, for the state test we named the deviate "L2NoiseB," and the SAS code used to generate this random deviate is:

L2Noise2=normal(-1)*sqrt(*L2Var.ModB.uc -L2Var.PredB.uc*)

In cases were the level-2 variance of the predicted values was greater than the level-2 variance from the unconditional model of the observed data we set the noise to zero.

- (i.e. when *L2Var.PredA.uc* > *L2Var.ModA.uc*) we set L2Noise A to zero.
- (i.e. when *L2Var.PredB.uc* > *L2Var.ModB.uc*) we set L2NoiseB to zero.

After one deviate for each class and each of the two test scores was created, the class-level file was merged back to the student-level data. Thus, each of the students within a class had the same value on each of these randomly generated level-2 deviates.

Generate Level-1 Random Deviates

Finally, we needed to generate level-1 random deviates in creating simulated values for each student's score on both the two post-tests, the evaluator test (Model A) and the state test (Model B). Our goal was to ensure that when the level-1 and level-2 random deviates were added to the predicted values of the model, the following conditions held for the simulated outcomes for Models A and B, which we call *SimA* and *SimB*, respectively*:*

[40] The argument "-1" to the SAS "normal" function instructs the program to generate a random seed as the starting value for the function.

1. The correlation between the simulated outcomes for the two tests (*SimA* and *SimB*) equals the correlation between the actual outcomes for the two tests ($Yevaluator_{post.ij}$ and $Ystate_{post.ij}$).
2. The mean and standard deviation of the simulated evaluator post-test scores (*SimA*) equal the mean and standard deviation of the actual evaluator post-test scores ($Yevaluator_{post.ij}$).
3. The mean and standard deviation of the simulated state post-test scores (*SimB*) equal the mean and standard deviation of the actual state post-test scores ($Ystate_{post.ij}$).
4. The unconditional and conditional level-1 variances of the simulated evaluator post-test scores should equal the unconditional and conditional level 1 variances of the actual evaluator post-test scores.
5. The unconditional and conditional level-1 variances of the simulated state post-test scores should equal the unconditional and conditional level-1 variances of the actual state post-test scores.

To meet the first condition, we assume a particular model structure with a common component of the level-1 random deviate that is common to both tests, as suggested in the previous section. For a more general discussion of this approach, see Appendix G. For the specifications we used based on this approach, without all of the algebra, see below.

To implement this approach, and satisfy the first condition listed above, we needed to estimate the variance of the common level-1 error component (see Appendix G for the algebra justifying this formula):

$$\text{var}(L1CommonNoise) = corr(Yevaluator_{post.ij}, Ystate_{post.ij}) * (\sigma_{SimA}\sigma_{SimB}) - \text{cov}(\tilde{Y}evaluator_{post.ij}, \tilde{Y}state_{post.ij})$$

The second and third conditions require that $\sigma_{SimA} = \sigma_{Yevaluator}$ and $\sigma_{SimB} = \sigma_{Ystate}$.

To estimate var(*L1CommonNoise*), we use the following equation:

$$\text{var}(L1CommonNoise) = corr(Yevaluator_{post.ij}, Ystate_{post.ij}) * (\sigma_{Yevaluataor}\sigma_{Ystate}) - \text{cov}(\tilde{Y}evaluator_{post.ij}, \tilde{Y}state_{post.ij})$$

To satisfy the fourth condition above, we needed to generate the unique random deviate for the simulated evaluator score such that the variance of the unique random deviate would be the difference between the level-1 variance of the observed data (*L1Var.ModA.uc*) and the level-1 variance of the predicted values (*L1Var.PredA.uc*) minus the variance of the common noise deviate (var(*L1CommonNoise*)). The SAS code we used for the evaluator test is given below:

L1UniqueNoiseA=normal(-1)*sqrt((*L1Var.ModA.uc* - *L1Var.PredA.uc*) - var(*L1CommonNoise*));

The SAS code we used for the state test is given below:

L1UniqueNoiseB=normal(-1)*sqrt((*L1Var.ModB.uc* - *L1Var.PredB.uc*) - var(*L1CommonNoise*));

In cases where sqrt((*L1Var.ModA.uc* - *L1Var.PredA.uc*) is less than

var(*L1CommonNoise*)), we set *L1UniqueNoiseA* to zero.

Likewise, in cases where sqrt((*L1Var.ModB.uc* - *L1Var.PredB.uc*) is less than var(*L1CommonNoise*)), we set *L1UniqueNoiseB* to zero

6. Create Simulated Outcome Scores

After generating the random deviates as described above, we created the simulated values of the two post-test scores as follows:

SimA = PredA2 + L2NoiseA + L1CommonNoise + L1UniqueNoiseA

and

SimB = PredB2 + L2NoiseB + L1CommonNoise + L1UniqueNoiseB

For each model, we estimate the following impact models (as we did for the actual data):

$$SimA = \beta_0 + \alpha_{0.j} + \beta_1 (Yevaluator_{pre.ij}) + \beta_2 (Trt_j) + \sum_{m=1}^{M} \beta_{m+2} (X_{m.ij}) + \varepsilon_{ij}$$

$$SimB = \beta_0^* + \alpha_{0.j}^* + \beta_1^* (Ystate_{pre.ij}) + \beta_2^* (Trt_j) + \sum_{m=1}^{M} \beta_{m+2}^* (X_{m.ij}) + \varepsilon_{ij}^*$$

Using the results from the 1,000 replicates, we verify that all of the required characteristics of the simulated values described in Section 1 are satisfied. Then, using the 1,000 estimates of $\hat{\beta}_2$ and 1,000 estimates of $\hat{\beta}_2^*$, we calculate the covariance between the estimates over the 1,000 replications.

7. Simulations for Other Pairs of Models

The expository example used for the previous sections was focused on the problem of obtaining the covariance of the impact estimates between Models A and B. In this section we describe in more general terms how we adapted this approach to bootstrapping for the other model comparisons.

Comparison of Model A to Model C (Question 2). These models specify the same post-test outcome (evaluator or study-administered test) but a different pre-test. In generating simulated values of the outcome variable, we relied on Model E, which specified the same post-test outcome as Models A and C, but included both pre-test variables as covariates. (For a description of the models, see Table 4 in the text of the report.)

Comparison of Model B to Model D (Question 2). These models specify the same post-test outcome (state test) but a different pre-test. In generating simulated values of the outcome variable, we relied on Model F, which specified the same post-test outcome as Models A and C, but included both pre-test variables as covariates.

Comparison of Model A to Model E (Question 3). These models specify the same post-test outcome (evaluator or study-administered test). However, while Model A includes a matched pre-test (evaluator or study-administered test), Model E includes both pre-tests. In generating simulated values of the outcome variable, we relied on Model E because Model A could be thought of as a restricted version of Model E.

Comparison of Model B to Model F (Question 3). These models specify the same post-test outcome (state test). However, while Model A includes a matched pre-test (state test), Model E includes both pre-tests. In generating simulated values of the outcome variable, we relied on Model F because Model B could be thought of as a restricted version of Model F.

Comparison of Model E to Model G (Question 4). These models specify the same pre-test variables (both tests) but a different post-test variable (evaluator or study-administered test for Model E, the simple average between the two post-test scores for Model G). In generating simulated values of the outcome variables, we relied on Models E and F to generate simulated scores for each of the two tests, and then we averaged the two scores to create the simple average composite outcome in each of the bootstrap samples.

Comparison of Model B to Model F (Question 4). These models specify the same pre-test variables (both tests) but a different post-test variable (state test for Model F, the simple average between the two post-test scores for Model G). In generating simulated values of the outcome variables, we relied on Models E and F to generate simulated scores for each of the two tests, and then we averaged the two scores to create the simple average composite outcome in each of the bootstrap samples (as for the comparison between Models E and G).

8. Bootstrap Estimates

The bootstrap estimates of variances and covariances are presented in Appendix H.

Appendix G: Conceptual Approach to Generating Correlated Residuals for the Parametric Bootstrap

This appendix describes our approach to generating correlated student-level residuals as part of the bootstrapping procedures described in Appendix F. While Appendix F is designed to help other researchers implement the bootstrapping approach we used, this appendix is designed to help technical readers understand our conceptual approach to generating correlated error terms. As a result, the notation used in this appendix is more general than the notation used in Appendix F.

In general, suppose we have one linear model for each of two correlated outcome variables (y_1 and y_2), as shown below:

(1) $y_1 = X_1'B_1 + u_1 + \varepsilon_1$

(2) $y_2 = X_2'B_2 + u_2 + \varepsilon_2$

where X_1 is a vector of independent variables that influence y_1, X_2 is a vector of independent variables that influence y_2, u_1 and u_2 are independent classroom-level errors with a mean of zero and variance of σ_u^2, and ε_1 and ε_2 are student-level errors with a mean of zero, variances of σ_1^2 and σ_2^2, respectively, and a covariance of σ_{12}^2.

By estimating equations (1) and (2), we can estimate the fixed effects (B_1 and B_2) and the three variances (σ_u^2, σ_1^2, and σ_2^2). The challenge is obtaining an estimate of the covariance between the two student-level errors (σ_{12}^2). Overcoming this challenge is critical to randomly generating values of the outcome variables (i.e., the study-administered post-test score and the state post-test score) with the right correlation.

To show how we can estimate the covariance between the two student-level errors, we begin by showing that the covariance between the two outcome variables can be expressed as in equation (3) below:

(3)
$$\text{cov}(y_1, y_2) = \text{cov}(X_1'B_1 + u_1 + \varepsilon_1, X_2'B_2 + u_2 + \varepsilon_2) = \text{cov}(X_1'B_1, X_2'B_2) + \text{cov}(u_1, u_2) + \text{cov}(\varepsilon_1, \varepsilon_2)$$

However, since the classroom-level errors are independent, and the covariance of the student-level errors is σ_{12}^2, the covariance between the two outcomes can be expressed as in equation (4):

(4) $$\text{cov}(y_1, y_2) = \text{cov}(X_1'B_1 + u_1 + \varepsilon_1, X_2'B_2 + u_2 + \varepsilon_2) = \text{cov}(X_1'B_1, X_2'B_2) + \sigma_{12}^2$$

Therefore, we can rearrange terms and express the covariance between the two student-level errors the difference between the covariance between the two outcomes and the covariance between the expected values of those outcomes:

(5) $\sigma_{12}^2 = \text{cov}(y_1, y_2) - \text{cov}(X_1'B_1, X_2'B_2)$

From the data, we can obtain estimates of $\text{cov}(y_1, y_2)$ and $\text{cov}(X_1'B_1, X_2'B_2)$: the latter can be estimated by taking the covariance between the predicted values from the regression model. In this way, we can estimate the covariance between the two student-level error terms.

To generate correlated student-level residuals for the two test score outcomes in this study, we assume that the student-level errors can be expressed as the sum of a common component of achievement (ε_c) and test-specific components of achievement (ε_1^* and ε_2^*), where the common and test-specific error components are assumed to be independent:

(6) $\varepsilon_1 = \varepsilon_c + \varepsilon_1^*$

(7) $\varepsilon_2 = \varepsilon_c + \varepsilon_2^*$

Under this model, the two student-level errors are correlated through the common component of achievement (ε_c), which we assume to have a mean of zero and a variance of σ_c^2. Under these assumptions, the covariance between the two error terms (σ_{12}^2) equals the variance the common component (σ_c^2).

Appendix H: Results from Bootstrapping and Hypothesis Testing

In this appendix, we provide supplementary tables of estimates for both the full sample and the common sample. As described in the text, the common sample excludes all of the students with missing values for one or more of the following four test scores: (1) study-administered post-test, (2) study-administered pre-test, (3) state post-test, and (4) state pre-test.

This appendix includes bootstrap estimates of the variances and covariances required to conduct the hypothesis tests described in Section C and reported in Section D. In addition, the appendix provides more detail on the hypothesis test results for differences between models than provided in Section D.

Note that given time and resource constraints, bootstrapping was conducted *only* for the common sample. However, we used the results to conduct hypothesis tests for both samples. To use bootstrap estimates from the common sample in conducting tests for the full sample, we had to make some assumptions. For the impact estimates, we assumed that the correlation between the estimates would be the same for the full sample as for the common sample. For the standard errors, we assumed that the variance of the difference between the two standard error estimates was the same for the full sample as for the common sample. While these assumptions may not hold, since the common sample served as the basis for the confirmatory hypothesis tests, violations of these assumptions would not affect the study's conclusions.

Finally, we note that the *p*-values presented Tables H.2 – H.5 have been corrected for multiple comparisons—that is, we multiplied the unadjusted *p*-value by the number of comparisons. For more details on this approach, see the earlier discussion in Section C.3, as well as the table notes to Tables 7 – 10. In some instances, this approach leads to an adjusted *p*-value that is greater than one. In these instances, the *p*-value is displayed in the exhibits below as 1.000.

Table H.1: Variance and Covariance Estimates from the Bootstrapping Procedure (Common Sample)

Question	Comparison	State	Bootstrap Standard Error 1	Bootstrap Standard Error 2	Correlation Between Impact Estimates	Correlation Between SE Estimates
Does the choice of test matter?	Model A vs. Model B	Arizona	0.169	0.184	0.477	0.184
		California	0.077	0.093	0.819	0.672
		Missouri	0.096	0.098	0.681	0.474
Does the mismatched pre-test matter?	Model A vs. Model C	Arizona	0.181	0.206	0.789	0.681
		California	0.073	0.177	0.404	0.237
		Missouri	0.093	0.122	0.839	0.742
	Model B vs. Model D	Arizona	0.186	0.201	0.876	0.709
		California	0.093	0.143	0.496	0.293
		Missouri	NA	NA	NA	NA
Does a second pre-test help?	Model A vs. Model E	Arizona	0.172	0.173	0.996	0.993
		California	0.077	0.077	0.995	0.985
		Missouri	0.097	0.097	0.999	0.996
	Model B vs. Model F	Arizona	0.188	0.185	0.963	0.922
		California	0.119	0.085	0.699	0.439
		Missouri	0.103	0.088	0.902	0.829
Does averaging two post-tests help?	Model E vs. Model G	Arizona	0.164	0.134	0.753	0.597
		California	0.074	0.065	0.828	0.603
		Missouri	0.086	0.072	0.795	0.675
	Model F vs. Model G	Arizona	0.179	0.134	0.804	0.609
		California	0.081	0.065	0.857	0.657
		Missouri	0.087	0.072	0.836	0.689

Table H.2: Testing For Differences in Impacts Between Models (Full Sample)

Question	Comparison	State	Impact Estimate 1	Impact Estimate 2	Difference in Impact Estimates	P-Value	Minimum Detectable Difference
Does the choice of test matter?	Model A vs. Model B	Arizona	-0.041	-0.094	0.053	0.808	0.512
		California	0.070	-0.089	0.159	0.017	0.153
		Missouri	-0.019	0.060	-0.079	0.243	0.157
		Pooled	0.005	-0.012	0.043	0.348	0.129
Does the mismatched pre-test matter?	Model A vs. Model C	Arizona	-0.041	0.109	-0.150	0.221	0.286
		California	0.070	-0.053	0.123	0.267	0.257
		Missouri	-0.019	-0.047	0.028	0.673	0.154
		Pooled	0.005	-0.034	0.016	1.000	0.144
	Model B vs. Model D	Arizona	-0.094	-0.188	0.094	0.462	0.299
		California	-0.089	0.083	-0.172	0.227	0.330
		Missouri	NA	NA	NA	NA	NA
		Pooled	-0.090	0.008	-0.024	1.000	0.266
Does a second pre-test help?	Model A vs. Model E	Arizona	-0.041	0.003	-0.044	0.005	0.035
		California	0.070	0.041	0.029	0.609	0.131
		Missouri	-0.019	-0.078	0.059	0.000	0.011
		Pooled	0.005	0.004	0.049	0.000	0.013
	Model B vs. Model F	Arizona	-0.094	-0.142	0.048	0.484	0.160
		California	-0.089	0.009	-0.098	0.199	0.177
		Missouri	0.060	0.025	0.035	0.330	0.083
		Pooled	-0.012	0.011	0.018	1.000	0.082
Does averaging two post-tests help?	Model E vs. Model G	Arizona	0.003	-0.042	0.045	0.728	0.303
		California	0.041	0.024	0.017	0.674	0.094
		Missouri	-0.078	-0.042	-0.036	0.501	0.124
		Pooled	0.004	-0.011	0.000	1.000	0.088
	Model F vs. Model G	Arizona	-0.142	-0.042	-0.100	0.511	0.356
		California	0.009	0.024	-0.015	0.784	0.127
		Missouri	0.025	-0.042	0.067	0.108	0.096
		Pooled	0.011	-0.011	0.031	0.686	0.091

Table H.3: Testing for Differences in Impact Between Models (Common Sample)

Question	Comparison	State	Impact Estimate 1	Impact Estimate 2	Difference in Impact Estimates	P-Value	Minimum Detectable Difference
Does the choice of test matter?	Model A vs. Model B	Arizona	-0.126	0.001	-0.128	0.537	0.587
		California	-0.053	-0.154	0.101	0.103	0.173
		Missouri	-0.028	0.004	-0.032	0.719	0.248
		Pooled	-0.050	-0.071	0.047	0.340	0.138
Does the mismatched pre-test matter?	Model A vs. Model C	Arizona	-0.126	0.092	-0.218	0.133	0.408
		California	-0.053	-0.093	0.040	0.776	0.396
		Missouri	-0.028	-0.048	0.021	0.768	0.198
		Pooled	-0.050	-0.042	-0.015	1.000	0.162
	Model B vs. Model D	Arizona	0.001	-0.187	0.188	0.096	0.317
		California	-0.154	-0.144	-0.009	0.933	0.310
		Missouri	NA	NA	NA	NA	NA
		Pooled	-0.033	-0.153	0.089	0.182	0.147
Does a second pre-test help?	Model A vs. Model E	Arizona	-0.126	-0.055	-0.072	0.000	0.051
		California	-0.053	-0.038	-0.015	0.085	0.024
		Missouri	-0.028	-0.061	0.033	0.000	0.020
		Pooled	-0.050	-0.047	0.007	0.387	0.015
	Model B vs. Model F	Arizona	0.001	-0.059	0.061	0.299	0.165
		California	-0.154	-0.122	-0.032	0.680	0.214
		Missouri	0.004	-0.009	0.013	0.789	0.141
		Pooled	-0.071	-0.076	0.021	0.726	0.095
Does averaging two post-tests help?	Model E vs. Model G	Arizona	-0.055	-0.050	-0.005	0.974	0.395
		California	-0.038	-0.080	0.042	0.304	0.115
		Missouri	-0.061	-0.035	-0.026	0.664	0.169
		Pooled	-0.047	-0.063	0.019	1.000	0.092
	Model F vs. Model G	Arizona	-0.059	-0.050	-0.009	0.943	0.365
		California	-0.122	-0.080	-0.042	0.321	0.119
		Missouri	-0.009	-0.035	0.026	0.645	0.156
		Pooled	-0.071	-0.063	-0.017	1.000	0.092

Table H.4: Testing for Differences in Standard Errors Between Models (Full Sample)

Question	Comparison	State	Standard Error of Estimate 1	Standard Error of Estimate 2	Difference in Standard Error Estimates	P-Value	Minimum Detectable Difference
Does the choice of test matter?	Model A vs. Model B	Arizona	0.171	0.238	-0.067	0.091	0.111
		California	0.115	0.093	0.022	0.147	0.043
		Missouri	0.086	0.083	0.003	0.848	0.044
		Pooled	0.106	0.109	0.007	0.494	0.029
Does the mismatched pre-test matter?	Model A vs. Model C	Arizona	0.171	0.195	-0.024	0.361	0.074
		California	0.115	0.081	0.034	0.399	0.113
		Missouri	0.086	0.119	-0.033	0.008	0.035
		Pooled	0.106	0.128	-0.027	0.026	0.030
	Model B vs. Model D	Arizona	0.238	0.263	-0.025	0.306	0.068
		California	0.093	0.163	-0.070	0.032	0.091
		Missouri	NA	NA	NA	NA	NA
		Pooled	0.118	0.211	-0.041	0.071	0.055
Does a second pre-test help?	Model A vs. Model E	Arizona	0.171	0.174	-0.003	0.431	0.011
		California	0.115	0.059	0.056	0.000	0.007
		Missouri	0.086	0.088	-0.002	0.096	0.003
		Pooled	0.106	0.082	0.007	0.000	0.003
	Model B vs. Model F	Arizona	0.238	0.253	-0.015	0.242	0.036
		California	0.093	0.102	-0.009	0.736	0.075
		Missouri	0.083	0.072	0.011	0.202	0.024
		Pooled	0.109	0.100	0.002	0.001	0.019
Does averaging two post-tests help?	Model E vs. Model G	Arizona	0.174	0.190	-0.016	0.555	0.076
		California	0.059	0.072	-0.013	0.257	0.032
		Missouri	0.088	0.073	0.015	0.152	0.029
		Pooled	0.082	0.083	0.001	1.000	0.021
	Model F vs. Model G	Arizona	0.253	0.190	0.063	0.011	0.070
		California	0.102	0.072	0.030	0.012	0.033
		Missouri	0.072	0.073	-0.001	0.920	0.028
		Pooled	0.100	0.083	0.016	0.052	0.020

Table H.5: Testing for Differences in Standard Errors Between Models (Common Sample)

Question	Comparison	State	Standard Error of Estimate 1	Standard Error of Estimate 2	Difference in Standard Error Estimates	P-Value	Minimum Detectable Difference
Does the choice of test matter?	Model A vs. Model B	Arizona	0.202	0.200	0.002	0.956	0.111
		California	0.077	0.106	-0.029	0.054	0.043
		Missouri	0.103	0.116	-0.013	0.406	0.044
		Pooled	0.101	0.126	-0.020	0.060	0.029
Does the mismatched pre-test matter?	Model A vs. Model C	Arizona	0.202	0.230	-0.028	0.282	0.074
		California	0.077	0.153	-0.076	0.059	0.113
		Missouri	0.103	0.129	-0.026	0.033	0.035
		Pooled	0.101	0.152	-0.030	0.010	0.030
	Model B vs. Model D	Arizona	0.200	0.230	-0.031	0.210	0.068
		California	0.106	0.113	-0.007	0.830	0.091
		Missouri	NA	NA	NA	NA	NA
		Pooled	0.139	0.170	-0.022	0.515	0.055
Does a second pre-test help?	Model A vs. Model E	Arizona	0.202	0.206	-0.005	0.228	0.011
		California	0.077	0.073	0.004	0.135	0.007
		Missouri	0.103	0.098	0.005	0.000	0.003
		Pooled	0.101	0.096	0.004	0.000	0.003
	Model B vs. Model F	Arizona	0.200	0.214	-0.015	0.251	0.036
		California	0.106	0.082	0.024	0.362	0.075
		Missouri	0.116	0.101	0.015	0.080	0.024
		Pooled	0.126	0.105	0.007	0.614	0.019
Does averaging two post-tests help?	Model E vs. Model G	Arizona	0.206	0.182	0.025	0.360	0.076
		California	0.073	0.064	0.009	0.436	0.032
		Missouri	0.098	0.086	0.012	0.265	0.029
		Pooled	0.096	0.085	0.012	0.243	0.021
	Model F vs. Model G	Arizona	0.214	0.182	0.033	0.187	0.070
		California	0.082	0.064	0.018	0.132	0.033
		Missouri	0.101	0.086	0.014	0.146	0.028
		Pooled	0.105	0.085	0.017	0.035	0.020

Appendix I: Differences in Sample Size Requirements

In this project, we are interested in estimating the consequences of choosing different pre-tests or post-tests on the precision of the impact estimates and ultimately on the sample size requirements of the study. This appendix presents a simple approach to comparing the sample size requirements from two impact models based on different tests. This approach relies on standard variance formulas, with some algebra, to show how much larger the sample size needs to be for a model that produces larger standard errors than for a model that produces smaller standard errors.

The remainder of this appendix includes three sections. The first presents standard formulas for minimum detectable effect size (MDES) given the design of the three RCTs that we chose to reanalyze. The second shows how we can calculate the sample size implications of choosing different models, based on different tests, to estimate impacts. The third provides additional details on the calculations we conducted for this project to generate the estimates in Table 10 presented earlier in the report.

Formula for the Minimum Detectable Effect

The minimum detectable effect size (MDES) of a design in effect size units can be calculated using:

$$(1) \qquad MDE = [T^{-1}(\frac{\alpha}{2}) + T^{-1}(\beta)] \times \sqrt{Var(impact)} / \sigma$$

where α is the significance level, β is the statistical power, $T^{-1}(.)$ is the inverse of the student's t distribution function with df degrees of freedom, $Var(impact)$ is the variance of the impact (or coefficient) estimate of interest, and σ is the standard deviation of the outcome measure (normalized to 1). In the current analyses, power and significance level will be held constant (0.8 and 0.05 respectively). We assume that df equals the number of teachers minus the number of randomization blocks minus one (Schochet 2008b).

The variance of the impact estimate depends on the design of the evaluation. In the three RCTs that we reanalyzed for this project, classrooms were randomly assigned to treatment and control conditions. More specifically, classrooms of students in the same school and grade level were paired, and then one member of each pair was randomly assigned to the treatment group and the other to the control group (hence, essentially blocking on teacher pairs). For this design, variance of the impact estimate can be calculated using:

$$(2) \qquad Var(impact) = \frac{\sigma_b^2 \times (1 - R_{BC}^2)}{c \times P \times (1 - P)} + \frac{\sigma_w^2 \times (1 - R_{WC}^2)}{c \times n \times P \times (1 - P)}$$

where:

σ_b^2: the variance of the outcome that lies between classrooms

σ_w^2: the variance of the outcome that lies within classrooms

R_{BC}^2: the proportion of the between-classroom variance explained by covariates and blocking variables

R_{WC}^2: the proportion of the within-classroom variance explained by covariates and blocking variables

c: the number of classrooms

n: the average number of students per classroom

P: the proportion of classrooms assigned to the treatment

Choosing the Sample Size to Equalize the Minimum Detectable Effect for Two Different Models

In this section, we show that given an actual sample size for the evaluation and estimates of the standard error of the impact estimate for two different models, we can compute the sample size required for the second model to match the MDES produced by the first model with the evaluation's actual sample. Below, we refer to the first model as Model A and the second model as Model B.

We begin by defining notation for this exercise:

c': the actual sample size in the evaluation (number of classrooms)

c'': the sample size required for Model B to achieve the same MDES as Model A with the actual sample (with c' classrooms)

S_A: the true standard error of the impact estimate from estimating Model A in the actual sample, which equals the square root of the variance of the impact estimate

S_B: the true standard error of the impact estimate from estimating Model B in the actual sample, which equals the square root of the variance of the impact estimate

$\hat{S}_A$: the estimated standard error of the impact estimate from estimating Model A in the actual sample

$\hat{S}_B$: the estimated standard error of the impact estimate from estimating Model B in the actual sample

In addition, for our purposes, it is helpful to express the MDES as the product between two functions: (1) f, which is a function of the number of classrooms, and (2) g, which is a function of the vector Z, which captures all factors that affect the MDES other than the number of classrooms:

(3) $$MDES = f(c) \times g(Z),$$

where:

$$f(c) \equiv \frac{1}{\sqrt{c}}\left(T^{-1}(\frac{\alpha}{2}) + T^{-1}(\beta)\right), \text{ and}$$

$$g(Z) \equiv \sqrt{\frac{\sigma_b^2 \times (1-R_{BC}^2)}{P \times (1-P)} + \frac{\sigma_w^2 \times (1-R_{WC}^2)}{n \times P \times (1-P)}} = \sqrt{\frac{\text{var}(impact)}{c}}.$$

To define the object of interest, c'', we simply set the MDES for Model B with sample size c'' to equal the MDES for Model A with sample size c':

(4) $$MDES_A = f(c') \times g(Z_A) = f(c'') \times g(Z_B) = MDES_B,$$

We can rearrange equation (4) to produce equation (5) below:

(5) $$\frac{f(c')}{f(c'')} = \frac{g(Z_B)}{g(Z_A)}$$

Equation (6) indicates that this ratio equals the ratio of the standard errors from the two models in the study sample (S_B/S_A):

(6) $$\frac{g(Z_B)}{g(Z_A)} = \frac{\sqrt{\text{var}(impact_B)/c'}}{\sqrt{\text{var}(impact_A)/c'}} = \frac{S_B}{S_A}$$

The first equality in equation (6) follows from the definition of the function g(Z); the second equality in equation (6) follows from the definition of a standard error, which is the square root of the variance. Combining equations (5) and (6), we can implicitly identify c'':

(7) $$f(c'') = f(c')\frac{S_A}{S_B}$$

If we made the simplifying assumption that the changes in the degrees of freedom has zero effect on the MDES, which is approximately true in large samples (Schochet 2008), equation (7) reduces to equation (8) below:

(8) $$c'' = c' \times \left(\frac{S_B}{S_A}\right)^2$$

Equation (8) expresses c'' *explicitly* as a function of the ratio of the estimated standard errors of the two models in the actual data, where the number of classrooms equals c'.

Because we never know the true standard errors, we have to rely on sample estimates. As a result, the ratio of the two standard errors is measured with sampling error. Fortunately, standard textbooks provide an approximate formula for the variance of the ratio of two random variables (e.g., see

Kendall, Stuart, & and Ord 1998). Equation (9) uses this formula to express the variance of the estimated standard errors in terms of parameters that we can estimate from the data:

$$(9) \qquad \text{var}\left(\frac{\hat{S}_B}{\hat{S}_A}\right) \approx \left(\frac{S_B}{S_A}\right)^2 \left[\frac{\text{var}(S_B)}{(S_B)^2} - 2\frac{\text{cov}(S_A, S_B)}{S_A S_B} + \frac{\text{var}(S_A)}{(S_A)^2}\right]$$

The standard errors in equation (9) were estimated as described earlier in Section C.3. The variances and covariances in equation (9) were estimated via bootstrapping, as described in Appendix F.

Estimating the Sample Size Required Under the Alternative Model

In this exercise, we used the available data to estimate the number of classrooms required under an alternative model (e.g., Model B) to achieve the same Minimum Detectable Effect Size as the primary model (e.g., Model A). To compute a point estimate of the sample required under the alternative model, we used equation (7), inserted our estimates of the standard error of each model, and implemented an iterative procedure to solve for an estimate of the number of classrooms required ($\hat{c}''$).

To compute a 95 percent confidence interval around our point estimate, we took the four steps. First, we used equation (9) to estimate the variance of the ratio of the two standard error estimates ($\hat{S}_B/\hat{S}_A$). Second, we assumed that the distribution of the ratio of the two standard error estimates was approximately normal, and we computed a 95 percent confidence interval around the ratio of the two standard error estimates. Third, we computed the lower bound of the 95 percent confidence interval around $\hat{c}''$ by estimating the sample size required under the assumption that S_B/S_A equals the lower bound of the 95 percent confidence interval around $\hat{S}_B/\hat{S}_A$. Fourth, we computed the upper bound of the 95 percent confidence interval around $\hat{c}''$ by estimating the sample size required under the assumption that S_B/S_A equals the upper bound of the 95 percent confidence interval around $\hat{S}_B/\hat{S}_A$.

The results of our sample size calculations are presented in Section D, Table 10.

Appendix J: Correlations between Scores on State and Study-Administered Tests

For this study, we selected three studies that all used study-administered tests developed by the Northwest Evaluation Association (NWEA). For simplicity, we refer to all of these tests as the Measures of Academic Progress (MAP). An important issue in assessing the generalizability of the study results is to assess whether the MAP is more or less well-aligned with the state test than other study-administered tests. This appendix provides some empirical evidence that users may find useful in reaching their own conclusions. In particular, this appendix provides estimates of the correlation between scores on study-administered tests and scores on state tests. Some of these estimates were computed from the study sample; other estimates were obtained from the literature.

Correlations can be used to help us assess how similar the MAP test is to other tests used by researchers to measure student achievement in the context of educational evaluations. If the MAP test is similar to other study-administered tests, we would expect:

- The correlation between MAP scores and scores from other study tests taken by the same students to be high, and
- The correlation between MAP scores and scores from state assessment tests to be similar to the correlation between scores from other study tests and scores from the same state assessment tests.

However, it is possible that there are substantial differences between the MAP and other commonly used tests. Unlike the Stanford 10 (Harcourt Assessment, Inc. 2004) or the TerraNova 3 (CTB/McGraw-Hill 2008a, 2008b), the MAP is a computer adaptive test. If this or some other feature of the MAP makes it perform differently from the tests that are more commonly used in educational evaluations, then we might expect low correlations between the MAP scores and scores on these other tests. In addition, the MAP is used as a formative assessment to predict students' scores on the state assessment. Therefore, we might be concerned that the MAP was designed to be well-aligned with state assessments, and that the correlation with state test scores would be higher for the MAP than for other commonly used tests that were not designed to align with state assessments.

To allow the reader to assess the seriousness of these concerns, we have computed correlations from the study data and reported correlations obtained from the literature. Note that we have not conducted statistical tests of the differences in correlations, and that comparisons between the correlations are based merely on visual inspection. Therefore, any differences observed may be due to chance. In addition, we are not able to pre-specify a criterion for concluding that the MAP is "similar enough" to other study tests. Therefore, the analysis provided in this appendix should be treated as descriptive and for the reader's information. At the same time, we expect that most readers will interpret the results as we have: that the differences are small and suggest that the MAP performs in a similar manner to other study tests that are more commonly used in educational evaluations.

The correlation between MAP scores and scores from other study tests. While we did not conduct an exhaustive review of the evidence on these correlations, a report published by the NWEA,

the publisher that developed the MAP, shows that MAP scores are highly correlated with scores from at least one more commonly used standardized test (NWEA 2004). This report shows correlations between MAP reading scores and reading scores from the Stanford 9, the predecessor to the Stanford 10, of 0.86 or 0.87 in elementary and middle school grades. These correlations are of the same order of magnitude as the correlations between different forms of the MAP (NWEA 2004). This evidence suggests that the MAP test measures similar content to at least one other standardized test that has been commonly used in education evaluations.

The correlation between MAP scores and scores from state assessment tests. To examine the correlation between MAP scores and scores from state assessments, we computed these correlations in the three studies used in our analysis. Table J.1 presents correlation coefficients between student reading scores from the MAP and student reading scores from state or district tests in each of the three experiments. Correlation coefficients are computed separately by group (treatment and control) and measurement period (pre-test and post-test). In summary, we found a moderate to strong relationship between scores on the MAP standardized test and scores on the state or district tests. For example, the post-test correlations for the treatment group were 0.69 for Arizona, 0.82 for California, and 0.69 for Missouri.

An external study conducted in Pennsylvania found correlations between the MAP and the Pennsylvania state assessments that were higher than the correlations from the Arizona and Missouri studies, but comparable to the correlations from the California study. For the one school district that administered the MAP, the authors reported correlations between MAP reading scores and reading scores from the state test in the 0.84 to 0.86 range (Thacker, Dickinson, & Koger 2004).

The key question is whether these correlations are higher than for other study-administered tests, which would suggest that the MAP is more highly aligned with state assessments than these other tests. Evidence for other commonly used tests can be used to address this question. In 2004, students in California in grades 2 through 11 were required to take both the California state test for English-Language Arts and the Terra Nova CAT/6. Evidence reported in technical documents on the California Department of Education website indicates that the correlations between these scores ranged from 0.75 to 0.80 (Educational Testing Service 2009).[41] These correlations are slightly lower but similar in magnitude to the 0.82 correlation that we computed and reported from the California experiment. This finding is similar to the difference found in the previously referenced study in Pennsylvania: the correlation with state reading scores is slightly higher for the MAP than for the CAT/5, the previous version of the CAT/6 (Thacker et al. 2004).[42]

Given that we have not conducted any formal hypothesis tests, we cannot say with confidence whether the MAP is more closely aligned with state assessments than other commonly used study-administered tests, or whether there is no actual difference and the observed differences are due entirely to random chance. Additional research would be necessary to determine whether the MAP falls with the range of other commonly used terms in terms of its alignment with state assessments.

---

[41] See p. 105 at www.cde.ca.gov/ta/tg/sr/documents/csttechrpt08.pdf.

[42] Tables 82 – 89 of Thacker, Dickinson & Koger (2004) present correlations for the MAP in the 0.83 – 0.87 range and correlations for the CAT-5 in the 0.76 – 0.80 range.

Table J.1: Correlations between MAP Reading Test Scores and State Reading Test Scores

State	Pre-test Scores			Post-test Scores		
	Control Group	Treatment Group	Pooled	Control Group	Treatment Group	Pooled
Arizona	.60 (n=41)	.57 (n=32)	.57 (n=73)	.53 (n=41)	.69 (n=32)	.60 (n=73)
California	.70 (n=127)	.76 (n=151)	.76 (n=278)	.84 (n=127)	.82 (n=151)	.83 (n=278)
Missouri	.52 (n=186)	.55 (n=175)	.55 (n=361)	.73 (n=186)	.69 (n=175)	.71 (n=361)

Notes: This table is identical to Table 3 in the body of the report. The analysis included all students with non-missing values for all four test scores (referred to as "the common sample" in this report). In Missouri, the study relied on pre-test scores from the district test instead of the state test. The *p*-value for each of these correlations is less than .01.

To this point, we have used correlational evidence to assess the likelihood that the study-administered test used in the analysis presented in this report was in some way unusual, which would call into the question the generalizability of our results. Another possible concern is that the set of state tests used in this analysis are in some way unusual. To address this possible concern, we examined the correlations between state test scores and the MAP. If this correlation were either stronger or weaker in Arizona, California, and Missouri than in other states, the results from this study might be misleading for evaluations that draw their samples from a broader range of states. Fortunately, NWEA (2004) provides estimates of the correlation between MAP scores and scores from state assessments in Arizona, Colorado, Illinois, Indiana, Minnesota, Nevada, South Carolina, Texas, Washington, and Wyoming. The correlations reported for Arizona were comparable to the correlations reported for Texas and South Carolina but lower than the correlations reported for most other states.[43] For example, the correlations in grade 5 scores were reported for seven states: Arizona (r=0.69), Colorado (r=0.87), Illinois (r=0.80), Minnesota (r=0.83), Nevada (r=0.83), South Carolina (r=0.70), and Texas (r=0.70). While the correlation is not reported for Missouri or California, estimates from an unpublished dissertation based on data from over 800 students in grades 6 – 8 in one rural school district in Missouri suggest that the correlation between MAP reading scores and reading scores from the state assessment is approximately 0.82, which is roughly comparable to the correlations for many other states (Shields 2008,[44] NWEA 2004). In summary, these estimates show that there is non-trivial variation across states in how highly correlated state reading test scores are with MAP reading scores. However, it is comforting that the three states chosen for our analysis are not clustered at either the high or low end of the distribution across states with regard to the correlation between state reading scores and MAP reading scores.

[43] This could be because the Arizona test includes language arts as part of its reading test, so the scope of the state reading test may be broader in Arizona than in many other states (see Cronin 2004.)

[44] See http://edt.missouri.edu/Spring2008/Dissertation/ShieldsJ-042908-D9955/research.pdf.

Appendix K: Estimates of Key Statistical Power Parameters

In Section C, we described the two-level models that were estimated for the analysis. This appendix provides estimates of the variance components at both levels—the cluster or classroom level and the student level. It also provides estimates of the intra-class correlation from the models used to estimate impacts, which included the covariates described in Section C. Finally, this appendix provides estimates of the variance components and intra-class correlations from the unconditional models, without the covariates, for readers interested in assessing the increase in statistical power from the inclusion of these covariates.

To interpret the magnitude of the variance components, it is important to understand the scale of the dependent variable. Student test scores were normalized so that the control group has a mean of zero and variance of one. However, Table K.1 provides estimates of the pooled variances, including both treatment and control cases. As a result, the total unconditional variance for the pooled sample is not equal to one by construction, but it tends to be close to one. For example, Table K.1 shows that for Model A, the cluster or classroom-level variance equals 0.03 and the student-level variance equals 0.90, which suggests that the total pooled variance equals 0.93.

Estimates of the intra-class correlations are provided to inform the design of future studies. To properly interpret these estimates, it is important to understand the research design of these studies. Each of these studies created matched pairs of classes, and randomized one class per pair to the treatment group and the other class to the control group.[45] If the matching were successful in creating blocks of similar classrooms, we would expect the cluster-level variance and the intra-class correlation in the unconditional models—models that include indicator variables for the blocks but no other covariates—to be low. Note that we would not expect the estimates presented in Tables K.1 to generalize to other research designs.

[45] Indicator variables for the randomization block were included in the model.

Table K.1: Estimates of the Variance Components for the Unconditional and Conditional Regression Models

Model	Cluster- and Student-Level Variance Estimates and Corresponding ICCs					
	Unconditional Models			Conditional Models		
	Cluster Variance	Student Variance	ICC	Cluster Variance	Student Variance	ICC
Arizona Study						
A	0.03	0.90	0.031	0.05	0.41	0.109
B	0.00	0.85	0.000	0.00	0.62	0.000
C	0.03	0.90	0.031	0.06	0.56	0.097
D	0.00	0.85	0.000	0.04	0.65	0.058
E	0.03	0.90	0.031	0.06	0.38	0.136
F	0.00	0.85	0.000	0.02	0.57	0.034
G	0.01	0.67	0.012	0.05	0.29	0.147
California Study						
A	0.10	1.02	0.089	0.00	0.28	0.000
B	0.06	1.04	0.058	0.01	0.42	0.016
C	0.10	1.02	0.089	0.03	0.48	0.067
D	0.06	1.04	0.058	0.01	0.41	0.026
E	0.10	1.02	0.089	0.00	0.26	0.000
F	0.06	1.04	0.058	0.00	0.33	0.000
G	0.08	0.93	0.079	0.00	0.20	0.000
Missouri Study						
A	0.11	0.93	0.103	0.02	0.57	0.038
B	0.07	1.10	0.062	0.04	0.59	0.062
C	0.11	0.93	0.103	0.06	0.56	0.102
D	0.07	1.10	0.062	0.03	0.65	0.039
E	0.11	0.93	0.103	0.02	0.46	0.051
F	0.07	1.10	0.062	0.03	0.49	0.051
G	0.09	0.85	0.093	0.02	0.32	0.066

Note: The conditional regression model includes one indicator variable for each matched pair of classrooms, and it includes all of the covariates described in Section C. The unconditional model excludes the covariates but includes the indicator variables for matched pairs of classrooms.


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA