NCEE 2012-4015

U. S. DEPARTMENT OF EDUCATION

Whether and How to Use State Tests to Measure Student Achievement in a Multi-State Randomized Experiment: An Empirical Assessment Based on Four Recent Evaluations

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

Whether and How to Use State Tests to Measure Student Achievement in a Multi-State Randomized Experiment: An Empirical Assessment Based on Four Recent Evaluations

October 2011

Marie-Andrée Somers
Pei Zhu
Edmond Wong
MDRC

NCEE 2012-4015
U.S. DEPARTMENT OF EDUCATION

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

Disclaimer
The Institute of Education Sciences at the U.S. Department of Education contracted with MDRC to develop a report documenting the analytical implications of using state test scores to measure student achievement in multi-state and multi-grade randomized experiments. 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

Somers, Marie-Andrée, Pei Zhu, and Edmond Wong. "Whether and How to Use State Tests to Measure Student Achievement in a Multi-State Randomized Experiment: An Empirical Assessment Based on Four Recent Evaluations." NCEE Reference Report 2012-4015. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2011.

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

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

Executive Summary

This study examines the practical implications of using state tests to measure student achievement in impact evaluations that span multiple states and grades. In particular, the study examines the sensitivity of impact findings to (1) the type of assessment used to measured achievement (state tests or an external assessment administered by the evaluators) and (2) to analytical decisions about how to pool state test data across states and grades. These questions are examined using data from four recent IES-funded randomized experiments where student achievement was measured using both state tests and a test administered by evaluators for the purposes of the study ("study-administered test"). These studies span multiple states – 8 to 10 states depending on the experiment.

The findings from this analysis are most applicable to *large-scale studies with multiple states*. When few states are involved in a study (say, 2 or 3), states that are "outliers" in terms of the content or quality of their state assessment may have a bigger influence on the overall finding when estimating the program's pooled effect on state test scores.

Are state tests suitable for use in an impact evaluation?

To answer this question, we examine whether state tests meet some of the necessary criteria for use in an impact evaluation. Specifically, we look at whether the content of state tests is aligned with the intervention (which affects the *validity* of causal inferences about program impacts) and we examine whether there are floor effects in state test scores (which affects the reliability of test scores, and by extension, the *precision* of impact estimates).[i] We also examine whether impact findings are sensitive to the type of assessment used to measure student achievement. To do so**,** we compare the pattern of estimated impacts on state tests and impacts on the study-administered test, to see whether impacts findings (with respect to inferences and precision) differ between the two types of assessment, and if so, whether these differences make sense given differences in the content and reliability of the two test types.

The findings from these analyses suggest that state tests are suitable for measuring impacts on general achievement in these four large-scale evaluations, but that they may not be suitable for measuring impacts on some of the more specific achievement outcomes targeted by the programs. The key findings are the following:

- *When the primary outcome is general achievement:* Two of the experiments reanalyzed in this paper are evaluations of an intervention that aims to improve general achievement in math or reading. For these two studies, our findings suggest that the broad content of state tests makes them suitable for evaluating the effect of the intervention on a policy-relevant measure of general achievement. We find that the magnitude of program impacts – as well as

[1] A *floor effect* occurs when many students *incorrectly* answer every (or most) test items. Because most low-performing students have the same score (i.e., zero), this makes it difficult to differentiate between the achievement levels of these students. This in turn reduces the reliability of the test at the lower end of the achievement distribution (May et. al, 2009).

inferences about program effectiveness (as measured by the p-value) – do not differ by a statistically significant amount across the two test types. Both types of test are intended to measure "general achievement", so this is the pattern of findings that one would expect to see if state tests were indeed aligned with the outcome of interest. For both of these studies, the standard error of the estimated impact on state test scores is larger than the standard error for the estimated effect on the study-administered test; moreover, in one of the studies, we find evidence of floor effects in state test scores. This suggests that the reliability of state tests in these two studies is less than that of the study-administered test. However, the resulting difference in the precision of impact estimates across assessment type is not sufficient to lead to differences in inferences about program impacts, as measured by the p-value, which does not differ across the two types of test. This suggests that the reliability of states tests in these studies is not so low as to make them unsuitable for use in an impact evaluation.

- *When the primary outcome is a specific skill:* The other two experiments reanalyzed in this paper are evaluations of interventions that target a more specific reading or math skill. In these two studies, it is not possible to use state tests as a measure of the targeted outcome, because subtest scores for the specific skill are not consistently available for all states. However, our findings suggest that state tests are suitable for measuring these programs' impact on general achievement, which is the longer-term goal of both interventions. In one of the two studies, the program has a positive impact on students' scores on the study-administered test (which is used to measure the specific skill targeted by the program) as well as on state test scores (which measure general achievement). In the other experiment, we find that the program does not improve students' performance on either the study-administered test (the targeted skill) or state tests (general achievement). This pattern of findings is consistent with what one would expect to see, given the fact that state tests measure a less proximal outcome than the study-administered test. For these two studies, we also find that the precision of the estimated impact on test scores is similar across the two types of test; furthermore, there is no evidence of floor effects in the state test scores. This suggests that in these studies the two tests are comparable with respect to the reliability of the outcome that they are intended to measure.

Overall, these findings indicate that in the four re-analyzed experiments, state tests can be a useful *complement* to a study-administered test: they provide a policy-relevant measure of achievement and they can be used to measure impacts on longer-term achievement outcomes. In these experiments, state tests could not have been used as a *substitute* for the study-administered test, however. In three of the experiments, state tests were not available for all grades or states; in the fourth study, the targeted outcome is a specific skill that cannot be measured using state tests. Therefore, using state tests as a substitute would not have been possible.

Should impact findings be pooled across states and grades?

Typically in IES studies, conclusions about program effectiveness are based on whether or not an intervention has a statistically significant impact on average across all students in the study sample (that is, on average across all states and grades in the study). Combining the results across states

and/or grades improves the generalizability of the findings to a broader range of contexts. It can also allow to the study to achieve sufficient power to detect meaningful impacts on the outcome of interest (as is true for the four experiments in this paper).

An important challenge when combining impacts on state tests, however, is that their content differs across states. As a result of these differences, program impacts may be larger in states or grades where the assessment is better aligned with the outcome targeted by the intervention. In this situation, the "average" impact of the program would mask meaningful variation in program effectiveness, which could complicate the interpretation of the aggregated finding. However, for the four experiments in this paper:

- Findings from the analysis suggest that it is acceptable to combine impact findings. Although the content of state tests does differ, these differences are not sufficiently large to lead to variation in estimated program effects across states or grades in the four experiments.

Is the pooled impact on state test scores sensitive to decisions about how to combine sores across states?

The most appropriate strategy for combining impact findings across states and grades is to use a "meta-analytic" approach.[ii] As indicated by its name, this approach consists of estimating the impact of the program for each state and grade, and then taking a weighted average of these estimates to obtain the average impact of the program.

A key question here is whether the average impact on state test scores is sensitive to choices about how to rescale state test scores to a common metric and how to weight each state- and grade-specific impact in the overall finding. In general, we find that for all four experiments, conclusions about the statistical significance of the impact estimate are robust to rescaling/aggregation decisions. Specific results are discussed below.

Sensitivity to rescaling method

When rescaling state test scores to a common metric, researchers must choose between a *linear* transformation (which produces traditional z-scores) or a *non-linear* transformation (which produces rank-based z-scores). Using a non-linear transformation makes the distribution of test scores normal, so the decision between linear and nonlinear rescaling is likely to matter most when state test data are not normally distributed.[iii] Descriptive analyses show that in the studies reanalyzed in this paper, state test scores are not consistently normally distributed in all study states.[iv] However, the sample size in this study may be sufficiently large that violations of non-normality do not matter. Indeed, we

[2] There is also another approach for combing impact findings across state, but it requires that state test scores be interchangeable across states and grades. This condition is not met for any of the four experiments used in this paper.

[3] If test scores are normally distributed, the two methods are equivalent.

[4] When data are non-normal, the impact estimate and its standard error are unbiased, but the distribution of the estimated impact may be non-normal. This means that p-values based on T and F statistics may be inaccurate. This is less of an issue when sample sizes are large.

find that conclusions about the statistical significance of the impact estimate are robust to the choice of functional form used for rescaling:

- *Linear rescaling (traditional z-scores) vs. non-linear rescaling (rank-based z-scores):* In the four experiments re-analyzed in this paper, impact findings are not sensitive to the choice between linear and non-linear rescaling methods. This suggests that in studies with large sample sizes, impacts on state test scores are less sensitive to violations of non-normality, and that simple linear rescaling (traditional z-scores) is an acceptable approach for converting test scores to a common metric. However, if state test scores appear to be non-normally distributed – and if the sample size is small – then researchers may prefer the rank-based method. Regardless of which method is chosen, researchers may want to also try the other approach as a sensitivity test.

If linear rescaling is chosen, researchers must also decide what to use as a "reference population", that is, whether to rescale students' test scores relative to that of other students in the *sample* who took the same test, or relative to *all students in the state* who took the test. The distribution of test scores for the sample is more homogeneous than the distribution for the state as a whole. Therefore, the magnitude of the average impact estimate should be larger when "sample-based" rescaling is used. However, because the standard error is also larger, the p-value should not differ substantially between the two reference populations. Findings from the four experiments confirm this result:

- *Choice of reference population for linear rescaling:* Inferences about program effectiveness (based on the p-value) are not sensitive to the choice between using the sample or the state distribution to create z-scores. This suggests that researchers should use the reference population that provides the desired interpretation of the impact estimate. Because the standard error for the estimated impact does differ (by a statistically significant amount) depending on which reference population is chosen, researchers should make this decision during the study design phase, so that the minimum sample size for detecting a meaningful effect can be correctly determined.

Sensitivity to weighting method

The next step when combining impact findings across states is to decide how to weight the impact estimate for a given state and grade when pooling the results. This decision depends on the type of inference that researchers want to make. If the goal is to estimate the average impact of the program for the *states in the study sample*, then precision weights should be used (also called "fixed effects" weights). Conversely, if the goal is to estimate the average impact for some *larger population of states*, then "random effects" weights should be used.

To use random-effects weighting, however, two further conditions must be satisfied. First, states in the study must be a representative sample of some larger identifiable population of states. Second, the estimated variation in impacts across states should be statistically significant, which is less likely in studies with a small number of states (since statistical power is limited).

- *Precision (fixed-effects) weighting vs. random-effects weighting:* For all four experiments in this analysis, we conclude that impact findings should be combined using precision weights, because the conditions for using random-effects weighting are not met: (i) study sites are not randomly selected and (ii) the estimated variation in impacts across states is not statistically significant (most likely because there are too few states in the evaluation to reliably detect impact variation).

In other studies, however, using random-effects weighting may be appropriate if the relevant conditions are met.

Is the precision of impact findings sensitive to the type of assessment used to measure student achievement at baseline?

The final topic addressed in this paper pertains to the implications for precision of using state tests from prior grades to measure student achievement at baseline (rather than a study-administered pretest). We find that:

- Using state tests to measure baseline achievement improves the precision of impacts estimates, with reductions in the standard error ranging from 8 percent to 45 percent, relative to not controlling for state tests). These precision gains are at least as large as when a study-administered pretest is used to measure baseline achievement.

Overall, these findings suggest that state tests can provide a cost-effective means of improving the precision of the impact findings for a given sample size.

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 a broad range of issues that emerge when state tests are used in impact evaluations that span multiple states and multiple grades. In particular, the study examines the sensitivity of impact findings to (1) the type of assessment used to measure achievement (state tests or study-administered tests) and (2) analytical decisions about how to pool state test data across states and grades. These questions are examined using data from four recent IES-funded randomized experiments where student achievement was measured using both state tests and a test administered by evaluators for the purposes of the study. These studies span multiple states – 8 to 10 depending on the experiment. Based on their analyses, study authors conclude that state tests were suitable for measuring impacts on *general achievement* in these four large-scale evaluations, but were not suitable for measuring impacts on the more specific achievement outcomes targeted by the individual programs. State tests can also serve as a useful *complement* to study-administered tests, as they provide a policy-relevant measure that can be used to measure impacts on longer-term achievement outcomes. Study authors note that their findings are most applicable to *large-scale studies with multiple states*. When fewer states are involved in a study, states that are "outliers" in terms of the content or quality of their assessments may have a bigger influence on findings when estimating the program's pooled effect on state test scores.

Acknowledgements

The authors would like to thank the National Center for Education Evaluation and Regional Assistance (Institute of Education Sciences) for supporting this work. We are also indebted to several individuals whose advice greatly improved the report's usefulness and readability. Fred Doolittle (MDRC), Henry May (University of Pennsylvania) and Irma Perez-Johnson (Mathematica Policy Research) provided guidance on the scope of the report. We also received useful comments from reviewers of our drafts –Howard Bloom and Michael Weiss (MDRC), Irma Perez-Johnson and other staff at Mathematica Policy Research Inc., as well as several IES staff members. We also extend thanks to Asya Magazinnik and Ezra Fishman for their assistance with report production. Finally, we are grateful to the program developers, districts, schools, and teachers who participated in the four multi-state experiments that were used in our analyses. This paper would not have been possible without the time and energy that these individuals contributed to their respective studies and to the field of educational evaluation.

Disclosure of Potential Conflicts of Interest

There are three authors for this report – Marie-Andrée Somers, Pei Zhu and Edmond Wong – all of whom were employees of MDRC during the preparation of this report. The authors and other staff do not have financial interests that could be affected by the content in this report.

Contents

List of Tables and Figures

Figures

1 Introduction

The past six years have seen a large and growing number of large-scale randomized field trials of interventions whose goal is to identify promising methods of improving students' academic achievement. This growth is based in large part on sponsorship from the Institute of Education Sciences (IES); Spybrook (2008), for example, identifies 55 randomized studies that evaluate a broad range of educational programs and interventions.

An important question for the educational evaluators of these studies has been how best to measure students' academic performance, which is the outcome of greatest interest in these studies. Several competing factors must be considered, including the assessments' alignment with the academic outcomes of interest, their suitability for the study participants, and the cost of data collection. In large-scale studies that span multiple states, the trade-off between these factors is especially evident: the geographical breadth of the samples – while advantageous in terms of statistical power and external validity – complicates the task of finding a consistent yet cost-effective measure of student achievement.

Thus far, the most commonly used approach for measuring student achievement in large-scale evaluations has been for evaluators to administer a common standardized test to students in the study (a "study-administered" test). This approach allows the same test to be used across the full study sample, thereby yielding consistent measures of student achievement across the study sites (and therefore states). It also allows evaluators to choose a test that is closely aligned with the specific outcomes that are targeted by the intervention and that is suitable to the population of students participating in the study (typically low-performing students in IES studies). On the other hand, administering a standardized test to students in the sample has several drawbacks. The cost of administering these tests can be high – sometimes even prohibitive – and it imposes additional testing burden on students and school staff who must already contend with the testing requirements mandated as part of their state's accountability system. Also, because tests administered by evaluators are not linked to school accountability and student progression, there may be little incentive for students to perform well on the test.

These limitations have led to growing interest in an alternative approach for measuring student achievement – that is, using students' scores on the tests administered by their state. In the past, state assessments were not administered by all states and in all grades, which precluded the use of state test scores for the purposes of large-scale program evaluation. In the era of *No Child Left Behind* (NCLB), however, all states now administer yearly assessments in reading and mathematics to students in grades 3 to 8 and in one grade level in high school, thus making state tests an increasingly viable source of information on student achievement, especially at the elementary and middle school level.

On a practical level, there are several advantages to using state test scores to measure student achievement. In particular, making use of already-existing data can yield substantial cost savings for the study, and it can eliminate the burden of additional testing on students, teachers, and school staff.

State test scores are also considered policy relevant, because they are used by districts and states to make decisions about individual students and schools.

However, the practical advantages of using state tests must be weighed against several important concerns related to their use. A first set of issues relates to whether state tests are in fact suitable for evaluating the intervention. State assessments typically cover a broad range of content areas, which may or may not be the focus of the program; thus, the content of state tests may be a poor measure of the outcome targeted by the intervention. State tests may also be a less reliable measure of student achievement, especially for the low-achieving students who are the target population in most IES studies; this in turn can decrease the power of the analysis to detect program effects if they exist.

Assuming that state tests are suitable for the evaluation, researchers must then decide whether to combine impact findings across states and grades. Because state tests differ in terms of their content and scale, they do not provide a standardized measure of student achievement across the entire study sample. This lack of standardization can complicate the interpretation of the "average" impact of the program. For example, program impacts may be larger in states or grades where test content is better aligned with the outcomes targeted by the intervention. The average impact finding would mask these differences in program effectiveness, in which case it may be preferable to present impact findings separately for each state and grade.

If combining the impact findings is deemed appropriate, then a further complication relates to how to combine the results across states and grades. One possible strategy is to use a "meta-analytic" approach.[5] As indicated by its name, this approach consists of estimating the impact of the program for each state and grade, and then taking a weighted average of these estimates to obtain the overall impact of the program. A key question here is whether the average impact of the program on state test scores is sensitive to decisions about (a) how to rescale test scores to a common metric (that is, what type of *linking* method to use to place the scores from different assessments on the same scale) and (b) how to weight each state- and grade-specific impact in the overall finding.

Finally, in addition to deciding whether and how to use state tests to measure student achievement at follow-up, researchers must also decide whether to collect data on state tests scores *at baseline* (prior to random assignment). Even if state tests are not a feasible or suitable choice of outcome measure, it may still be worthwhile to collect data on students' scores on state tests in prior school years. These data can be used as a covariate in the impact model to improve the precision of the impact estimates, thereby reducing the number of students that need to be recruited into the study to achieve a given level of statistical power (and by extension, the cost of the study).

Given the potential and possible pitfalls of using state tests in an evaluation, it is important to take a closer look at the factors that affect whether and how to use these data. A recent discussion paper by Henry May and colleagues (2009) – in the *NCEE Reference Report* series – is an important

[5] There is also another approach for combing impact findings across state, but it requires that rescaled test scores provide an equated measure of achievement across states and grades. Appendix C presents descriptive analyses that indicate that this requirement is not met for the four experiments. We therefore focus on the meta-analytic approach in this paper.

first step in this direction. The paper provides a clear overview of the factors that researchers need to consider when using state tests in their evaluation, and puts forth a set of recommendations and best practices to help guide evaluators' decisions in this area.

Guided by the framework laid out in May *et al.* (2009), the purpose of the present paper is to bring actual data to bear on the factors that affect whether and how to use state tests in a large-scale multi-state (and multi-grade) randomized experiment. These factors are examined by re-analyzing data from four recent IES-funded evaluations where student achievement was measured using both state tests *and* a study-administered test. The four studies re-analyzed in this paper are typical of large-scale IES studies and represent a broad range of grade levels, subject areas, and states.

Based on these four experiments, we examine several issues related to using state tests in an impact evaluation:

- We look at whether *inferences about program impacts* are sensitive to the type of assessment that is used to measure student achievement (state tests or a study-administered test).
- We examine whether the *precision* (standard error) of estimated impacts is sensitive to the type of assessment (state tests or study-administered test).
- We examine whether the "pooled" impact on state test scores is sensitive to different ways of *linking* test scores across different assessments and *aggregating* these rescaled scores across states and/or grades.

The answer to these questions depends in large part on the psychometric properties of state tests. To yield *precise* impact estimates, state tests must reliably measure the achievement of students in the target population (reliability). To provide *valid* inferences about impacts, state tests must yield valid inferences about student achievement (validity). Similarly, for the *pooled* impact to be precise and causally valid, then the reliability and validity of state test scores must hold even after scores from different assessment have been linked and aggregated across states and/or grades.

An in-depth analysis of the psychometric properties of state tests is beyond the scope of this paper, in part because the item-level test information (scores, difficulty, etc.) necessary for such an analysis are not available. However, we do supplement the impact findings in this paper with a descriptive analysis of test characteristics in the four re-analyzed studies. We use published information on state tests to examine whether the content of these tests is aligned with the outcomes targeted by the intervention being evaluated (which is one aspect of test validity) and we look for floor effects in the state test scores (which affects reliability).[6] Although they do cover some aspects of validity and reliability, these analyses are incomplete – we cannot assess the reliability of the tests for the low-achieving students in our samples, nor can we formally examine important aspects of validity such as item depth, and how much overlap there is between state tests at the item level.

[6] A *floor effect* occurs when many students incorrectly answer every (or most) test items because the assessment is too difficult for them. This reduces the reliability of the test at the lower end of the achievement distribution. Floor effects will be discussed in detail later in this paper.

Given these limitations, the descriptive analyses of state tests in this paper are not a formal assessment of validity or reliability. Rather, the analysis is simply used to examine whether the state tests in the four experiments meet some necessary conditions for use in an evaluation; they are also used to provide context for interpreting the impact findings, and to inform our decisions about how to link and aggregate test scores across states.

The unavailability of item-level information also affects the scope of the methods examined in this paper. As noted earlier, the scale of state tests differs across states, so these scores must be converted to a common metric to make pooling possible. If item-level scores were available, then it would be possible to use sophisticated methods such as item response theory (IRT) for this purpose. With only total scores at our disposal, however, only a handful of linking methods are feasible. Thus, this paper focuses on these basic methods and does not delve into more complex approaches.

The limitations of our analysis – such as the fact that item-level information is not available – reflect what researchers are likely to encounter in the field. When deciding whether to use state tests in the planning phase, evaluators typically only have access to published information about tests from state websites, and not item-level information. Similarly, in the analysis phase, item-level scores is unlikely to available for all states in a study, therefore making it unlikely that they will be able to use complex methods (like IRT) to rescale test scores. Thus, the results of this analysis – which are limited by the same practical considerations – are likely to be useful to evaluators.

For all of these reasons, this paper is best characterized as a study of the practical implications of using state tests for evaluation purposes. In general, the findings in this paper are most applicable to large-scale educational studies spanning *multiple states*. When few states are involved in a study (say, 2 or 3), states that are "outliers" in terms of the content or quality of their state assessment will have a bigger influence on the overall finding when estimating the program's pooled effect on state test scores. Therefore, with very few states, impact findings may be more sensitive to using state tests to measure achievement, as opposed to using a study-administered assessment. Similarly, with small sample sizes, the impact analysis state test scores could be more sensitive to decisions about how to link scores and aggregate them across states.

The remainder of this section discusses the research questions that guide the analyses in this paper; this is followed by a brief overview of the content of the paper.

Research Questions

In this paper, we examine several questions related to whether state tests are suitable for use in an impact evaluation, and the implications of using these tests to measure achievement:

- **Are state tests available for all states/grades in the evaluation?** This question is examined by gauging the extent to which state test data are available for students in the four experiments, and in particular whether these data are consistently available for all grades and states (Section 3.1).

- **Do state tests meet minimal criteria for use in the evaluation?** To answer this question, we examine the characteristics of state tests – as well as the study-administered test – in each of

the four experiments (Section 3.2). We conduct descriptive analyses to help us better understand whether the content of state tests is aligned with the intervention (which affects the validity of causal inferences about program impacts) and to verify that there are no floor effects in the test scores (which affects the reliability of the test and the precision of impact estimates). These results also provide context for interpreting the impact estimates that are at the core of this paper.

- **Are impact findings sensitive to the type of assessment used to measure student achievement?** To answer this question, we compare the pattern of estimated impacts on state tests and impacts on the study-administered test, to see whether impact findings (with respect to inferences and precision) differ between the two types of assessment, and if so, whether these differences make sense given differences in the content and reliability of the two test types (Section **Error! Reference source not found.**1).

We also look at several questions related to the analysis of state tests in a multi-state experiment:

- **Can impact findings be combined across states and grades?** We conduct two types of analyses to answer this question. First, we look at the extent to which state tests differ in terms of their content, for each of the four experiments. We then examine whether in practice, these differences in test content are sufficiently large to lead to different impact findings across the study states (Section 4).

- **How should state test scores be converted to a common metric? What weight should be attributed to the impact for each state and grade in the overall finding?** This question is examined in two ways. First, we use descriptive information from the randomized experiments to examine whether a given rescaling or aggregation approach may be more appropriate than another given the characteristics of the data or the study design (Section 0). Second, we estimate the impact of the program on state test scores using different combinations of rescaling and weighting strategies, to examine whether impact findings are sensitive to these decisions (Section 6.2).

The last research question pertains to using state tests scores *at baseline* (prior to random assignment) as a means of improving the precision of the impact estimates. In a previous paper in this series, Deke et al. (2010) compare the precision gains from adjusting for study-administered pretest scores with the gains from using publically available school-level proficiency data. They conclude that on average, adjusting for school-level proficiency does not increase statistical precision as much as controlling for study-administered pretest scores. However, part of this result may be due to the fact that proficiency rates are measured at the *school level*, which limits how much these data can improve precision.[7] In this paper, we examine the precision gains from controlling for *student-level* state test scores:

[7] Specifically, school-level proficiency rates can reduce the amount of unexplained between-school variability in achievement, but not within-school variability, thus limiting the precision gains to be had.

- **Is the precision of estimated impacts on achievement sensitive to the type of assessment used measure student achievement at baseline?** To answer this question, we examine the precision gains from using students' state test scores from prior school years as a covariate in the impact model, relative to (i) no covariates, and (ii) using a study-administered pretest instead. We also look at the extent to which these precision gains depend on the outcome measure (that is, whether achievement at follow-up is measured using a state test or a study-administered test) (Section 7).

Overview of the Paper

The remainder of the paper is structured as follows:

- **Section 2** provides further information on the four studies that are re-analyzed in this paper, including a description of the treatment, the study design, the assessment measures, and the sample of students from each study that are included in our analysis.

Sections 3 to 5 present the findings from our descriptive analyses:

- **Section 3** discusses the factors that affect *whether* to use state tests. We present descriptive information on the content and reliability of state tests for the four experiments re-analyzed in this paper.

- **Section 4** looks at the factors that affect *whether to combine* impact findings across states and grades. This includes, in particular, an assessment of how much state tests in the four experiments differ in terms of their content, and whether these differences in content are sufficiently large to lead to differences in program effectiveness across state and grades.

- **Section 5** discusses issues relevant to deciding *how to combine* impact findings across states and grades. Specifically, we look at the different options for rescaling test scores and weighting the findings, and we use descriptive information from the four experiments to examine whether one approach may be more appropriate than another.

Sections 6 and 7 present the results of the impact analyses for the four experiments:

- **Section 6** presents impact findings on state tests and the study-administered test in the four experiments, to examine whether results are sensitive to the type of assessment used to measure achievement. It also looks at whether the estimated impact of the program on state test scores is sensitive to decisions about rescaling and aggregation.

- **Section 7** looks at whether the type of assessment used to measure student achievement at *baseline* (prior state test scores or a study-administered pretest) affects the precision of the impact findings.

The paper concludes with a summary of the findings and their implications for educational evaluation.

2 Data and Samples

This section provides an overview of the four IES-funded randomized studies that are re-analyzed in this paper. These four studies were chosen primarily because they use both a study-administered test and state tests to measure student achievement. They also represent useful variation that can be used to examine the factors that can affect whether and how state tests should be used in an evaluation. The studies differ with respect to grade level (elementary school, middle school, and high school) and subject area (reading and math). They also represent a continuum in terms of the primary outcome targeted by the intervention: two of the interventions aim to improve achievement more broadly (math achievement, reading achievement), while the other two programs primarily focus on skills that are more specific (reading comprehension, rational numbers). Overall, these four studies span 22 states. All state assessments are high-stakes tests, either for students or for the school. Further information on each study is provided in the next section.[8]

For each of the four experiments, the analysis sample used in this paper is restricted to students who are part of the first cohort of study participants, and who have both a study-administered test score and a state test score at follow-up.[9] Even with these restrictions, the analysis sample for each study is sufficiently large to detect program impacts of reasonable magnitude. This is an important criterion when comparing impact findings across test type (state tests, study-administered test) and rescaling/aggregation method.[10]

It is important to note that the samples of students used in this methodological report differ from the samples used in the official evaluation reports for these four studies, because more restrictive criteria are used to define the samples in the present analysis. Thus, the impact findings that will be presented in this report should not be compared to the impact findings from the official reports for these four studies. To dissuade readers from making such comparisons, the studies are not identified by name in this report; instead they are denoted as Studies A through D. In this report, all references to the "impact analysis sample" or the "analysis sample" refer to the samples of students used in this methodological study, rather than the analysis samples used in the official impact evaluation reports.

[8] Databases for each study were made available by the evaluation contractor (MDRC), because restricted use files for two of the experiments were not yet available at the time of analysis.

[9] For Studies A, B and D, the analysis sample is further limited to states where the *state-wide* mean and standard deviation in test scores on the state test is known. As will be described in a later section, this restriction is imposed to make it possible to compare two rescaling methods (z-scoring based on the sample distribution *vs.* the state-wide distribution in achievement). This restriction results in the loss of 1 state in Study A and 2 states in Study B. For Study C, this additional sample restriction is not imposed because the state-wide mean and standard deviation in state test scores is only available for one of the study states.

[10] With small sample sizes, estimated impacts would not be statistically significant using either approach, and so we would conclude that findings are not sensitive to these analytical decisions even though they in fact could be. The minimum detectable effect size for the impact on achievement (as measured by the study-administered test) is 0.12, 0.13, 0.16 and 0.11, respectively, for each of the four experiments.

Overview of the Experiments and Samples

Table 2.1 summarizes the key features of each study and, from each of these studies, the analysis sample that is used for the purposes of this report. In the first two studies, the intervention being tested aims to improve general achievement in either math or reading. A standardized assessment was administered by evaluators to provide a consistent measure of general achievement. State test scores were also collected to measure the program's impact on helping students meet state standards:

- **Study A:**[11] The treatment in Study A is an after-school reading program that aims to improve students' reading achievement. Twenty-five after-school centers across 10 states implemented the program for two school years. Students in each center were randomly assigned to the enhanced program or to the regular after-school program offered by the center. The target population consists of students in 2nd to 5th grade who are behind grade level in reading based on the assessment of their regular school-day teachers. Students' reading achievement at the end of the program was measured using the Stanford Achievement Test 10th edition (SAT-10) abbreviated reading battery.[12] Additionally, state test scores in ELA/reading for the spring were also collected for these students. In total, 1,032 students are included in the analysis sample used in this report.[13] As expected given the eligibility criteria, these students have characteristics associated with low academic achievement: 86 percent of students are eligible for free/reduced price lunch and 26 percent of the students are overage for grade.

- **Study B:** The treatment in Study B is an after-school math program that aims to improve students' mathematics achievement, and is similar in design to Study A. Twenty-five after-school centers across 9 states implemented the program for two school years. Students in each center were randomly assigned to the enhanced program or to the regular after-school program offered by the center. The target population is students in 2nd to 5th grade who are behind grade level in math. Students' math achievement at the end of the program was measured using the Stanford Achievement Test, 10th edition (SAT-10) abbreviated mathematics battery.[14] Additionally, state test scores in math for the spring were collected for these students. In total, 944 students are included in the analysis sample.[15] On average, 74 percent of students in the analysis sample are eligible for free/reduced price lunch and 22 percent are overage for grade.

[11] Studies A and B are part of a larger project that evaluates two after-school programs. See Rebeck Black, Doolittle, Zhu, Unterman, & Baldwin Grossman (2008); Rebeck Black, Somers, Doolittle, Unterman, & Baldwin Grossman (2009)

[12] This assessment consists of three subtests: reading comprehension, vocabulary, and word study skills.

[13] Among all students in the study, 92 percent took the SAT-10 assessment while 75 percent have state test data (this difference is due to the fact that 2nd grade students in the study are not tested by their states).

[14] This assessment consists of two subtests: problem solving and procedures.

[15] Among all students in the study, 94 percent took the SAT-10 assessment while 75 percent have state test data (this difference is due to the fact that 2nd grade students in the study were not tested by their states).

Table 2.1
Features of the Four Randomized Experiments Used in this Paper

	Study A	Study B	Study C	Study D
Features				
Unit of randomization	Student	Student	Student	School
Subject area	Reading	Math	Reading	Math
Outcome targeted by program	Reading achievement	Math achievement	Reading comprehension[a]	Rational numbers
Study-administered test	SAT 10 Abbreviated	SAT 10 Abbreviated	GRADE reading comprehension subtest	NWEA computer adaptive test on rational numbers
Program implementation	2 school years (2005-06 and 2006-07)	2 school years (2005-06 and 2006-07)	2 school years (2005-06 and 2006-07)	1 year (2007-08)
Grade levels	Grades 2-5	Grades 2-5	Grade 9	Grade 7
States in the study	10	9	8	9
Availability of baseline test scores				
State tests from prior school year	No	No	Yes	Yes
Study pretest	Yes	Yes	Yes	Yes
Sample used in this paper				
Target population	First cohort (grades 3-5)[b]	First cohort (grades 3-5)[b]	First cohort	All
Impact analysis sample[c]	1,032 (9 states)	944 (7 states)	1,065 (4 states)[d]	4,387 (9 states; 77 schools)

NOTES:
[a] The program in Study C also targeted reading vocabulary, but this outcome is not examined in this paper.
[b] State test scores are not available for students in 2nd grade.
[c] For all studies except Study C, the primary impact analysis sample includes students who: (1) have both a state test score and a study-administered test score and (2) who live in states where the state-wide mean and standard deviation in scores for the state test are known. For Study C, the second restriction is not applied, because the state-wide mean and standard deviation for the state test are known for only 1 of 4 states in the study. Note that for Studies A and B, the second sample restriction (i.e., that the state-wide mean and standard deviation are known) results in the loss of 1 and 2 states from the analysis sample, respectively; Appendix D presents impact findings for Studies A and B when the second sample restriction is not applied.
[d] Only 4 states have an annual assessment for 9th grade reading.

Note that because Studies A and B span multiple grades, all descriptive and quantitative analyses of state test scores for these studies occurs at the *grade-by-state* level (because the content of state assessments differs across grades within a given state).

The next two experiments used in this paper evaluate the impact of an intervention whose short-term goal is to improve a specific skill, with the broader goal of also improving students' general achievement. In these experiments, a study-administered test was used to measure the specific outcome directly targeted by the program. State test data were also obtained to examine whether the program also improved student achievement more broadly:

- **Study C:**[16] The treatment in Study C consists of year-long supplemental literacy programs that aim to improve the reading comprehension skills and school performance of struggling adolescent readers. These programs were implemented in 34 high schools for two school years. In each school, students were randomly assigned to either enroll in the supplemental reading class or to remain in a regular ninth-grade elective class. The target population for this study consists of ninth-grade students whose reading skills are two to five years below grade level as they enter high school. The Group Reading Assessment and Diagnostic Evaluation (GRADE) was used to assess student reading achievement (reading comprehension and vocabulary) in the spring of ninth grade. Data were also collected on students' test scores on their state's ninth grade ELA/reading assessment (if administered by the state). In total, 1,065 students are included in the analysis sample that is used in this report.[17] On average, 77 percent of these students were eligible for free/reduced price lunch and 22 percent were overage for grade at the start of the study.

- **Study D:**[18] The treatment in Study D is a year-long math teacher professional development intervention focused on rational numbers. The goal of the intervention is to improve students' math achievement, specifically their knowledge of rational numbers. The target population for this study are 7th grade math teachers in high-poverty middle schools. The study was implemented in 77 schools in 12 districts, with approximately equal numbers of schools randomly assigned in each district to receive the treatment provided by the study, or a "business as usual" group, which participated only in the usual professional development offered by the district. A customized computer-adaptive assessment on rational numbers was used to measure students' knowledge of rational numbers.[19] The test was administered to a random sample of students in the study schools in the spring of 2008. State test records for the spring were also obtained from school districts. In total, 4,387 students are included in the analysis sample.[20] Of these students, 65 percent were eligible for free/reduced price lunch.

[16] See Kemple, Corrin, Nelson, Salinger, Herrmann, & Drummond (2008); Corrin, Somers, Kemple, Nelson, & Sepanik (2009); Somers, Corrin, Sepanik, Salinger, Levin, & Zmach (2010).

[17] Overall, only 40 percent of students in the study have a state test score, while 83 percent of students have a GRADE score (this difference is due to the fact that only 4 of 8 states administer and ELA/reading assessment in ninth grade).

[18] See Garet *et. al.* (2010).

[19] This test was constructed by the Northwest Evaluation Association (NWEA).

[20] On average, 90 percent of students have a score on the study-administered test, while 97 percent of students have a state test score.

In these two studies, students' *baseline* achievement was also measured using both types of assessment: a study-administered pretest was given to students before random assignment, and data were obtained on students' scores on state tests in the grade prior to random assignment (8th grade reading/ELA tests for Study C and 6th grade math tests for Study D). In Section 7, these two studies are used to assess the precision gains from using these two types of test as covariates in the impact analysis.

3 Whether to Use State Tests in a Multi-State Experiment

As discussed in May *et al.* (2009), researchers should consider two key factors when deciding whether to use state tests to measure student achievement at the end of the intervention: (i) whether state tests are *available* for the relevant subject areas and grades, and (ii) whether state tests can provide *valid* inferences about program effects and *reliable* measurements for the target population of students.

This section uses descriptive data from the four randomized experiments to examine these issues more closely. As noted in the introduction, a formal assessment of test properties is beyond the scope of this paper, because we are only able to examine certain aspects of validity and reliability. In other words, we can examine whether state tests in the four experiments satisfy certain *necessary* conditions for use in an evaluation, but not whether they are in fact *sufficiently* valid and reliable. In practice, however, evaluators are similarly limited in what they can find out about state tests, so the findings of this exercise remain informative. Because the content and availability of state tests varies across states, the bulk of the descriptive analysis focuses on state tests; however, we also examine the characteristics of the study-administered test to provide context for interpreting the impact findings to be presented later.

3.1 Availability

Under NCLB, states must test their students annually in reading and math in grades 3 to 8. However, testing in earlier grades and high school can be more sporadic, thereby making state tests a less consistent source of achievement data. For example, in 3 of the 4 randomized experiments re-analyzed in this paper, state tests are not consistently available for all students:

- **Elementary grades:** In Studies A and B, state test data are not available for 2nd grade students in the study, which means that inferences about program impacts on state test are limited to students in 3rd to 5th grade.
- **High school:** In Study C, which targets ninth grade students, only 4 of 8 states in the study administer an ELA/reading assessment in the relevant grade, which means that conclusions about program impacts on state test scores are confined to a less diverse set of states.

Despite the limited availability of state test data in these studies, they are used in these evaluations as a *complement* to the study-administered test. In the study reports, state tests are characterized as a "secondary" outcomes and their limited generalizability is clearly acknowledged.[21]

[21] Another issue to consider related to feasibility is whether state test scores can be obtained from states or school districts, and the level of effort that this would require (see May et al., 2009). The availability of test data in all 4 experiments re-analyzed in this paper confirms that such data *can* be collected; however, obtaining test data may be impossible or more complicated in other study contexts.

3.2 Descriptive Analysis of State Tests

As described in May *et al.* (2009), it is important to consider two properties of state tests when deciding whether to use them in an evaluation:

a. *Validity:* Whether state tests can be used to provide valid inferences about the achievement outcomes of interest (and by extension, valid inferences about program impacts on these outcomes).

b. *Reliability:* Whether state tests can be used to obtain an accurate measure of the outcomes of interest for students in the study (and by extension, precise estimates of program impacts).

The remainder of this section provides a brief discussion of these two test properties. We also present a descriptive assessment of some aspects of these properties for the four randomized experiments, based on published information and simple descriptive analyses of test scores.

A. Validity

Several factors can affect whether or not state tests yield valid inferences about program effectiveness (May *et al.*, 2009):

(1) Stakes of the testing

(2) Participation rates

(3) Testing accommodations

(4) Breadth and depth of test items

(5) Alignment with the outcome(s) of interest in the evaluation.

The first three factors – testing stakes, participation rates, and accommodations are important because "treatment-control" differences with respect to these factors could compromise the causal inferences that come from a randomized experiment (and therefore causal validity). In the four experiments re-analyzed in this paper, these factors are not a cause for concern. All state tests are high stakes (see Section 2), so we expect both the treatment and control group to be equally and highly motivated. Nor were there treatment-control differences in response rates on the state tests, which suggests that differential participation is unlikely to compromise causal inferences. Finally, with respect to testing accommodations, students requiring accommodations were not given the study-administered test and are therefore excluded from the analysis; therefore, treatment-control differences in testing accommodations are not a relevant source of bias.

The last two factors – breadth/depth and content alignment – are important because they affect whether state tests adequately capture the outcomes targeted by the program. Unfortunately, the breadth and depth of test items cannot be examined for the four experiments, because item-level data are not available for the studies that we reanalyze in this paper. However, we can examine the

alignment of state tests with the outcomes of interest, which is the focus of the remainder of this section.

Alignment between state tests and the outcome of interest

In order to provide valid inferences about program impacts, state tests must be aligned with the outcome that the program is expected to improve.[22] If they are poorly aligned with this outcome, then the evaluation may incorrectly conclude that the program is ineffective. In other words, the magnitude of the impact estimate could be statistically indistinguishable from zero.

The criteria for assessing whether state tests are well aligned with the outcomes of interest depends in part on the intervention's theory of action, and in particular, whether the intervention's immediate goal is to improve general achievement or a more specific skill.

If the goal of the program is to improve student achievement on a broad range of skills ("general achievement"), then state tests are likely to provide valid inferences about program impacts on the targeted outcome. State assessments are typically general achievement tests that cover a broad range of skills in a given subject area, especially in elementary and middle school. In this paper, for example, two of the four experiments target general achievement – the intervention in Study A targets general reading achievement, while the intervention in Study B targets general math achievement. Information on the content of state assessments in these two studies reveals that the state assessments for these two studies are general achievement tests (reading tests in Study A; math tests in Study B). Therefore, one would expect state tests to be aligned with the outcome targeted by the intervention in these experiments.

In other interventions – like the ones evaluated in Studies C and D – the goal is to improve a specific skill in the hopes of ultimately improving students' general achievement. In studies where the targeted outcome is more specific, determining the validity of state tests is a more nuanced exercise because there are two outcomes of interest in the program's theory of action: the first is the specific skill on which the program is focused (the targeted outcome), and the second is students' general achievement (the broader outcome that the program aims to improve). The program's impact is expected to be greater for the targeted outcome given its closer "proximity" to the intervention, but both outcomes are expected to improve.

In this context, the extent to which state tests can provide valid inferences about program effectiveness depends on which of these two outcomes they are intended to measure. To provide valid inferences about impacts on the *targeted outcome*, all state assessments must include a *subtest* that measures that outcome. If this condition is met, then these subtest scores can be used as a measure of the targeted outcome that is also policy relevant.[23] State tests can also be used to measure

[22] Because educational programs often aim to improve multiple outcomes, in this paper we use the term "targeted outcome" to refer to the most proximal student outcome in the intervention's theory of change (that is, the student outcome that is expected to be *most* affected by the program).

[23] An advantage of using state subtest scores (rather than a study-administered test) is that they are not "overly aligned" with the targeted outcome. Overalignment is a common criticism of study-administered tests when they are used to measure a specific skill, because the test chosen by the evaluator may be too closely aligned with the experiences of the treatment group to be fair to the control group. Indeed, a study may fail to meet the *What Works*

the second outcome of interest, *general achievement*. When the targeted outcome is a specific skill, state tests must meet two criteria to be used as a measure of general achievement. Similar to the situation where the targeted outcome is general (see previous section), all state tests in the study must be "general achievement" tests. When the targeted outcome is specific, however, meeting this condition is not sufficient. In addition, each test must include a set of items that either measure the specific outcome targeted by the intervention, or a set of items that are affected by the specific outcome.[24] If this second condition is not met, then state tests are not aligned with the targeted outcome, so the program cannot be expected to improve general achievement.

In this analysis, two of the experiments target a specific skill – the intervention in Study C targets a specific reading outcome, while the intervention in Study D targets a specific math outcome. Information on the content of state tests in these two studies reveals that subtest scores on the specific targeted outcome are not consistently available in all study states, so state tests cannot be used to measure the shorter-term targeted outcome; instead, a study-administered assessment is used to measure the specific skill. On the other hand, state tests do meet the conditions for using them to measure student's general achievement: in all of the study states, the state test is a general achievement test and includes items on the specific targeted outcome.[25] Therefore, in these studies, total scores on state tests can be used to measure the program's impact on general achievement, which is an outcome of interest in both evaluations.

Table 3.1 takes a closer look at these issues, by looking at the degree of similarity between the content of state tests and the study-administered test, for each of the four studies. The table presents information on the amount of overlap between the content of the study-administered test and state tests. Test overlap is measured based on the percentage of state tests' content that covers the domains in the study-administered test (on average across states and grades).[26]

Clearinghouse standards for evidence if its measures are overaligned (What Works Clearinghouse, 2008). See Slavin (2008) for a discussion.

[24] For example, assume that a program aims to improve students' general vocabulary, with the ultimate goal of improving their achievement in all subject areas. If state tests are to be used to measure general science achievement, then state tests would either have to include a set of vocabulary items on the definition of different science terms, *and/or* include a set of science-related content items that a student could more easily answer if their general vocabulary was improved.

[25] At the domain level, the specific targeted outcome represents 29 percent of the state test's content on average in Study C, and 21 percent in Study D. The content of these tests will be examined in greater detail in Section 4.

[26] For example, consider Domain A, the first domain in the study-administered test. If a state test includes 10 different domains, and Domain A is one of them, then a value of 10 percent will be assigned to this state test for Domain A. Such a percentage is assigned to every state test in the study, and the average across states is the "overlap" for Domain A. This is repeated for each domain in the study-administered test. The overlap value for each domain is then averaged, and this overall average is the value presented in the table. This value represents the percentage of domains in the state tests (averaged across states and grades) that cover the domains in the study-administered test. A 100 percent total would indicate complete content overlap between these two types of tests; the further away the measure is from 100 percent, the less overlap there is between the two types of test.

Table 3.1
Content Overlap (at the domain level)
Between State Tests and the Study-Administered Test

Study	Study test's content as a percentage of state test content[a]	Study/State test correlation	
		Average	Range
Study A (n = 9)	41.9	0.441	0.194 - 0.801
Study B (n = 7)	96.0	0.628	0.541 - 0.743
Study C (n = 4)	29.2	0.453	0.424 - 0.535
Study D (n= 9)	20.7	0.687	0.572 - 0.833

SOURCE: State Department of Education websites.
NOTES: [a]Percentages are calculated at the domain level.

The key points to note form the table are the following:

- **Studies A and B (targeted outcome is general achievement):** For Study B, which targets general math achievement, the amount of overlap between state tests and the study test is 96 percent. However, for Study A, which targets reading, overlap is 42 percent, even though both types of test measure of "general reading achievement".
- **Studies C and D (targeted outcomes is a specific skill):** The overlap between the study test state tests is 29 percent and 21 percent, respectively for these two studies, which confirms that state tests are broader in scope than the study-administered test, and that they are best used to measure impacts on general achievement (as opposed to impacts on the targeted skill).

B. Reliability

One of the key factors affecting the precision of impact estimates is the reliability of the outcome measure. When an assessment has low reliability, it measures student achievement with a substantial amount of error. Measurement error inflates the variance in test scores across study participants, which increases the standard error of the impact estimate, and reduces the study's power to detect program effects. Reliability is also a prerequisite for validity: if a test does not reliably measure the outcome of interest, then it cannot provide valid inferences about this outcome.

In general, the overall reliability of state tests based on all students in the state is quite high. For the four randomized experiments used in this paper, for example, the lowest value for any given state in the analysis sample is 0.81. Also, while there is some variation in the reliability of

assessments across states, the reliability for state tests is approximately the same on average as the overall reliability of the study-administered test.[27]

It is important to note, however, that the reliability of an assessment depends on the *value* of the test score. The reliability of state tests is usually maximized for scores around the state proficiency cut-off[28] and it decreases for scores that are further away from the cut-off. This means that state assessments are least reliable for the highest- and lowest-performing students. For low-performing students, the items in the assessment are too difficult, so many students will have the same (low) score. This makes it more difficult to differentiate (discriminate) between the proficiency of low-performing students, which in turn reduces the reliability of the test for these students. For high-achieving students, the test is too easy, which through a similar process also leads to a lowered reliability.[29]

Therefore, when deciding whether to use state tests in an evaluation, the most important factor is the *conditional reliability* of these assessments – that is, their reliability at a *particular value* of the test score or *for a particular subgroup of students*, in this case students participating in the study. Most IES studies, including the ones used in this paper, target low-performing students, so the conditional reliability of the tests for these students is likely to be lower than the reported average test reliability (which is based on all students in the state). However, for the four randomized experiments in this analysis, the conditional reliability of state tests for low-achieving students is not consistently available. Some states do not publish this information on their education department websites, while others report this information but they do so based on non-continuous metrics (e.g., classification errors for proficiency levels).

Given the absence of such information, we can instead examine whether the reliability of state tests is so low that they should not be used at all. Specifically, we can obtain a lower bound for conditional reliability by investigating whether there is a floor effect in the distribution of state test scores. A *floor effect* occurs when many students *incorrectly* answer every (or most) test items. Because most low-performing students have the same score (i.e., zero), this makes it difficult to differentiate between the achievement levels of these students. This in turn reduces the conditional reliability of the test at the lower end of the achievement distribution (May et. al, 2009).[30] Floor effects can be identified by visual inspection of state test scores. If the lowest-scoring students in a state are clustered or "piled up" at the bottom of the test score range, then this would indicate that the test is too difficult for these students, causing a floor effect. In turn, this would suggest that the conditional reliability of the test may be so low that it should not be used in the evaluation.

[27] The reliability of the study-administered tests for national samples ranges from 0.89 to 0.93. See Appendix A for detailed information on the published reliability of the states tests and the study-administered test in each experiment.

[28] Test items are typically chosen to maximize the reliability of the test around the state proficiency cut-offs. Because these cut-offs are used to make decisions about students, it is important that the achievement of students around the cut-off be accurately measured.

[29] See Hambleton, Swaminathan, and Rogers (1991) for a general discussion of reliability.

[30] Conversely, a *ceiling effect* occurs when many students correctly answer every item; this makes it difficult to distinguish between high-performing students, and reduces the conditional reliability of the test at the upper end of achievement distribution.

Accordingly, Figure 3.1 shows the density of state test scores for each of the four randomized experiments in this study.[31] In order to make it possible to show the distribution for all states in the same figure, these densities are estimated based on test scores that have been converted to z-scores using the sample mean and standard deviation (which does not affect the distribution of scores). As shown in this figure:

- **For Studies B, C and D:** There does not appear to be a floor effect in any of the study states. The distribution of state tests is approximately normal or bell-shaped, such that students in the sample are concentrated in the mid-point of the test score range. (If there were a floor effect, the maximum point of the density function would be at the lower end of the range). For these three studies, the reliability of state tests does not appear to be so low that state tests should not be used at all.

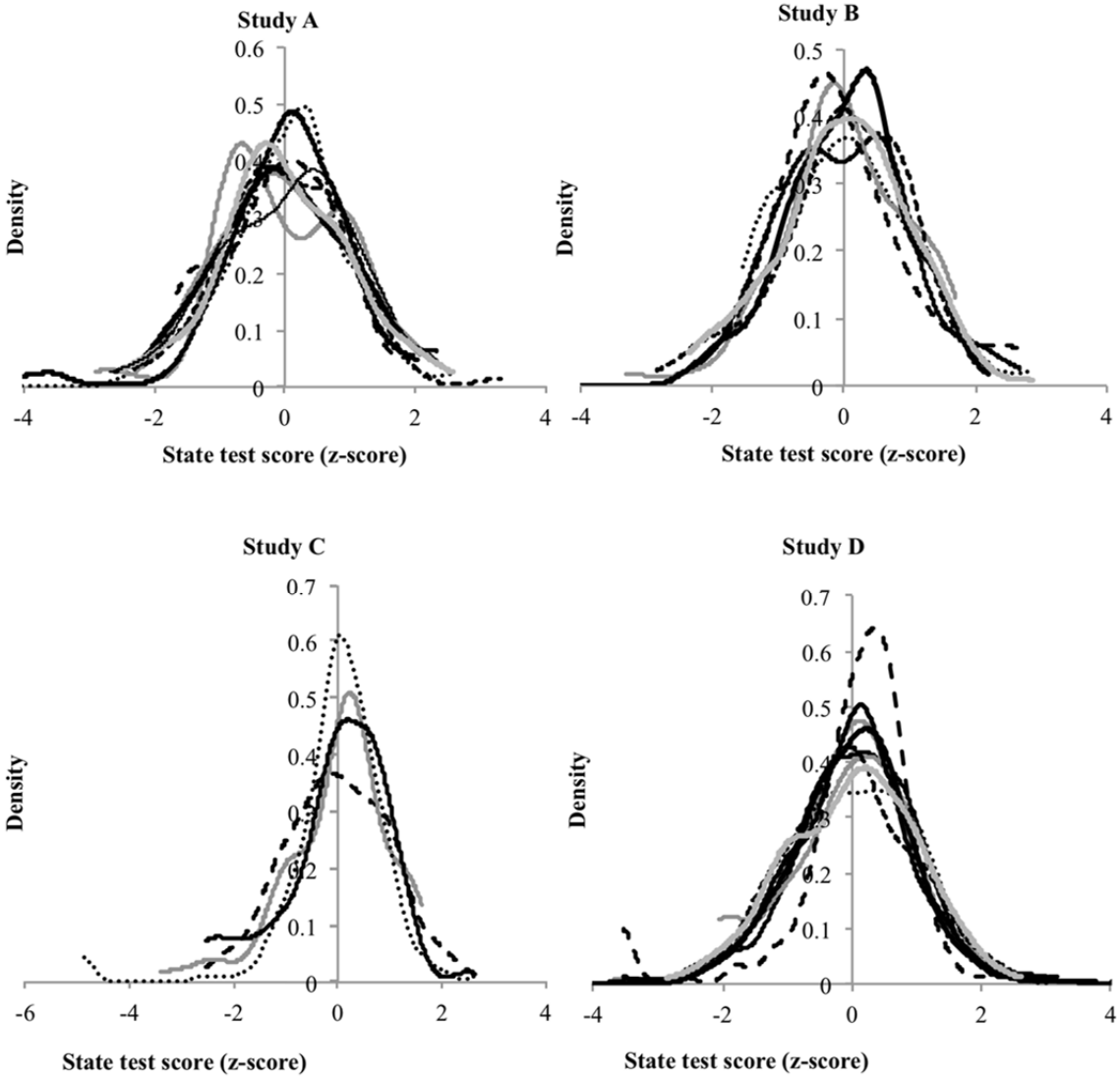
- **In Study A:** One of the study states has a bi-modal distribution, and most students are concentrated in the lower "bump" at the end of the test score range. This suggests that there may be a floor effect in test scores in this particular state, which could compromise the reliability of state tests in Study A. The extent to which this affects the precision of the estimated impact on state test scores in Study A will be examined later in this paper.

Taking a step back, it appears that state tests in all four experiments meet an important condition for validity – their content is aligned with one of the outcomes of interest (general achievement). With respect to reliability, however, only three of the studies meet the necessary condition that there be no floor effects; the conditional reliability of state test scores in Study A is questionable.

If the content alignment of state tests and their reliability is deemed suitable, the next step for evaluators is to decide *how* to use state tests in the evaluation. Two questions are especially relevant in a multi-state experiment: (1) *whether* to combine impact findings across states, and if so, (2) *how* to combine them. These questions are examined in the following sections.

[31] These figures are based on the students in the impact analysis sample that are used in this report.

Figure 3.1
Density of State Test Scores in Each Study, by State



SOURCE: State test scores obtained from school districts in the study.
NOTES: Each density curve in the figure represents a state in the study. Kernel density estimates are based on students with both a state test score and a study test score. The bandwidth used for kernel density estimates is based on Silverman's rule of thumb (Silverman, 1986). Test scores are rescaled (converted to z-scores) by state and grade using the mean and standard deviation of the sample.

4 Whether to Combine Findings across States and Grades

Typically in IES studies, conclusions about program effectiveness are based on whether or not an intervention has a statistically significant impact *on average* across all students in the study sample (that is, on average across all states and grades in the study). Combining the results across states and/or grades improves the generalizability of the findings to a broader range of contexts. It can also allow the study to achieve sufficient power to detect meaningful impacts on the outcome of interest.

State tests pose an important challenge in this regard, because they do not provide a *consistent* measure of achievement across the study sample due to differences in their content. In some states, for example, "reading proficiency" as measured by the state test may be more heavily weighted towards vocabulary, while in other state tests grammar may be the primary focus. This means that the impact of the program – in terms of the types of skill or knowledge gains that it represents – does not have a consistent interpretation across states and/or grades.

These differences in test content have important implications for the interpretation of the "combined" impact finding. In the first instance, the average impact of an intervention across states and grades must be interpreted as the impact of the program on the *skills measured by state tests,* rather than on a consistent measure of achievement. In other words, differences in state tests must be accepted as reflecting intended variation in state standards.

All else equal, this interpretation is typically not an undesirable one. Many would argue that impacts on "state tests" are an important and policy-relevant outcome, because districts and states use these assessments to make decisions about individual students and schools. Indeed, given its policy relevance, students' performance on state tests is often an outcome of interest in an evaluation (including the four experiments examined in this study). If it were not, then this would raise questions about why evaluators had decided to collect these data in the first instance. As long as researchers are careful to acknowledge that their "combined" finding represents the effect of the program on the skills measured by state tests, then the overall impact can be a meaningful indicator of program effectiveness.

A second (and competing) concern, however, is that differences in test content may lead to differences in the estimated effectiveness of the program across states and grades. Impacts may be larger in the states/grades whose assessment is better aligned with the outcome targeted by the intervention, in which case the average impact of the program would mask meaningful differences in program effectiveness. This, in turn, would complicate the interpretation of the "average" impact of the program.

Variation in program effectiveness across states or grades is therefore an important factor to consider when deciding whether it is acceptable to combine impact findings. Specifically:

- If impacts are *consistent* across states and grades, then combining the impact findings is probably acceptable because it provides information on the overall effect of the program on a policy-relevant aggregate outcome.

- If impacts *differ* across states or grades, however, the policy relevance of the combined impact finding must be weighed against the "utility" of presenting the overall impact, since in this case the pooled result would mask either variation in program effectiveness across states/grades and/or differences in the ability of the state test to measure the outcomes affected by the intervention. In this case, it may be preferable to present the impact findings separately for each state/grade, assuming that the sample size is large enough to detect meaningful impacts at this disaggregated level.

Accordingly, the remainder of this section examines whether differences in test content are sufficiently large to cause impacts to differ across states or grades. In the first instance, we take a closer look at the content of state tests in the four randomized experiments, to assess the extent to which their content differs. We then formally test whether there is variation in impacts across states in these studies.

Variation in test content and impacts

A program's effect on student achievement may differ across states or grades for several reasons, including differences in the strength of program implementation and the service contrast across states or grades, and differences in the characteristics of students in the study sample.[32] These factors are relevant to all randomized experiments, regardless of whether achievement is measured using a study-administered test or a state test.

When achievement is measured using *state tests*, however, this introduces an additional driver of impact variation: differences in the "validity" of the tests across states and grades. Even if all state tests provide a valid measurement of the outcomes of interest, it is not necessarily true that they are *equally* valid for the evaluation. Some state tests may be better aligned with the outcome targeted by the intervention than others. All else equal, one would expect estimated impacts to be larger in states or grades whose assessments are better aligned with the targeted outcome.

A relevant question here is whether the dissimilarity among state tests is so great that alignment with the targeted outcome can truly be expected to differ. After all, state assessments are typically general achievement tests, so in practice there will be at least *some* overlap between states in terms of the content of their assessments. This is especially true for subjects like mathematics, where the topics included in curricula – and the order in which these topics are taught and tested – are more standardized.

In order to get a better sense of the degree to which state tests differ in practice, we collected information on the content of state assessments for the four experiments used in this paper. When looking for information on test content, we focused in particular on the *domains* covered by these tests (in mathematics, for example, domains include areas such as geometry, algebra, while in

[32] All else equal, impacts are expected to be greater in states/grades that implemented the program with a higher degree of fidelity, and in states/grades where there is a greater contrast between the services received by the treatment group and those received by the control group. Differences in the program's impact may also occur if the program has a differential effect on certain types of student – and if some states/grades in the study have a higher concentration of such students.

reading or ELA, domains include broad topics such as reading comprehension, and literary analysis).[33] As explained earlier, state tests in all four experiments are general achievement tests.

Table 4.1 summarizes the results of this data collection exercise.[34] The first column presents a comprehensive list of the domains included in the state assessments, while the second column then shows the percentage of state tests that include a set of items on that particular domain. As a reference point, consider that if all state tests included *exactly the same set of domains*, then the value associated with each domain would be 100 percent. Therefore, the extent to which the percentages in Column 2 differ from 100 percent (and differ across domains) provides a measure of the degree of dissimilarity in the content of state assessments.[35] The more these percentages diverge from 100 percent, the greater the variation in test content.[36]

The findings in this table confirm that the content of state assessments differs across states and grades, and that these tests measure somewhat different overall constructs:

- For each of the four experiments in this paper, there is less than perfect overlap in the content of state assessments at the domain level. Though the content of math assessments is reasonably similar (at least at the domain level), the content of reading assessments appears to differ by a notable degree for the two relevant experiments in this paper.[37]

That said, these differences in test content may not lead to variation in the impact of the program across states or grades. Whether they will or not partly depends on whether the outcome targeted by the intervention is general achievement or a specific skill.

First consider an evaluation where the targeted outcome is general achievement. In this situation, even though the content of state tests differs to some extent (as confirmed by Table 4.1), all state tests measure a broad set of skills (general achievement), which is what the intervention aims to improve. Therefore, state tests' *alignment* with the targeted outcome is consistent across states. In this situation, one might not expect variation in test content to lead to differences in impacts across states or grades (at least not by a substantial amount). For example, even though there is variation in test content across states in Studies A and B, all state tests measure "general achievement" which is the outcome targeted by the intervention. All else equal, impacts are not expected to vary due to between-state differences in alignment with the state test.

In contrast, consider an evaluation where the targeted outcome is a *specific skill* and where state tests are used to measure the program's longer-term impact on general achievement. In this

[33] See Appendix A for detailed information on the content domains that are covered in each state assessment.

[34] The results in this table are based on the states represented in the impact analysis samples.

[35] In Studies A and B, which span multiple grades (grades 3-5), the state test for each grade is treated as a different "state test" in this analysis.

[36] To have a more accurate representation of test overlap, it would have been preferable to weight each content domain by the number of test items that it includes. However, this information was not consistently available for all state assessments.

[37] For the two *math* studies, four of seven content domains in Study D (and five of seven for Study B) are included in all state tests in the study, and six of the seven content domains are included by most state tests (that is, by at least 70 percent of states). In contrast, for the two *reading* studies (Studies A and C), only one content domain is included in all state tests, and all other domains are included in no more than half of the state tests.

Table 4.1
Overlap in the Content of State Tests,
States in the Impact Analysis Sample

	% of states that test domain
Study A	
All state tests (n = 9)	
Word Analysis	33.3
Reading Comprehension	100.0
Writing	33.3
Vocabulary	22.2
Information and Media Literacy	33.3
Literary Analysis	44.4
Reference and Research	11.1
Study B	
All state tests (n = 7)	
Number Sense	100.0
Algebra	100.0
Geometry	100.0
Measurement	100.0
Data Analysis	71.4
Probability	100.0
Mathematical processes	14.3
Study C	
All state tests (n = 4)	
Word analysis	25.0
Reading Comprehension	100.0
Writing	50.0
Vocabulary	25.0
Information Literacy	50.0
Literary Analysis	25.0
Study D	
All state tests (n = 9)	
Number Sense	100.0
Algebra	100.0
Geometry	100.0
Measurement	100.0
Data Analysis	77.8
Probability	88.9
Mathematical processes	11.1

SOURCE: State Department of Education websites.

situation, all state tests would include at least some items on the targeted outcome (otherwise, state tests would not have been deemed valid for the evaluation). However, some tests may include a *greater percentage* of items on the targeted outcome than others, and will therefore be better aligned with the intervention. For these types of studies, differences in test content are more consequential, because such differences could lead to variation in the "depth" of the targeted outcome in the state assessment.

For Studies C and D, we therefore examined whether there is variation in the depth with which state assessments measure the targeted outcome. Specifically, we looked at the percentage of test content that measures the targeted outcome, at the domain level.[38] Based on this analysis, we find that the "depth" with which the specific outcome is measured in state tests does differ across states. For Study C, the depth of the specific targeted outcome in state assessments ranges from 25 to 33 percent of the state test at the domain level; in Study D, the depth ranges from 17 to 25 percent of the state test. This suggests that test alignment with the targeted outcome may differ across states.

This means that all else equal, impacts are most likely to differ across states in Studies C and D, which target a specific outcome, and least likely to vary for Study B, which targets general math achievement. Whether these hypotheses bear out in practice is an empirical question that we can examine using the four randomized experiments.

For each experiment, therefore, we tested whether program impacts on state tests differ by a statistically significant amount across the study states and grades.[39] We find that for three of the studies (A, C, and D), variation in impacts across states is not estimable, or in other words that the estimated cross-state variation in impacts in zero. For Study B, there is a small amount of variation in impacts between states and grades, but the estimated variation is not significantly different from zero.[40] The key point here is the following:

- Differences in test alignment are not sufficiently large to lead to detectable variation in estimated program impacts across states or grades, for any of the four experiments (including Studies C and D, which target a specific skill).

It is worth noting that the statistical power to detect variation in *true* program impacts is low in these four experiments, because they each include fewer than ten states. However, the primary concern here is that pooling the results will mask variation in *estimated* impacts rather than true impacts, since the alternative to pooling the results is to present the estimated impacts separately for each state. The variance tests reported above indicate that the variability in *estimated* impacts is not so large that the pooled impact would mask differences in the estimated impacts across states.

[38] For example, if the targeted outcome represents 1 of 10 domains included in the state test, then the "depth" of the targeted outcome is 10 percent.

[39] For Studies A and B, we tested whether the variation in impacts across *grades-by-state* is statistically significant.

[40] Variance in impacts = 0.0178, p-value = 0.4429. Because the targeted outcome in Study B is general math achievement, this small amount of variation in impacts is likely due to factors other than differences in test content (such as variation in implementation fidelity or service contrast across states).

Therefore, for these four experiments, pooling the results is acceptable.[41] Although this result may not apply in all contexts and studies, it does suggest that in some cases, it is reasonable to expect estimated impacts on state tests to be consistent across states and grades (even when the targeted outcome is a specific skill). In the event that there *is* variation in impacts across states, researchers would have to decide whether it is still useful to present a pooled result. If so, then they must be especially careful about which type of pooling approach to use, given that impacts vary across states. These issues will be discussed in greater detail in the next section.

[41]Consider the alternative scenario, in which impacts *do* differ across states. In this situation, researchers could choose to present the impact findings separately for each state or grade in the study sample. It may be the case, however, that the sample size is not large enough to reliably detect state-specific or grade-specific impacts. Large-scale randomized experiments funded by IES often span multiple states, each with a small sample size; for the four experiments used in this paper, for example, the minimum detectable effect sizes (MDES) for state-specific impacts are 0.44, 0.36, 0.33, and 0.34, respectively. If the sample size is too small to allow impact findings to be presented by state and grade – and impacts are expected to differ due to variation in test content – then a "compromise" approach would be to combine the impact findings across states and grades, but to also conduct an exploratory analysis of the factors that mediate impacts. In order to model associations between mediators and impacts, however, the study must include a reasonably large number of study states (Study C, which includes only 4 states, would not meet this criteria). In this situation, some would argue that state tests should not be used in the evaluation at all (see May *et. al.*, 2009, for a discussion).

5 How to Combine Impact Findings across States

If researchers decide to combine impact findings, the next question is how to go about estimating the *average* or *pooled* impact of the program across states and grades. For the purposes of this discussion, we start by defining some key terms that are used in this section:

- **Scaling of test scores:** This is the process by which raw test scores are converted to a different metric, typically for the purpose of giving the test scores a more meaningful interpretation. One of the simplest scales is the percentage of test items that are correctly answered by a student – a scale often used in teacher-developed formative assessments. For more complex assessments, however, such as state tests, it is also important to account for the difficulty of test items that are correctly answered. In this case, raw test scores can be scaled using item response theory (IRT). In this paper, the state education authority has already scaled the raw test scores and the scale of these tests differs across states. Thus, in order to be able to pool impacts on state tests across states and/grades, we must convert them to a common metric using a linking method.

- **Linking of test scores:** When tests differ with respect to their content and/or difficulty, a linking method must be used to establish a relationship between the test scores so that they can be placed on the same scale. There are two types of linkage: (1) *equating*, which is used to link scores on tests that have the same content but that differ in difficulty and (2) a *concordance*, which is used to link scores across assessment that do *not* measure the same construct (Kolen and Brennan, 2004).

- **Rescaling of test scores:** In this paper, we use the term "rescaled scores" to refer to test scores that have been converted to a common metric using one of the two linking methods. Traditional methods can be used to convert scores to a common metric (this includes linear and non-linear linking) or IRT methods can be used.

It is important to note that the distinction between the two linking methods – "equating" and "creating a concordance" – is *not* the type of rescaling method that is used to link scores, because rescaling strategies are applicable to both types of linkage. (These conversion methods will be described in detail later.)

Rather, the main distinction between equating and creating a concordance is in the *interpretation* of the rescaled scores. Because equating is used on tests that measure the same construct, this process yields rescaled scores that are *interchangeable* or *exchangeable*: once scores have been equated, score x on one assessment represents the same latent proficiency as score x on the other assessment. Viewed otherwise, it should not matter which test a student takes – their equated score on either assessment should be the same. In contrast, scores that are linked using a concordance are not interchangeable – linked scores do not represent the same latent proficiency because the content of the tests is different. The typical example of a concordance is the relationship between ACT and SAT scores. Using a concordance table, scores on one test can be mapped onto the scale of the other test. However, the content of the SAT and the ACT are different, so it cannot be assumed

that linked scores are interchangeable. When creating a concordance, the goal is to place scores on a common metric; when equating scores, the goal is to also make these rescaled scores interchangeable.

This distinction is important, because the interchangeability of rescaled scores affects the way in which state test scores should be pooled across states and/or grades. The following section discusses these two linking methods in greater detail. As will be argued in this section, test scores can only be equated across states under very specific conditions – none of which apply to the four experiments reanalyzed in this paper – and so a concordance must often be used to link scores instead.

5.1 Equating State Test Scores *vs.* Creating a Concordance

As noted above, equating is used to link scores across assessments that measure the same construct but that differ with respect to their level of difficulty. Scores are equated to adjust for differences in difficulty across the two tests and to place the two sets of test scores on the same scale, so that students' scores can be compared across the two assessments. In this context, the two assessments measure the same construct, but their original scale happens to be different due to differences in difficulty; equating is used to rescale the test scores from each assessment and to put them on a common scale. Because the tests measure the same construct, scores on the two forms are *interchangeable* after equating has been used to put scores on the same scale – that is, a given rescaled score on one test should represent the same latent proficiency as the rescaled score for another test. There are two types of equating: horizontal and vertical. Horizontal equating is used to link scores across different tests for a given grade level, while vertical equating is used to link scores for a given test across different grades. Equating is used, for example, when different forms of the same test are administered – test forms measure the same content but differ in difficulty due to differences in the test items included in each form.

In order to equate test scores, researchers need to know the "link" between the scale of the tests. Once this relationship is known, it can be used to convert the original scores from each assessment onto a common scale. The linking relationship between two assessments can be stated in terms of a difference in means ("mean link", which assumes that the two tests differ only in terms of the mean score of each assessment); a difference in means and variance ("linear link", which assumes that the two tests differ only with respect to the mean and standard deviation of scores); a difference in distribution ("equipercentile or rank-based link", which assumes that the two tests differ with respect to their percentile ranks). These links fall under the category of "traditional" equating functions. Modern methods such as Item Response Theory (IRT) can also be used to model the relationship between test scores; however, this method is more data intensive because it requires information on item-level characteristics.

To determine which type of relationship to use to equate scores, researchers can estimate the equating link between assessments using two data collection designs. The first design is to administer both tests to the same group of students (single group design). The second design is to randomly select two groups of students from a population and to administer one of the tests to each group

(equivalent group design). Importantly, in both designs, the distribution of *latent proficiency* (the construct measured by the tests) among students who took Test A is the same as for student who took Test B. In the single group design this is achieved by having students take both tests; in the equivalent group design, this is achieved by randomizing students to the two tests. Because of this expected equivalence in proficiency across students who take Test A and Test B, any difference in the distribution of test scores between the two assessments is entirely due to a difference in their scale. Thus, the appropriate equating link can be determined by comparing the distribution of scores across the two tests – that is, whether the scores differ with respect to their means (in which case mean linking should be used), their mean and standard deviation (which would require a linear link), or their percentile ranks (which implies an equipercentile link).[42]

Therefore, we can see that two necessary conditions must be met in order to be able to equate test scores: (1) the assessments must measure the same construct and (2) it must be possible to establish the linking relationship between test scores using a rigorous design.[43] Having defined these general requirements, we can address the conditions under which it would be possible to equate scores on *state assessments*:

- *Tests measure the same construct:* State tests are aligned with state standards, which vary across the country; this means that nationally, state assessments measure different constructs. This is especially true at the high school level: even within a given subject area (reading, math, science), the content of state tests can vary with respect to the specific domains and items included in the assessment. That said, if states in the study are purposefully selected and have a similar set of standards and assessments, then it is possible that state tests in the study measure the same construct.

- *Equating relationship can be determined:* Neither of the two equating designs described earlier applies to state tests. Each student takes only one state test, so the single group design cannot be used to establish the linking relationship. Nor are students randomized across states (randomization always happens within states), so strictly-speaking the equivalent group design does not apply either. However, we can potentially consider state test administration as a special case of the second design, if we assume that the sample of students from each state are from a common reference population and that the distribution of "proficiency" is similar across the sample of students from each state. In theory, this condition could be met if students are chosen based on a pretest score on a standardized assessment (thereby ensuring that students are similar across states).

[42] Once the linking relationship has been estimated, it can also be applied to a different sample of students (all test takers), assuming that this relationship is population invariant (see footnote 43).

[43] There are four other conditions for equating test scores (Kolen & Brennan, 2004): (1) the tests must have the same reliability; (2) the linking relationship between their raw scales must be symmetrical (the linking function for equating scores from Test A to Test B must be the inverse of the function for mapping scores from Test B to Test A); (3) the tests must be equitable (the test-taker should be indifferent to which assessment is used to measure their achievement); and (4) the linking relationship between the test should be population invariant (the population used to establish the equating function between the scores should not matter).

In summary, two conditions must hold to equate state test scores: (1) states in the study must have common standards and their assessments must measure the same construct, and (2) the sample of students from each state must come from a common reference population and be similar on average.

If these requirements are met, then the average impact of the program across all states can be estimated simply by comparing the average rescaled (equated) scores of students in the treatment and control group. The difference in rescaled scores between the two groups represents the estimated impact of the program.[44] In this analysis, students are treated the same regardless of their state or their grade. This explains why it is important that scores be interchangeable across states and grades. A student with a rescaled score of -0.20 in State A must have the same "proficiency" or "achievement" as a student with the same rescaled score in State B (horizontally equated). Or similarly, a student in grade 3 with a rescaled score of -0.20 must have the same achievement level as a student in grade 5 with that same score (vertically equated).

However, state test scores can only be equated under very specific circumstances, which may not be encountered very often in practice. Most relevant for the purposes of this paper, they are not met in any of the four experiments re-analyzed: state tests in these four studies do not measure the same construct, nor can students in the study be assumed to come from the same reference population (see Appendix B for a more detailed examination of this issue).[45]

In this situation, test scores can be linked using a *concordance*. In practice, a concordance can be created using the same linking relationships that are used for equating (i.e., linear linking, equipercentile linking). However, rescaled scores cannot be used interchangeably: a given rescaled score on Test A does not represent the same "proficiency" as the same rescaled score on Test B, because the two assessments measure different constructs. The concordance simply makes the distribution of scores the same across tests, so that they can be pooled for analytical purposes. The scores of students who wrote the same assessment can be compared; however, scores cannot be compared across different assessments.

This implies a different approach for estimating the overall impact of a program on rescaled state test scores, which is referred to as the "meta-analytic" approach in May *et al.* (2009). As indicated by its name, this approach treats the estimated impact for each state and grade as a separate "study". Analytically, the approach consists of:

a) Estimating the impact of the program for each grade and state (within-grade/state impact), and then
b) Calculating the average program impact by taking a weighted average of these estimates.

In practice, this means that the estimated impact is obtained by comparing rescaled student scores *within states* but not *across states*, which is necessary because test scores cannot be

[44] If a blocked random assignment design was used, then sampling weights also need to be used to account for the fact that a student's probability of being assigned to the program differs by block.

[45] As discussed in May *et al.* (2009), these conditions are not likely to be met in any multi-state experiment.

considered equated (interchangeable) across state assessments. Also, because impact findings for each grade and state are treated and analyzed as separate studies, the approach implicitly allows for differences in testing standards and difficulty across states and grades. At this stage of the analytical process, researchers have presumably decided that it makes sense to combine impact findings (see Section 4.1), so this variation in standards is acceptable to them.

Because the conditions for equating scores are not met in the four experiments, the remainder of this paper focuses on the meta-analytic approach for estimating impacts on state tests. From this point onwards, any reference to "linking" test scores should be understood as creating a concordance.

5.2 Using the Meta-Analytic Approach

When using the meta-analytic (concordance) approach, researchers must make two types of analytical decision:

- **Rescaling (linking):** Researchers must decide how to rescale state test scores – that is, which linking relationship to use to establish a concordance. In some ways, this decision is much easier when the goal is to create a concordance rather than to equate scores. When the objective is to equate scores, the choice of linking (rescaling) method is crucial, because the correct link must be used to make the rescaled scores interchangeable. When creating a concordance, however, the objective is simply to convert scores to a common metric, so the "right" rescaling method primarily depends on the preferred interpretation of the combined impact estimate.

- **Weighting:** Researchers must also decide what weight to attribute to the impact estimate for each grade/state when aggregating the findings. Several options exist, including weighting each impact estimate by its corresponding sample size, weighting by its precision, or using weights that are adjusted for variation in impacts (random-effect weights).

This section supplements the description of the meta-analytic approach in May *et. al.* (2009) in two ways. First, it provides greater technical detail on the analytical decisions to be made with respect to rescaling test scores and aggregating the impact findings across states and grades. Second, this section uses descriptive data from the four experiments to bring data to bear on the factors that influence these decisions.

5.2.1 The Choice of Linking Function

This section reviews the different linking functions that can be used to rescale state test scores to a common metric, which is the first step when using the meta-analytic approach to combine the impact findings. The terms "linking" and "rescaling" are used interchangeably in this discussion, as well as the terms "relationship", "function", and "transformation".

As noted earlier, there are two broad classes of linking functions from which researchers can choose:

- ***Traditional:*** This includes more conventional linking functions, such as linear rescaling (z-scoring) and non-linear (rank-based) rescaling.

- ***Item Response Theory (IRT):*** This method transforms state test scores onto a scale that measures a student's latent proficiency, accounting for the characteristics of the test items that a student answered correctly or incorrectly (for example, item difficulty).[46]

In multi-state evaluations, it is typically not possible or practical to use IRT methods, because the characteristics of each item in each state test must be known. In theory, this information could potentially be obtained from technical manuals, or else estimated using item-level test data. In evaluations involving multiple states, however, such information is typically not available for all states, and therefore "traditional" rescaling methods must be used instead. This is certainly true of all four experiments reanalyzed in this paper, and for this reason, the remainder of this section focuses on conventional rescaling (linking) functions.

Among the set of conventional functions, researchers must make two further analytical decisions:

- ***Choice of reference population:*** As explained earlier, when a concordance is used to link test scores, the rescaled scores of students who wrote the same assessment can be compared. That is, rescaled scores are "relative" scores: they represent a student's score relative to that of other students *who took the same test*. Researchers must decide whether scores should be rescaled relative to that of other students in the sample who took the test, or relative to all students in the state who took the test.

- ***Functional form of the transformation:*** If scores are rescaled relative to students in the sample (as opposed to all students in the state), researchers must also decide whether to rescale state test scores using a *linear* transformation (z-scores) or a *non-linear* transformation (rank-based rescaling).

As discussed in this section, the "appropriate" method depends on the desired interpretation of the overall impact estimate, as well as the shape of the distribution of test scores. These factors are discussed below using data from the four randomized experiments. We begin by discussing simple linear transformations, because this type of transformation can be applied with either of the two reference populations (sample or state). We then discuss non-linear transformations, which can be used when the reference population is the sample.

A. Linear transformation: Z-Scores

Linear rescaling converts test scores to a common scale by making the *mean* and *standard deviation* of test scores the same across assessments. The easiest way to make this happen is to convert test scores into z-scores using a simple linear transformation, as follows:

[46] For an introduction to IRT scaling, see Harris (1989) and Hulin, Drasgow, and Parson (1983).

$$Z_{isg} = \frac{Y_{isg} - \mu_{sg}}{\sigma_{sg}}$$

where for student *i* in state *s* and grade *g*:

Y_{isg} = The "raw" score on the g^{th} grade state test (i.e. the score on the original scale used by the state);

μ_{sg} = The mean score of students who took the same test for grade *g* in state *s*

σ_{sg} = The standard deviation of scores on the state test for grade *g* in state *s*;

And therefore:

Z_{isg} = Student *i*'s rescaled score on the state test (z-score).[47]

This transformation forces the *mean* and *standard deviation* of scores to be the same in each state and/or grade.

Using simple mathematical manipulation, it can be shown that the estimated impact of the program on z-scores in grade *g* and state *s* (ES_{sg}) is the estimated impact of the program on raw scores (I_{sg}), divided by the standard deviation used for z-scoring:

$$ES_{sg} = \frac{I_{sg}}{\sigma_{sg}} \qquad (1a)$$

This means that the estimated impact of the program on z-scores is scaled as an effect size and represents the effect of the program on raw (original) scores *as a proportion of the variance in state test scores*.

As seen above, the rescaled scores (*Z*) and the estimated impact (*ES*) both depend on the choice of the mean and standard deviation that are used to rescale the scores (μ_{sg}, σ_{sg}). Two types of "reference population" can be used for this purpose:

- *Sample distribution:* Raw scores can be rescaled relative to the distribution of test scores among students in the *study sample* who took the test. In this case, [μ_{sg}, σ_{sg}] are the mean and standard deviation for students in the sample who are in grade *g* and state *s*. With this

[47] Specifically, the mean and standard deviation of z-scores for students in grade *g* in state *s* ($\overline{Z}_{sg}$ and sd_Z_{sg}) is equal to:

$$\overline{Z}_{sg} = \overline{Y}_{sg} - \mu_{sg}$$

$$sd\_Z_{sg} = \frac{sd\_Y_{sg}}{\sigma_{sg}}$$

Where:

$\overline{Y}_{sg}$ = The mean of *raw* test scores on the g^{th} grade test in state *s*;

sd_Y_{sg} = The standard deviation of *raw* test scores on the g^{th} grade test in state *s*;

approach, a student with a z-score of -0.25 would be one quarter of a standard deviation below the mean test score of students in the *sample* who took the test. The mean of z-scores is 0 and the standard deviation is 1.

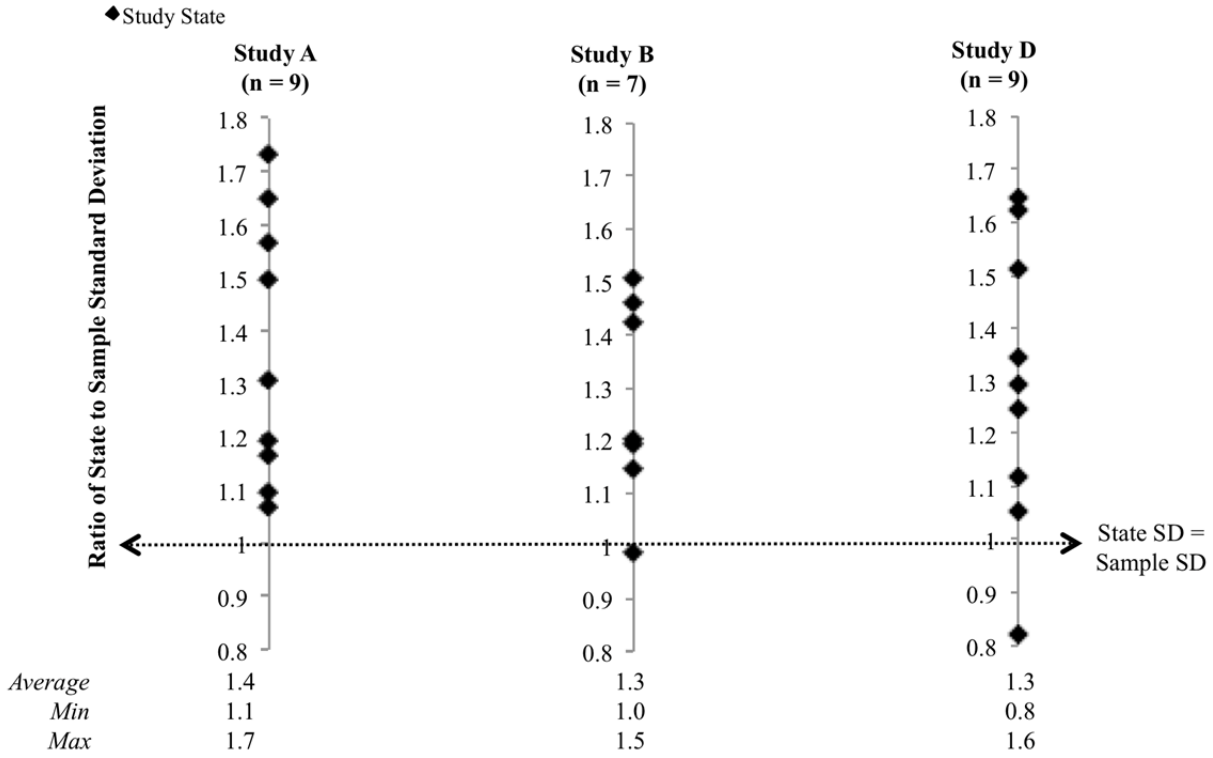
- *State distribution:* Raw scores can also be rescaled relative to the state-wide distribution of test scores. In this case, $[\mu_{sg}, \sigma_{sg}]$ are the mean and standard deviation for all students in grade *g* in state *s* who took the test (and not just students in the study sample). In this case, a student with a z-score of -0.25 is one quarter of a standard deviation below the *state-wide* mean for all students in their state (as opposed to the mean for students in the study sample). Z-scores have *non-zero* mean and a standard deviation *less than 1*, because the study sample is a subset of the state-wide student population. [48]

Because educational programs typically target a particular subset of students (rather than all students in a state), the distribution of test scores in the sample is likely to be quite different from the state-wide distribution. For example, consider programs that target low-achieving students. In these studies, the *mean* score for the sample will be lower than the state-wide mean score; the *standard deviation* of test scores – which affects the magnitude of the effect size (*ES*) as shown in (1a) – will also be smaller for the sample than for the state. To illustrate, Figure 5.1 compares the sample and state-wide standard deviations, for the states in Studies A, B and D for which the state-wide standard deviation is known.[49] For all but one state (in Study D), the standard deviation in test scores for the state is larger than the standard deviation for the sample, which is to be expected given that these studies target low-performing students.

[48] If the sample is lower (higher) performing than the overall state population, then the mean of rescaled scores will be less than (greater than) zero. The standard deviation of z-scores will also be less than 1, because students in the sample are a targeted group of students (whether high or low-achieving) rather than a random sample of students in the state. The extent to which the standard deviation of z-scores is below 1 depends on the degree to which the study sample is more homogeneous than the statewide student population; if the standard deviation of scores for the sample is much smaller than for the state as a whole, then the standard deviation of z-scores rescaled based on the state distribution will be much lower than 1. See also footnote 47.

[49] Study C is not shown because information on the state-wide mean and standard deviation is only known for one state. For Study A and B, Figure 5.1 shows the standard deviations for students in grade 3 only; findings are similar for other grades in the study (grades 4 and 5).

Figure 5.1
State vs. Sample Standard Deviation



SOURCE: State test scores obtained from school districts in the study and state department of education websites.
NOTES: Values in the figure are the state standard deviation in test scores divided by the sample standard deviation in test scores. Sample standard deviations are based on students with both a state test score and a study test score, who live in states where the state-wide mean and standard for the state assessment are known. For Studies A and B, ratios in the figure focus on third grade students.

This means that in IES studies, which usually target low-performing students, the estimated impact (*ES*) will be larger in magnitude when z-scoring is based on the sample distribution in test scores than when the state-wide distribution is used. In Study A and B, for example, estimated program impacts on z-scores will be approximately 40 percent larger in magnitude on average when the sample distribution is used rather than the state distribution. In Study D, the magnitude of impacts on z-scores will be about 30 percent larger when the sample standard deviation is used.[50]

In practice, however, these differences in the magnitude of the estimated impact on z-scores do *not* affect conclusions about whether estimates of the impact are statistically significant for a given grade/state. This is due to the fact that z-scoring affects the impact estimate and its standard error by the same amount (that is, they are both divided by the same standard deviation). Mathematically, the

[50] In Study A and B, the state standard deviation is 1.4 times greater than the sample standard deviation on average (range = 1.1 – 1.7 for Study A and 1.0 – 1.5 for Study B). In Study D, the state standard deviation is 1.3 times greater than the sample standard deviation (range = 0.8 to 1.6).

standard error of the impact on z-scores in grade *g* in state *s* [se_ES_{sg}] is simply the standard error of the impact on raw scores [se_I_{sg}], divided by the standard deviation used for z-scoring (σ_{sg}): [51]

$$se\_ES_{sg} = \frac{se\_I_{sg}}{\sigma_{sg}} \quad (1b)$$

Therefore, the T statistic (and p-value) associated with the impact estimate for a given grade and state is not affected by the choice of standard deviation.[52]

On the other hand, the choice between state-based or sample-based z-scoring *does* affect the interpretation of the impact estimate (*ES*). As seen in (1a), the magnitude of the estimated impact on z-scores represents the extent to which the intervention moved students along the variability in student achievement. In turn, expectations about student achievement depend on what reference population is used for creating z-scores. Because the standard deviation for the sample is smaller, it may provide a more realistic reference point for expected growth among low-achieving students. On the other hand, if the goal of the program is to bring students up to grade level, then it may be more policy-relevant to scale the magnitude of the impact estimate by the variability in student achievement for all students in the state. This suggests that the choice of z-scoring approach depends partly on the standard against which researchers want to judge the effectiveness of the program.

In addition to the desired interpretation of the impact estimate, there are two other factors that researchers may want to consider when deciding what reference population to use for creating z-scores. First, on a practical level, the state-wide standard deviation of test scores must be actually

---

[51] For a student-level randomized experiment:

$$se(impact) = \frac{\sigma}{\sqrt{np(1-p)}}$$

where:
σ = Standard deviation of test scores for students in the sample
n = Number of students in the sample
p = Proportion of students assigned to the treatment group (random assignment ratio)

For a school-level randomized experiment:

$$se\ (impact\ ) = \frac{1}{\sqrt{P\ (1\ -\ P\ )}} \sqrt{\frac{\tau^{2}}{J} + \frac{\sigma^{2}}{JN}}$$

Where:
P = the proportion of schools assigned to the treatment group
J = the number of schools in the sample
N = the number of students per school
τ^2 = the school-level variance in the outcome

σ^2 = the student-level variance in the outcome.

See Bloom et al. (2007) for details.

[52] This is demonstrated in Appendix C for a student-level randomized experiment.

available. For Study C, for example, state-wide standard deviations could not be obtained for 3 of 4 states.[53] Therefore, in Study C, z-scores based on the sample distribution is the only viable option.

Second, researchers may also want to consider the interpretation of *levels* (as opposed to impacts). When z-scores are based on the *sample* mean and standard deviation, the average z-score will be zero by definition.[54] However, when z-scores are based on the *state-wide* mean and standard deviation, the average z-score will represent the average amount by which students in the sample scored below (or above) other students in the state. Thus, the latter method is more useful for understanding and describing the sample of students in the study. For example, Figure 5.2 looks at the average state test score of students in Studies A, B and D, relative to the average score for all students in the state. As expected, students in the study samples have lower scores on average than students in the state as a whole (with the exception of one state in Study D, where the sample is higher-performing than the state average). Therefore, in these experiments, rescaling test scores based on the state-wide distribution could potentially provide useful descriptive information on the relatively lower level of achievement of students in the study.[55]

Therefore, in general, the choice between state-based and sample-based z-scores depends primarily on how researchers want to contextualize the magnitude of achievement levels and impact estimates for students in the study. This is due to the fact that, as noted earlier, inferences about program effectiveness (p-values) in a given state and/or grade are not sensitive to the choice of z-scoring approach.

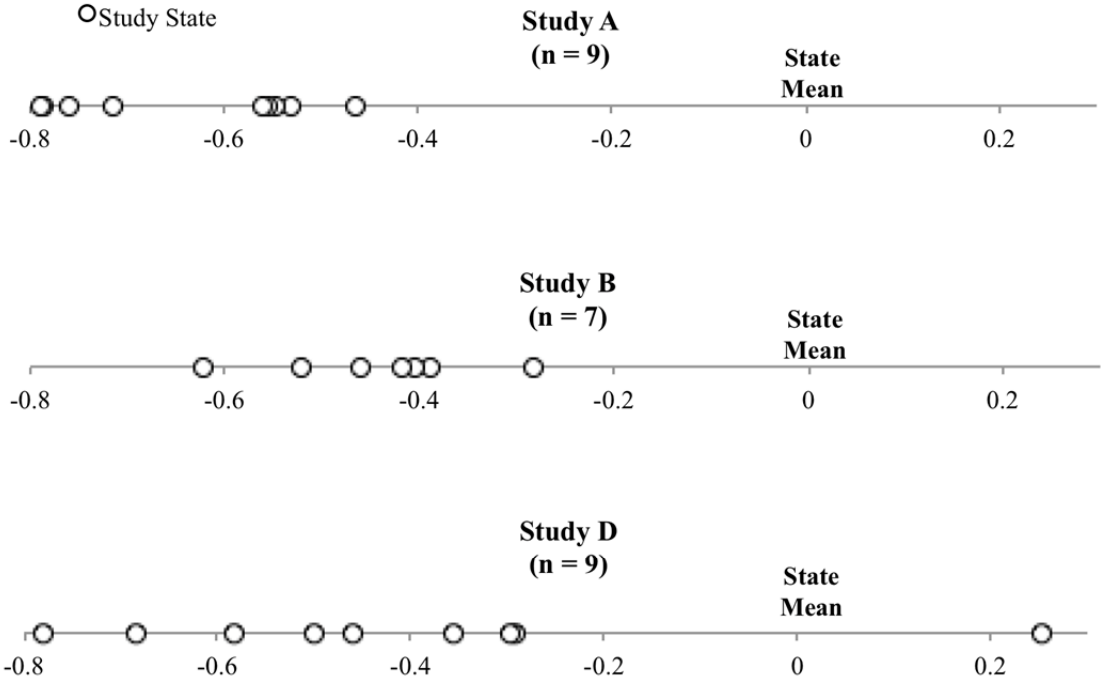
However, it is also important to note that while this conclusion may hold for *state-specific* impact findings, it may not be true of the pooled program impact across all states and/or grades. As will be discussed in Section 5.2.2, the relative weight of a grade/state in the "combined" impact estimate depends on the precision (standard error) of its impact estimate relative to that of other states and grades in the study, which *does* depend on the rescaling method. Therefore, inferences about the average impact of the program may be affected by the choice of rescaling approach. The extent to which the pooled result is sensitive to the choice of rescaling method will be examined empirically in Section 6.2.

[53] This could be due to the fact that these states administer "end-of-course" high school tests (EOCTs). Students take EOCTs once they have completed a course sequence. This means that the grade of test taking differs across students: most students take the EOCT in a given grade (in this case, grade 9), but some students may take it in a later grade. Such variation in the timing of test-taking may explain why states do not publically issue state-wide standard deviations on these tests.

[54] And the standard deviation will be equal to 1.

[55] Another rationale for state-level z-scoring is that it may come closer to achieving z-scores that are "equated" across grades and states. As shown in Appendix B, however, neither sample-based nor state-based z-scoring is capable of producing equated test scores in the four experiments reanalyzed in this paper.

Figure 5.2
State Test Scores of Students in the Sample Relative to All Students in the State (State Mean = 0)



SOURCE: State test scores obtained from school districts in the study and state department of education websites.

NOTES: The values in the figure represent the difference in mean test scores between the sample of students in a given state and all students in the state, divided by the standard deviation of test scores for all students in the state. For Studies A and B, standardized differences are first calculated by grade and state, and then averaged across grades for a given state to obtain the mean standardized difference for the state.

Sample means are based on students with both a state test score and a study test score, who live in states where the state-wide mean and standard for the state assessment is known. The sample sizes are: 1,032 students across 9 states for Study A, 944 students across 7 states for Study B, and 4,387 students across 9 states for Study D. Study C is not shown in this figure because the state mean and/or standard deviation is not known for most states in this study.

B. Nonlinear Transformation: Rank-Based Z-scores

Rank-based rescaling takes traditional z-scores a step further, by making the *distribution* of test scores the same across assessments (not just the mean and standard deviation). Specifically, this type of rescaling makes the distribution of test scores normal (Gaussian) for each assessment. For each state assessment:

i. The first step is to calculate students' percentile rank relative to other students in the sample who took the same test

ii. The second step is to convert these percentile ranks into z-scores based on a standard normal distribution.

There are two key features to note about this method:

- Due to the combination of Steps 1 and 2, the distribution of rank-based z-scores is standard normal. Hence, if the distribution of raw scores is *non*-normal, then rescaling scores using the rank-based method will have the effect of normalizing the distribution of test scores.[56] In contrast, linear rescaling does *not* affect the shape of the distribution of test scores, only the mean and standard deviation.[57]

- The interpretation of rank-based z-scores is relative to the *sample* (as opposed to the state). In theory, one could obtain information from school districts about a student's percentile rank in the *state* distribution, and use these percentile ranks in Step 2. However, this would compromise the second feature (and advantage) of the nonlinear rank-based method, which is that it makes the distribution of test scores the same in each state (Gaussian).

If the distribution of raw scores is *non-normal*, then using the rank-based method to rescale scores may actually yield more accurate inferences about program impacts. Recall that violations of normality can affect inferences about program impacts. When data are non-normal, the impact estimate and its standard error are unbiased, but the distribution of the estimated impact may be non-normal. Therefore, using T and F statistics to make inferences about program impacts may not be correct (i.e., p-values based on these statistics may not be accurate). If state test scores are not normally distributed based on their original metric, then making the distribution of rescaled scores normal may affect inferences about program impacts. That said, even if the raw test scores are *not* normally distributed, sample sizes in multi-state studies are so large that violations of normality may not be relevant.[58]

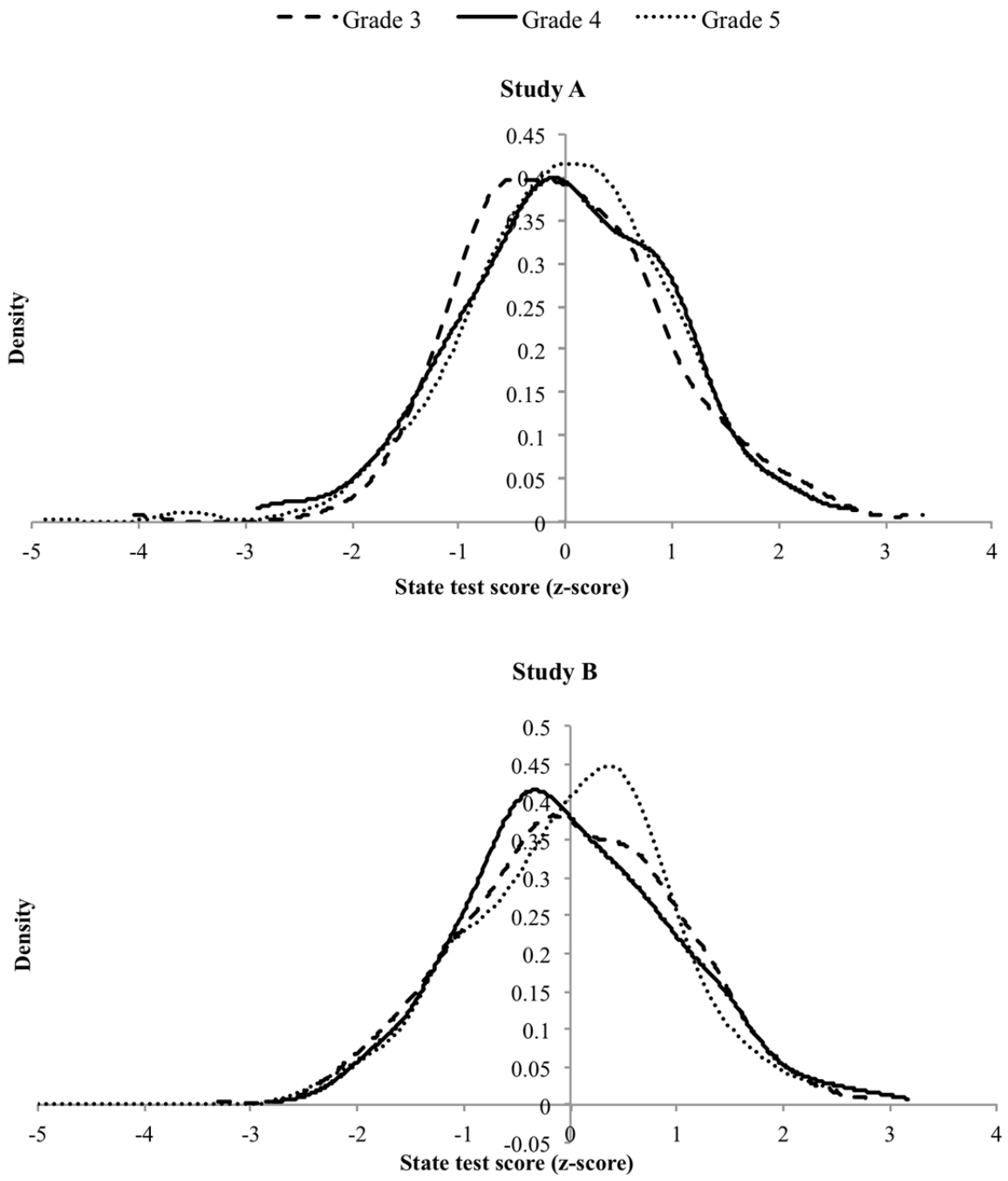
The density plots in Figure 3.1 show the distribution of scores for Studies A to C. Visually, the distribution of test scores looks approximately normal in all studies, though a couple of states in Studies A and B have a bimodal distribution. As shown in Figure 5.3, these bi-modal results are not driven by the pooling of grade levels in these two studies; the figure shows that the shape of test score distributions is similar across the three grade levels.

[56]The rank-based rescaling method also has property that it gives greater weight to test score differences in the tails of the distribution than in the middle of the distribution. This could be important in studies that focus on low-performing students, as in the case of Study D (Nataraj Kirby, McCaffrey, Lockwood, McCombs, Naftel, & Barney, 2002).

[57] See footnote 47.

[58] If state test scores are approximately normally distributed in their original metric, then the rank-based rescaling method and traditional z-scoring will produce a similar result.

Figure 5.3
Density of State Test Scores by Grade
Study A and B



SOURCE: State test scores obtained from school districts in the study.
NOTES: Each density curve in the figure represents a grade level in the study. Kernel density estimates are based on students with both a state test score and a study test score. The bandwidth used for kernel density estimates is based on Silverman's rule of thumb (Silverman, 1986). Test scores are rescaled (converted to z-scores) by state and grade using the mean and standard deviation of the sample.

In order to formally test whether these distributions are normal, we also conducted normality tests for each state; Table 5.1 shows the results of these tests. Overall, the results indicate that the distribution of state tests scores is more consistently normal in some studies than others. In Studies A and B, the distribution of test scores is statistically normal in at least half of the study states (5 of 9 states in Study A and 5 of 7 states in Study B based on the Shapiro-Wilk test). In contrast, the distribution of test scores is normal in only a handful of states in Study C (1 of 4 states) and Study D (2 of 9 states). Therefore, the rank-based rescaling method is *most* likely to yield different inferences about program impacts in Studies C and D. In practice, however, the sample sizes in these studies are large enough that violations of the normality assumption may be less important.

Table 5.1
Normality of State Test Scores
(Number of States with a Normal Distribution, Based on Four Tests)

Study	Shapiro-Wilk	Kolmogorov-Smirnov	Cramer-von Mises	Anderson-Darling
Number of states with normal distribution				
Study A (n = 9 states)	5	7	6	6
Study B (n = 7 states)	5	5	5	5
Study C (n = 4 states)	1	1	1	1
Study D (n = 9 states)	2	2	2	2

SOURCE: State test scores obtained from school districts in the study.
NOTES: Normality tests are based on students with both a state test score and a study test score. Test scores are standardized (z-scored) by state and grade using the mean and standard deviation of the sample.

Section 6.2 will examine whether, for the four experiments in this paper, the combined impact estimate is sensitive to the choice between linear and nonlinear (rank-based) rescaling methods. However, before proceeding to these findings, the next section reviews another important analytical question that pertains to estimating the combined impact of the program: how to aggregate impact findings across states.

5.2.2 The Choice of Aggregation Weights

Once state test scores have been rescaled, the next step is to estimate the effect of the program in each state, and to then average these estimates to obtain the average effect of the program across states and grades. In doing so, researchers must decide how to weight the impact estimates when calculating the average impact. The pooled impact ($\overline{ES}$) and its standard error ($se\_\overline{ES}$) are a weighted average of the site-specific results, as shown in the following general notation:[59]

$$\overline{ES} = \sum_{S,G} w_{sg} ES_{sg} \qquad (2a)$$

[59] See also Bloom (2002, Section 2.4).

$$se\_\overline{ES} = \sqrt{\sum_{S,G} w_{sg}^2 (se\_ES_{sg})^2} \quad (2b)$$

Where:

ES_{sg} = the estimated impact of the program for grade *g* in state *s*

se_ES_{sg} = the standard error of the estimated impact for grade *g* in state *s*

w_{sg} = the weight of the estimated impact of the program for grade *g* in state *s*

The "appropriate" choice of weight w_{sg} depends on whether researchers want to estimate the pooled impact of the program for the *states in the study sample* (the local impact), or for some *larger population of states* from which states in the study are assumed to be a representative sample (the generalized impact).[60]

These two types of inference are conceptually different. If researchers want to generalize their impact findings to a broader sample of states, then they must build additional sampling error (or uncertainty) into the standard error of the pooled impact estimate, in order to account for the fact that the estimated pooled impact of the program would have been different had the program been implemented in a different set of study states. However, if researchers do *not* want to draw inferences about the impact of the program in a different set of study states – and simply want to confine their inferences to the states in the study – then they do not need to account for this extra uncertainty.

These two types of inference are obtained by using different kinds of aggregation weights. The "local" impact for the study sample can be estimated using *fixed-effects* (FE) weights, while the "generalized" impact for the larger population can be estimated using *random-effects* (RE) weights.

In practice, one can use meta-analysis to apply either of these weights and obtain the pooled impact finding. The application of meta-analytic techniques depends on the level at which data are available:

- *When only aggregate data is available:* Meta-analysis was developed to aggregate impact findings across independent published studies; in this context, typically the only data available to researchers are the impact estimates from each study and their variance. In this context, the appropriate FE and RE weights are calculated, and these weights are then used in equations (2a) and (2b) to obtain the weighted pooled impact finding. In practice, the pooled impact can also be obtained by fitting a regression model where the impact estimates are the

[60] In the meta-analytic literature, these two types of inference are reflected by the difference between "fixed-effect meta-analysis" (for estimating the impact for the study sample) and "random-effects meta-analysis" (for estimating a more generalized impact). See Lipsey and Wilson (2001) and Hedges and Olkin (1985) for a general discussion. See also Bloom (2002) for a discussion in the context of educational evaluation.

dependent variable, and by specifying an error structure where these impact estimates are the second level of a multi-level framework (observations nested within studies). The latter approach is called a meta-regression.

- *When person-level data is available:* If person-level data are available (as opposed to group-level data), researchers have two options for aggregating the impact findings. The first option is to estimate the impact of the program for each group, and then weight the impact estimates like in a classical meta-analysis. A second option is to estimate and aggregate the impacts in a one step, by fitting a single FE or RE statistical model to the person-level dataset.[61] Both of these approaches – the classical (two-step) approach or the one-step regression – should yield the same result and are considered "meta-regression" models (May *et al.* 2009, p.39).

In multi-state educational evaluations, student-level data are typically available, so either approach can be used. We focus primarily on the one-step approach given its greater ease, but we will also look at the classical approach, since state test data may sometimes only be available at the state level. As will be shown, the two approaches produce very similar results, as expected. Therefore, the most important decision for researchers is *not* how to meta-analyze the results, but rather how to weight the impact findings (that is, whether to use fixed-effects or random-effects weights/modeling).

The remainder of this section describes these meta-analytic approaches – as well as the distinction between FE and RE weighting – in more detail. The section begins by describing the weights that should be used to estimate the "local" impact of the program on the study sample (fixed-effects weighting). This is followed by a discussion of random-effects weights, which are used to estimate the "generalized" impact of the program.

A. "Local" Impact: Average Impact for States in the Study Sample

In many evaluations, states in the study sample are chosen based on convenience, and are not a representative sample of states from some larger population. In this situation, researchers must confine their inferences about the average impact of the program to *states in the study sample* – that is, the local impact of the intervention. To estimate this impact, the state/grade- specific impact estimates must be weighted and pooled in one of the following ways: weight by precision (also called "fixed effects" weighting), weight by sample size, or weight equally. Each of these weighting approaches is described below.

Fixed-Effects Weighting (Weight by Precision)

The first approach is to weight each impact estimate based on its precision (i.e., the inverse of the variance of the impact estimate):

[61] See Wooldridge (2002) for a general treatment. Also note that random-effects one-step regression models are multi-level or hierarchical models, which have been extensively discussed in the field of sociology and educational research (Raudenbush & Bryk, 2002).

$$w_{sg} = \frac{\frac{1}{(se\_ES_{sg})^2}}{\sum_S \frac{1}{(se\_ES_{sg})^2}} \qquad (3a)$$

The weights w_{sg} can be used in the general notation (2a and 2b) to calculate the weighted pooled impact and its standard error. These weights maximize the precision of the pooled impact finding, because less precise estimates are not weighted as heavily when calculating the standard error of the average estimate (in Equation (2b)). This is also called "fixed effects" weighting. As described earlier, there are two ways of implementing this weighting scheme:

- ***Classical (two-step) approach:*** The impact of the program is first estimated separately for each grade and state in the sample for a given study. The pooled impact of the program and its standard error are then obtained by substituting the precision weight (3a) and the relevant impact estimates into Equations (2a) and (2b).

- ***One-step regression approach:*** In this approach, data from all states and grades in a given study are pooled together into one dataset, and the precision-weighted average impact of the program is estimated "automatically" using the a single statistical model:[62]

$$Z_i = \beta\, T_i + \sum_{S,G} \lambda_{sg} ST_{si} * GR_{gi} + \sum_{S,G} \sum_{M} \delta_{msg} X_{mi} * ST_{si} * GR_{gi} + \varepsilon_i$$

Where:

Z_i = Rescaled test score (z-score) for student *i*.

T_i = Indicator of treatment group membership (treatment status).

X_{mi} = Set of *m* pre-random assignment characteristics and prior achievement outcomes for student *I*, whose effect is allowed to vary across states and grades.

ST_{si} = Set of S indicators for state, equal to 1 if student *i* is in state *s* and zero otherwise.

GR_{gi} = Set of G indicators for grade, equal to 1 if student *i* is in grade *g* and zero otherwise (only relevant for multi-grade studies).

[62] Many educational studies are based on a randomized block design. Blocks are typically schools (in student-level experiments) or school districts (in school-level experiments). In Studies A and B, for example, the blocks are grades within schools; in Study C, the blocks are schools, and in Study D they are school districts. The blocking of random assignment also needs to be accounted for in the analysis (regardless of whether the classical or one-step regression approach is used). This can be done either by using weights in the analysis, or by including dummy indicators of block in the impact model (in the fixed effects model, for example, the second term would include interactions between state, grade, and *school* or *district*, depending on the type of block).

The interaction between *ST* and *GR* are state-grade fixed-effects, i.e. a set of intercepts for each state-by-grade. Therefore:

β = The average impact of the program on state test scores *in the sample of states in a given study*. Because the impact model controls for state and grade, β represents the average impact of the program within states and grades, weighted by precision.

As noted earlier, the two-step classical approach and the one-step regression approach should produce very similar results, but the latter may be more convenient to implement in practice, because pooling is accomplished simultaneously.

Weighting by Sample Size

The second approach for pooling the impact findings is to weight the impact estimates by the sample size in each state and grade:

$$w_{sg} = \frac{n_{sg}}{N} \qquad (3b)$$

In a *student-level random assignment study* (like Studies A, B and C), n_{sg} is the number of students in the sample who are in grade *g* in state *s*, and *N* is the total number of students in the study across all state/grades. In this type of study, sample size weighting should generate a pooled impact estimate (magnitude, standard error, and p-value) that is similar to the one produced by precision weighting, because precision is a function of the number of students in the study.[63] In both cases, the pooled impact represents the estimated impact of the program for the average *student* in the study sample.

In a *school-level random assignment study* (like Study D), weighting by sample size is somewhat more complex because there are two ways in which impacts can be weighted to obtain the pooled impact. The first is to weight the impact estimates for each state/grade by the number of students in the sample, as described above. The second option is to weight the impact estimates by the number of *schools* in the sample; in this case, the numerator in (3b) is the number of schools in the study that are in state *s* (n_s), and N is be the number of schools in the study. Let's consider the differences between these two weighting options, and how they compare to precision weighting:

- *Number of students vs. number of schools:* When choosing between these two weighting strategies, an important factor to consider is the desired interpretation of the average impact estimate. When impacts are weighted based on the number of students, the pooled result represents the estimated effect of the program for the average *student* in the study sample. Conversely, when impacts are weighted by the number of schools, the pooled result represents the estimated effect of the program for the average *school* in the study sample. When the number of students differs across schools, these two types of inference can be quite different. This means the magnitude of the pooled impact estimate may be sensitive to the

[63] See Appendix C (last section) for a discussion of the specific conditions under which sample size weighting and precision weighting are exactly equivalent.

choice of weight (students *vs.* schools). The precision of the pooled finding may also differ between approaches. In this case, weighting by the number of schools may produce a larger pooled standard error, because there are fewer schools in the study than students.

- *Sample size weighting vs. precision weighting:* In a school-level experiment, sample size weighting (whether based on students or schools) can produce different results compared to precision weighting. This happens because the precision of a state-specific impact estimate is a function of both the number of schools *and* the number of students in the study sample.[64] In contrast, sample size weighting is based on the number of students *or* schools. This means that sample size weighting and precision weighting are not equivalent and can yield different pooled impact findings. Of particular interest, the *precision* of the pooled impact estimate may differ: precision weighting minimizes the pooled standard error by definition, which means that sample size weighting may yield a less precise pooled impact estimate.

It is also worth noting that when precision weighting is used in a school-level experiment, the pooled impact estimate represents (approximately) the impact of the program for the average *student* in the sample, which is also the inference obtained from weighting by the number of students.[65] This means that if researchers want to estimate the impact of the program for the average student in the sample, then precision weighting may be preferable to weighting by the number of students, because precision weighting is expected to produce a more precise estimate of the pooled impact. However, if researchers want to estimate the impact of the program for the average *school* in the study sample, then weighting by the number of schools in the sample may be preferred because it will yield the desired inference. These issues will be examined empirically using data from the four experiments.

Equal Weighting

A third approach for obtaining the pooled impact finding is to give each state- and grade-specific impact estimate an equal weight in the pooled result, regardless of how many schools or students are in each state and grade. With this type of weighting scheme, the pooled impact estimate represents the impact of the program for the average state/grade in the study.

B. "Generalized" Impact: Average Impact for a Broader Population of States

If the study states are a representative sample of some larger identifiable population of states, then researchers also have the option of estimating the average impact of the program for this broader population (rather than the impact for states in the study only).

[64] See footnote 51 for the standard error of the pooled impact in a school-level experiment. This formula is based on the assumption that a hierarchical linear model is used to estimate the impact of the program (students nested within schools).

[65] Assuming that impacts are estimated using a multilevel model, the strict interpretation of the pooled estimate lies somewhere between the impact for the average school and the impact for the average student, because precision is a function of (1) the between-state variation in the outcome and (2) the variation between students within the state. However, the latter is often the larger of the two components, and it is a function of the number of students in the state.

If researchers choose to generalize their pooled finding in this way, then they must account for an additional factor when averaging impacts across states: the extent to which the impact of the program varies across states. Conceptually this makes sense, because the assumption here is that the study states were randomly sampled from some larger population of states. Had a different set of states been sampled for the study – and assuming that the impact of the program varies by state – then the estimated impact would be different. The greater the variation in impacts across state, the more the pooled impact depends on which states are included in the study, and the greater the uncertainty about whether the estimated pooled impact is an accurate estimate of the true average impact of the program in the broader population of states to which we want to generalize. Hence, in order to correctly infer whether the program is effective for the larger population of states, variation in the impact of the program must be built into the standard error of the pooled impact estimate.

In this context, the optimal aggregation weight is a "random effects" weight that is based on a combination of precision as well as the amount of true variation in program impacts across states:[66]

$$w_{sg} = \frac{\frac{1}{(se\_ES_{sg})^2 + V}}{\sum_S \frac{1}{(se\_ES_{sg})^2 + V}} \qquad (4)$$

where V is the true variation of the program's impact across states.[67]

The average "random-effects" impact and its standard error can be estimated manually, by substituting the random-effects weight (4) into Equations (2a) and (2b). However, estimating V is labor-intensive,[68] so in practice, it is simpler to estimate the average impact by fitting a multi-level one-step regression model where the effect of the treatment is allowed to vary across the states in a study:

[66] Huedo-Medina *et. al.* (2006).

[67] In studies that span multiple grades (like Studies A and B), *in theory* one could also generalize the results to a broader sample of *grades*, in which case V would be the variation in impacts across grades and states. In practice, however, grade levels cannot be assumed to be a representative sample from some larger population of grades. For multi-grade studies, the weight in (4) applies a regular precision-weight to states

[68] A commonly used estimator for the between-state variance is:

$$V = \frac{Q - (k-1)}{\sum w_s - \frac{\sum w_s^2}{\sum w_s}}$$

where w_s is the fixed-effects weight for state s, and where Q is the homogeneity statistic, defined as:

$$Q = \sum_S w_s ES_s^{\,2} - \frac{\sum (w_s ES_s)^2}{\sum w_s}$$

In this context, Q measures the amount of variation in effects across states, where each state's impact estimate is weighted by its precision (Lipsey & Wilson, 2001).

$$Z_i = \beta T_i + \sum_{S,G} \lambda_{sg} ST_{si} * GR_{gi} + \sum_{S,G} \sum_{M} \delta_{msg} X_{mi} * ST_{si} * GR_{gi}$$
$$+ u_s * T_i + \varepsilon_i$$

Where:

$u_s * T =$ A state-specific error term for students in the treatment group. This error term represents the random treatment effect for each state

Therefore:

$\beta =$ The average impact of the program on state test scores *in the broader population of states from which states in a given study have been sampled.*

Having laid out the options, now consider the difference between standard precision weights (3a) and random-effects weights (4):

- ***Random-effects weighting* (4):** Given its more generalized inference, the random-effects approach incorporates two sources of uncertainty about the pooled impact of the program: (1) student-level sampling error[69] (as measured by the standard error of each impact estimate) and (2) uncertainty arising from state-level sampling (as represented by the amount of true variation in program impacts across states).

- ***Precision or fixed-effects weighting* (3a):** When researchers limit their inferences about program effectiveness to sites *in the study states*, then the only relevant source of uncertainty about the pooled impact estimate is student-level sampling error. Therefore, it is not necessary to account for *state-level* sampling error, because inferences about program impact are not generalized beyond the study sample.[70]

The extent to which these two aggregation approaches (and types of inference) will yield different impact findings depends on the amount of variation in program impacts. If there is no true variation in the impact of the program across states (V=0), then the average impact of the program is the same regardless of which states are included in the study. Hence, the average impact of the program *for the study states* will be the same as the average impact for the *broader population of states* (in other words, the two weighting methods will produce the same findings). Conversely, if there is true variation in program impacts across states (V is larger than zero), then one would expect the standard error of the pooled impact estimate for the broader population (random-effects

[69] Student-level sampling error includes (i) uncertainty due to the random assignment of students to the treatment/control group, and (ii) uncertainty due to the subsample of students for whom outcome data are available (less than perfect response rates).

[70] Viewed otherwise, standard precision-weighting assumes that variation in impacts across states is irrelevant (whether because inferences about program impacts are confined to the states in the study, or because impacts are homogeneous).

weighting) to be larger than the standard error for the pooled impact estimate for the study states (precision weighting), since the former incorporates two sources of error.[71]

Following from these points, there are several precision-related issues to consider when random-effects weighting is used. In particular, it is important to think carefully about the issue of sample size:

- **Number of students:** Generalized inferences are less precise than local inferences. Therefore, if researchers decide to estimate the "generalized" impact of the program (that is, use random effects weights), then a larger number of observations (students or states) will be required to achieve equivalent statistical power to that for a study aimed at estimating "local" impacts (that is, using fixed-effect weights).

- **Number of states/grades:** The conceptual difference between precision weighting and random-effects weighting is grounded in there being true variation in impacts across states/grades (that is, V is not equal to 0). In practice, however, the statistical power for detecting true impact variation is dependent on the number of states/grades in the study (that is, the number of impact estimates being aggregated). If the study includes few grades/states, then the study will only be able to detect true variation in impacts of large magnitude. This means that in practice, random-effects weighting is less likely to be usable in studies with only a few states/grades, even if generalized inferences are warranted in theory.[72] A practical way to determine whether random-effects weighting is a feasible option is to test whether there is statistically significant variation in impacts across states and grades. If a statistical test indicates that estimated variation V is not reliably different from zero, then this indicates that either (a) true impacts do not vary across states and grades or (b) that given the number of states it is only possible to detect large impact variation. In either case, the use of fixed-effects weighting is warranted.

Thus, there are two conditions for using random-effects weighting: (i) the study states must be representative of some larger identifiable population, and (ii) there should be statistically significant variation in estimated impacts across states/grades (which is less likely when few states are included in the study).

These two conditions are not met in the four experiments used in this paper, which means that precision weights are preferred. Later in this paper, we use data from the four experiments to compare the two approaches.

[71] See Appendix C for further discussion of differences between precision weighting and random-effects weighting.

[72] There is no clear guidance in the meta-analytic literature about the minimum number of impact estimates required for random-effects weighting, because it depends on the amount of true variation in impacts that can reasonably be expected (which in turn depends on which states are included in the study and the nature of the intervention).

6 The Sensitivity of Impact Findings to Using State Tests

In this section, we present impact findings for each of the four randomized experiments, to see what they can tell us about: (i) the sensitivity of impact findings to the type of assessment used to measure student achievement (state tests or study-administered test), and (ii) the extent to which the overall impact of the program in state tests is sensitive to the choice of rescaling function and/or aggregation weights.

For each of the four experiments, we estimated the average impact of the program on state test scores and study-administered test scores, based on different combinations of rescaling approaches and aggregation weighting. This means that the impact model used for the analysis varies by rescaling/weighting method, as well as by study.[73] However, the impact models all share the following features:

- *Rescaling method:* State test scores are rescaled based on the linking functions described in Section 5.2.1, so all impact estimates are represented as effect sizes. In order to make the results comparable across test type, estimates for the study-administered test are also rescaled in effect size units, based on the mean and standard deviation of scores for students in the sample.[74]

- *Covariates:* All impact models control for random assignment blocks (schools or districts) to reflect the design features of each study. The impact model also controls for student achievement at baseline (as measured by the study-administered baseline test), to improve the precision of the impact estimate. For Studies A, B, and C, student-level baseline pretest scores are used, while for Study D, both school-level and student-level baseline pretest scores are included.

- *Missing data:* Observations with missing outcomes are dropped from the analysis. Missing values for the covariates (pretest scores) were imputed using the "dummy variable" approach.[75]

The analysis sample used to estimate impacts is restricted to students in each of the four experiments who have both a study administered test score *and* a state test score at follow-up. For Studies A, B and D, the analysis sample was further limited to states where the *state-wide* mean and standard deviation in test scores on the state test is known. This restriction makes it possible to use a consistent sample of students to examine the sensitivity of impact findings to z-scores based on the

[73] Appendix C provides the model specification for each study.

[74] Information on the state-wide standard deviation of tests scores is not available for the study-administered tests.

[75] This involves a) creating a missing indicator that equals 1 if the value of the variable is missing and 0 otherwise; b) creating a new covariate that equals the value of the original variable if the value is not missing and equals zero (or any constant) if the value is missing; and c) including the missing indicator and the new covariate (instead of the original covariate) in the impact model. See Puma *et al.* (2009) for a more detailed discussion of this approach and its usage in randomized experiments.

sample *vs.* the state-wide distribution in scores.[76] For Study C, however, this additional sample restriction is not imposed because the state-wide mean and standard deviation in scores is only available for one of the study states. Hence, for Study C, we cannot compare impact findings across these two z-scoring reference populations.

The remainder of this section discusses the general pattern of impact findings, in terms of what they tell us about whether and how to use state test scores in multi-state experiments. (Detailed impact tables are located in Appendix D.[77]) As explained in Section 2, the impact results presented in this section should not be compared to impact findings reported in the four studies' official reports. Both the samples and the impact models used in this analysis have been simplified to make it easier to compare findings across the four experiments, and therefore differ from those used in the official reports.

6.1 Sensitivity to Assessment Type

In this section, we examine the sensitivity of impact findings in each of the four experiments to the type of assessment, by comparing the estimated impact of the program on *state test scores* to its estimated impact on *study-administered test scores*. We compare the results for the two types of assessment, to see whether the observed pattern of impact findings conforms to what we would expect to see based on differences (or similarities) between the two types of test.

The discussion in this section is based on Figure 6.1, which compares estimated impacts on the study-administered test to the estimated impact on state test scores, for each of the four experiments. The figure compares the magnitude, standard error, and p-value of the impact estimates across the two types of test. To simplify the comparison across tests and to deduce general patterns, the findings in this figure focus on a specific aggregation method, precision/fixed-effects weighting (based on a one-step regression approach). As discussed in the previous section, this is the most appropriate weighting method for the four experiments.[78] Impacts on the study-administered test are converted to effect sizes by dividing the impact on raw scores by the standard deviation in scores for students in the sample. This means that the relevant comparison for estimated impacts on the study test is against the estimated impact on state test scores rescaled using *sample-based z-scores*.[79]

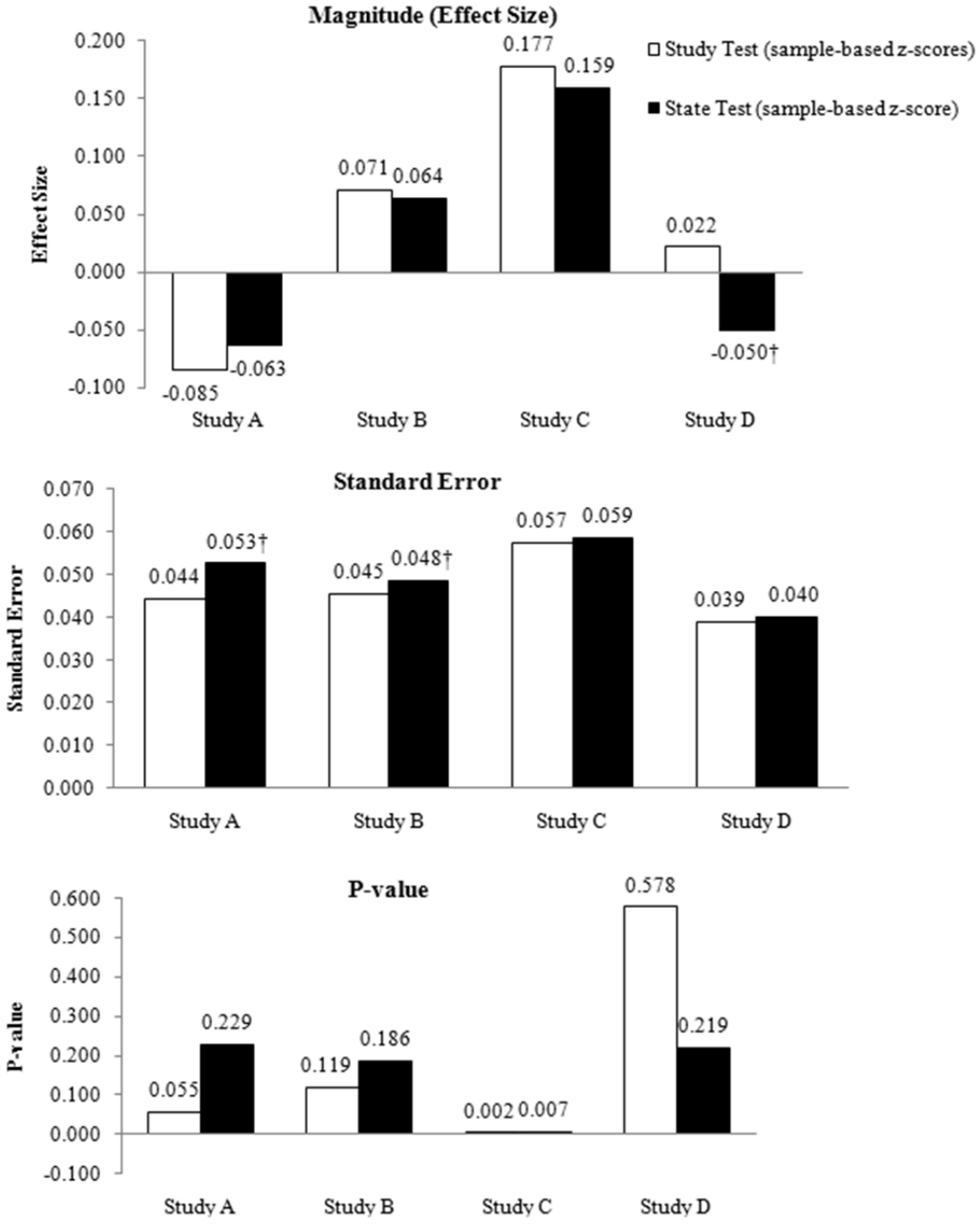
[76] In practice, this additional restriction does not affect the sample for Study D, because state-wide means/SDs are known for all 9 states in the study. In Studies A and B, however, this restriction limits the number of states in the analysis sample: state-wide means/SDs are known for 9 states (of 10) in Study A and 7 states (of 9) in Study B. Appendix D presents impact findings for all states in Studies A and B, when the second sample restriction is not imposed; the general conclusions are similar to the ones discussed in this section.

[77] See also Appendix E for correlations between the student achievement measures used in this paper.

[78] The pattern of results discussed in this section is consistent across aggregation/weighting methods (see Appendix D).

[79] The estimated impact on *state-based* z-scores is not comparable to the estimated impact on study-administered test scores because these two sets of estimates have a different interpretation. Similarly, impacts on test scores rescaled using the rank-based method are also not comparable, because even though this rescaling method is sample-based, it uses a non-linear transformation that affects the distribution of test scores.

Figure 6.1
Impact Findings by Type of Assessment
(Study-Administered Test or State Tests)



SOURCE: Authors' analysis based on data from four experiments.
NOTES: All impact estimates are pooled using fixed-effects weighting (one-step regression approach). The sample sizes used in the analyses are: 1,032 students across 9 states for Study A, 944 students across 7 states for Study B, 1,065 students across 4 states for Study C, and 4,387 students across 9 states for Study D. The statistical significance is indicated (†) when the p-value for the *difference* in the parameter of interest (magnitude, standard error, p-value) between the two types of assessment is less than or equal to 5 percent.

As will be discussed in greater detail in this section, in some cases one would expect impact findings to differ across the two types of assessment; in other cases, the results should be similar, and in yet other situations we have no prior expectation about comparability. Therefore, to formally assess whether the results are different or not, statistical tests were conducted to examine whether the magnitude, standard error and p-value of program impacts differs across outcome measures by a statically significant amount. The results of these tests are reported in the figures in this section and in Appendix F.

A. Precision

As discussed in Section 3, if state tests are to be used in an evaluation, then it is important that they provide a reliable measure of the outcome of interest for the population of students targeted by the intervention, because reliability affects the precision of estimated impacts. From the perspective of evaluators, an important concern is that using state tests to measure achievement could compromise the statistical power of the analysis. If state tests are less reliable, then the estimated impact will be less precisely estimated, which will decrease the study's ability to reliably detect an impact of given magnitude. All else equal, one might expect the conditional reliability of the study-administered test to be higher, because evaluators have the flexibility of choosing an assessment that is reliable for low-performing students.[80]

To examine whether these concerns bear out in practice, we can compare the *standard error* of the estimated impact across the two types of test. Although this comparison does not provide information on the conditional reliability of the tests, it does yield information about how reliability may affect the relative precision of estimated impacts on the two types of test, which is an important question for evaluators.

Figure 6.1 shows that the precision of estimated impacts on state tests is not necessarily lower than that of a study-administered test. As seen in the middle panel of Figure 6.1, the estimated impact on state test scores is less precise in two of the four studies. While the standard error for the study-administered test is numerically smaller for all four studies, these differences in precision are statistically significant for Studies A and B only. For Studies C and D, differences in the standard error are not statistically significant.

Interestingly, the impact findings Figure 6.1 also corroborate the descriptive findings on floor effects reported in Section 3.2. Recall that a floor effect was observed for Study A, but not for Studies B, C, and D:

- **Floor effect in the distribution of test scores (Study A):** The standard error of the impact estimate is 19 percent greater for state tests than for the study-administered test (a statistically significant difference).
- **No floor effect in the distribution of test scores (Studies B, C, D):** The standard error of the impact estimate is 7 percent, 3 percent, and 4 percent greater for state tests than for the

[80] In Study D, for example, the study-administered test is computer-adaptive.

study-administered test, respectively.[81] In Studies C and D, this difference in precision is not statistically significant.

These findings suggest that precision is not necessarily lower for estimated impacts on state tests (as shown in Studies C and D). The findings also suggest that in cases where precision *is* substantially lower (such as Study A), a descriptive analysis of state test scores can be used to find floor effects and to identify instances in which state tests may be insufficiently reliable for the evaluation.

B. Inferences about program impacts

Having examined the precision of estimated impacts, we next examine the sensitivity of inferences to the type of assessment. As explained in Section 3, state tests can provide valid inferences about program impacts if they are aligned with the outcomes that the intervention is meant to affect. If state tests are not aligned with the outcomes interest, then the evaluation may incorrectly conclude that the program does not improve student outcomes, or it may underestimate the magnitude of true impacts.

Inferences about program effectiveness depend on the magnitude of the estimated impact, as well as its standard error. The size of the impact is estimated with error and so in practice, inferences about program effectiveness are based on whether or not the estimated impact is statistically different from zero. Thus, the precision (standard error) of the impact findings for state tests – compared to the study-administered test – is also an important factor when comparing their magnitudes. As seen in the previous section, precision does not differ across test types in Studies C and D, but it does differ in Studies A and B.

For this reason, we focus on the p-value of the impact estimates because it takes into account both the magnitude and precision of the findings. Figure 6.1 compares the p-value of the estimated impact on state test scores and study-administered test scores. It is important to compare the p-values directly, rather than looking at whether both p-values are lower than 5 percent. This is because the choice of a fixed alpha level (say 5 percent) can create the illusion of a difference between state tests and the study test where there is none. The estimated impact on state test scores is less precise, so one can imagine a situation where researchers could conclude that program impacts on state tests are *not* statistically significant at the 5 percent level, but that impacts are significant based on the study-administered test. However, because some of the difference in the p-value across test types is due to random noise, differences in conclusions about program effectiveness between the two test types would simply be due to the fact that researchers have to choose a fixed alpha level for deciding whether or not the program is effective. It would not be due to a meaningful difference between the suitability of the two types of test and their ability to yield valid inferences about program impacts.

We can start by looking at Studies A and B. Recall that in these experiments, the study-administered test and state tests both measure "general achievement" and should produce similar

[81] These percentages are based on the tables in Appendix D. Discrepancies with Figure 6.1 are due to rounding error.

impact findings, all else equal. In the previous section, however, we found that for these two studies, the estimated impact on state test scores is less precise than the estimated impact on study-administered test scores. This suggests that even though state tests measure general achievement on average, they may do so with more error than the study-administered test.

However, the results in Figure 6.1 indicate that in Studies A and B, the lesser precision of state tests does not lead to a meaningful difference in the p-value between the two test types:

- **Studies A and B:** Inferences about program impacts on general achievement are not sensitive to the type of test, despite differences in the precision of the impact estimates. Turning next to Studies C and D, recall that the study-administered test in these two experiments is used to measure the specific skill targeted by the program, while state tests are used to measure general achievement (a less proximal outcome). In the previous section, it was found that in Studies C and D, the precision of impacts on the study-administered test and state tests is similar. This means that even though the two types of test measure different skills, they appear to do so with similar amounts of error.

Therefore, for these two studies, we can focus on comparing the magnitude of estimated impacts on state tests and the study-administered test. In particular, we look at whether the pattern of findings across the two test types conforms to what one would expect to see given the difference in their content. That is, the estimated impact on state test scores (which measure general achievement) should be equal or smaller in magnitude than the estimated impact of the program on study-administered test scores (which measure the targeted skill).

The results in Figure 6.1 show that the impact findings for these two studies do indeed conform to the expected pattern:

- **Study C:** For Study C, the program has a statistically significant impact on both the study-administered test (targeted outcome) *and* on state tests (general achievement). The latter estimate is smaller in magnitude, but not by a statistically significant amount.

- **Study D:** For Study D, the program does not have a statistically significant effect on the study-administered test (targeted outcome), nor does it improve performance on state tests (general achievement), as one would expect given the theory of action. Although the program did not affect either outcome, the magnitude of the estimated impact on state test scores is statistically smaller than the estimated impact on study-administered test scores, as one would expect since the latter assessment is more general.

This general pattern of findings is consistent across all aggregation methods (see Appendix D for detailed impact tables).

6.2 Sensitivity to Linking Function and Aggregation Weights

In this section, we examine whether the estimated impact of a program on state test scores is sensitive to choice of rescaling method and/or aggregation weights.

A. Linking Function

Following the presentation order in Section 5.2.1, we begin by focusing on linear rescaling function, and in particular we look at whether the average impact estimate is sensitive to the choice between converting state test scores to z-scores using the *state-wide* distribution *vs.* the *sample* distribution in scores. We then look at the sensitivity of findings to the choice between linear *vs.* non-linear rescaling functions (that is, between traditional z-scores and rank-based z-scores).

Linear Rescaling: Z-scores Based on the State vs. sample Distribution

As discussed in Section 5.2.1, the decision between sample-based and state-based z-scores does not affect the p-value of the impact estimate for each state/grade, because the estimate and its standard error are rescaled by the same amount. When combining impact estimates across states and grades, however, this conclusion may not hold, because the choice of rescaling method can affect the relative weight of each state/grade in the combined result. Therefore, when comparing impact findings across these two rescaling options, we mainly focus on the difference between their *p-values*. We also look at the magnitude and standard error of the impact estimates, but only to better understand any difference between the p-values.

Figure 6.2 compares the estimated impact on sample-based z-scores and state-based z-scores. As in the previous section, we simplify the comparison by focusing on the precision-weighted result (fixed-effects one-step regression). Recall that Study C is excluded from this analysis because state-wide test score information is only known for one state in the study. The main finding is that:

- **Sample-based *vs.* state-based z-scores:** For the studies examined in this paper, inferences about program impacts are not sensitive to the choice of reference population used for z-scoring.

The top panel of Figure 6.2 shows that, relative to the findings for state-based z-scoring, the magnitude of the impact estimate is larger when z-scoring is based on the sample distribution, as expected, though these differences are not statistically significant. The standard error of the impact estimate is also larger as expected, in this case by a statistically significant amount. However, this difference in the standard errors is not sufficiently large to lead to different conclusions about program effectiveness.[82]

It is important to note the p-values do not differ by a statistically significant amount between the two rescaling approaches. Because the *size* of the p-values is larger, in practice conclusions about whether the program has a statistically significant impact on state test scores *could* be sensitive to the choice of reference population, were p-values for the four studies closer to the margin (5 percent). In particular, we could conclude that the program improves state test scores when state-based z-scoring is used, but not when sample-based z-scoring is used. However, because differences in the p-values are due to random noise (as indicated by the finding that they are do not differ statistically between the two rescaling methods), any difference in conclusions about program effectiveness would simply

[82] This result is consistent across aggregation methods (see Appendix D for detailed impact tables).

be due to the fact that researchers have to choose an specific alpha level for deciding whether or not the program is effective. It could not be attributed to a meaningful difference between the inferences produced by the two rescaling approaches.

Linear Rescaling (Z-scores) vs. Non-Linear Rescaling (Rank-Based Z-scores)

As described in Section 5.2.1, the nonlinear rank-based linking function yields rescaled scores that are normally distributed, while the "traditional" linear method of z-scoring preserves the original distribution of raw scores. Therefore, one would expect the estimated impact on state test scores to be most sensitive to the choice of rescaling method when the distribution of state test scores is non-normal.[83] As discussed in Section 5.2.1, the distribution of state test scores is not consistently normal in the states included in the four experiments. On the other hand, sample sizes are large enough that violations of non-normality may not matter.

To examine this issue further, Figure 6.3 compares the estimated impact on sample-based z-scores (linear rescaling) and rank-based z-scores scores (non-linear rescaling). As in the previous section, we focus on the precision-weighted results (fixed-effects) for simplicity. Note that sample-based z-scores are the appropriate reference point here (as opposed to state-based z-scores), because rank-based method also rescales scores relative to the sample distribution.

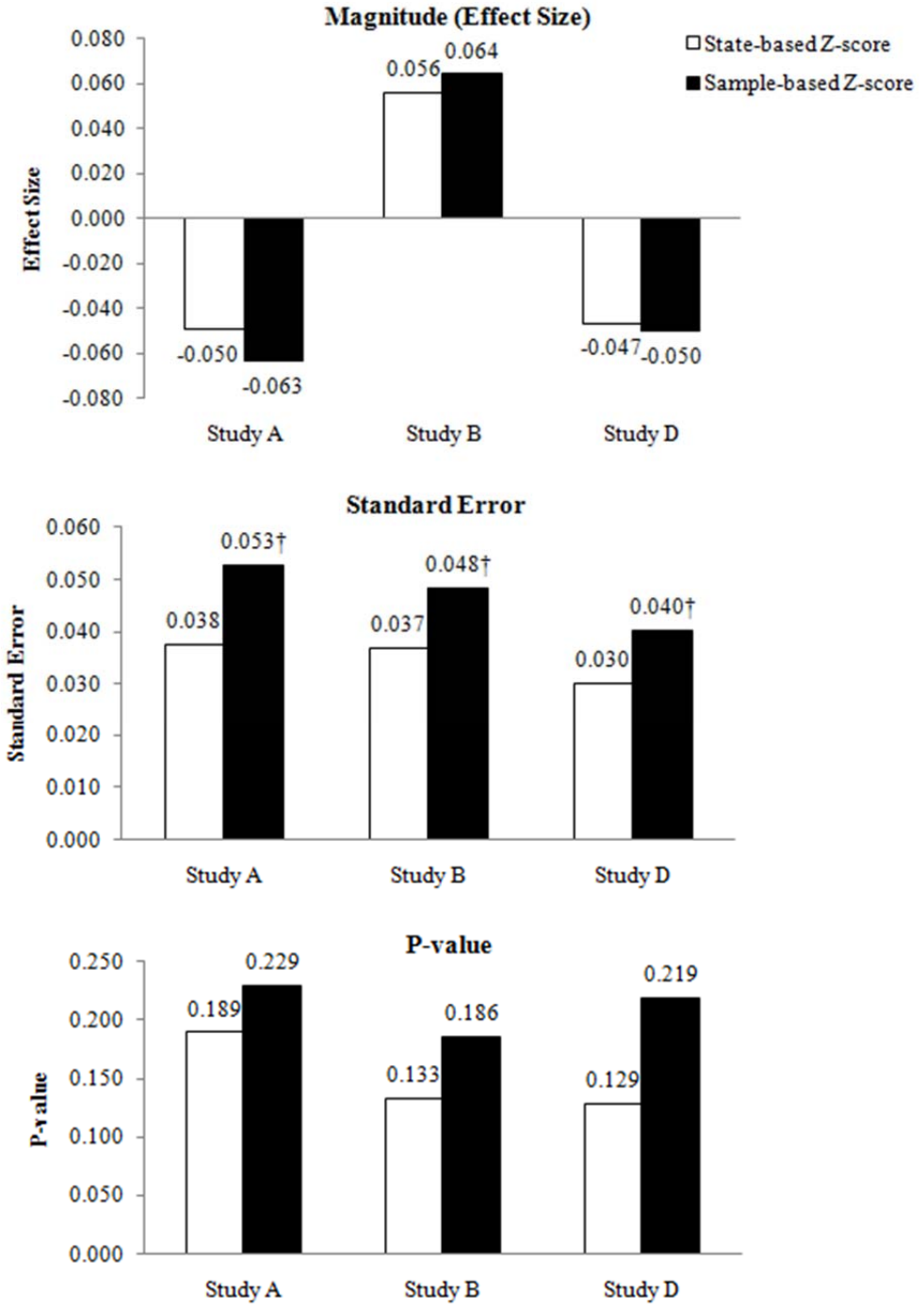
When comparing the two sets of results in this figure, attention should be drawn in particular to the *p-value* of the impact estimate. When data are non-normal, the impact estimate and its standard error are unbiased, but the distribution of the estimated impact is non-normal. This means that using T and F statistics to make inferences about program impacts may not be correct (i.e., p-values based on these statistics may not be accurate). The general pattern of results is the following:

- **Linear vs. non-linear rescaling:** For the four studies examined in this paper, inferences about program impacts are not sensitive to the decision to rescale scores using a linear or non-linear function. This conclusion also holds for the studies where the distribution of scores appeared to be least normal. This result may be due to the larges sample sizes in these studies, which make the results robust to non-normality.[84]

As seen in Figure 6.3, the magnitude of impacts estimates does not differ by a statistically significant amount across the two rescaling methods. In Studies A and C, the standard error of the estimated impact is smaller for rank-based (non-linear) z-scores than traditional (linear) z-scores (by a statistically significant amount). However, in all four experiments, the p-value does not differ by a statistically significant amount between rescaling functions. This means that even though in practice, the two rescaling methods *could* yield different conclusions about whether or not the estimated effect is statistically significant, these differences would simply be due to chance and the fact that decisions about statistical significance are based on a set cut-off (5 percent). It would not be due to a meaningful difference between the results of the two approaches.

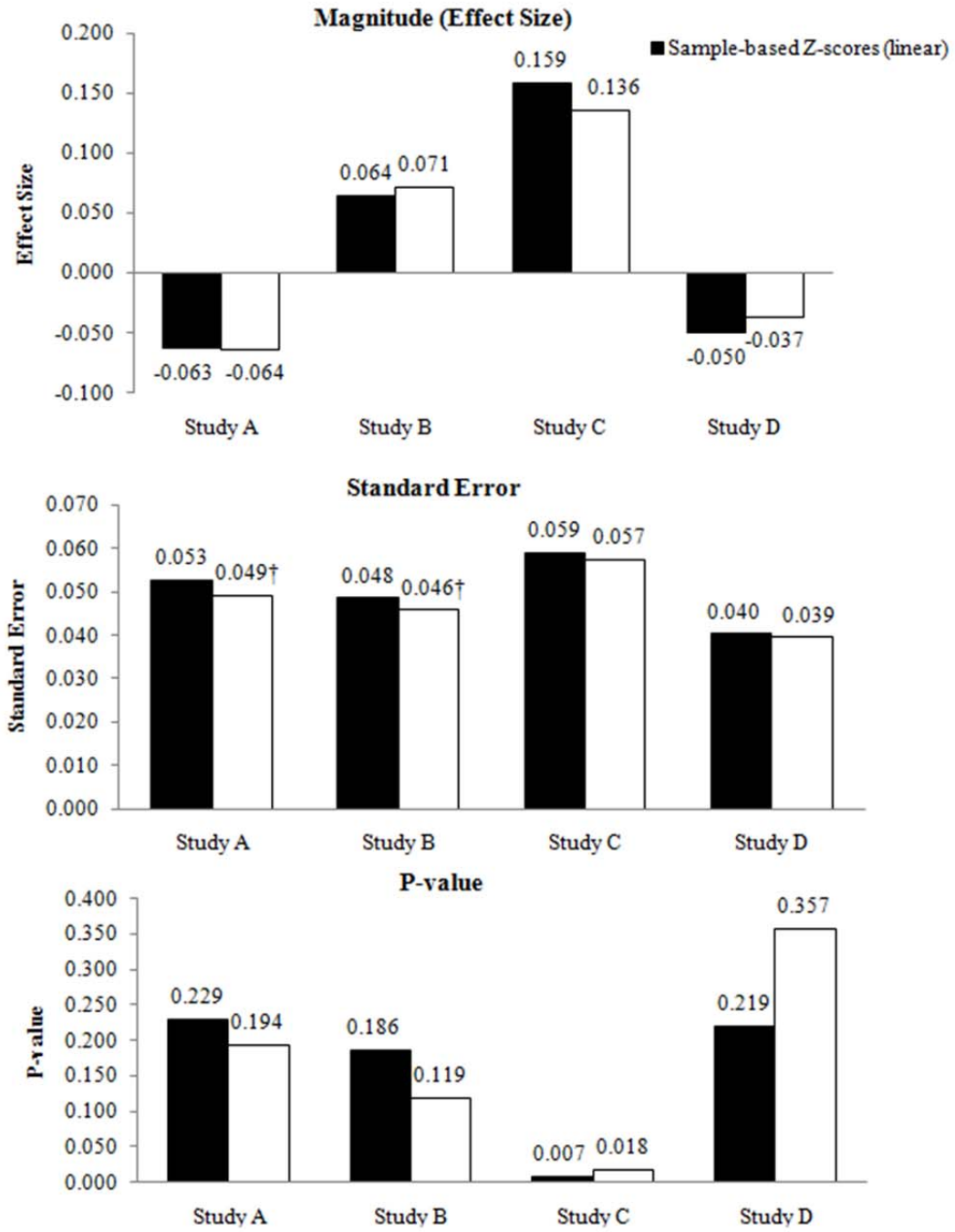
[83] When test scores are normally distributed, the two methods are equivalent.
[84] These results are consistent across aggregation methods (see Appendix D for detailed impact tables).

Figure 6.2
Impact on State Test Scores by Reference Population Used for Linear Rescaling (State or Sample)



SOURCE: Authors' analysis based on data from four experiments.
NOTES: All impact estimates are pooled using fixed-effects weighting (one-step regression approach). The sample sizes used in the analyses are: 1,032 students across 9 states for Study A, 944 students across 7 states for Study B, and 4,387 students across 9 states for Study D. The statistical significance is indicated (†) when the p-value for the *difference* in the parameter of interest (magnitude, standard error, p-value) between the two rescaling methods is less than or equal to 5 percent.

Figure 6.3
Impact on State Test Scores by the Functional Form of the Rescaling Method (Linear or Non-linear)



SOURCE: Authors' analysis based on data from four experiments.

NOTES: All impact estimates are pooled using fixed-effects weighting (one-step regression approach). The sample sizes used in the analyses are: 1,032 students across 9 states for Study A, 944 students across 7 states for Study B, 1,065 students across 4 states for Study C, and 4,387 students across 9 states for Study D. The statistical significance is indicated (†) when the p-value for the *difference* in the parameter of interest (magnitude, standard error, p-value) between the two rescaling methods is less than or equal to 5 percent.

It is important to point out, however, that these results may not be applicable to one-state studies – or to studies with few students per state. In these situations, the impact findings could be more sensitive to violations of the normality assumption.

B. Aggregation Weights

We next examine the sensitivity of the average impact finding to the choice of aggregation weight. Following the discussion in Section 5.2.2, we begin by comparing the weighting approaches that are used to estimate the impact of the program for *states in the study*. We then examine the impact findings when random-effects weights are used to generalize the findings to a *broader population of states*. Note that formal statistical tests of differences in impact findings across aggregation methods were not conducted; as explained in Section 5.2.2, the choice of aggregation method affects the type of inference that is made, which means that impact findings are not directly comparable across weighing strategies.

Average Impact for States in the Study Sample

If researchers want to estimate the impact of the program for the study sample, their first option is to weight the impact estimates for each state/grade by their precision (inverse of the squared standard error). As explained in Section 5.2.2, this weighting strategy can be implemented in two ways: the two-step (classical) approach or a one-step regression. The empirical results confirm that these two implementation approaches produce similar findings (see Appendix D for detailed findings):

- **Precision (fixed-effects) weighting:** For all four experiments, the classical (two-step) approach and the one-step regression approach generate impact estimates that are similar in terms of their magnitude, standard error, and p-value, as expected.

The second strategy for aggregating the impact estimates and obtaining the pooled impact is to weight by the sample size in each state/grade. As explained in Section 5.2.2, the implementation of this weighting strategy depends on the type of study design that is used:

- **Sample size weighting (*Student-level random assignment*)***:* In this type of design, the precision of impact estimates depends on the number of students in the analysis. This means that weighting the impact estimates by the number of students in each state should yield pooled findings that are very similar to those obtained from precision weighting. The empirical results confirm that for Studies A, B and D, the two approaches produce similar pooled estimates, in terms of their magnitude, standard error, and p-value. In practice, however, weighting by precision may be preferred because it can be easily implemented using the one-step regression approach.

- **Sample size weighting (*School-level random assignment*):** In this type of design, there are two options for sample size weighting: impacts can be weighted by the number of *students* or the number of *schools* in the state and/or grade. As explained in Section 5.2.2, the choice between these two approaches depends primarily on the type of inference that researchers

want to make – weighting by the number of schools produces the estimated impact for the average school in the sample, while weighting by the number students produces the impact for the average student in the sample. However, researchers should also consider how the *precision* of the pooled estimate (and therefore the minimum impact that can be reliably detected) is affected by their decision. In this regard, the empirical findings from Study D show that, as expected, the standard error of the pooled impact estimate is smaller when impacts are weighted by the number of students than when they are weighted by the number of schools; the standard error is even smaller when precision weighting is used. This suggests that precision weighting may be preferable to sample size weighting, but only if the study's goal is to estimate the impact for the average *student* in the sample; if researchers want to estimate the effect for the average school, then weighting by the number of schools is preferable because it will provide the desired inference. It is also worth noting that in practice, differences in the impact findings across weighting methods (precision weighting, sample size weighting) are not sufficiently large to lead to different conclusions about program effectiveness in Study D.

The results summarized above are consistent across rescaling methods (linear and non-linear functions).

Average Impact for a Broader Population of States

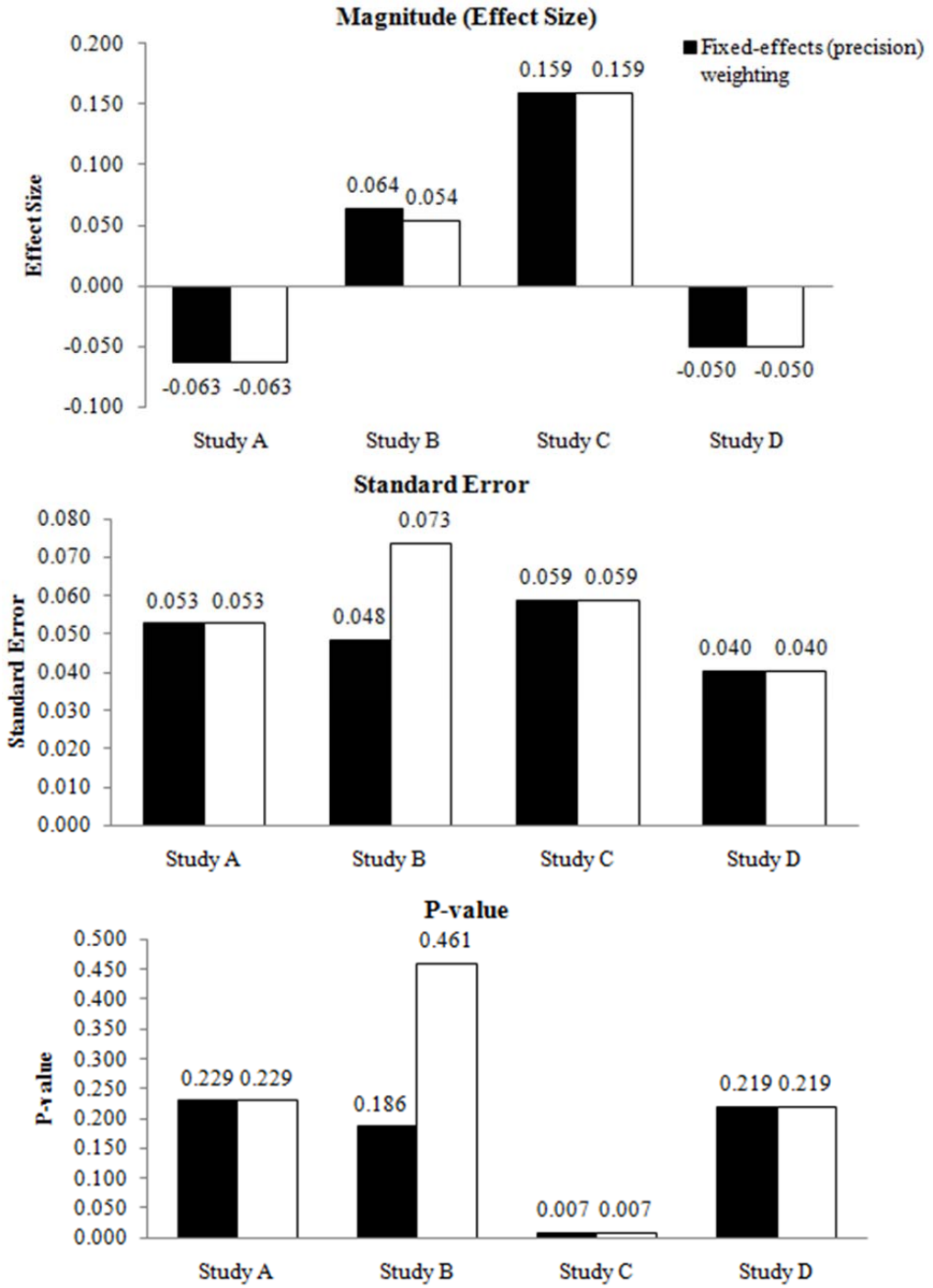
We next compare precision (fixed-effects) weighting and random-effects weighting. As explained in Section 5.2.2, the latter type of weight is used when researchers wish to "generalize" the findings to a broader sample of states. As explained earlier, none of the four experiments used in this analysis meet the criteria for using random-effects weights. Not only is there no statistically detectable variation in impacts across states (perhaps due to the small number of states in each study), but also these states are not a representative sample of some larger population of states.

However, as an exercise, we can compare the random-effects results to the precision-weighted (fixed-effects) results, to confirm that the standard error is indeed larger when random-effects weighting is used (as one would expect, see Section 5.2.2). We can also examine whether for the four experiments, conclusions about their "generalized" impact differs from conclusions about their "local" impact. This is especially relevant for Study C, where the local (fixed-effect) impact estimate is statistically significant, but where the generalized (random-effect) impact estimate may not be due to its expected lower precision. The overall finding is the following:

- **Fixed-effects *vs.* random-effects weighting:** For the four experiments in this paper, inferences about the generalized impact of the program (using random-effects weighting) do not differ from inferences about the local impact of the program (using precision-weighting). In other contexts, however, most notably in situations where program impacts *do* differ by a statistically significant amount across states, the two types of inference could differ.

Figure 6.4 compares the fixed-effects and random-effects estimates (both obtained using the one-step regression approach), focusing in particular on estimated impacts on sample-based z-scores.

Figure 6.4
Impact on State Test Scores by Weighting Strategy
(Fixed-effects weighting or random-effects weighting)



SOURCE: Authors' analysis based on data from four experiments.
NOTES: All test scores are rescaled as sample-based z-scores. Weighting is conducted using the one-step regression approach. The sample sizes used in the analyses are: 1,032 students across 9 states for Study A, 944 students across 7 states for Study B, 1,065 students across 4 states for Study C, and 4,387 students across 9 states for Study D. Statistical tests of the difference in the parameter of interest (magnitude, standard error, p-value) between weighting methods were not conducted.

We are mainly interested in the second panel of this figure, which compares the standard errors. Figure 6.4 shows that for Studies A, C, and D, the random-effects impact estimate is exactly the same as the precision-weighted estimate. This is due to the fact that variation in impacts across states (V) is not estimable for these studies, which is usually a sign that there is no true cross-state variation in impacts. For Study B, there is a small amount of variation in impacts across states. Although this variation is not statistically significant, it is used to calculate the standard error of the impact estimate. As a result, the standard error for the random-effects impact estimate is 51 percent larger than the standard error for the precision-weighted estimate.[85] However, the resulting increase in the p-value does not affect conclusions about whether the program's impact on state test scores is statistically significant. This is to be expected – the local (fixed-effect) impact estimate is not statistically significant, and therefore the generalized (random-effects) impact is not either, because of its lesser precision.

Because the difference in the standard error is quite large, however, in other contexts or studies, it is possible that one could find that the program has a "local" impact, but not a "generalized" one. This difference does not represent a flaw in either of the two weighting methods, simply the fact that they yield different types of inference.

[85]This percentage is based on the tables in Appendix D. Discrepancies with Figure 6.4 are due to rounding error.

7 The Precision of Estimated Impacts: Sensitivity to the Type of Assessment Used to Measure Student Achievement at Baseline

Even though in a randomized experiment it is not necessary to control for students' baseline achievement to obtain an unbiased estimate of program impacts, most studies do so to improve the statistical precision of the impact estimates.[86] Therefore, researchers face an important choice when it comes to deciding how to measure student achievement at baseline: should they administer their own pretest, or should they use students' state test scores from prior grades?

For researchers, the choice of baseline achievement measure has important cost implications for the study. On the one hand, using state tests at baseline would lower the cost of the study by not having to administer an additional test to students in the sample. On the other hand, if using state tests at baseline were to reduce the precision of impact estimates as compared with using a study-administered pretest, then a larger sample size would be needed to compensate for this reduction in precision, and as a result, the cost of the study could increase.

A key factor affecting the choice between a study-administered pretest and using students' prior state test scores is therefore the precision of the impact estimates. The precision that each of the test options can provide depends on how well the baseline test measure explains the variation in the outcome measure. To a large extent, this in turn depends on the correlation between the baseline and outcome measures.

Hence, in the remainder of this section, we examine the implications for the precision of impact estimates (and possibly inferences) of using state test scores rather than a study-administered test to measure achievement in the *baseline* period. Two of the four studies used in this paper (Studies C and D) collected baseline achievement information using both a study-administered pretest and prior state test scores, providing an opportunity to empirically explore the implications for precision of using state test scores or study-administered test to measure achievement in the baseline period.

In what follows, we provide a description of the analytic approaches employed in this section, followed by the empirical findings. The section concludes with a discussion of the findings.

7.1 Analytic Approach

The analysis focuses on two scenarios: one in which achievement at follow-up is measured using a study-administered test, and one in which achievement outcomes are measured using state tests. For each scenario, two kinds of baseline achievement measures were used as covariates—the study-administered pretest and the prior state test scores. The latter test scores are in different scales; therefore, in order to pool these scores across states, we standardized (z-scored) the baseline state test scores based on the sample mean and standard deviation for each state.[87] Since the study-

[86] See Bloom, Richburg-Hayes, and Rebeck Black (2007) for a discussion of using covariates to improve precision in randomized experiments.

[87] For Studies A and B, scores are standardized (z-scores) by state and grade. An alternative model specification would be to interact the original baseline state test scores with state/grade indicators to allow the coefficients for the baseline test to vary by state/grade. The results for this approach are very similar to the ones reported here.

administered pretest is uniformly administered across state/grade, the original scale of the pretest is used.

The impact analysis in this section has the following features:

- *Sample:* Like the other analyses in this paper, the sample is limited to students with both a study-administered test score and a state test score at follow-up (the impact analysis sample);
- *Outcome measures:* For this analysis, we focus on impacts on state test scores rescaled using *sample-based* z-scoring, since state-based z-scoring is not possible for Study C.
- *Estimation models:* Impacts are pooled across states using fixed-effects weighting as described in Section 5.2.2 (that is, weighted by precision), which is more appropriate than random-effects weighting for these studies. Estimates are obtained using the one-step regression approach.
- *Covariates:* All models control for random assignment block in order to account for the random assignment study design. In addition, the study pretest score or the (z-scored) state test score at baseline are included as a covariate in the model. No other covariates are used;[88]
- *Missing data*: Missing pretest values are imputed using the "dummy variable" approach.

7.2 Findings

The primary purpose of controlling for students' achievement at baseline is to improve the precision of impact estimates. Accordingly, Figure 7.1 compares the standard error of the estimated impact from three different models (see Appendix D for more detailed impact tables). The first bar presents the standard error from a model that only controls for random assignment blocks (that is, it does not control for student achievement at baseline); these results serve as a "benchmark" against which to compare the precision of impact estimates when baseline measures of achievement are used. The second and third bars present the standard error from a model that controls for the study pretest and the z-scored state test score at baseline, respectively. Each panel in the figure represents a different outcome measure: the first panel is for the study-administered test as the outcome, while the second row is for state tests as the outcome (using sample-based z-scoring). Formal statistical tests were also conducted to assess whether the precision of estimated impacts differs across baseline measures (see Appendix F).

- Using state tests as a baseline achievement measure yields precision gains that are at least as high as using a study pretest as a baseline measure.

Deke et al. (2010) conclude that on average, adjusting for *school-level* proficiency does not increase statistical precision as much as controlling for study-administered pretest scores. However, our results indicate that *student-level* state test scores can provide similar precisions gains as a

[88] As discussed in May *et. al* (2009), baseline state test scores can be used in two ways: they can either be used as covariates to improve precision, or they can be used to construct gain scores (in which case they are part of the outcome measure). We focus here on the former approach, since this is the primary purpose of baseline scores in an experiment.

baseline pretest.[89] For Study C and D, the realized precision gains from controlling for the study pretest or state test scores at baseline are similar in magnitude. Using state tests as a baseline covariate improves precision by 8 percent to 45 percent, relative to the unadjusted benchmark level). These precision gains compare favorably to those obtained when using a study-administered pretest (where precision gains range from 4 percent to 38 percent relative to the benchmark).[90] The greatest difference in the standard error of the impact estimate between these two types of baseline test is 0.008 (or 14 percent of the unadjusted benchmark precision level),[91] and the smallest difference is 0.002 (or 3 percent of the benchmark).[92]

Statistical tests provide further support that precisions gains are similar across the two types of baseline measure (see Appendix F). For Study D, the precision of the impact estimates does not differ by a statistically significant amount across baseline achievement measures (regardless of how the outcome is measured). For Study C, the same result holds when the outcome measure is a study-administered test. When state tests are the outcome measure, the precision of the impact estimate is statistically *greater* when state tests are used as the baseline measure rather than a study pretest.

This leads to another key finding:

- Precision gains are not necessarily greater when the follow-up test and the baseline test are the same type of measure.

Because correlations are typically higher across similar types of measures, one might expect precision gains to be larger when the baseline test and the follow-up test are of the same type (that is, when both the outcome measure and the baseline measure are state tests, or when both are the same study-administered test). However, the findings from Studies C and D do not uniformly bear out this expectation. In Study D, the reduction in the standard error is indeed greater when the follow-up test and the baseline test are of the same type, but these differences in precision are not statistically significant, and so it cannot be concluded that precision gains are greater when there is alignment between the pretest and the posttest. In Study C, precision gains are greater when both the outcome and baseline achievement measure are state tests. As noted above, when the outcome measure is state test score, precision gains are statistically greater when also using state test scores as a baseline covariate (a reduction of 10 percent in the standard error, compared to 4 percent when a study pretest is used). However, when the outcome is the study-administered test, the precision of the impact estimate is not statistically greater when the baseline measure is a study pretest (and in fact, the standard error is actually smaller when state test scores are used as the baseline covariate, though not by a statistically significant amount).

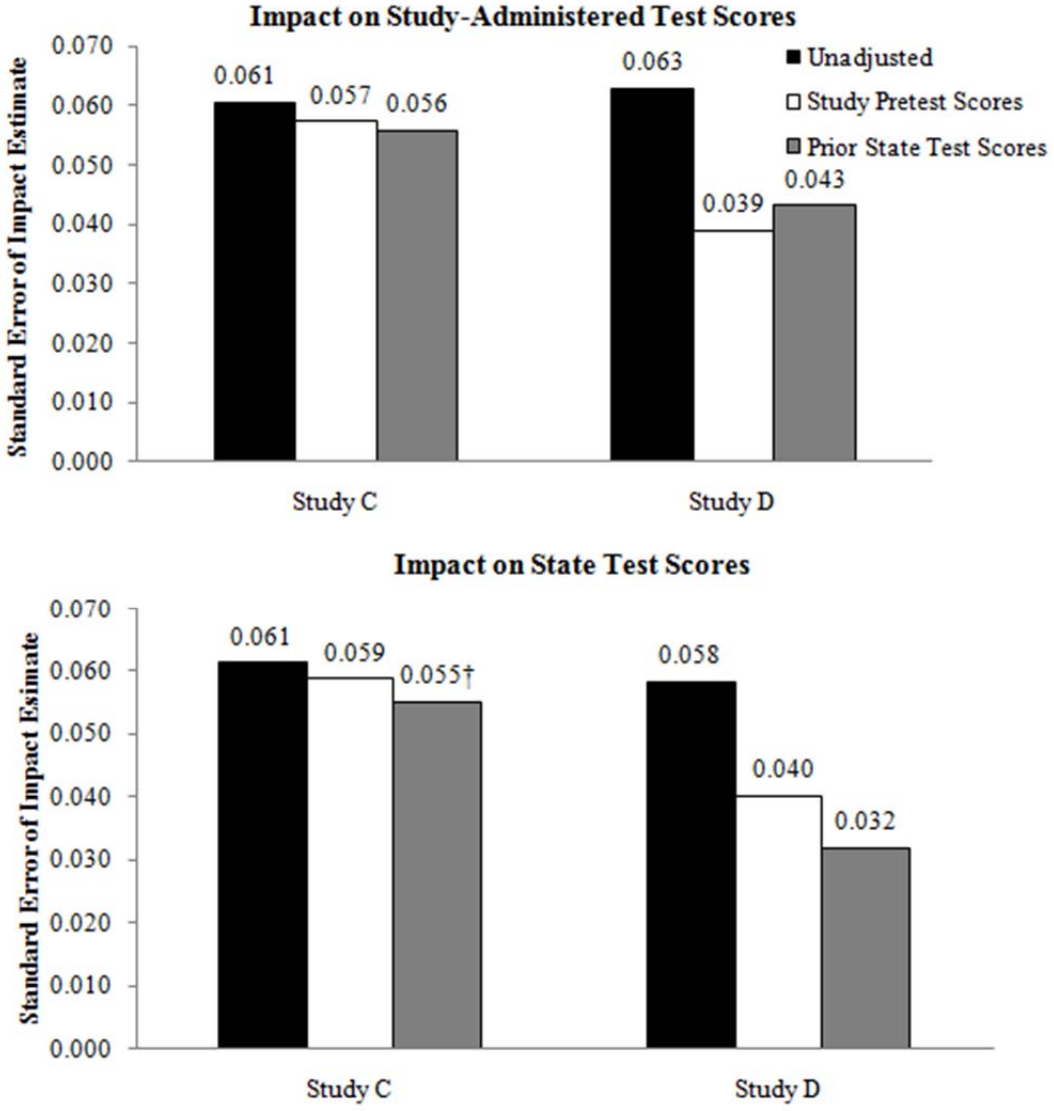
[89] This is because student-level state test scores can explain outcome variation both within and between schools, whereas school-level proficiency rates can only explain variation between schools.

[90] These percentages are based on the tables in Appendix D. Discrepancies with Figure 7.1 are due to rounding error.

[91] This is for Study D (state test as outcome).

[92] This is for Study C (study test as outcome).

Figure 7.1
Standard Error of the Estimated Impact, by Type of Baseline Assessment Scores Used as a Baseline Covariate (Study Pretest or Prior State Tests)



SOURCE: Authors' analysis based on data from four experiments.
NOTES: Impact estimates are pooled across states using fixed-effects weighting (one-step regression approach). The sample sizes used in the analyses are 1,065 students across 4 states for Study C and 4,387 students across 9 states for Study D. The statistical significance is indicated (†) when the p-value for the *difference* in the standard error across the two baseline measures is less than or equal to 5 percent.

Ultimately, the amount of precision that can be gained from controlling for a baseline achievement measure depends on how well this baseline measure explains variation in the outcome measure. Table 7.1 provides the correlation coefficients between the baseline measures and the follow-up measures of student achievement for Studies C and D.[93] These correlations are a useful summary measure of whether scores on one test are a good predictor of scores on another test.

For Study C, baseline state test scores are more highly correlated with both types of follow-up test than the study pretest. For example, the correlation between the study pretest and the study-administered follow-up test is 0.32, while the correlation between the baseline state test scores and the study-administered follow-up test is 0.41. This indicates that controlling for the state test scores at baseline helps to explain more of the variation in the outcome measure than the study pretest, which leads to smaller standard errors and better precision, regardless of how the outcome is measured.

For Study D, on the other hand, we observe that, in most cases, the correlation between the baseline and follow-up test is larger when both tests are of the same type. For instance, the correlation is 0.76 between the study pretest and the study follow-up test, while it is 0.64 between the state test scores at baseline and the study-administered follow-up test. This explains why, in Figure 7.1, we see that the precision gains is larger when baseline and follow-up achievement are measured using the same type of test.[94]

In general, one can also see that correlations between baseline and follow-up achievement measures are lower in Study C than for Study D, regardless of how achievement is measured at a given time point. This explains why overall precision gains are uniformly smaller for Study C (4 percent to 10 percent) than for Study D (31 percent to 45 percent).

This result may be due to range restriction. In Study C, program eligibility was limited to students 2 to 5 years below grade level, as opposed to Study D, where all students at the school were eligible. As a result of the eligibility criteria in Study C, student achievement is more homogeneous in this study, which means that there is less variation in the outcome for the baseline measures to explain. From a statistical perspective, this gives rise to range restriction, which reduces the correlation between baseline and outcome measures.

[93] More detailed correlation coefficient tables for these two studies are available in Appendix E.

[94] In addition to how student achievement is measured in the follow-up period, other factors may also affect the relative precision gains of the two types of baseline measure. These factors include the timing of the baseline test. In practice, the study pretest is usually administered at the beginning of the same school year in which the follow-up test is administered, while state baseline data are most likely collected in the spring of the previous school year. Therefore, the study pretest is closer to the follow-up test in terms of timing than prior state test scores. Another factor at play here is the sample of students that each type of baseline test is likely to capture. Study pretest scores are collected at the start of the school year and therefore cover a sample of students who are present in the study schools at that point in time. On the other hand, baseline state test scores are collected in the spring of the previous school year. Given that the student turnover rate is usually higher between school years than within a given school year, it is likely that the missing rate for baseline state test scores will be higher than for the study pretest. (We do not observe this for Study D because that study used a special sampling procedure to randomly select students for the study test at baseline and follow-up).

Table 7.1
Correlation between Student Achievement Measures (Baseline and Follow-up)
Studies C and D

Type of baseline test	Follow-Up Measures: Study-administered test scores		Follow-Up Measures: State test scores (sample-based z-scores)		Study-administered pretest scores	
Study C						
Study-administered pretest scores	0.323	*	0.283	*	1.000	
Prior state test scores (sample-based z-scores)	0.414		0.467	*	0.224	*
Study D						
Study-administered pretest scores	0.761	*	0.657	*	1.000	
Prior state test scores (sample-based z-scores)	0.644	*	0.680	*	0.665	*

SOURCE: Authors' analysis based on data from four experiments.
NOTES: The sample sizes used in the analyses are: 1,065 students across 4 states for Study C, and 4,387 students across 9 states for Study D. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

8 Conclusion

As noted in the introduction, the goal of this paper is to explore the practical implications of using state tests in an impact evaluation. We now consider the findings in light of our research questions, to see what lessons can be gleaned in terms of best practices for using state tests in a multi-state impact evaluation. When reviewing these results, it is important to remember that the findings from this analysis are applicable to *large-scale studies with multiple states*. In smaller studies with only 2 or 3 states, states that are outliers in terms of their assessment will have a greater influence on the overall results; therefore, in this situation impact findings may be more sensitive to using state tests to measure achievement.

Are state tests suitable for use in an impact evaluation?

In this paper, we started by examining whether state tests meet some of the necessary criteria for use in an evaluation. Specifically, we looked at whether the content of state tests is aligned with the intervention (which affects the validity of causal inferences about program impacts) and we examined whether there are floor effects in state test scores (which affects the precision of impact estimates). We then examined whether impact findings are sensitive to the type of assessment used to measure student achievement. In particular**,** we compared the pattern of estimated impacts on state tests and impacts on the study-administered test, to see whether impacts findings (with respect to inferences and precision) differ between the two types of assessment, and if so, whether these differences make sense given differences in the content and reliability of the two test types. The following key findings emerged:

- **Studies A and B (targeted outcome is general achievement):** For studies where the outcome of interest is general achievement (whether reading or math), our descriptive analysis suggests that the broad content of state tests makes them suitable for evaluating the effect of the intervention on a policy-relevant measure of general achievement. We find that for both of these studies, the standard error of the estimated impact on state test scores is larger than the standard error of the estimated effect on the study-administered test, which suggests that the reliability of state tests in these studies is less than that of the study-administered test. However, this difference in the precision of impact estimates is not sufficient to lead to differences in inferences about program impacts, as measured by the p-value, which does not reliably differ across the two types of test. This suggests that the reliability of states tests in these studies is not so low as to make them unsuitable for use in an impact evaluation.

- **Studies C and D (targeted outcome is a specific skill):** For studies where the targeted outcomes is a more specific skill, our descriptive analysis of state tests' content indicates that state tests are a good complement to a study-administered test, because they make it possible to look at the longer-term impact of a program on general achievement. In Study C, for example, using state tests as part of the evaluation makes it possible to show that the program improved students' general achievement, and not just the specific reading skill targeted by the program. Moreover, in Studies C and D, the standard error of the estimated impact for

state is not statistically greater than for the study-administered test, which suggests that the two tests are comparable with respect to the reliability of the outcome they are intended to measure.

Taking a step back, two key lessons emerge from this analysis. The first is that state tests can be a useful *complement* to a study-administered test: they can provide a policy-relevant measure of achievement, or they can be used to measure impacts along an intervention's entire theory of action. Whether state tests can be used as a *substitute* for a study-administered test is still an open question, however, and is likely to depend on the context. Based on the four experiments re-analyzed in this report, there are several obstacles to relying solely on state tests in an evaluation:

- *State tests are not available for all states and grades.* Annual state testing is not mandatory in early grades and in high school. For example, in Studies A and B, state tests data are not available for second grade students, while in State C, test data are not available for in states that do not test ninth grade students in English Language Arts.

- *Subtests are not always available to measure the targeted outcome.* For studies where the targeted outcome is a specific skill (Studies C and D), the findings from this analysis suggest that state tests are most feasibly used as a complement to a study-administered test. In these studies, state tests cannot be used to measure the specific skill targeted by the program (since subtest scores are not available for that skill), which means that state tests can only be used to measure the impact of the program on general achievement. In this situation, using *only* state tests scores in the evaluation would overlook a core piece of the program's theory of action. For example, had Study D only looked at impacts on state tests, it would not have been possible to determine whether this occurred because the program did not have an impact on the specific targeted math skill, or because gains on the targeted skill do not "translate" to gains on state test performance.[95]

- *Lower precision.* When state tests are being considered as a substitute for a study-administered test – which is most relevant in evaluations of programs that target general achievement – it is important to consider that state tests *may* lead to less precise impact estimates, as found in the four experiments reanalyzed in this paper. In Studies A and B, conclusions about program impacts are the same regardless of which type of test is used, which suggests that state tests could be used as a substitute. On the other hand, the findings also show that estimated impacts on state tests can be *less precise* than estimated impacts on a study test (Studies A and B). In these two studies, the lower precision of the state test findings is not consequential, as indicated by the finding that p-values do not differ across the two types of test. However, this result may not generalize to other experiments. It is possible that in other contexts, the reliability of state tests could be so low as to make them unsuitable.

[95] If state tests are used as a complement to a study-administered test, appropriate allowances have to be made for dealing with the multiple hypothesis testing problem (the greater the number of outcome measures, the greater the probability of making a Type I error, i.e., concluding that an impact estimate is statistically significant when in fact the program is not effective). One approach is to adjust the p-values for multiple testing; another approach would be to characterize state tests as a "secondary" measure of student achievement (Schochet, 2008).

Related to this latter point, the suitability of state tests cannot be implicitly assumed, because the results of this analysis may not generalize to all experiments. Thus, researchers must thoroughly investigate the suitability of state tests for their particular study, using the descriptive methods described in May *et al.* (2009) and illustrated in this paper. On this point, this paper arrives at a reassuring finding: descriptive analyses of state tests can provide useful information for assessing whether state tests satisfy basic requirements for use in an evaluation. For example, our descriptive analysis of floor effects correctly identified that the conditional reliability of state tests in Study A could be quite low. The impact findings supported this conclusion, showing that the relative precision of estimated impacts on state tests was lowest for Study A.

This highlights the importance of carefully examining the characteristics and content of state tests in the study design phase, when researchers are trying to decide whether and how to use state test data. In addition to the analyses conducted in this paper, researchers should also carefully review technical manuals on all state tests in their study, review the process used to construct the assessments, and undertake a thorough examination of the content and difficulty of the assessments.

Should impact findings be pooled across states and grades? Is the pooled impact on state test scores sensitive to decisions about how to combine scores across states?

With respect to the issue of *whether* to combine the findings, we find that even though the content of state tests differs across states and grades, program impacts do *not* differ (that is, the variation in impacts across states is not statistically significant), which makes it easier to interpret the average impact of the program.

In terms of *how* to combine the findings, our analyses indicate that the conditions for equating state test scores are not met, and that for this reason, impact findings should be combined using a "meta-analytic approach", as recommended by May *et al.* (2009). The meta-analytic approach requires making decisions about how to rescale test scores and weight the impact findings for each state and grade. An important question here is whether the average impact finding is sensitive to choices about how to rescale test scores to a common metric and how to weight each state's impact in the overall finding. The standard error and/or the p-value of the impact estimate in particular could be affected by these decisions.

In general, we find that for the four experiments, conclusions about the statistical significance of the impact estimate (at the 5 percent level) are robust to rescaling/weighting methods. Specifically:

- ***Choice of reference population for linear rescaling (z-scores):*** Findings from the analysis show that p-values (inferences) do not differ by a statistically significant amount across these two rescaling methods. Therefore, the decision about which reference population to use for z-scores can be based on the desired interpretation of the impact estimate. However, because the standard error of the impact estimate is larger when sample-based z-scoring is used, researchers should choose the reference population during the study design phase, so that the minimum sample size for achieving a given minimum detectable effect size (MDES) can be correctly determined.

- ***Linear scaling (traditional z-scores) vs. non-linear rescaling (rank-based z-scores):*** Findings from the analysis show that p-values (and inferences) do not differ by a statistically significant amount when a traditional (linear) z-score transformation is used as opposed to a (non-linear) rank-based function. This result also holds for the studies where the distribution of state test scores was non-normal in some states. This suggests that in studies with large sample sizes, impacts on state test scores are less sensitive to violations of non-normality, and that simple linear rescaling (traditional z-scores) is an acceptable approach for converting test scores to a common metric. However, if state test scores appear to be non-normally distributed – and if the sample size is small – then researchers may prefer the rank-based method of rescaling. Regardless of which method is chosen, researchers may want to also try the other approach as a sensitivity test.

- ***Precision (fixed-effects) weighting vs. random-effects weighting:*** For all four experiments in this analysis, precision weights are the preferred method for combining the findings, because the conditions for using random-effects weights are not met: (i) study sites are not randomly selected and (ii) the estimated variation in impacts across states is not statistically significant (most likely because there are too few states in the evaluation to reliably detect impact variation). In other studies, however, using random-effects weighting may be appropriate if the relevant conditions are met.

Is the precision of impact findings sensitive to the type of assessment used to measure achievement at baseline?

The final topic addressed in this paper pertains to the consequences of using of state tests from prior grades – rather than a study-administered pretest – to measure student achievement at baseline. Based on the two randomized experiments used to answer this question, we find that:

- The precision gains from using state tests to measure baseline achievement improves precision are at least as large as the gains from using a study pretest, in some cases. This result holds regardless of whether the outcome measure is a study-administered test or a state test.

These findings suggest that state tests can provide a cost-effective means of improving the precision of the impact findings, thereby making it possible to recruit a smaller number of students or schools into the study.

References

Bloom, H. S. (2002). *Measuring the Impacts of Whole School Reforms: Methodological Lessons from an Evaluation of Accelerated Schools.* Washington, DC: U.S. Department of Education (Doc #2002-10).

Bloom, H. S., Richburg-Hayes, L., & Rebeck Black, A. (2007). Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis , 29* (1), 30-59.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management , 27* (4), 724-750.

Corrin, W., Somers, M.-A., Kemple, J., Nelson, E., & Sepanik, S. (2009). *The Enhanced Reading Opportunities Study: Findings from the Second Year of Implementation (NCEE 2009-4036).* Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Deke, J., Dragoset, L., and Moore, R. (2010). *Precision Gains from Publically Available School Proficiency Measures Compared to Study-Collected Test Scores in Education Cluster-Randomized Trials* (NCEE 2010-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Walters, K., Song, M., et al. (2010). *Middle School Mathematics Professional Development Impact Study: Findings After the First Year of Implementation (NCEE 2010-4009).* Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Eductional Measurement and Practice , 8* (1), 35-41.

Hedges, L., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* San Diego CA: Academic Press.

Hill, C. J., Bloom, H. S., Rebeck Black, A., & Lipsey, M. W. (2008). Emprirical benchmarks for interpreting effect sizes in education research. *Child Development Perspectives , 2* (3), 172-177.

Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychological Methods , 11* (2), 193-206.

Hulin, C., Drasgow, F., & Parson, C. (1983). *Item Response Theory: Application to Psychological Measurement.* Homewood, IL: Dow Jones-Irwin.

Kemple, J., Corrin, W., Nelson, E., Salinger, T., Herrmann, S., & Drummond, K. (2008). *The Enhanced Reading Opportunities Study: Early Impact and Implementation Findings (NCEE 2008-4015).* Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Kolen, M., & Brennan, R. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* New York NY: Springer-Verlag.

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

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

Nataraj Kirby, S., McCaffrey, D. F., Lockwood, J. R., McCombs, J. S., Naftel, S., & Barney, H. (2002). Using state school accountability data to evaluate federal programs: A long uphill road. *Peabody Journal of Education , 77* (4), 122-145.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (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.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Thousand Oaks: Sage.

Rebeck Black, A., Doolittle, F., Zhu, P., Unterman, R., & Baldwin Grossman, J. (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.

Rebeck Black, A., Somers, M.-A., Doolittle, F., Unterman, R., & Baldwin Grossman, J. (2009). *The Evaluation of Enhanced Academic Instruction in After-School Programs: Final Report (NCEE 2009-4077).* Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London UK: Chapman and Hall.

Slavin, R. E. (2008). Evidence-based reform in education: Which evidence counts? *Educational Researcher , 37* (1), 47-50.

Somers, M.-A., Corrin, W., Sepanik, S., Salinger, T., Levin, J., & Zmach, C. (2010). *The Enhanced Reading Opportunities Study Final Report: The Impact of Supplemental Literacy Courses for Struggling Ninth-Grade Readers (NCEE 2010-4021).* Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Spybrook, J. (2008). Are power analyses reported with adeqaute detail: Findings from the first wave of group randomized trials funded by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness , 1* (3).

What Works Clearinghouse. (2008). *Procedures and Standards handbook (Version 2.0).* Washington DC: Institute of Education Sciences.

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

Appendix A: Descriptive Information on Study and State Tests

This appendix presents descriptive information on the state tests and the study-administered test in each of the four randomized experiments. Each table includes information on (i) their content; (ii) their scale (state-wide range, mean, and standard deviation), and (iii) their reliability. The first row of each table presents similar information for the study-administered test.

Appendix Table A.1
Study and State Test Descriptions: Study A

State	Test Name	Subject	Year	Grade	Test Content	Range	State Average	State Standard Deviation	Reliability (Cronbach alpha)
	SAT 10	Reading	2005-06	3	Word Study Skills; Reading Vocabulary; Reading Comprehension	402-782	621	38.1	0.90 - 0.93
				4	Same as above	417-800	630.4	39.2	
				5	Reading Vocabulary; Reading Comprehension	454-805	644.2	36.7	
CA	CST: California Standards Test	ELA	2005-06	3	Word Analysis, Fluency, and Systematic Vocabulary Development; Reading Comprehension; Literary Response and Analysis; Writing Strategies; Writing Applications	150-600	331	62	0.93
				4	Same as above	150-600	351	59	0.94
				5	Same as above	150-600	342	57	0.94
FL	FCAT: Florida Comprehensive Achievement Test	Reading	2005-06	3	Words and Phrases in Context; Main Idea, Plot Purpose; Comparison and Cause/Effect; Reference and Research	86-2514	1381.94	344.89	0.920
				4	Same as above	295-2638	1546.69	313.39	0.915
				5	Same as above	474-2713	1618.59	334.14	0.905
GA	CRCT: Criterion-Referenced Competency Tests	Reading	2005-06	3	Reading Skills and Vocabulary Acquisition; Literary Comprehension; Reading for Information	650-900	828	32	not avail.
				4	Reading Skills and Vocabulary Acquisition; Literary Comprehension; Information and Media Literacy	650-900	827	30	not avail.
				5	Same as 4th grade	650-900	822	26	not avail.
LA	LEAP: Louisiana Educational Assessment Program	ELA	2005-06	4	Read, comprehend and respond to a range of materials; Write competently; Use conventions of languages; Locate, select, and synthesize information; Read, analyze and respond to literature; Apply reasoning and problem-solving skills	100-500	310.91	59.75	0.91

Appendix Table A.1
Study and State Test Descriptions: Study A

State	Test Name	Subject	Year	Grade	Test Content	Range	State Average	State Standard Deviation	Reliability (Cronbach alpha)
LA	iLEAP: Integrated Louisiana Educational Assessment Program	ELA	2005-06	3	Same as above	100-500	290.82	60.1	0.93
				5	Same as above	100-500	295.16	55.9	0.92
MI[a]	MEAP: Michigan Educational Assessment Program	Reading	2005-06	3	Word Recognition and Word Study (Phonemic Awareness, Phonics, Word Recognition, Vocabulary); Fluency; Vocabulary; Narrative Text; Informational Text; Comprehension; Metacognition; Critical Standards; Reading Attitude	186-471	327	25.1	0.83
				4	Same as above	269-622	424	25.2	0.86
				5	Same as above	379-647	521	25.1	0.87
NM	NMSBA: New Mexico Standards Based Assessments	Reading	2005-06	3	Reading and listening for comprehension; Literature and Media	423-819	624	31.97	0.90
				4	Same as above	425-834	642	34.53	0.91
				5	Same as above	428-861	660	35.56	0.91
NY	NYSTP: New York State Testing Program	Reading	2005-06	3	Reading, Writing, Listening, and Speaking for: Information and Understanding; Literary Response and Expression; Critical Analysis and Evaluation	475-780	668.79	40.91	0.85
				4	Same as above	430-775	665.73	40.74	0.88
				5	Same as above	495-795	662.69	41.17	0.81
PA	PSSA: Pennsylvania System of School Assessment	Reading	2005-06	3	Comprehension and Reading Skills; Interpretation and Analysis of Fiction; and Nonfiction Text	300-1999	1329.47	233.58	0.92
				4	Same as above	700-2302	1338.01	218.32	0.90
				5	Same as above	700-2236	1309.42	234.36	0.91

Appendix Table A.1
Study and State Test Descriptions: Study A

State	Test Name	Subject	Year	Grade	Test Content	Range	State Average	State Standard Deviation	Reliability (Cronbach alpha)
TX	TAKS: Texas Assessment of Knowledge and Skills	Reading	2005-06	3	Basic understanding; Literary elements; Analysis using reading strategies; Analysis using critical thinking skills	1416-2606	2311.69	183.52	0.891
				4	Same as above	1313-2653	2226.85	154.14	0.849
				5	Same as above	1187-2733	2228.19	189.19	0.870
WI	WKCE: Wisconsin Knowledge and Concepts Examination	Reading	2005-06	3	Determine the meaning of words and phrases in context; Understand text; Analyze text; Evaluate and extend text	220-630	458.53	38.93	0.93
				4	Same as above	240-650	477.33	45.17	0.91
				5	Same as above	270-680	484.76	46.96	0.91

SOURCE: Corresponding state Department of Education websites. Stanford Achievement Test Series, Tenth Edition – Technical Data Report.
NOTES:
[a]Study data for this state's standardized test was not available.

Appendix Table A.2
Study and State Test Descriptions: Study B

State	Test Name	Subject	Year	Grade	Test Content	Range	State Average	State Standard Deviation	Reliability (Cronbach alpha)
	SAT 10	Math	2005-06	3	Number Sense and Operations; Patterns, Relationships,and Algebra; Data, Statistics, and Probability; Geometry and Measurement Computation with Whole Numbers and Decimals	414-756	607.6	40.8	0.89 - 0.92
				4	same as above and Computation with Fractions	436-791	619.0	39.7	
				5	same as above	465-812	640.9	35.9	
AL	ARMT: Alabama Reading and Mathematics Test	Math	2005-06	3	Number and Operations; Algebra; Geometry; Measurement; Data Analysis and Probability	<535, >628	not avail.	not avail.	not avail.
				4	Same as above	<548, >640	not avail.	not avail.	not avail.
				5	Same as above	<561, >652	not avail.	not avail.	not avail.
CA	CST: California Standards Test	Math	2005-06	3	Number Sense (Place value, fractions, and decimals, addition, subtraction, multiplication); Algebra and Functions; Measurement and Geometry; Statistics, Data Analysis, and Probability	150-600	369	84	0.95
				4	Number Sense (Decimals, fractions, and negative numbers, operations and factoring); Algebra and Functions; Measurement and Geometry; Statistics, Data Analysis, and Probability	150-600	361	74	0.94
				5	Number Sense (Estimation, percents, and factoring, operations with fractions and decimals); Algebra and Functions; Measurement and Geometry; Statistics, Data Analysis, and Probability	150-600	356	90	0.94

Appendix Table A.2
Study and State Test Descriptions: Study B

State	Test Name	Subject	Year	Grade	Test Content	Range	State Average	State Standard Deviation	Reliability (Cronbach alpha)
CT	CMT: Connecticut Mastery Test	Math	2005-06	3	Algebraic Reasoning: Patterns and Functions; Numerical and Proportional Reasoning; Geometry and Measurement; Working with Data: Probability and Statistics	100-400	248.90	52.66[a]	not avail.
				4	Same as above	100-400	252.60	50.55[a]	not avail.
				5	Same as above	100-400	256.40	52.39[a]	not avail.
FL	FCAT: Florida Comprehensive Achievement Test	Math	2005-06	3	Number Sense, Concepts and Operations; Measurement; Geometry and Spatial Sense; Algebraic Thinking; Data Analysis and Probability	375-2225	1408.97	300.30	0.927
				4	Same as above	581-2330	1533.57	265.83	0.923
				5	Same as above	569-2456	1648.66	239.82	0.947
GA	CRCT: Criterion-Referenced Competency Tests	Math	2005-06	3	Number and Operations; Measurement; Geometry; Algebra; Data Analysis and Probability	300-350	335	26	not avail.
				4	Same as above	300-350	323	30	not avail.
				5	Same as above	300-350	335	30	not avail.
KS[b]	Kansas General Assessments	Math	2005-06	3	Number and Computation; Algebra; Geometry; Data	0-70	55.91	11.36	0.93
				4	Same as above	0-73	54.36	11.60	0.92
				5	Same as above	0-73	53.37	11.79	0.92
PA	PSSA: Pennsylvania System of School Assessment	Math	2005-06	3	Numbers & Operations; Measurement; Geometry; Algebraic Concepts; Data Analysis & Probability	200-1999	1396.30	237.02	0.91
				4	Same as above	700-2282	1401.12	221.22	0.92
				5	Same as above	700-2293	1421.90	238.52	0.93

Appendix Table A.2
Study and State Test Descriptions: Study B

State	Test Name	Subject	Year	Grade	Test Content	Range	State Average	State Standard Deviation	Reliability (Cronbach alpha)
TX	TAKS: Texas Assessment of Knowledge and Skills	Math	2005-06	3	Number, operations, and quantitative reasons; Patterns, relationships, and algebraic reasoning; Geometry and spatial reasoning; Measurement; Probability and Statistics; Mathematical processes and tools	1228-2733	2255.61	200.97	0.881
				4	Same as above	1281-2680	2267.51	192.10	0.886
				5	Same as above	1111-2792	2292.90	235.09	0.896
WI	WKCE: Wisconsin Knowledge and Concepts Examination	Math	2005-06	3	Mathematical Process; Number Operations and Relationships; Geometry; Measurement; Statistics and Probability; Algebraic Relationships	220-630	434.33	46.81	0.93
				4	Same as above	240-650	466.31	43.12	0.91
				5	Same as above	270-680	489.39	44.09	0.91

SOURCE: Corresponding state Department of Education websites. Stanford Achievement Test Series, Tenth Edition – Technical Data Report

NOTE:

[a] Standard deviaton from 2007-08

[b] Study data for this state's standardized test was not available.

Appendix Table A.3
Study and State Test Descriptions: Study C

State	Test Name	Subject	Year	Grade	Test Content	Range	State Average	State Standard Deviation	Reliability (Cronbach alpha)
	GRADE	Reading	2005-06	9	Reading Comprehension; Reading Vocabulary	<55-145	99.72	14.83	0.93
GA	EOCT: End of Course Test	ELA	2005-06	9	Reading and Literature; Reading, Listening, Speaking and Viewing Across the Curriculum; Writing; Conventions	<400, >450	414	not avail.	not avail.
SC	EOCEP: End of Course Examination Program	ELA	2005-06	9	Reading Comprehension (Literary and informational text); Word Analysis and Vocabulary; Writing; Access and use information	0-100	76.1	not avail.	not avail.
TX	TAKS: Texas Assessment of Knowledge and Skills	Reading	2005-06	9	Basic Understanding,; Literary Elements and Techniques; Analysis and Evaluation	1342-3628	2247.25	171.28	0.880
UT	English Language Arts Criterion-Referenced Test	ELA	2005-06	9	Reading Comprehension; Writing; Seeking information in text	not avail.	not avail.	not avail.	not avail.

SOURCE: Corresponding state Department of Education websites. Group Reading Assessment and Diagnostic Evaluation: Technical Manual.

Appendix Table A.4
Study and State Test Descriptions: Study D

State	Test Name	Subject	Year	Grade	Test Content	Range	State Average	State Standard Deviation	Reliability (Cronbach alpha)
	NWEA	Math	2007-08	7	Fractions and Decimals; Ratios, Proportions, and Percents	not avail.	not avail.	not avail.	not avail.
CO	CSAP: Colorado Student Assessment Program	Math	2007-08	7	Number Sense (fractions, decimals, percent); Algebraic Methods; Geometry; Data Analysis and Probability	280-860	548	73.4	0.93
CT	CMT: Connecticut Mastery Test	Math	2007-08	7	Algebraic Reasoning: Patterns and Functions; Numerical and Proportional Reasoning (fractions, decimals, percents, operations scientific notation); Geometry and Measurement; Working with Data: Probability and Statistics	100-400	260.71	47.77	0.973
FL	FCAT: Florida Comprehensive Achievement Test	Math	2007-08	7	Number Sense, Concepts and Operations; (decimals, percents, problem solving) Measurement; Geometry and Spatial Sense; Algebraic Thinking; Data Analysis and Probability	100-500	315	58.8[a]	not avail.
KY	KCCT:Kentucky Core Content Test	Math	2007-08	7	Number Properties and Operations (integers, fractions, percents, decimals, estimation, ratios operations, proportions) Measurement; Geometry; Data Analysis and Probability; Algebraic Thinking	700-780	744	21.57	0.882
LA	iLEAP: Integrated Louisiana Educational Assessment Program	Math	2007-08	7	Number and Number relations (fractions, rates, ratios, proportions, percents, decimals) Algebra; Measurement; Geometry; Data Analysis, Probability, and Discrete Math;	100-500	297.61	63.48	0.89

Appendix Table A.4
Study and State Test Descriptions: Study D

State	Test Name	Subject	Year	Grade	Test Content	Range	State Average	State Standard Deviation	Reliability (Cronbach alpha)
					Patterns, Relations, and Functions				
NY	NYSTP: New York State Testing Program	Math	2007-08	7	Number Sense and Operations (number systems, number theory, rational numbers, exponents, operations, estimation) Algebra; Geometry; Measurement; Statistics and Probability	500-800	674.20	38.27	0.90
OH	OAT: Ohio Achievement Test	Math	2007-08	7	Data Analysis and Probability; Geometry and Spatial Sense; Measurement; Number Sense (solve expressions with integers, exponents, proportions, fractions, decimals, and percents); Patterns, Functions and Algebra	275-569	416.95	32.22	0.89
OK	OCCT: Oklahoma Core Curriculum Tests	Math	2007-08	7	Algebraic Reasoning; Number Sense (Integers, Ration/Proportion/ Percent, Exponents); Geometry; Measurement; Data Analysis and Statistics	400-990	744.7	71.8	0.88
WI	WKCE: Wisconsin Knowledge and Concepts Examination	Math	2007-08	7	Mathematical Process; Number Operations and Relationships (fractions, decimals, number theory, percent, proportions) Geometry; Measurement; Statistics and Probability; Algebraic Relationships	330-710	535.86	45.77	0.93

SOURCE: Corresponding state Department of Education websites.
NOTES:
[a] Standard deviation from 2005-06 year

Appendix B: Assessing the Conditions for Equating State Test Scores

This appendix uses descriptive data from the four experiments to examine whether rescaled test scores provide an equated measure of achievement. Test scores are equated if they are *exchangeable* -- once scores have been equated, score x on one assessment represents the same latent proficiency as score x on the other assessment. For reasons explained in Section 5.1, two conditions must be met for rescaled state test scores to be equated:

1. State assessment must measure the same construct

2. The sample of students in each state must come from a common reference population (or in other words, students from each grade and state are a random sample from some larger common reference population of low-performing students.)

The findings in Section 4 indicate that, for the states in the four experiments, the content domains of state assessments does differ, which means that the first condition is not met. Thus, the remainder of this appendix focuses on the second condition – that the sample comes from a common reference population. To satisfy this condition, either the sample or the state distribution of achievement must be the same across states. Below we examine each of these scenarios in turn.

A. The sample distribution of achievement is the same for each state

Depending on which linking function is used to convert scores to a common metric (mean scaling, linear scaling, or percentile-based scaling), the mean, standard deviation, and/or distribution of achievement must be the same in each state and grade.

For the four experiments used in this paper, this condition can be examined using data from the study-administered pretest for each study, which provides a consistent measure of achievement for all students in the sample. Figures B.1 and B.2 show the mean and standard deviation of pretest scores for students in the sample, by state. For each

experiment, we also tested whether the cross-state variation in means/SDs in these figures is statistically significant.[96]

The key findings are that:

- **Average score:** In Studies A and D, the average pretest score of students differs by a statistically significant amount across states. In these studies, the range of test scores between the lowest-performing and the highest-performing state is 0.94 and 0.67, respectively (in effect size).[97] In Studies B and C, however, the range of scores across states is not statistically significant.[98]

- **Standard deviation in scores:** The standard deviation of pre-test scores does not differ by a statistically significant amount across states in any of the four experiments.[99]

Based on these findings, in Studies B and C it can be assumed that students in each state come from the same reference population, because the mean and the standard deviation of achievement are similar across states. This means that the second condition for equating state test scores is satisfied for these two studies.[100]

B. The state-wide distribution is the same across states

Even if the achievement of students in the *sample* differs across states, students in the study may still come from a common reference population, as long as the *state-wide* distribution of achievement is the same across states. If the latter holds true, then differences in *sample* distributions are simply due to the fact that students are drawn from different parts of their state's distribution of achievement. Ultimately, however, students are

[96] For Studies A and B, another relevant question is whether there is variation in achievement across *grades*. However, the fact that there is variation in achievement across *states* for these two studies (as seen in the figures) is sufficient for showing that rescaled scores are not equated.

[97] The effect size is based on the mean and standard deviation of scores in the national norming sample.

[98] The similarity of student achievement across states in Study C may be due to the fact that eligibility for the program was explicitly based on the pretest (students in Study C had to be more than 2 years below grade level on the reading pretest). Conversely, eligibility for the programs in the other studies was locally-determined, which may have resulted in greater variation in the achievement of recruited students across sites and therefore states.

[99] This may be due to the fact that the target population for all of these experiments is low-performing students. This targeted recruitment of students may have made the sample of students in each state more homogeneous in terms of their achievement, which in turn may have made within-state variability in achievement more similar across states.

[100] In addition to having the same *mean* and *standard deviation*, the distribution of the samples should also have the same *shape* in all states. However, as explained in May *et al.* (2009), this assumption is easily satisfied by using a nonlinear transformation to make the distribution of test scores normal in each state.

drawn from the same common population of low-achieving students, which means that this condition for equating test scores would be met.

We can examine whether the state-wide distribution of achievement is the same for the four experiments by looking at state-level scores on the National Assessment of Educational Progress (NAEP), which provides a consistent measure of achievement across states. Figure B.3 and B.4 show the mean and standard deviation in NAEP scores by state, respectively.[101] We also test whether the mean and SD in NAEP scores differ by a statistically significant amount across states. As seen in these figures:

- **Average score:** The average NAEP score differs by a statistically significant amount across states for Studies A, B and D.[102] In these studies, the test score range between the lowest-performing and highest-performing state is 0.53, 0.62, and 0.39, respectively (in effect size).[103] In Study C, the range of scores across states is 0.14 and is not statistically significant.

- **Standard deviation in scores:** The standard deviation of NAEP scores differs by a statistically significant amount across states for Studies A, B and D, but not for Study C.[104]

Only Study C satisfies the second condition for equating state test scores (that is, z-scores based on the state distribution are equated, assuming that test content is the same in all states).

[101] These figures show scores for the NAEP assessment that is most closely aligned with the subject area and grade level of the randomized experiment. For example, for Study A – which examines the effect of the treatment on reading achievement in grades 3 to 5 – the figure shows the standard deviation in NAEP scores on the 4th grade reading assessment.

[102] This test conducted using a "v-known" analysis of variation in impacts, based on the standard error of the NAEP standard deviations (Raudenbush & Bryk, 2002).

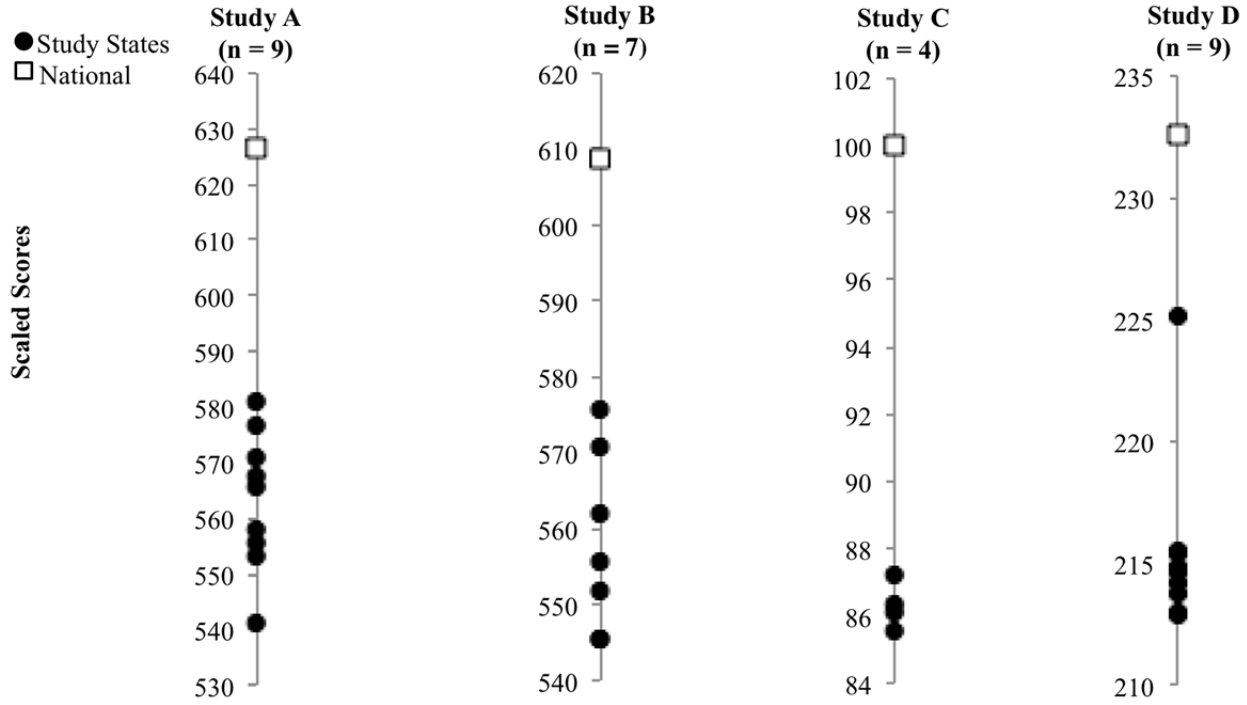
[103] The effect size is based on the national mean and standard deviation of the relevant NAEP assessment.

[104] This finding for Study C – which focuses on high school students – may be due to sample selection arising from the fact that testing is not mandatory for all grades in high school. While *NCLB* requires that states assess their students in reading and math in one grade level in high school, states can choose the grade level in which to administer the test. Importantly, states' decision about when to test a given subject area may be related to the achievement of their students. As explained in Section 3.1, for example, only four of the eight states that participated in Study C test their students in reading/ELA in ninth grade, so only these four states can be included in the analysis sample for Study C. If these states' decision to test their students in 9th grade is related to the distribution of achievement in their state (and in particular its heterogeneity), then this could give rise to the pattern of findings observed in the figure, whereby heterogeneity is similar across the four states.

Summary

None of the four experiments meet both conditions for equating state test scores. In Studies B and C, the second condition for equating may be met, but the first condition is not satisfied because achievement tests in the study states differ in terms of their content. This means that the first approach for combining impact findings across states should not be used (and that the meta-analytic approach is preferable).

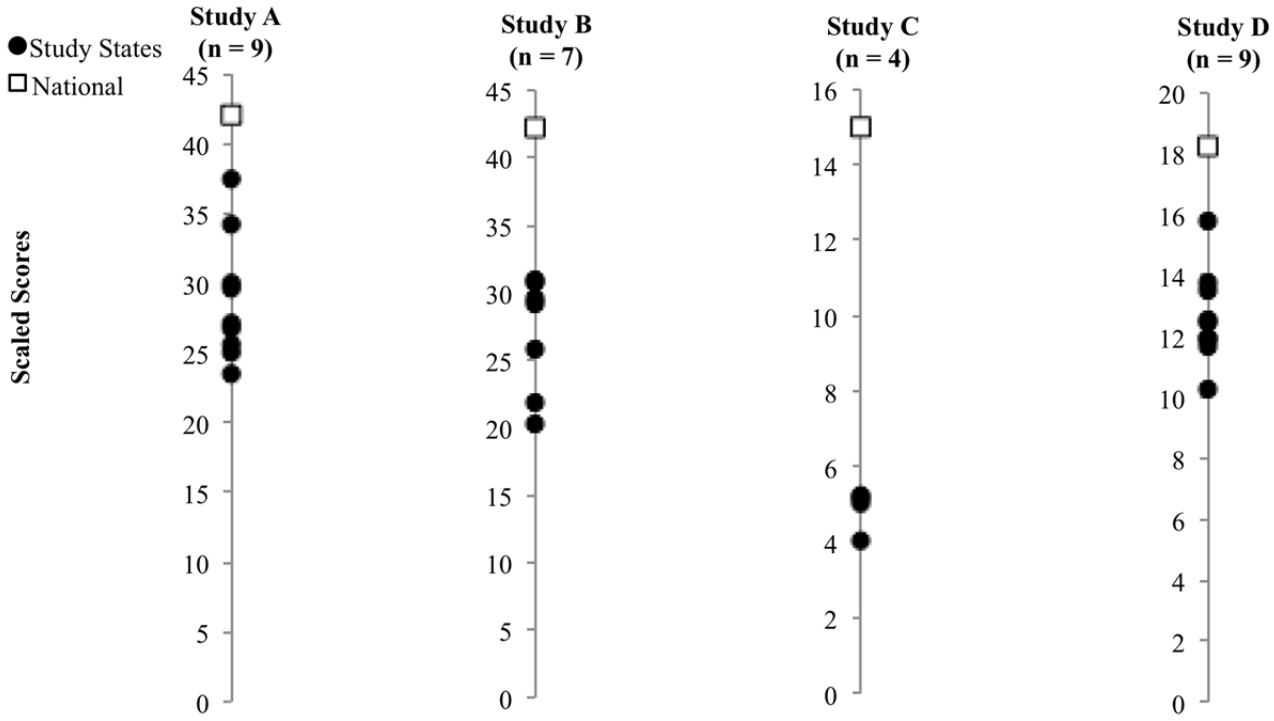
Appendix Figure B.1
Study-Administered Pretest Scores
Means by State



	Study A (n = 9)	Study B (n = 7)	Study C (n = 4)	Study D (n = 9)
Test of variation across states (p-value)	p = 0.048	p = 0.075	p = 0.163	p = 0.033
Min (in effect size)	-2.02	-1.51	-0.96	-1.08
Max (in effect size)	-1.08	-0.79	-0.85	-0.41
Range	0.94	0.72	0.12	0.67

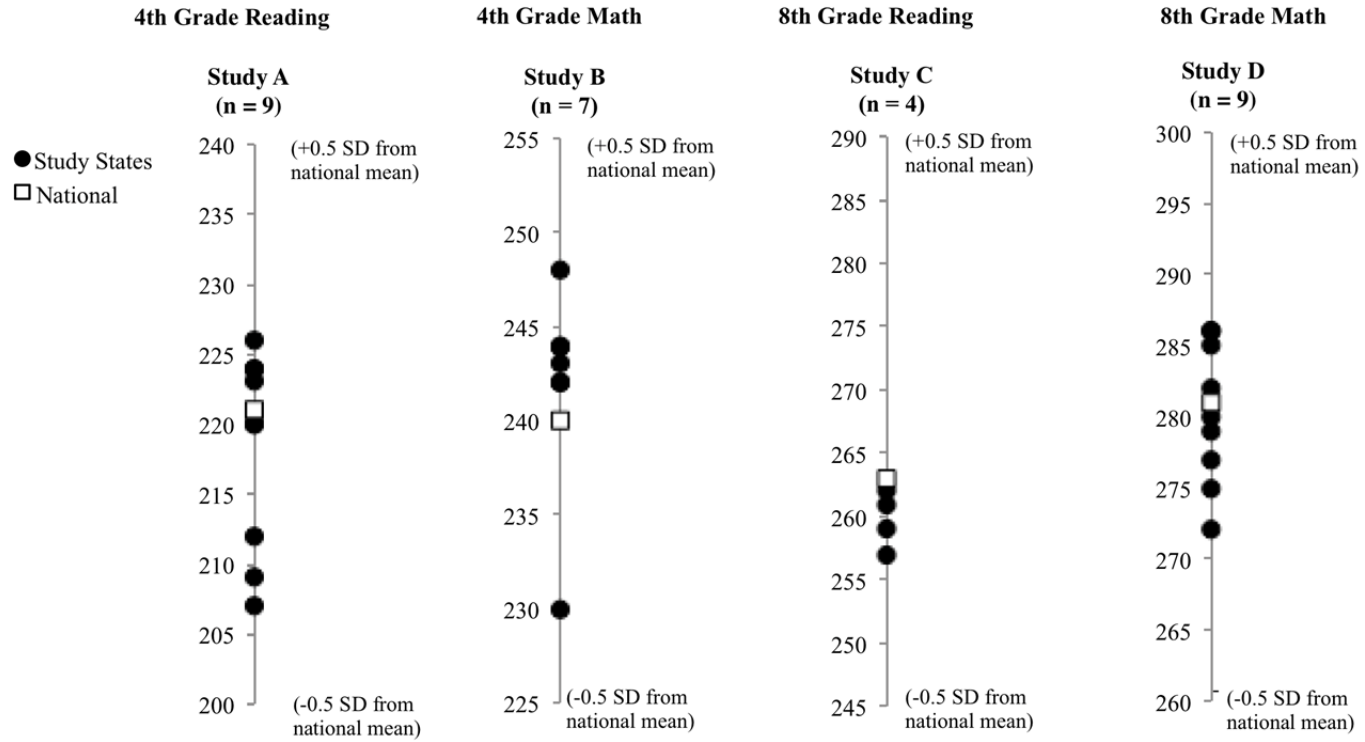
SOURCE: Authors' analysis based on data from four experiments.
NOTES: Sample means are based on students with both a state test score and a study test score. Values in Studies A and B are for 3rd grade students. Effect sizes are based on the national mean and standard deviation.

Appendix Figure B.2
Study-Administered Pretest Scores
Standard Deviation by State



SOURCE: Authors' analysis based on data from four experiments.
NOTES: Sample means are based on students with both a state test score and a study test score. Values for Studies A and B are for 3rd grade students. Effect sizes are based on the national mean and standard deviation.

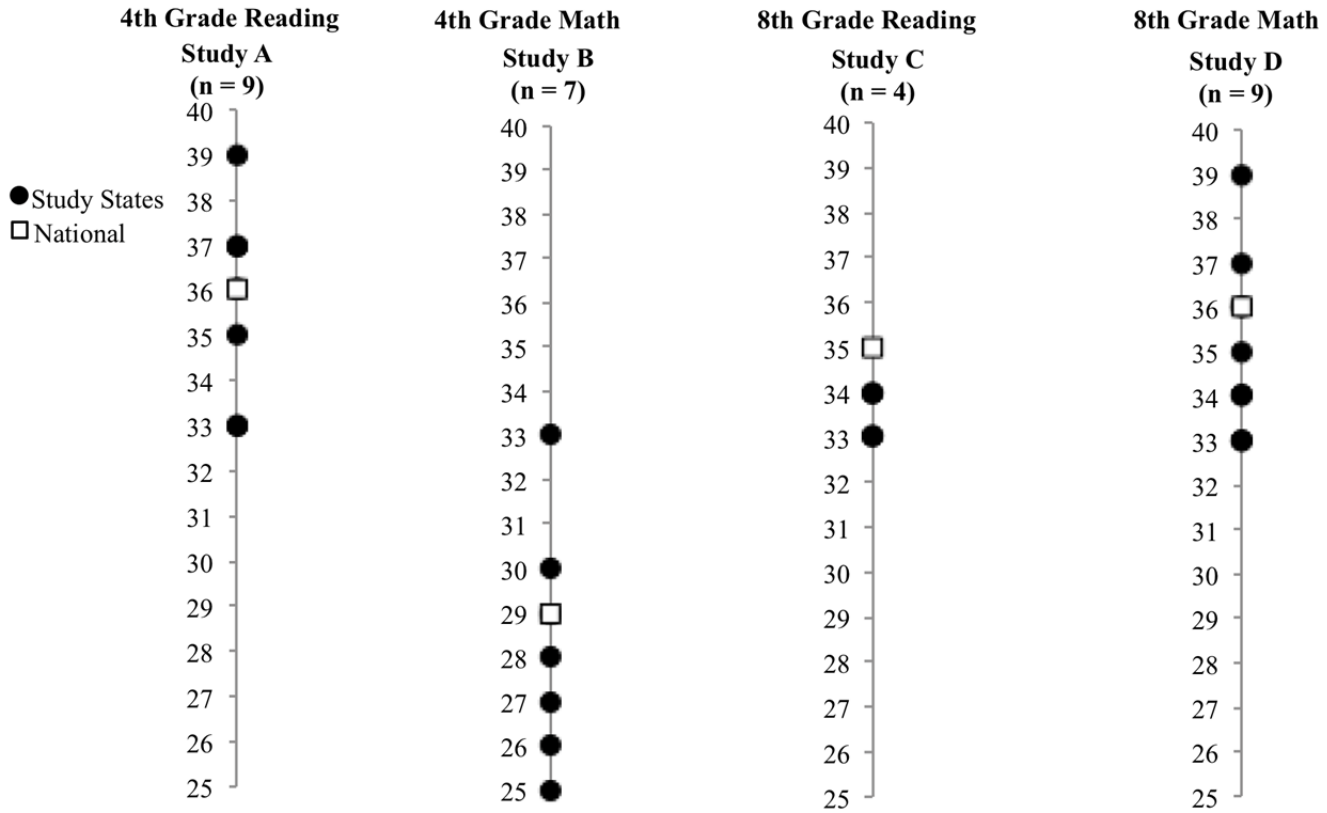
Appendix Figure B.3
National Assessment of Educational Process (NAEP)
Mean Scaled Score



	4th Grade Reading	4th Grade Math	8th Grade Reading	8th Grade Math
Test of variation across states (p-value)	p = 0.026	p = 0.044	p = 0.157	p = 0.028
Min (in effect size)	-0.39	-0.34	-0.17	-0.25
Max (in effect size)	0.14	0.28	-0.03	0.14
Range	0.53	0.62	0.14	0.39

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2007 Reading Assessment. Effect sizes are based on the national mean and standard deviation.

Appendix Figure B.4
National Assessment of Educational Process (NAEP)
Standard Deviation



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2007 Reading and Mathematics Assessment. Effect sizes are based on the national mean and standard deviation.

Appendix C: Technical Notes

This appendix describes the model specifications used in the impact analyses and provides further discussion of the rescaling and aggregation methods, and how these two factors interact.

Impact Model Specification

The statistical model used in the impacts analyses differs based on the weighting strategy that is used to combine the within-state/grade estimates. Below we provide details of the general model specification used for each of the different aggregation methods.

Precision/fixed-effects weighting (classical two-step approach) and sample size weighting

Both of these two weighting approaches require first estimating the impact of the program for each state and grade in the study, and then calculating a weighted average of the estimates. The weight can be the sample size of each state and grade (for sample size weighting) or the inverse of the estimated variance of the impact estimate for each state and grade (for precision weighting). Therefore the estimation model used for these two approaches is the same. Specifically, for each study separately, the following model is fit to a pooled dataset that includes all students in the study:

$$Z_i = \sum_{S,G} \beta_{sg} T_i * ST_{si} * GR_{gi} + \sum_{K} \lambda_k B_{ki} + \sum_{S,G} \delta_{sg} X_i * ST_{si} * GR_{gi} + \varepsilon_i$$

Where:

Z_i = Rescaled test score for student *i*.

T_i = Indicator of treatment group membership (treatment status).

B_{ki} = Random assignment block indicators, equal to 1 if student *i* is in random assignment block *k* and zero otherwise.[105]

X_i = Score for student *i* on a study-administered pre-test, whose effect is allowed to vary across states and grades.

[105] In Studies A and B, students were randomly assigned within grades and by school, so the blocks are grade-by-school dummy indicators. In Study C, the random assignment of students happened within high schools, so the blocks are school dummy indicators. In Study D, entire schools were randomly assigned within school districts, so the blocks are district indicators.

ST_{si} = Set of S indicators for state, equal to 1 if student *i* is in state *s* and zero otherwise.

ST_{gi} = Set of G indicators for state, equal to 1 if student *i* is in grade *g* and zero otherwise (relevant only for Studies A and B).

Therefore:

β_{sg} = The estimated program impact on state test scores for a given study, for grade *g* in state *s*.

In order to estimated the average or "combined" program impact, the within-state and within-grade estimates (β_{sg}) are average across the grades and states, using either the sample size of each grade in the state, n^{gs}, or the inverse of the variance of each of these estimates, , $[\text{se}(\beta)_{sg}]^2$, as the weight.

Precision/fixed-effects weighting (one-step regression)

This approach estimates a precision-weighted average impact estimate by fitting the following model to the pooled dataset for a given study:

$$Z_i = \beta T_i + \sum_K \lambda_k B_{ki} + \sum_{S,G} \delta_{sg} X_i * ST_{si} * GR_{gi} + \varepsilon_i$$

All variables are defined as before, and β is the average program impact across states and grades, weighted by precision.

This is also the model that is used in Section 7 to examine the precision gains of using state tests to measure achievement at *baseline* (except that here *X* is defined as students' prior state test scores).

Random-effects weighting (one-step regression)

This approach uses a pooled two-level regression model to estimate the overall treatment impact, which is allowed to vary across states. Specifically, it estimates the following model:

$$Z_i = \beta T_i + \sum_K \lambda_k B_{ki} + \sum_{S,G} \delta_{sg} X_i * ST_{si} * GR_{gi} + u_s * T_i + \varepsilon_i$$

Where:

$\mu_s T_i$ = a state-level error term (that varies across states and grades) for students in the treatment group.

All other variables are defined as before. Therefore, β is the average program impact for the larger population of states from which the study states are drawn.

A Note about Study D

Even though these model specifications are general across all four studies, sometimes they need to be adjusted to reflect the specific design of each study. In particular, Study D is a school-level randomized experiment. For this study:

- A multi-level regression model, instead of an individual-level model, was used to account for the fact that the treatment status is determined at the school level and students in the sample are clustered within schools.
- The model controls for student-level *and* school-level pretest scores (rather than just student-level scores, X_i).

Technical Notes About Rescaling and Aggregation Options

Rescaling and Aggregation Methods: How They Interact

As discussed in Section 5.2, inferences about the *within-state* estimated impact of the program are not affected by the choice between using the sample or the state-wide distribution in test scores when z-scoring to a common metric. This is due to the fact that the magnitude of the estimated impact and its standard error are rescaled by the same amount. Therefore, the T statistic for the within-state impact estimate (and its p-value) is not affected. Appendix Table C.1 demonstrates this identity for a student-level randomized experiment.

However, inferences about the *average* impact of the program across states (i.e, its p-value) may differ depending on the choice of rescaling method. This is because the rescaling method – and in particular whether to use the state or the sample standard deviation to rescale test scores – affects the standard error of the impact estimate for a given state, and in particular may affect its precision *relative to other states*. This, in turn, affects a state's relative weight in the combined impact estimate when aggregation is based on precision weighting or random-effects weighting.

Appendix Table C.1
Z-scoring Method:
Magnitude, Standard Error, and T statistic of Within-State Impact Estimate
(For a Student-level Randomized Experiment)

		Choice of Reference Population	
Impact Parameter	Raw Scores	Z-scores based on sample	Z-scores based on state
Magnitude	$\overline{Y}_{Ts} - \overline{Y}_{Cs}$	$\frac{\overline{Y}_{Ts} - \overline{Y}_{Cs}}{\sigma_Y}$	$\frac{\overline{Y}_{Ts} - \overline{Y}_{Cs}}{\sigma_{ST}}$
Standard Error	$\frac{\sigma_Y}{\sqrt{n_s p_s (1-p_s)}}$	$\frac{1}{\sqrt{n_s p_s (1-p_s)}}$	$\frac{\sigma_Y}{\sigma_{ST}\sqrt{n_s p_s (1-p_s)}}$
T-statistic	$\sqrt{n_s p_s (1-p_s)}\,\frac{\overline{Y}_{Ts} - \overline{Y}_{Cs}}{\sigma_Y}$	$\sqrt{n_s p_s (1-p_s)}\,\frac{\overline{Y}_{Ts} - \overline{Y}_{Cs}}{\sigma_Y}$	$\sqrt{n_s p_s (1-p_s)}\,\frac{\overline{Y}_{Ts} - \overline{Y}_{Cs}}{\sigma_Y}$

NOTES:

$\overline{Y}_{Ts}$ = Average raw score for students in treatment group in state *s*

$\overline{Y}_{Ts}$ = Average raw score for students in control group in state *s*

σ_Y = Standard deviation of raw test scores for students in the sample in state *s*

σ_{ST} = Standard deviation of raw test scores for all students in the state *s*

n_s = Number of students in the sample in state *s*

p_s = Proportion of students assigned to the treatment group (random assignment ratio) in state *s*

To illustrate, Appendix Table C.2 shows the precision weight and the random-effects weight for the within-state impact estimate, by rescaling method, for a student-level randomized experiment. The values in this table are obtained by substituting the standard error of the impact estimate for each rescaling method (from Table C.1) into the weight formulas (Equations (3a) and (4)).

As seen here, for a given weighting approach, the choice of rescaling method affects the relative weight of an impact estimate except under one condition: the ratio of state-to-sample standard deviation (σ_{ST}/σ_Y) is the same across all states. In this scenario, the relative precision of impact estimates will be the same regardless of whether the sample or state standard deviation is used to rescale test scores, and by extension, the relative weight

of each estimate when pooling across states will be the same. As seen in Section 5.2.1, however, that the ratio of state-to-sample standard deviation *does* differ across states, and hence, different rescaling methods can potentially lead to different p-values for the average impact of the program. This is confirmed by the impact findings in Section 6.2.

Random-Effects vs. Precision (Fixed-Effects) Weighting

These two weighting approaches differ most in terms of the weight that they attribute to the *least reliable* impact estimates (typically, states with the smallest sample sizes). When precision-weighting is used, *true* variation in program impacts across states is assumed to be zero or inconsequential, in which case any difference in the *observed* impact estimates across states is assumed to be due to differences in reliability (sample size). From this perspective, the most optimal approach is to weight impacts estimates based on their precision. With random-effects weighting, however, variation in the impact estimates across states is assumed to reflect both sampling error *and* true variation in impacts. As such, less reliable estimates cannot be "discounted" to the same degree when calculating the pooled estimate, since the difference between these impact estimates and others may reflect true variation in the effect of the program (and not just sampling error). This means that less precise impact estimates are given a relatively greater weight in a random-effects approach than in a standard precision-weighting scheme.

This point is illustrated in Table C.2. As shown here, as sample sizes for each state get larger (go to infinity), each state's impact is weighted increasingly equally. This happens because with larger sample sizes, the relative precision of impact estimates becomes more similar, so "less reliable" estimates are no longer penalized at all, and are therefore given the same weight as other impact estimates when calculating the average impact of the program.

Precision (Fixed-Effects) Weighting vs. Sample Size Weighting

Table C.2 also shows that in a student-level experiment, weighting by precision is the same as weighting by sample size, when: (i) test scores are rescaled based on the *sample* mean and standard deviation, and (ii) the random assignment ratio is the same in all states.

Appendix Table C.2
Weights for Combining Impacts Estimates across States (Normalized Weights), By Rescaling and Aggregation Method
(For a Student-level Randomized Experiment)

	Aggregation (Weighting) Approach					
	Impact for States in the Study Sample				*Impact for Broader Population of States*	
		Fixed Effects (By Precision)			Random-Effects	
Rescaling method	By Sample Size	Ω_s and p_s vary across states	p_s constant	Ω_s and p_s constant	Standard	Large n_s
Standardization based on sample (Z, EQ)	$\frac{n_s}{N}$	$\frac{n_s p_s(1-p_s)}{\sum_S n_s p_s(1-p_s)}$	$\frac{n_s}{N}$	$\frac{n_s}{N}$	$\frac{\frac{n_s p_s(1-p_s)}{1+n_s p_s(1-p_s)V}}{\sum_S \frac{n_s p_s(1-p_s)}{1+n_s p_s(1-p_s)V}}$	$\frac{1}{S}$
Standardization based on state (Z)	$\frac{n_s}{N}$	$\frac{\Omega_s n_s p_s(1-p_s)}{\sum_S \Omega_s n_s p_s(1-p_s)}$	$\frac{\Omega_s n_s}{\sum_S \Omega_s n_s}$	$\frac{n_s}{N}$	$\frac{\frac{\Omega_s n_s p_s(1-p_s)}{1+\Omega_s n_s p_s(1-p_s)V}}{\sum_S \frac{\Omega_s n_s p_s(1-p_s)}{1+\Omega_s n_s p_s(1-p_s)V}}$	$\frac{1}{S}$

NOTES:

n_s = number of students in the sample in state *s*

N = total number of students in the sample (all states)

Ω_s =ratio of the variance in raw scores in state *s* and the variance among students in the sample for state *s* (usually, $\sigma^2_{ST} / \sigma^2_Y > 1$)

p_s = proportion of students assigned to the treatment group (random assignment ratio)

V = estimated variance in impacts across states

S = number of states

Appendix D: Impact Tables

This appendix presents impact estimates for the four randomized experiments. Tables D.1 to D.6 present impact findings related to whether and how to use state tests to measure student achievement (Section 6). In these tables:

- Each *column* corresponds to a different outcome measure. The first column of findings presents the estimated impact of the program on study-administered test scores. The remaining columns present the estimated impact on state test scores rescaled using different types of linking function (linear and non-linear).[106]

- Each *row* in the table corresponds to a different weighting strategy for aggregating impact findings: weighting by sample size,[107] precision weighting (classical (two-step) and one-step regression approach), and random-effects weighting (one-step regression approach).

Note that for Studies A and B, two sets of tables are included: one for the primary analysis sample (which excludes states where information on the state-wide mean and standard deviation in state test scores was not available) and another for the full sample of states in these studies.

Table D.7 presents impact findings related to whether to use state tests as a baseline covariate in the impact analysis to improve precision (Section 7), for the two studies where prior achievement is measured using both state tests and a study test. In these tables:

- Each *column* corresponds to a different type of baseline covariate. Columns 2 and 3 show results from models that control for the study pretest and the z-scored state pretest, respectively. Column 1 presents results from a model that only controls for random assignment blocks (that is, it does not control for pretest scores). Results reported in this latter column serve as a benchmark for the precision levels of impact estimates when pretests are used.

- Each *row* in a given panel represents a different outcome measure: the first row is for the study-administered test as the outcome, while the second row is for state tests as the outcome (sample-based z-scores).[108]

[106] For Study C, it was not possible to look at z-scores based on the state-wide distribution, because state-wide means and or standard deviations were only available for one state in the study.

[107] Because Study D is a school-level random experiment, the table examine weighting by the number of students in each state , as well as the number of *schools* in each state (the latter reflects the cluster randomization of the study).

[108] Findings are similar for other methods of rescaling state test scores at follow-up.

Appendix Table D.1
Impact Estimates by Assessment Type and Rescaling/Aggregation Strategy: Study A

Aggregation/weighting strategy	Study-administered test (sample-based z-scores)	State Test Scores by Rescaling Method: Z-score using state mean and standard deviation (sample-based)	Z-score using sample mean and standard deviation (state-based)	Z-score using percentile ranks (rank-based)
Weight by number of observations				
Estimated impact effect size	-0.0865	-0.0474	-0.0612	-0.0611
Standard error	(0.0443)	(0.0378)	(0.0528)	(0.0492)
P-value	0.0510	0.2100	0.2471	0.2147
Precision (fixed-effects) weighting -- Classical (two-step)				
Estimated impact effect size	-0.0797	-0.0529	-0.0677	-0.0700
Standard error	(0.0451)	(0.0384)	(0.0537)	(0.0501)
P-value	0.0770	0.1693	0.2082	0.1623
Precision (fixed-effects) weighting -- One-step regression				
Estimated impact effect size	-0.0850	-0.0495	-0.0633	-0.0638
Standard error	(0.0442)	(0.0377)	(0.0527)	(0.0490)
P-value	0.0550	0.1892	0.2294	0.1935
Random-effects weighting -- One-step regression				
Estimated impact effect size	-0.0850	-0.0495	-0.0633	-0.0638
Standard error	(0.0442)	(0.0377)	(0.0527)	(0.0490)
P-value	0.0550	0.1892	0.2294	0.1935
Variance in impact	N/E	N/E	N/E	N/E
P-value	N/E	N/E	N/E	N/E

SOURCE: Authors' analysis based on data from four experiments.

NOTES: The sample size used in the analyses is 1,032 students across 9 states. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent. N/E indicates values that are not estimable. These estimates cannot converge in maximum likelihood because the variance is zero.

Appendix Table D.2
Impact Estimates by Assessment Type and Rescaling/Aggregation Strategy: Study A (all states)

Aggregation/weighting strategy	Study-administered test (sample-based z-scores)	State Test Scores by Rescaling Method: Z-score using sample mean and standard deviation (state-based)	State Test Scores by Rescaling Method: Z-score using percentile ranks (rank-based)
Weight by number of observations			
Estimated impact effect size	-0.0261	-0.0394	-0.0376
Standard error	(0.0399)	(0.0484)	(0.0448)
P-value	0.5127	0.4155	0.4016
Precision (fixed-effects) weighting -- Classical (two-step)			
Estimated impact effect size	-0.0447	-0.0580	-0.0585
Standard error	(0.0416)	(0.0504)	(0.0466)
P-value	0.2828	0.2504	0.2098
Precision (fixed-effects) weighting -- One-step regression			
Estimated impact effect size	-0.0256	-0.0406	-0.0394
Standard error	(0.0401)	(0.0483)	(0.0447)
P-value	0.5229	0.4011	0.3775
Random-effects weighting -- One-step regression			
Estimated impact effect size	-0.0385	-0.0411	-0.0394
Standard error	(0.0574)	(0.0443)	(0.0453)
P-value	0.5026	0.3541	0.3842
Variance in impact	0.0158	-0.003	0.0005
P-value	0.2531	0.6815	0.9565

SOURCE: Authors' analysis based on data from four experiments.

NOTES: The sample size used in the analyses is 1,236 students across 10 states. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Appendix Table D.3
Impact Estimates by Assessment Type and Rescaling/Aggregation Strategy: Study B

Aggregation/weighting strategy	Study-administered test (sample-based z-scores)	State Test Scores by Rescaling Method: Z-score using state mean and standard deviation (sample-based)	Z-score using sample mean and standard deviation (state-based)	Z-score using percentile ranks (rank-based)
Weight by number of observations				
Estimated impact effect size	0.0744	0.0588	0.0683	0.0754
Standard error	(0.0452)	(0.0368)	(0.0482)	(0.0455)
P-value	0.1001	0.1108	0.1566	0.0977
Precision (fixed-effects) weighting -- Classical (two-step)				
Estimated impact effect size	0.0710	0.0375	0.0393	0.0526
Standard error	(0.0464)	(0.0378)	(0.0494)	(0.0467)
P-value	0.1261	0.3211	0.4267	0.2601
Precision (fixed-effects) weighting -- One-step regression				
Estimated impact effect size	0.0710	0.0557	0.0642	0.0714
Standard error	(0.0454)	(0.0370)	(0.0485)	(0.0457)
P-value	0.1185	0.1327	0.1857	0.1192
Random-effects weighting -- One-step regression				
Estimated impact effect size	0.0723	0.0467	0.0542	0.0618
Standard error	(0.0709)	(0.0525)	(0.0734)	(0.0628)
P-value	0.3081	0.3733	0.4605	0.3261
Variance in impact	0.0175	0.0079	0.0178	0.0105
P-value	0.3291	0.4882	0.4429	0.5127

SOURCE: Authors' analysis based on data from four experiments.

NOTES: The sample size used in the analyses is 944 students across 7 states. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Appendix Table D.4
Impact Estimates by Assessment Type and Rescaling/Aggregation Strategy: Study B (all states)

Aggregation/weighting strategy	Study-administered test (sample-based z-scores)		State Test Scores by Rescaling Method: Z-score using sample mean and standard deviation (state-based)	Z-score using percentile ranks (rank-based)
Weight by number of observations				
Estimated impact effect size	0.0752	*	0.0360	0.0548
Standard error	(0.0372)		(0.0416)	(0.0387)
P-value	0.0436		0.3878	0.1564
Precision (fixed-effects) weighting -- Classical (two-step)				
Estimated impact effect size	0.0620		0.0012	0.0263
Standard error	(0.0385)		(0.0431)	(0.0400)
P-value	0.1079		0.9774	0.5114
Precision (fixed-effects) weighting -- One-step regression				
Estimated impact effect size	0.0728		0.0321	0.0512
Standard error	(0.0374)		(0.0418)	(0.0388)
P-value	0.0521		0.4422	0.1872
Random-effects weighting -- One-step regression				
Estimated impact effect size	0.0708		0.0285	0.0471
Standard error	(0.0515)		(0.0548)	(0.0458)
P-value	0.1696		0.6036	0.3032
Variance in impact	0.0094		0.0094	0.0042
P-value	0.3890		0.4787	0.6197

SOURCE: Authors' analysis based on data from four experiments.
NOTES: The sample size used in the analyses is 1,307 students across 9 states. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Appendix Table D.5
Impact Estimates by Assessment Type and Rescaling/Aggregation Strategy: Study C

Aggregation/weighting strategy	Study-administered test (sample-based z-scores)		State Test Scores by Rescaling Method: Z-score using sample mean and standard deviation (state-based)		Z-score using percentile ranks (rank-based)	
Weight by number of observations						
Estimated impact effect size	0.1773	*	0.1588	*	0.1361	*
Standard error	(0.0573)		(0.0588)		(0.0572)	
P-value	0.0020		0.0070		0.0176	
Precision (fixed-effects) weighting -- Classical (two-step)						
Estimated impact effect size	0.1772	*	0.1586	*	0.1357	*
Standard error	(0.0573)		(0.0588)		(0.0572)	
P-value	0.0021		0.0071		0.0179	
Precision (fixed-effects) weighting -- One-step regression						
Estimated impact effect size	0.1772	*	0.1586	*	0.1357	*
Standard error	(0.0573)		(0.0588)		(0.0572)	
P-value	0.0020		0.0071		0.0178	
Random-effects weighting -- One-step regression						
Estimated impact effect size	0.1780	*	0.1586	*	0.1357	*
Standard error	(0.0580)		(0.0588)		(0.0572)	
P-value	0.0022		0.0071		0.0179	
Variance in impact	0.0002		N/E		N/E	
P-value	0.0211		N/E		N/E	

SOURCE: Authors' analysis based on data from four experiments.
NOTES: The sample size used in the analyses is 1,065 students across 4 states. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent. N/E indicates values that are not estimable. These estimates cannot converge in maximum likelihood because the variance is zero.

Appendix Table D.6
Impact Estimates by Assessment Type and Rescaling/Aggregation Strategy: Study D

Aggregation/weighting strategy	Study-administered test (sample-based z-scores)	State Test Scores by Rescaling Method: Z-score using state mean and standard deviation (sample-based)	Z-score using sample mean and standard deviation (state-based)	Z-score using percentile ranks (rank-based)
Weight by number of students				
Estimated impact effect size	0.0097	-0.0611	-0.0709	-0.0562
Standard error	(0.0413)	(0.0307)	(0.0405)	(0.0397)
P-value	0.8150	0.0533	0.0879	0.1644
Weight by number of schools				
Estimated impact effect size	0.0012	-0.0566	-0.0635	-0.0467
Standard error	(0.0437)	(0.0334)	(0.0433)	(0.0424)
P-value	0.9783	0.0978	0.1501	0.2772
Precision (fixed-effects) weighting -- Classical (two-step)				
Estimated impact effect size	0.0221	-0.0466	-0.0503	-0.0369
Standard error	(0.0405)	(0.0300)	(0.0398)	(0.0390)
P-value	0.5889	0.1293	0.2130	0.3494
Precision (fixed-effects) weighting -- One-step regression				
Estimated impact effect size	0.0218	-0.0466	-0.0502	-0.0367
Standard error	(0.0389)	(0.0301)	(0.0403)	(0.0394)
P-value	0.5777	0.1289	0.2190	0.3569
Random-effects weighting -- One-step regression				
Estimated impact effect size	0.0218	-0.0466	-0.0502	-0.0367
Standard error	(0.0389)	(0.0301)	(0.0403)	(0.0395)
P-value	0.5778	0.1229	0.2192	0.3569
Variance in impact	N/E	N/E	N/E	N/E

SOURCE: Authors' analysis based on data from four experiments.

NOTES: The sample size is 4,387 students across 9 states. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent. N/E indicates values that are not estimable. These estimates cannot converge in maximum likelihood because the variance is zero.

Appendix Table D.7
Impact Estimates by Type of Baseline Achievement Measure
(Study pretest scores or prior state test scores)
Studies C and D

Outcome	Unadjusted (no covariates)		Covariates: Study pretest		Covariates: Prior state tests[a]	
Study C						
Study-administered test (sample-based z-score)						
Estimated impact effect size	0.1621	*	0.1772	*	0.1247	*
Standard error	(0.0606)		(0.0573)		(0.0557)	
P-value	0.0076		0.0020		0.0253	
State tests (sample-based z-score)						
Estimated impact effect size	0.1450	*	0.1586	*	0.1012	
Standard error	(0.0613)		(0.0588)		(0.0551)	
P-value	0.0182		0.0071		0.0668	
Study D						
Study-administered test (sample-based z-score)						
Estimated impact effect size	0.0632		0.0218		0.0170	
Standard error	(0.0628)		(0.0389)		(0.0431)	
P-value	0.3183		0.5777		0.6945	
State tests (sample-based z-score)						
Estimated impact effect size	-0.0167		-0.0502		-0.0543	
Standard error	(0.0582)		(0.0403)		(0.0319)	
P-value	0.7755		0.2190		0.0950	

SOURCE: Authors' analysis based on data from four experiments.

NOTES: Impact estimates are pooled across states using fixed-effects weighting (one-step regression approach). The sample sizes used in the analyses are 1,065 students across 4 states for Study C and 4,387 students across 9 states for Study D.

Missing data (Study C): 1 student was missing the study pretest covariate. 83 students were missing the state pretest covariate. Missing data (Study D): 1,676 students were missing the study pretest covariate. 633 students were missing the state pretest covariate. Test scores for these students were imputed with a value of zero, and an indicator for missing data is included in the model (dummy variable approach).

[a]Z-scores based on sample mean and standard deviation.

Appendix E: Correlation between Student Achievement Measures

This appendix presents correlations between the follow-up and baseline achievement measures, for each of the four randomized experiments (Tables E.1 to E.4).

Appendix Table E.1
Correlation between Student Achievement Measures: Study A

Type of test score	Study-administered follow-up	State test follow-up (sample-based z-score)	State test follow-up (Rank-based z-score)	State test follow-up (state-based z-score)	Study-administered pretest
Study-administered follow-up	1.000	--	--	--	--
State test follow-up (sample-based z-score)	0.441 *	1.000	--	--	--
State test follow-up (rank-based z-score)	0.453 *	0.974 *	1.000	--	--
State test follow-up (state-based z-score)	0.436 *	0.937 *	0.913 *	1.000	--
Study-administered pretest	0.697 *	0.392 *	0.405 *	0.394 *	1.000

SOURCE: Authors' analysis based on data from four experiments.

NOTES: The sample size used in the analyses is 1,032 students across 9 states.

The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Appendix Table E.2
Correlation between Student Achievement Measures: Study B

Type of test score	Study-administered follow-up		State test follow-up (sample-based z-score)		State test follow-up (rank-based z-score)		State test follow-up (state-based z-score)		Study-administered pretest
Study-administered follow-up	1.000		--		--		--		--
State test follow-up (sample-based z-score)	0.628	*	1.000		--		--		--
State test follow-up (rank-based z-score)	0.626	*	0.984	*	1.000		--		--
State test follow-up (state-based z-score)	0.690	*	0.954	*	0.939	*	1.000		--
Study-administered pretest	0.719	*	0.531	*	0.527	*	0.590	*	1.000

SOURCE: Authors' analysis based on data from four experiments.

NOTES: The sample size used in the analyses is 944 students across 7 states.

The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Appendix Table E.3
Correlation between Student Achievement Measures: Study C

Type of test score	Study-administered follow-up		State test follow-up (sample-based z-score)		State test follow-up (rank-based z-score)		Study-administered pretest		State test at baseline (sample-based z-score)
Study-administered follow-up	1.000		--		--		--		--
State test follow-up (sample-based z-score)	0.453	*	1.000		--		--		--
State test follow-up (rank-based z-score)	0.486	*	0.939	*	1.000		--		--
Study-administered pretest	0.323	*	0.283	*	0.306	*	1.000		--
State test at baseline (sample-based z-score)	0.414	*	0.467	*	0.471	*	0.224	*	1.000

SOURCE: Authors' analysis based on data from four experiments.

NOTES: The sample size used in the analyses is 1,065 students across 4 states.

The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Appendix Table E.4
Correlation between Student Achievement Measures: Study D

Type of test score	Study-administered follow-up	State test follow-up (sample-based z-score)	State test follow-up (rank-based z-score)	State test follow-up (state-based z-score)	Study-administered pretest	State test at baseline (sample-based z-score)
Study-administered follow-up	1.000	--	--	--	--	--
State test follow-up (sample-based z-score)	0.687 *	1.000	--	--	--	--
State test follow-up (rank-based z-score)	0.698 *	0.979 *	1.000	--	--	--
State test follow-up (state-based z-score)	0.692 *	0.940 *	0.917 *	1.000	--	--
Study-administered pretest	0.761 *	0.657 *	0.662 *	0.680	1.000	--
State test at baseline (sample-based z-score)	0.644 *	0.680 *	0.677 *	0.661 *	0.665 *	1.000

SOURCE: Authors' analysis based on data from four experiments.

NOTES: The sample size used in the analyses is 4,387 students across 9 states.

The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Appendix F: Statistical Tests of Differences in Impact Findings Across Achievement Measures

This appendix presents statistical tests for the difference in impacts findings across different achievement measures. Tables F.1 to F.4 present p-values for the difference in impact findings across different types of *outcome measure* (study tests and state tests rescaled using different functions). Table F.5 presents p-values for the difference in impact findings across different types of *baseline covariate* (study pretest *vs.* state test scores from an earlier school year). The tables compare impact findings with respect to three parameters: the point estimate (magnitude), the standard error, and the inference about program effectiveness (p-value). To minimize the number of comparisons, the comparisons in these tables focus on the fixed-effects (one-step regression) aggregation strategy. P-values for differences in impact findings across achievement measures were obtained using a bootstrapping procedure:

1. For each of the four studies, 5,000 bootstrap datasets were generated by randomly sampling students with replacement from the original dataset.[109]

2. For each bootstrap dataset, impacts on the study-administered test and on state tests (based on different rescaling methods) were estimated using the same statistical model that produced the original estimates. Then, the difference in impact parameters (point estimate, standard error, and p-value) was computed for each pair of achievement measures (e.g., study test *vs.* state tests linearly rescaled as sample-based z-scores, latter *vs.* state tests rescaled using rank-based method, etc.), This process resulted in 5,000 sets of difference estimates for each pair-wise comparison of impact findings.

3. The p-value for the difference in a given parameter (point estimate, standard error, or p-value) is equal to the percentage of differences across all bootstrap datasets that are less than or equal to 0 (when the median difference is positive) or the percentage of differences that are greater than or equal to 0 (when the median difference is negative).

[109] For Study D, which is based on school-level random assignment, these bootstrap datasets were generated by randomly sampling schools with replacement.

Appendix Table F.1
Comparison of Impact Findings on Different Measures of Achievement
Study A

Impact on (row) minus impact on (column)	State test (state-based z-score)		State test (sample-based z-score)		State test (rank-based z-score)	
	Parameter difference	P-value for difference	Parameter difference	P-value for difference	Parameter difference	P-value for difference
Study-administered test (sample-based z-score)						
Impact Estimate	-0.036	0.432	-0.022	0.664	-0.021	0.667
Standard Error	0.007	0.000	-0.008	0.000	-0.005	0.000
Inference (p-value)	-0.134	0.592	-0.174	0.531	-0.139	0.584
State test (state-based z-score)						
Impact Estimate	--	--	0.014	0.469	0.014	0.462
Standard Error	--	--	-0.015	0.000	-0.011	0.000
Inference (p-value)	--	--	-0.040	0.652	-0.004	0.974
State test (sample-based z-score)						
Impact Estimate	--	--	--	--	0.000	0.948
Standard Error	--	--	--	--	0.004	0.000
Inference (p-value)	--	--	--	--	0.036	0.787

SOURCE: Authors' analysis based on data from four experiments.
NOTES: The sample size used in the analyses is 1,032 students across 9 states. Values in the "Difference" column are the difference between the impact on the row outcome (point estimate, standard error, and p-value) minus the impact on the column outcome. The impact findings being compared are for the precision (fixed-effects) weighting strategy for aggregation (one-step regression). Standard errors for differences are obtained using bootstrapping.

Appendix Table F.2
Comparison of Impact Findings on Different Measures of Achievement
Study B

Impact on (row) minus impact findings on (column)	State test (state-based z-score)		State test (sample-based z-score)		State test (rank-based z-score)	
	Parameter difference	P-value for difference	Parameter difference	P-value for difference	Parameter difference	P-value for difference
Study-administered test (sample-based z-score)						
Impact Estimate	0.015	0.815	0.007	0.965	0.000	0.909
Standard Error	0.008	0.000	-0.003	0.000	0.000	0.108
Inference (p-value)	-0.014	0.986	-0.067	0.859	-0.001	0.982
State test (state-based z-score)						
Impact Estimate	--	--	-0.008	0.530	-0.016	0.283
Standard Error	--	--	-0.011	0.000	-0.009	0.000
Inference (p-value)	--	--	-0.053	0.332	0.013	0.854
State test (sample-based z-score)						
Impact Estimate	--	--	--	--	-0.007	0.546
Standard Error	--	--	--	--	0.003	0.000
Inference (p-value)	--	--	--	--	0.067	0.459

SOURCE: Authors' analysis based on data from four experiments.
NOTES: The sample size used in the analyses is 944 students across 7 states. Values in the "Difference" column are the difference between the impact on the row outcome (point estimate, standard error, and p-value) and the impact on the column outcome. The impact findings being compared are for the precision (fixed-effects) weighting strategy for aggregation (one-step regression). Standard errors for differences are obtained using bootstrapping.

Appendix Table F.3
Comparison of Impact Findings on Different Measures of Achievement
Study C

Impact on (row) minus impact on (column)	State test (sample-based z-score)		State test (rank-based z-score)	
	Parameter difference	P-value for difference	Parameter difference	P-value for difference
Study-administered test (sample-based z-score)				
Impact Estimate	0.019	0.793	0.041	0.503
Standard Error	-0.001	0.578	0.000	0.945
Inference (p-value)	-0.005	0.732	-0.016	0.512
State test (state-based z-score)				
Impact Estimate	--	--	0.023	0.283
Standard Error	--	--	0.002	0.401
Inference (p-value)	--	--	-0.011	0.373

SOURCE: Authors' analysis based on data from four experiments.
NOTES: The sample size used in the analyses is 1,065 students across 4 states. Values in the "Difference" column are the difference between the impact for the *row* outcome (point estimate, standard error, and p-value) and the impact on the *column* outcome. The impact findings being compared are for the precision (fixed-effects) weighting strategy for aggregation (one-step regression). Standard errors for differences are obtained using bootstrapping.

Appendix Table F.4
Comparison of Impact Findings on Different Measures of Achievement
Study D

Impact on(row) minus impact on (column)	State test (state-based z-score)		State test (sample-based z-score)		State test (rank-based z-score)	
	Parameter difference	P-value for difference	Parameter difference	P-value for difference	Parameter difference	P-value for difference
Study-administered test (sample-based z-score)						
Impact Estimate	0.068	0.028	0.072	0.046	0.058	0.074
Standard Error	0.009	0.080	-0.001	0.800	-0.001	0.912
Inference (p-value)	0.449	0.830	0.359	0.915	0.221	0.984
State test (state-based z-score)						
Impact Estimate	--	--	0.004	0.780	-0.010	0.672
Standard Error	--	--	-0.010	0.000	-0.009	0.001
Inference (p-value)	--	--	-0.090	0.711	-0.228	0.492
State test (sample-based z-score)						
Impact Estimate	--	--	--	--	-0.013	0.151
Standard Error	--	--	--	--	0.001	0.484
Inference (p-value)	--	--	--	--	-0.138	0.463

SOURCE: Authors' analysis based on data from four experiments.
NOTES: The sample size used in the analyses is 4,387 students across 9 states. Values in the "Difference" column are the difference between the impact on the *row* outcome (point estimate, standard error, and p-value) and the impact on the *column* outcome. The impact findings being compared are for the precision (fixed-effects) weighting strategy for aggregation (one-step regression). Standard errors for differences are obtained using bootstrapping.

Appendix Table F.5
Comparison of Impact Findings
Study-Administered Pretest *vs.* Prior State Tests as a Baseline Covariate
Studies C and D

Outcome measure	Study pretest vs. state test scores as a baseline covariate: Parameter Difference	Study pretest vs. state test scores as a baseline covariate: P-value for difference
Study C		
Study-administered test (sample-based z-score)		
Impact Estimate	0.052	0.066
Standard Error	0.002	0.095
Inference (p-value)	-0.023	0.108
State test (sample-based z-score)		
Impact Estimate	0.057	0.049
Standard Error	0.004	0.000
Inference (p-value)	-0.060	0.112
Study D		
Study-administered test (sample-based z-score)		
Impact Estimate	0.005	0.731
Standard Error	-0.004	0.456
Inference (p-value)	-0.117	0.796
State test (sample-based z-score)		
Impact Estimate	0.004	0.859
Standard Error	0.008	0.321
Inference (p-value)	0.124	0.821

SOURCE: Authors' analysis based on data from four experiments.
NOTES: The sample size used in the analyses is 1,065 students across 4 states for Study C and 4,387 students across 9 states for Study D. Values in the "Difference" column are the difference between the impact findings when a study pretest is used as a baseline covariate (point estimate, standard error, and p-value) compared to the impact findings when state tests are used as a baseline covariate. The impact findings being compared are for the precision (fixed-effects) weighting strategy for aggregation (one-step regression). Standard errors for differences are obtained using bootstrapping.


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA