NCEE 2009-013

U. S. DEPARTMENT OF EDUCATION

Using State Tests in Education Experiments: A Discussion of the Issues

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

Using State Tests in Education Experiments: A Discussion of the Issues

November 2009

Henry May
University of Pennsylvania,
Consortium for Policy Research in Education

Irma Perez-Johnson
Joshua Haimson
Samina Sattar
Phil Gleason
Mathematica Policy Research, Inc.

Abstract

Securing data on students' academic achievement is typically one of the most important and costly aspects of conducting education experiments. As state assessment programs have become practically universal and more uniform in terms of grades and subjects tested, the relative appeal of using state tests as a source of study outcome measures has grown. However, the variation in state assessments—in both content and proficiency standards—complicates decisions about whether a particular state test is suitable for research purposes and poses difficulties when planning to combine results across multiple states or grades. This discussion paper aims to help researchers evaluate and make decisions about whether and how to use state test data in education experiments. It outlines the issues that researchers should consider, including how to evaluate the validity and reliability of state tests relative to study purposes; factors influencing the feasibility of collecting state test data; how to analyze state test scores; and whether to combine results based on different tests. It also highlights best practices to help inform ongoing and future experimental studies. Many of the issues discussed are also relevant for nonexperimental studies.

NCEE 2009-013
U.S. DEPARTMENT OF EDUCATION

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

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

Disclaimer
The Institute of Education Sciences (IES) at the U.S. Department of Education contracted with Mathematica Policy Research to develop a discussion paper on the issues that researchers should consider when making decisions about whether and how to use state test data in education experiments. The views expressed in this report are those of the author and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

November 2009

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

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

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

Disclosure of Potential Conflicts of Interest

There are five authors for this report with whom IES contracted to develop the discussion of issues presented. Dr. Henry May is an employee of the University of Pennsylvania, and Drs. Irma Perez-Johnson, Joshua Haimson, Phillip Gleason, and Ms. Samina Sattar are employees of Mathematica Policy Research, Inc. (Mathematica). The authors and other staff of the University of Pennsylvania and Mathematica do not have financial interests that could be affected by the content in this report.

Foreword

The National Center for Education Evaluation and Regional Assistance (NCEE) 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.

Subjects selected for *NCEE Reference Reports* are those that examine and review rigorous evaluation studies conducted under NCEE to extract examples of good or promising evaluation practices. The reports present study information to demonstrate the possible range of "solutions" so far developed.

In this way, *NCEE Reference Reports* are aimed to promote cost-effective study designs by identifying examples of the use of similar and/or reliable methods, measures, or analyses across evaluations. It is important to note that *NCEE Reference Reports* are not meant to resolve common methodological issues in conducting education evaluation. Rather they present information about how current evaluations under NCEE have focused on an issue or selected measurement and analysis strategies. Compilations are cross-walks that make information buried in study reports more accessible for immediate use by the researcher or the evaluator.

This *NCEE Reference Report* is intended to help researchers evaluate and make decisions about whether and how to use data from state-administered proficiency assessments in randomized education studies. Securing data on students' academic achievement is typically one of the most important and costly aspects of conducting education research studies. As state assessment programs have become practically universal and more uniform in terms of grades and subjects tested, the relative appeal of using state tests as a source of study outcome measures has also grown. The variation in state assessments—in both content and proficiency standards—nevertheless complicates decisions about whether a particular assessment is suitable for research purposes and poses difficulties for combining results across multiple states or grades. This discussion paper outlines the issues that researchers should consider when deciding whether to use state test data for evaluation purposes and highlights best practices that can help inform ongoing and future experimental studies. Many of the issues discussed are also relevant for nonexperimental studies.

Acknowledgments

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

The authors would like to thank Mathematica Policy Research, Inc. staff members, including Dr. Melissa Clark and Dr. John Burghardt for their careful review of the draft document. The editing and production staff included Cindy George, John Kennedy, Cindy McClure, and Jane Nelson. We are also indebted to Dr. Frank Jenkins and Dr. Russell Cole for their thorough review of the draft document and their thoughtful and helpful comments. We thank all these individuals for their important contributions to this discussion paper, as well as any others we may have omitted unintentionally.

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

CONTENTS

I. INTRODUCTION

Securing data on students' academic achievement is often a central challenge faced by researchers conducting education experiments. These data are typically obtained by either (1) administering an assessment to students as part of the study or (2) collecting students' test scores on existing assessments administered by states or districts. Three trends have increased the relative appeal of the second strategy:

1. ***Growth in Statewide Proficiency Assessments.*** Requirements under the No Child Left Behind (NCLB) Act of 2001 and parallel standards-based state education reforms have led nearly all states to test students yearly in grades three through eight and in at least one grade in high school. Because of the adequate yearly progress (AYP) provisions in NCLB, these tests have significant stakes, leading school staff to encourage nearly all students to take the tests and to apply themselves, which increases the potential value of state tests as a comprehensive measure of students' academic achievement. Because educators and policymakers are mindful of student performance on these tests, the ability of a program to demonstrate impacts on state assessments also increases the likelihood that evaluation results may lead to changes in education policy and practice.
2. ***Declines in the Relative Costs of Using Test Data from States and Districts.*** As states and districts develop electronic databases with unique student identifiers, the costs of securing and using these data for research purposes have declined. Finding cost-effective ways to collect student achievement data is important because obtaining these outcome measures can be among the largest costs of an experiment.
3. ***Growing Demand to Minimize Testing Burden on Students and School Staff.*** With the growth of state proficiency tests and of formative assessments designed to prepare students for these tests or to help tailor instruction, some educators have become increasingly concerned about taking additional time out of the school day to administer a separate study test to students. Both the U.S. Department of Education (ED) and Office of Management and Budget (OMB) have sought to minimize burden on students and school staff. Recruiting districts and schools for experimental studies is sometimes easier if the study team commits to relying on existing assessments rather than adding a new assessment.

Although these trends have made state proficiency tests a more common and increasingly appealing source of outcome measures, the use of these assessments in education experiments

can also pose important challenges. Substantial variations in state assessments complicate the determinations about whether a specific assessment is suitable for research purposes and pose difficulties for combining results from multiple states and/or grades. States vary in both their academic content standards and their proficiency standards and, hence, in the focus of their assessments and the meaning of commonly reported results, such as proficiency rates. Even states with similar content standards can vary in the coverage of such standards within state assessments, the format of assessment tasks (for example, multiple choice versus constructed-response items), and the degree to which test items align with the standards (Achieve 2002). States also vary in the consequences or "stakes" attached to student, teacher, or school performance in these assessments. These and other differences among state assessments have raised questions about whether and how studies should use these assessments, and about whether findings based on distinct state assessments can be compared and/or synthesized.

In an effort to shed light on these issues and inform future study design decisions, the Institute for Education Sciences (IES) asked Mathematica Policy Research, Inc. (MPR) to examine the issues and tradeoffs that researchers should consider when deciding whether and how to use student test scores on state-administered assessments for experimental impact evaluations. This discussion paper aims to alert researchers to issues surrounding the use of state test data for evaluation purposes. It also seeks to identify best practices that can help inform ongoing and future experimental studies. Although our focus is on randomized control trials (RCTs), many of the issues discussed are also relevant for nonexperimental studies.

The remainder of this discussion paper is organized in three parts. Part II focuses on issues related to the decision of *whether* to use state tests in evaluation research. Part III focuses on issues related to *how* state tests are used in evaluations. In each part, we identify important questions and assumptions that researchers planning RCTs should contemplate. Throughout, we

provide specific advice for dealing with methodological caveats and point out instances in which decisions can influence study results. Part IV provides conclusions and recommendations.

In addition, two appendices provide important background information and additional context for the discussion paper. This information is provided especially for readers who may not be fully familiar with the current landscape and recent evolution of state assessment programs in the United States, or with how data from state proficiency tests are commonly used within education experiments. Appendix A highlights important characteristics of and recent trends in state assessment systems. Appendix B describes how state tests have been or were planned to be used in recent IES-funded studies; this information helped anchor our discussion of issues related to the use of state tests to those which various research teams called attention to.

The material in these appendices reveals a number of important themes about state proficiency assessment systems in the U.S. that researchers should bear in mind as they read Parts II and III of this discussion paper:

- ***State assessment programs have become practically universal and more uniform in terms of grades and subjects tested.*** All 50 states test students yearly in English/Language Arts (ELA) and math in grades 3 through 8, and at least once in grades 10 through 12. Most states also test students yearly in science, but such assessments are administered only in selected grades. Testing in other subjects and other grades is less prevalent.
- ***The design of state tests generally reflects their main purpose, to assess skills relative to state-specified proficiency standards.*** This objective is reflected in at least two important traits of state testing systems. First, there is notable diversity in the structure, content, and emphasis of tests across grades and states, which reflects the diversity in states' academic standards. Second, state tests consist primarily of multiple-choice items sampled broadly from states' many content and proficiency standards. Such broad sampling is consistent with a desire to assess proficiency relative to the entire set of standards. Furthermore, multiple-choice tests tend to produce highly reliable scores for the overall student population, which is desirable given the high stakes attached to proficiency determinations.
- ***The diverse content of state assessments complicates the task of determining whether a particular test is suitable for research purposes.*** It also poses important

challenges when deciding whether and how to combine evaluation results based on distinct assessments.

- ***The multiple-choice format of many assessments raises other important challenges for evaluations.*** For instance, the reliability of such tests tends to be highest around the cut point of interest—in this case, the scores that define proficiency—and can be much lower for students at the tail-ends of the test score distribution (that is, very high- or very low-performing students). Multiple choice tests might be relatively more prone to ceiling and floor effects and therefore of potentially limited value for evaluations examining the effects of interventions targeting high- or low-performing students. Another concern is that multiple-choice tests do not measure higher-order skills well. Thus, test scores on state assessments might not be appropriate for evaluations of interventions focused on such outcomes.

- ***A key advantage to using state assessments is that the cost of obtaining these data is typically much lower than the cost of administering new tests.*** Nevertheless, the process to gain access to state test data is not necessarily simple. Researchers intending to use state test data should therefore have a clear understanding of the steps necessary and allow sufficient time for data collection from the appropriate state and/or local education agencies.

- ***Many studies funded by the Institute of Education Sciences (IES) rely on state assessments as a source of outcome measures.*** Such studies tend to evaluate a diverse set of interventions generally focused on improving students' overall achievement (in one or more subject areas) and/or their ability to meet states' academic standards. Estimating program impacts using students' test scores seems appropriate in such contexts.

- ***Many of these studies are nevertheless conducted across multiple states and/or grades, and it is not always clear if the necessary assumptions to aggregate results have been met.*** Study reports do not make clear whether or how the research teams established that the state tests were sufficiently well aligned with key outcome objectives for the intervention. When results based on tests for different grades and/or states are combined, reports do not typically discuss whether the rescaling is appropriate given characteristics of the study sample, the different tests administered across grades or states, and the intervention's overall target population.

- ***Possible changes to relevant Federal and state legislation could prompt changes to state assessment policies.*** Such changes, in turn, would prompt changes in the types of state test data collected and potentially available to researchers for evaluation purposes. Researchers should therefore be mindful of the issues and assumptions in using state tests for education evaluations.

II. WHETHER TO USE STATE TESTS IN EDUCATION EXPERIMENTS

Researchers have numerous issues to consider when deciding whether or not to use state assessment data in an RCT. Nearly all of these issues can be thought of as related to either the suitability or feasibility of using state assessments. Suitability issues relate to whether the state assessment(s) will provide accurate and useful information about the effects of an intervention.[1] Feasibility issues focus on the practical aspects of obtaining access to the necessary state test data.

Our identification and discussion of issues associated with the suitability of state assessments in evaluation research is guided by two key concepts from basic measurement theory. The first concept is *validity*, which we define as the degree to which the state assessment adequately measures the outcomes targeted by the intervention. Validity issues include considerations about the relevance and appropriateness of the assessments for the intervention and its target population. The second concept is *reliability*, which we define as the degree to which the state assessment provides scores that are sufficiently free of random measurement error that they can be used to detect program effects.[2] Reliability issues concern the precision

[1] In judging the suitability of *any* assessment, it is clearly important to review the technical details for the test(s) under consideration. Researchers intending to use a state assessment in an evaluation should obtain the technical manual or report published by the test developer or the state department of education. In many cases, states have published these reports on their websites. In other cases, a researcher might need to contact the office of assessment within the state department of education to obtain the technical manual.

[2] Reliability is defined strictly in terms of random measurement error as a component of total variability in test scores. This definition does not take into consideration systematic error that would result in measurement bias. Although the presence of random measurement error can detrimentally affect the statistical power of a treatment-control comparison, its presence would not produce a systematic under- or overestimation of the treatment effect (that is, bias). On the other hand, measurement bias that operates differently for treatment and control groups (for example, treatment group scores biased upwards due to a Hawthorne effect) might result in biased impact estimates. Such measurement bias threatens the validity of an instrument for use as an outcome measure in an experiment (see section on "Assessing the Validity of State Assessments for Evaluation Purposes").

(and interpretation) of the scores produced and their sensitivity for detecting differences among groups and changes over time.

Our definitions of reliability and validity align closely with definitions from the psychometric literature (see AERA, APA, and NCME 1999; Lord and Novick 1968), while focusing explicitly on the suitability of state assessments as an outcome measure in randomized evaluations. Note also that both concepts are interdependent in that it is impossible to have a valid measure without sufficient reliability (that is, an unreliable score is too noisy to be a valid measure of anything) and it is not useful to have a reliable measure that is not a valid indicator of the outcome of interest.[3]

In the sections that follow, we discuss issues related to assessing the validity and reliability of state test scores as they relate to randomized evaluations of program effects. We also discuss feasibility issues, which concern the task of gaining access to state test scores and the ability to link individual-level data from one year to the next.

A. ASSESSING THE VALIDITY OF STATE ASSESSMENTS FOR EVALUATION PURPOSES

Perhaps the most important first step when evaluating the utility of state assessment data in an RCT is to identify the outcomes specified in the research questions. If the research questions focus on specific skills that are theorized to be influenced by the intervention, then the state assessment will be useful to the extent that it measures those specific skills. On the other hand, if the research questions focus on the ability of the intervention to improve overall performance on

[3] Some researchers might be used to thinking of suitability issues in terms of the statistical concepts of precision and accuracy. Precision refers to a relative lack of error variance and is largely redundant with the measurement concept of reliability. Accuracy refers to a relative lack of bias (i.e., the correct answer is the most likely result), which is a necessary but not sufficient condition for validity—both accuracy and precision are required for a test to be valid.

the state assessment, then the skills and knowledge measured by the state test comprise the de facto outcomes targeted by the intervention.

In evaluations in which the research questions focus on specific skills, the alignment of the state test can be established by determining the proportion of test items that measure skills and knowledge targeted by the intervention.[4] In evaluations in which the research questions focus on overall achievement or proficiency as defined by the state test, it would be important to justify the expectation that the intervention can have a significant impact on such broad measures of student performance.

Subject Area and Test Domain Alignment. A more detailed examination of the issue of alignment between the assessment and the intervention might be the next crucial consideration in establishing the validity of a state assessment as a source of outcome measures in an RCT. This is because, generally speaking, estimated impacts will typically be largest when the outcome measure aligns closely with the outcomes targeted by the intervention and smaller when outcome measures and the intervention's targeted outcomes are not closely aligned. In other words, the largest impact estimate would be expected when the test measures those aspects of student performance that the intervention is designed to affect.

The most obvious aspect of alignment concerns the subjects tested and the domains tested within each subject. Consider, for example, a study in which researchers are considering using the scores from a third-grade state assessment in English Language Arts (ELA) to evaluate the

[4] More sophisticated methods for evaluating alignment between an assessment and an intervention come from research on the alignment of curriculum standards and state assessments. Such studies of alignment generally focus on content match, breadth of knowledge, balance across standards, cognitive demand (that is, challenge level), and the inclusion of irrelevant material (AERA 2003). Norman L. Webb (2007) has developed a process to match curriculum standards and assessments along four criteria related to the categorical concurrence, depth of knowledge, range of knowledge, and balance in content coverage. Porter and colleagues (2008) describe an alignment index that can be used to describe alignment not only between standards and tests, but also for textbooks and even classroom instruction, and that can also be used in analyses.

impacts of an intervention that relies heavily on techniques of guided and shared reading in order to develop students' fluency and comprehension. The theory of action behind the intervention is that, in order for students to comprehend what they read, they must be able to read fluently. Hence, the intervention seeks to improve comprehension primarily through improvements in fluency. The state assessment uses multiple-choice items exclusively to measure ELA achievement. It produces an overall score and two subscale scores: (1) reading comprehension and (2) vocabulary. Without an additional measure of reading fluency, the evaluation risks failing to detect a program effect, because the state test will not provide valid information about the primary outcome targeted by the intervention—reading fluency.

This lack of alignment is even more obvious when the intervention of interest focuses on a subject that is not included in the battery of state assessments. This is a common problem for interventions in early literacy, social studies, science, and most high school subjects because a state assessment might not exist for the targeted subject and grade. In cases in which a sufficiently aligned test simply does not exist, the use of state assessments for RCT outcome measures might not be a viable option.

There are, nevertheless, instances in which the state test could be used despite imperfect alignment with the intervention. If the goal of the intervention is to improve reading or math skills in the context of other classes (for example, social studies or science), it might be defensible to use the state reading or math test scores that reflect general (rather than subject-specific) skills as an outcome measure. In the reading intervention example above, although the primary targeted outcome (reading fluency) was not captured by the state test, a secondary outcome (reading comprehension) was captured. It might therefore be argued that the reading comprehension scores can still serve as a useful outcome measure. However, the study's power for detecting a treatment effect might be lower if the size of the intervention's effect on reading

comprehension is smaller than for fluency. Perhaps the best situation involves multiple measures, including both direct and indirect outcomes as specified in the intervention's theory of action. Analyses would produce impact estimates for multiple outcomes (with appropriate adjustments for multiple statistical tests), thus providing a comprehensive test of the program theory of action.

Last, a well-aligned state test that is administered more than a year after intervention might still provide value in the context of an RCT. This situation is quite likely for high school interventions, in which the state assessment is typically administered in only one grade or as end-of-course tests. For example, May and Robinson (2007) conducted an RCT in which they found positive impacts of individualized assessment reports and feedback on student performance on the Ohio Graduation Test (OGT), a test which is administered in the 10^{th} grade, but also in 11^{th} and 12^{th} grades for students who do not pass on their first attempt.

Breadth and Depth of Test Items. Related to the issue of subject area and domain alignment are the specific knowledge and skills assessed by the state test. Most state tests rely on 40 to 50 multiple-choice items (see Appendix A). With this in mind, a researcher considering a state test as an outcome measure for an RCT should include in his or her appraisal of test content whether the types of knowledge and skills that the intervention is designed to foster are captured sufficiently well by the *items* on the state test. This assessment can be based on the proportion of test items that measure skills and knowledge targeted by the intervention. Note, however, that this simple approach does not take into account the *difficulty* of relevant test items.

Stakes of Testing. Researchers should also investigate the stakes or consequences attached to performance on the state assessments. Impact estimates based on low-stakes tests might be biased by motivational differences among students; those based on high-stakes tests (for example, state accountability tests) might be biased by cheating. Common sense suggests, and

research confirms, that performance on low-stakes or no-stakes assessments tends to be lower than on high-stakes tests (Wise and DeMars 2005; Segal 2006). If an RCT evaluation includes an administration of its own assessment, a lack of incentives for students to perform well might lead to biased results if treatment and control students do not put forth the same level of effort on the test.[5] This problem of differential motivation with low-stakes tests might be even more difficult to address than the problem of cheating associated with high-stakes tests (see Amrein and Berliner 2002). Although cheating might be kept to a minimum by implementing controlled testing procedures in high-stakes situations, ensuring that student motivation is consistently high in low-stakes situations might be impossible without external performance incentives (for example, rewarding students for high scores). Therefore, one potential advantage of using scores from a state assessment is increased validity due to the presence of high-stakes incentives and a tightly controlled testing environment. In addition, it is important that motivation and incentives, which are likely to vary across states and districts, be balanced across the treatment and control conditions, presumably through random assignment.

Participation Rates. Another important consideration for research teams evaluating whether to use state tests is participation rates. Fortunately, participation rates for state assessments are generally very high (Riddle 2005) and are almost always higher than the minimum participation rate set by RCT researchers. This is because federal and state accountability programs require very high participation rates in state testing programs (see Appendix A). Although there might be little incentive for students to participate in an assessment administered exclusively as part of a

[5] Although this type of differential bias in which the treatment and control group means are under- or overestimated by different degrees will translate into biased impact estimates, there might be instances in which measurement bias exists for both groups but does not translate into a biased impact estimate. For example, if the mean performance of both treatment and control groups is consistently underestimated due to lack of motivation to perform on a low-stakes test, the treatment impact estimate will be unbiased so long as the underestimation of the mean for both treatment and control groups is the same. However, the likelihood of differential bias between treatment and control groups makes this a tenuous assumption.

research study, participation in state assessments is typically mandatory. Furthermore, differential participation by the treatment and control groups might be more likely to occur when using an external assessment (for example, if control students become dispersed across many schools or districts and are difficult for the researchers to locate). This suggests that state assessments might be more attractive as outcome measures in RCTs simply because participation rates are so high.

Testing Accommodations and Exemptions. Researchers should pay attention to accommodation and related policies that might influence the quality of test data, participation rates, and/or test performance for the study population or important subgroups of interest (such as English language learner [ELL] students). For example, if an alternate version of the test is used or testing accommodations are common within the study population, the psychometric properties of the test used by the state might not meet the requirements of the RCT study. Furthermore, tests with and without accommodations and alternate versions of the tests could provide incomparable scores that cannot be analyzed without additional assumptions and adjustments to the statistical models. Perhaps the key issue for RCT evaluations is determining whether the prevalence of accommodations is different for the treatment versus the control group. If accommodations are relatively rare, and no more common among one group than the other, then they might not be much of a concern. If so, data from students taking alternative forms might simply be excluded from the impact analyses.

Comparability of Test Scores Across Grades and States. A final consideration affecting the validity of state assessments scores as outcome measures in RCT evaluations is whether the study involves more than one grade and/or more than one state. In such cases, the work of defining the targeted outcomes is complicated by the fact that each test emphasizes different outcomes (reflecting differences in both grade level and overall state standards). In turn, this

variation in outcome measures complicates the appraisal of the alignment between the tests and the intervention. For studies in which the research questions focus on specific skills targeted by the intervention, dissimilarities in the state tests raise serious concerns about the validity of using different tests, given that the objective is to estimate impacts on clearly defined outcomes. Alternatively, when the study involves an intervention intended to have effects on students' ability to meet state standards, variation in state standards and assessments might accurately represent the breadth of outcomes targeted by the intervention. We discuss issues related to multistate and/or multigrade RCTs in more detail later in this section.

B. ASSESSING THE RELIABILITY OF STATE ASSESSMENTS

As defined earlier, reliability is the degree to which an assessment provides scores that are sufficiently free of random measurement error that they can be used to detect program effects. One general mathematical representation of reliability is:

$$(1) \qquad r = 1 - \frac{\text{var}(\varepsilon)}{\text{var}(\text{Y})}$$

where var(Y) is the total variance in the outcome Y, and var(ε) is the error variability in Y.[6] In other words, reliability is the proportion of variance in an outcome that is not measurement error.[7]

[6] Classical true-score measurement theory states that observed scores comprise two components: (1) a true-score component, which reflects the true performance of the individual; and (2) a random measurement error component. Estimates of reliability seek to partition total variance in observed scores into true-score and error components.

[7] Common techniques for estimating reliability include Cronbach's alpha, split-half, and test-retest reliability. Each technique uses correlations between items and/or overall scores to estimate the proportion of observed score variance that is not attributable to measurement error. Note that reliability requires both precision and variability in scores. Correlation-based measures are undefined, corresponding to a total absence of reliability, if every student receives the same score (that is, var(Y) = 0). Thus, a test must be developmentally appropriate for participating students in order for the data to be reliable.

Reliability is important in the context of RCT studies because it influences the statistical power of treatment-control comparisons (Zimmerman and Williams 1986). For example, an experiment with 80 percent power to detect a 0.20 standardized mean difference in true scores is reduced to having power of 71 percent if the actual reliability of the outcome measure is 0.80 (that is, the uncorrected minimum detectable effect [MDE] is 0.20 while the MDE after adjusting for unreliability is 0.22).[8]

Nearly all standardized tests, including state assessments, have published estimates of test score reliability and/or standard errors of measurement. Both of these statistics are indicators of the precision of test scores—standard errors are the reciprocal of precision—and are usually found in the technical manuals published by the test developer or the state department of education.

What is not usually reported in technical manuals are *conditional* reliabilities or *conditional* standard errors of measurement, which show how the precision of test scores changes depending on the *value* of the score (Lord and Novick 1968; Hambleton and Swaminathan 1984). For most assessments, reliability is maximized near the average score on norm-referenced tests or near performance cut-scores on criterion referenced tests, with a downward curve as scores move away from the average or cut-scores (Hambleton, Swaminathan, and Rogers 1991). This suggests that the reliability of relatively high or relatively low scores can be much worse than the reliability of scores near the cut-points or the average score on a state assessment. When the

[8] Most power analyses for RCT studies do not explicitly account for unreliability in the outcome measure. Instead, the usual practice is to specify an MDE size relative to the total observed variance in an outcome measure. Assuming that the treatment has no effect on measurement error (which must be true if the measurement error is random noise), then the actual MDE of a study is equal to the unadjusted MDE divided by the square root of the reliability. It should also be noted that a few studies have argued that greater reliability can actually decrease power, but this is true only if there is a simultaneous increase in true score variability (Zimmerman and Williams 1986).

performance of students is high enough or low enough to produce ceiling or floor effects,[9] the reliabilities of assessments can be reduced dramatically. In fact, a test would have *no* reliability if every student in the sample got all items correct or incorrect, which also makes it impossible to detect treatment effects.[10]

This decreased reliability of high and low test scores has important implications for evaluations in which the intervention is focused on relatively high- or low-performing students. In these cases, the state assessment might be an ineffective measure of achievement for the population of students targeted by the intervention. For high-performing students, the state test *for their grade level* may be too easy, whereas for low-performing students it may be too difficult. Again, this drives down the reliability of the state test scores and reduces the study's power to detect an effect of the program.

Because conditional reliabilities or conditional standard errors of measurement are rarely published, it can be difficult to ascertain whether a state assessment can be expected to produce reliable scores for a particular study population or subpopulation of interest. To make this determination, researchers should consult with the state office of assessment or the test developer to determine whether the proposed test is unusually hard or easy for the population of students targeted by the intervention.[11]

---

[9] A ceiling effect occurs when many students get every item correct, so that it is difficult to make distinctions among high-performing students. Likewise, a floor effect occurs when many students get every item incorrect, making it difficult to distinguish among low-performing students.

[10] In this situation, total variance would be zero and the general formula for reliability and correlation-based measures of reliability would be undefined, as discussed above.

[11] If data are available prior to implementing the intervention, scatterplots of pretest data for the study population may be used to identify potential ceiling and floor effects. See Cronin (2005, p. 10) for an example of scatterplots showing ceiling and floor effects. If data are not available, a simulation study should be conducted based upon hypothetical distributions of scores and conditional reliabilities for the target population. Such a study would extend the typical Monte Carlo power simulation (May 2005; Muthén and Muthén 2002) to include (a) an error term whose variance increases as scores move away from the mean to represent variation in reliability, and (b) maximum and minimum values to represent ceiling and/or floor effects. Such a simulation would reveal power to detect effects

C. ASSESSING THE FEASIBILITY OF COLLECTING STATE TEST DATA

The single most-expensive aspect of an RCT is typically data collection. This is especially true when the evaluation calls for the administration of an external student achievement test, which can be expensive in terms of test materials and scoring, school participation incentives, and the effort required to recruit schools willing to administer yet another test. Therefore, the feasibility and cost of gaining access to state test data should be evaluated relative to the feasibility and cost of collecting external assessment data. In general, the cost of obtaining state test data can be expected to be far lower. In this section, we nevertheless highlight factors that can make collecting state test score data somewhat difficult and costly.

Gaining access to student-level test data is not an easy task, nor should it be given the need to protect students' personal information. The Family Educational Rights and Privacy Act of 1974 (FERPA) prohibits educational agencies from disclosing students' personal information to other agencies, except under specific conditions (U.S. Department of Education 2008). In some cases, written permission from the parent or guardian of each student is required to disclose personally identifiable data including individual test scores. The difficulty of accessing existing test scores depends on several factors, including (1) whether a state or school district sponsors the study, (2) the dispersion and mobility of students in the study sample, (3) the extent to which the data can be obtained from state education agencies rather than individual districts, and (4) the complexity of the state or district research application process.

Study Sponsorship. Securing school records without parental consent is sometimes possible when the study is sponsored by a state or district. FERPA allows data to be disclosed without

(continued)
given the conditional reliability of the test and the likely occurrence of ceiling or floor effects. Furthermore, grade equivalent scores or vertically scaled scores, if available, could be used to inform these analyses by pinpointing the likely performance of students in the target population (for example, students targeted by the intervention might be those who score between one and two grade levels below their current grade).

written permission from a parent so long as the evaluator is "conducting studies for, or on behalf of, educational agencies or institutions" (FERPA, 20 U.S.C. § 1232g; 34 CFR Part 99). The educational agencies covered by this provision include school districts, state education agencies, postsecondary institutions, and the U.S. Department of Education. However, states and districts have varying interpretations of this FERPA provision. Hence, even after addressing FERPA requirements, it is often necessary to satisfy additional state or district rules restricting the release of school records.

Dispersion and Mobility of the Study Population. For studies involving a highly dispersed or mobile student population it might be necessary to secure records from a larger number of districts and states, which can increase the costs of data collection. The dispersion of the study sample hinges on the intervention and study design. For example, an evaluation of charter schools in a suburban area that involves random assignment of applicants might result in a highly dispersed control group spread over multiple districts. By contrast, an experiment that randomly assigns an intervention to classrooms within a set of schools, or even schools within a district, is likely to have a more-concentrated sample. The mobility of the sample following the point of random assignment and the length of the study's follow-up period will also affect the dispersion of the sample.

Contact Point for Data Collection. It is generally easier and less costly to obtain records for a high proportion of the sample when these are available at the state level rather than from individual school districts; this is especially the case when the study sample is dispersed and mobile. Although many states are currently developing statewide databases with individual-level data, some states still cannot provide such data. In cases in which state databases do not include the data required for the study, a researcher will likely need to submit data requests to individual districts or schools participating in the research. However, even under this scenario, the process

of obtaining state test score data is likely to be less costly than administering an external assessment to individual students in an expanded sample of schools.

Research Application Process and Requirements. The cost and time involved in securing data from any state or district also depends on the research application process. Typically, districts and states require a written data request in which the researcher states what data are being requested, what research questions will be addressed through analyses of the data, and how data security and confidentiality will be ensured. Some states and/or districts require compensation to cover the costs of processing and filling the data requests. The length of time required for review and approval of the request and transfer of the data varies, but typically ranges between one and six months. In general, researchers should expect negotiations with individual states or districts to take at least three months. These negotiations often require the establishment of formal data use/sharing agreements and the establishment of mechanisms for the transfer and storage/backup of sensitive data.

III. HOW TO USE STATE TEST DATA IN EDUCATION EXPERIMENTS

When the determination is made to use state assessment data in an RCT—that is, once researchers have determined that gaining access to the state data is feasible and that state test scores can be expected to produce reliable and valid information about the impacts of the intervention being evaluated—additional issues related to analyzing the data must be considered. Our discussion does not cover all potential designs or analyses for studies that rely on state assessment data. Instead, we aim to highlight important considerations for researchers formulating their analysis plans and to provide recommendations for deciding among alternative methods.

This part is organized into four sections. The first two sections focus on the costs of obtaining baseline data and the methods for and benefits of using baseline data to increase statistical power. The third section discusses analysis of different types of test scores for estimating program impacts in an RCT. The final section examines issues and methods for combining impact estimates across multiple grades and multiple states, including the assumptions underlying different approaches and recommendations for choosing appropriate methods.

A. WHETHER TO SECURE BASELINE DATA

Random assignment in an RCT enables the construction of treatment and control groups that are statistically equivalent prior to the implementation of an intervention. This prior equivalence makes it possible to estimate the impact of an intervention using only posttest data collected from both groups after the intervention is completed. The primary advantage of this approach is that only one wave of data is collected, eliminating the need to link multiple waves of data and the potential errors associated with such linking. The primary disadvantage of such posttest-only

analyses is that analyses that utilize more than one wave of data (for example, covariance analyses[12] and repeated measures analyses[13]) typically have much greater statistical power (Shadish, Cook, and Campbell 2002).[14]

Increased Statistical Power. Bloom, Richburg-Hayes, and Black (2005) advocate the use of baseline covariates to improve power in multilevel RCTs. They show that even the use of aggregate school-level data (which are *very* easy to obtain from district and state websites) as a baseline covariate can dramatically increase the power of a *school-level* RCT. We agree that the benefits of getting baseline data (aggregate or individual-level) will generally outweigh the costs of obtaining such data. In particular, using prior state test results as baseline covariates could lead to a substantial decrease in overall study costs—the improvement in power from the baseline tests is so large in many contexts that the overall sample size can be greatly reduced (thus reducing other data collection costs) while maintaining the target level of statistical power. Baseline data can also help establish the equivalence of treatment and comparison groups and facilitate assessments of the potential for nonresponse bias in impact estimates.

Data Linking. Although efforts should be made to maximize statistical power within the available resources, researchers should recognize that substantial effort might be required to link longitudinal data from state assessments. Fortunately, in our experience, the effort and cost of linking multiple waves of data from state assessments are far less than the costs typically

[12] Covariance analysis includes two types of mathematically equivalent models: (1) analysis of covariance (ANCOVA), and (2) linear regression in which a pretest is included as a predictor (also known as a covariate) in the regression model. Here we use the term covariance analysis to refer to either model.

[13] Repeated measures analyses include several types of analyses in which multiple measurements are available for each individual in the analysis. Popular statistical models for repeated measures include multivariate analysis of variance (MANOVA), multivariate hierarchical linear modeling (MHLM), and growth curve models.

[14] Additional advantages of collecting baseline data include the ability to confirm equivalence of treatment and control groups on observable covariates, and analyses of potential sampling bias due to attrition.

associated with increasing statistical power by administering additional waves of external assessments.

Arguably, the benefits associated with one or two additional waves of prior assessment data outweigh the costs in even the worst case scenario. More specifically, the worst case scenario involves a state in which student identifiers are assigned at the school or district level, the study involves multiple schools or districts in a state, and student mobility across districts is common. This makes it impossible to link student records across waves using only a numeric identifier. The linkages must make use of additional identifiers, such as students' names, birthdates, and demographic characteristics. Fortunately, data-linking programs exist that implement probabilistic matching algorithms designed to deal with common database errors or inconsistencies, such as incorrectly keyed ID numbers or birthdates, transposed first and last names, and nicknames (for example, Jon instead of Jonathan).

Other states present the best case scenario, in which such linking problems are minimized because the state has taken great care to implement a longitudinal database, including the use of state-assigned student identifiers, in which multiyear histories of test scores are available for individual students. This variation in data quality across states and variation in sophistication of state databases means that the costs associated with linking multiple waves of state assessment data must be evaluated separately for each state. Fortunately, the current trend is toward more states developing longitudinal databases.[15]

In sum, for any study in which student achievement is an outcome, the cost savings associated with increased power are likely to far exceed the added costs of obtaining and linking

[15] An example of such efforts can be seen in the Statewide Longitudinal Data System (SLDS) grant program funded through the Institute of Education Sciences, which has provided funding to 41 states and the District of Columbia to develop or enhance statewide longitudinal data systems (see http://nces.ed.gov/Programs/SLDS/).

prior test score data. Therefore, we conclude studies should generally aim to collect and use baseline data. If the RCT involves student-level random assignment, efforts should be made to link pretest and posttest scores for individual students. If the RCT involves school- or district-level random assignment, linking individual and/or aggregate data can be used to increase power.

B. HOW TO USE BASELINE MEASURES

Baseline measures in an RCT can serve as a mechanism for blocking or stratifying subjects in an RCT, they can serve to confirm the equivalence of treatment and control conditions prior to an intervention, and they can greatly increase statistical power when they are used as covariates in impact analyses. However, pretest-posttest and longitudinal designs that capitalize on the availability of baseline data raise other issues that can influence how state test data are analyzed. Underpinning many of these issues is the comparability of assessment data across grades and across time. The comparability (or incomparability) of tests can guide a researcher to choose between two primary ways to use baseline scores in an RCT: (1) along with follow-up scores to measure students' explicit gains in achievement (which might then be used as an outcome); or (2) as a covariate to adjust statistically for baseline achievement in a regression or ANCOVA framework.

State assessments are generally designed to align with the state's curriculum and/or performance standards at each grade level. This goal is accomplished with varying success by different states (Rothman, Slattery, Vranek, and Resnick 2002). Furthermore, the clarity and quality of progression in the content of the state standards also varies across states (Finn, Petrilli, and Julian 2006; Schmidt, Wang, and McKnight 2005). The implications of this are that, although linking state test scores over time creates a multiyear assessment profile for each student, the achievement tests providing the scores change each year as students progress from one grade to the next.

In some states (for example, Florida), tests for adjacent grades are explicitly linked through psychometric equating (Kolen and Brennan 2004). This so-called vertical equating results in test scores that are on the same developmental scale across multiple grades.[16] Critics of vertical equating nevertheless argue that the shift in content taught and tested at each grade level makes it impossible to equate tests across several grades using a single scale. Under this argument, any attempt to use test scores from adjacent grades to measure absolute change is inherently flawed (Martineau 2006). If the goal is to produce an explicit measure of change in achievement (for example, gain scores), it is important to consider the similarity in what is being tested at each grade level (Linn 1993). We take a pragmatic stance and argue that if the knowledge and skills measured are consistent across grades, or if they exhibit a logical developmental progression over multiple grades, then analysis of vertically scaled scores to produce explicit measures of growth might be the best approach in terms of internal validity and interpretability. Invariably, the selection of statistical models used to analyze data from a multiyear study depends on the number of data points collected and whether the scores are vertically scaled.

In the simplest multiyear study—the pretest-posttest design that involves two years of state test data—there are only two general approaches to analyzing these data: (1) covariance analyses or (2) analyses of difference scores.

Covariance Analysis. The more-prevalent approach to analyzing pretest-posttest data when test content differs across grades involves the use of covariance analysis. Unlike the difference score approach, whose calculations might inappropriately imply learning gains,[17] the covariance

[16] Theoretically, vertically equated tests would produce the same score regardless of which version of the test is taken (Kolen and Brennan 2004; Holland and Dorans 2006). In other words, a fourth grader's score should remain the same, even if she took the fifth-grade version of the vertically scaled test.

[17] We use the term "difference score" here to signify a simple subtraction of the pretest score from the posttest score. Because state tests are seldom equated from one grade to the next, subtraction of these two scores might have little or no interpretation. In the case in which the tests are scaled to have the same mean score in every grade, the

analysis treats the pretest as a control variable to be held constant when estimating group differences on the outcome variable (Wildt and Ahtola 1978). The objective is not to estimate a pre-post gain, but to control for differences on the pretest. In fact, the pretest in covariance analyses need not be directly comparable to the posttest from a statistical standpoint. Scores from a completely different pretest assessment may work well as a covariate in analysis of data from an RCT in so much as that pretest data explains variability in the outcome, thus increasing statistical power to detect program impacts.

A general criticism of the covariance analysis approach is that it is prone to under-controlling, because the pretest regression slope is underestimated due to unreliability in the pretest scores (Sanders 2006).[18] Fortunately, this criticism is not such a serious issue in the context of RCTs, because random assignment usually eliminates the need to adjust for pre-existing differences. In the RCT context, the pretest is primarily a mechanism to reduce error variance and increase statistical power (Shadish, Cook, and Campbell 2002). However, the underadjustment of pretest scores in an RCT can make interpretation of impact estimates less straightforward because group differences are based on regression-adjusted posttest means (that is, residualized gain scores) instead of simple pre-post difference scores. In addition, under-adjustment due to imperfect reliability of measured covariates can diminish power to detect effects (Holland & Rubin, 1983).

(continued)
expected difference for the average student is zero. Thus, difference scores for tests that do not have a vertical scale cannot be used to reflect absolute annual learning gains.

[18] In regression and covariance models, parameter estimates for any predictor variable measured with error will be attenuated toward zero by an amount equal to one minus the reliability of that predictor (see Neter, Kuter, Nachtscheim, and Wasserman 1996, p. 164). Because achievement test scores always have less-than-perfect reliability, the slope estimate for the pretest covariate will be attenuated, resulting in underadjustment of pretest scores.

Difference Score Analysis. The second approach, using difference scores to analyze pretest-posttest data, involves calculating gains by subtracting each student's pretest score from his or her posttest score. Aside from criticisms focusing on the unreliability of difference scores (Cronbach and Furby 1970; see Rogosa and Willett 1983 for a counterargument), a conservative perspective would suggest that this is appropriate only when the tests from adjacent grades are vertically scaled and clearly measure very similar content from one year to the next. On the other hand, a different perspective might enable one to calculate difference scores even when the tests are not vertically scaled and even when they measure different content. After converting scores from each test to z-scores, the difference scores would show differences in performance relative to the average student in standard deviation units.

An important distinction between difference scores calculated using similar tests and those calculated using different tests is in how the difference scores are interpreted. Without vertical equating and similar content, the difference scores would not reflect differences in students' rates of learning. For example, subtracting a student's third-grade math score (which might focus mostly on whole numbers) from the student's fourth-grade score (which might focus mostly on fractions and decimals) will not necessarily reveal how much a student learned between the end of the third and fourth grades. To get that information, the student would have had to take a different pretest focusing primarily on fractions and decimals.

An alternative is to avoid interpreting difference scores as learning gains. Instead, the difference scores reflect only differences in relative performance from one year to the next. For example, a student might move from one standard deviation above the mean to 1.2 standard deviations above the mean. Although this change cannot be attributed solely to learning that occurred in the past year, such shifts in performance are equalized, on average, across treatment and control groups in an RCT. Therefore, any significant difference in the magnitude of such

relative shifts in test scores can serve as unbiased estimates of the impact of the intervention. For example, a significant positive difference between treatment and control groups from an analysis of z-score difference scores could be interpreted as implying that the average percentile ranking of subjects in the treatment group increased more over time than the average percentile ranking of subjects in the control group.

Repeated Measures Analysis. When more than two years of data are available, statistical power can be increased further through the use of repeated measures analyses or growth curve models (Allison, Allison, Faith, Paultre, and Pi-Sunyer 1997).[19] Moreover, growth curve models might improve interpretability when the state test is vertically scaled. This is because results from a growth curve analysis can be benchmarked against the average achievement growth trajectory for a district or a state, to determine the degree to which students are making or exceeding a year's worth of learning gains as a result of an intervention.

It can also be useful (and very inexpensive) to get baseline scores in other subjects for use in a repeated measures model. These additional variables can more fully account for treatment-control differences in baseline achievement and explain additional outcome variation, resulting in further increases to statistical power. This can be achieved by including prior achievement measures on these other subjects as covariates and/or by modeling the covariance structure of residuals for the multiple outcomes in the repeated measures model.

[19] When these models are estimated in an HLM or mixed model framework, they also have the additional advantage of handling missing data without the need for listwise deletion or imputation (Raudenbush and Bryk 2002; Singer and Willett 2003). In both the difference score models and the covariance models a missing pretest score or posttest score for a student means that student is excluded from the analysis unless alternative means for dealing with missing data are implemented such as multiple imputation. Repeated measures and growth curve models enable students with missing data to be included in the analysis, under the assumption that their data is missing at random (MAR) (see Rubin 1987 for a definition of MAR).

C. ANALYSIS OF SCALE, PROFICIENCY LEVEL, OR OTHER TEST SCORES

Most state tests produce at least two types of scores: scale scores and proficiency levels. The primary distinction between them is that scale scores are measured on a continuous scale[20] while proficiency level scores are measured on an ordinal scale. When considering their use as an outcome measure in an RCT, both scores have advantages and disadvantages.

The primary advantage of scale scores is that they provide greater precision (that is, the ability to distinguish the relative performance of students at the high and low ends of the same proficiency level), which translates to greater statistical power to detect program effects. Although proficiency level scores yield lower statistical power, they do support a more intuitive description of program effects. This is because proficiency levels are not just categorized continuous scores, but rather judgments about what cutoff points indicate substantively meaningful attainment of different levels of proficiency. For example, results from a logistic regression analysis of proficiency level scores might be interpreted as showing that "students who participated in the intervention were two times more likely to score proficient or above on the state test." Arguably, such descriptions of the effects of an intervention might be more easily understood than a mean difference in scale scores or a standardized effect size.[21]

It is important to note, however, that each state defines proficiency differently, because both the content of tests differs and states' proficiency cut scores vary (Porter, Polikoff, and Smithshon 2008; Petrilli 2008; NCES 2007). This complicates analyses when data come from

[20] Some measurement experts might be more specific and claim that scale scores are usually measured on an interval scale, which suggests that the intervals between scores are equivalent throughout the full range of scores. In other words, a difference of one point reflects the same degree of difference in knowledge or skills regardless of whether the difference is observed at the low end or the high end of a scale. In truth, scale scores are interval scaled in theory and might not actually be perfectly interval scaled in practice.

[21] Because proficiency level scores are simply a categorization of the scale scores, one analysis option involves a staged analysis in which the first stage of analyses uses scale scores. Then, if a significant program effect is revealed, the second stage of analyses uses proficiency level scores in order to improve interpretability of the results.

more than one grade or from multiple states. It could be argued, however, that effects on proficiency rates are still worth measuring across grades and states because proficiency rates are a key focus of federal, state, and district policy. Caution must nevertheless be exercised when interpreting these kinds of results. We discuss this issue further in the section entitled *Combining Results Across Tests for Different Grades and/or States*.

Some state tests produce additional scores such as normal curve equivalent (NCE) scores, z-scores, T scores, and percentile ranks (see Allen and Yen (1979) for a discussion of common measurement scales). Z-scores and T scores are simply linear transformations of the scale scores (that is, with a different mean and standard deviation), therefore the same issues and methods for scale scores apply to these scores. NCE scores are a non-linear transformation of scale scores that ensures the scores follow a normal distribution with a mean of 50 and a standard deviation of 21. A potential advantage of using one of these popular rescaled scores is that it might serve to place different tests on a common scale, so long as the tests measure the same or similar knowledge and skills. For example, an NCE of 50 on any test corresponds to the average score for the norming sample for that test.[22] Even if states do not provide these additional scores, researchers can typically convert scale scores to T or z-scores, as discussed later in this section. Analyzing scores from different tests on a common scale makes it possible to combine results across different grades and even different states *under certain assumptions.* We discuss these assumptions and additional considerations in combining results across states or grades later in the next section.

[22] A norming sample is the sample of students from the tested population that were included in the original calibration and scaling of the test. For state assessments, the norming sample is representative of the population of students in the state for whom that version of the test was written.

Notably, using percentile ranks to estimate treatment effects is usually not advisable, because these scores are on a cumulative scale such that the absolute size of a 10-point difference in percentile rank depends on its location on the scale (for example, moving from the 70th to the 80th percentile represents a larger shift in underlying ability than moving from the 50th to the 60th percentile). When percentile ranks are available, it is possible to convert these scores to z-scores, T scores, or NCEs based on the quantiles of the normal distribution (see Allen and Yen 1979).

D. COMBINING RESULTS ACROSS TESTS FOR DIFFERENT GRADES OR STATES

RCT evaluations often involve multiple grades or multiple states (see Appendix B). In such instances, researchers must consider whether to combine results across grades or states, and if so, they must determine the best methods for combining results. In some studies, it might be absolutely necessary to combine results across grades or states in order to achieve sufficient statistical power. In other studies, sample sizes might be sufficient to produce impact estimates separately by grade and state, but an overall estimate might also be desired. Whatever the case, the decision of whether or not to combine results should be made during the planning stages of the research design and analysis plan for estimating program impacts.

Combining results across grades might be an important goal for RCTs in which the intervention is designed to have broad impacts on performance across multiple grades or throughout an entire school or district. Combining results across states might be an important goal when the intervention is intended to have consistent impacts across states and an overall estimate of program impacts across the set of states participating in the study is desired. Whatever the circumstance, researchers must carefully consider differences in state tests and whether combining results across grades and states is appropriate. In this section, we present

several methods for combining results across grades or states, and we discuss when these methods are appropriate and when their use might be ill-advised.

Deciding Whether to Combine Impact Estimates. Again, in our view, the decision of whether to combine results across grades or states should be driven, first and foremost, by the research questions underlying the evaluation. These can be classified into two categories addressing different goals. First, if the goal of the research is to demonstrate that an intervention has an effect on students' abilities to meet state standards, then differences among the standards and assessments of different grades and states simply reflect intended variation in the targeted outcome (that is, proficiency on the standards). Arguably, a program that purports to have an impact on the broad set of knowledge and skills encompassed in state standards should be able to produce impacts on any state test, in any grade, and policymakers would want to know about these broad impacts.

Alternatively, if the goal of the research is to demonstrate that an intervention has an effect on specific skills or knowledge, then combining results across grades or states might be inappropriate unless each test provides valid, reliable, and comparable data on the targeted outcomes. More specifically, it is important to consider whether the standards and assessments are sufficiently similar across grades and states to support combining results.[23] If the tests differ substantially in terms of the knowledge and skills assessed or the difficulty of test items, estimated program impacts for one grade or in one state might be systematically different from estimated impacts in another grade or state. From a conservative viewpoint, differences among tests and the lack of formal psychometric equating may preclude the rescaling approaches described in this report. Under this perspective, results should never be combined across grades

[23] Helpful references addressing issues of linking, rescaling, and equating different assessments include Linn (1993); Mislevy (1992); and Feuer, Holland, Green, Bertenthal, and Hemphill (1999).

or states unless the tests can be formally equated using common items or common populations. Even when an argument can be made to support pooling data from different state tests, aggregating to produce an overall impact estimate could mask important variation in measured effects. Therefore, if there are concerns about such substantive differences in tests, and if statistical power allows, one should compare across grades and/or states the ability of *each* test to provide valid and reliable data on the impacts of the intervention on the targeted skills. Furthermore, a lack of alignment between a state test and targeted outcomes might be reason to exclude that state from the study or administer a different assessment.

Whether the data are similar enough to support combining results is subjective at this point, although studies with sample sizes large enough to produce separate estimates for each grade and state might shed light on the extent to which different standards and assessments are likely to moderate program impacts. In fact, because we know so little at this point about the influence of differences in state assessments on possible variations in estimated impacts on targeted outcomes, it might be argued that research focused on specific skills should only be conducted using an external assessment or, when state tests are used, with sufficient sample sizes to estimate impact estimates separately for each grade and state. For studies that involve a large number of grades and/or states (for example, four states with three grades each), moderator analyses might be conducted to explore how the alignment (or lack thereof) between the state assessments and the intervention influences impact estimates.[24]

Deciding How to Combine Impact Estimates. If a decision is made to combine results across grades or states, there are several analytic approaches to doing so. In our discussion of

[24] Moderator analyses utilize regression methods to examine systematic variation in the size of treatment effects as it relates to variation in context or conditions across sites (Baron and Kenny 1986). It is important to note that moderator analyses are exploratory, however. They cannot differentiate between differences in effect sizes due to differences in the tests versus differences in program implementation across states and/or grades.

these methods, we first focus on strategies for analysis of student-level data rescaled to a common metric. This is generally the most powerful and efficient way to combine results across grades and states; however, it also requires strong assumptions about the comparability of assessments and study samples across grades and/or states that should be tested explicitly. Following this, we discuss strategies for combining grade-specific and state-specific impact estimates within a meta-analytic framework. Generally speaking, meta-analytic strategies treat impact estimates based on different state tests as generated from separate, independent studies that jointly provide a distribution of treatment effects. While computationally more intensive, a strength of meta-analytic approaches is that they can explicitly test the tenability of assumptions necessary to generate estimates of average treatment effects.

Rescaling Individual-Level Scores. The simplest method for producing combined impact estimates involves running multigrade or multistate analyses after converting test scores in each grade and state to z-scores (or some other common scale). However, this approach also imposes the most stringent assumptions on the data. Implied in this z-score approach are two key assumptions. First is the assumption that differences in the knowledge and skills tested by the different assessments are inconsequential in the context of the particular evaluation. That is, either the tests measure the same content, or differences in content are accepted as reflecting intended variation in state standards and the desired impact estimate is one which is pooled across states, despite variation in standards. Second is the assumption that differences among the tests consist primarily of differences in the scale of the test scores. In other words, it is assumed (1) that the study sample from each grade and state represents a similar cross-section of the population of students targeted by the intervention and (2) that the underlying distributions of scores from each test are identical, except for differences in the means and standard deviations of the scale scores.

The plausibility of the first assumption can be evaluated by comparing the demographic characteristics of samples from different grades and states, and also by comparing the means and variances of pretreatment test scores from each grade and state to the respective statewide means and variances of test scores for each grade. For example, relative to the statewide distribution of scores, study samples from different grades and states might be found fairly consistently to have an average baseline test score that is one standard deviation below the state average and a variance that is one-half the magnitude of the statewide variance for that grade.

The simple approach of converting to z-scores separately for each grade and state can also be adapted to suit conditions in which study samples are heterogeneous, by standardizing using the statewide means and standard deviations instead of the sample means and standard deviations.[25] If the comparisons described above revealed that the achievement of the study sample was not comparable across grades or states, the statewide means and standard deviations for each grade could be used to rescale test scores by grade and state into a cross-state comparable z-score metric by subtracting the state mean and dividing by the state standard deviation for each grade and state. For example, the resultant z-scores for a study involving fourth graders from two states might then have a smaller range in one state, accurately reflecting the more homogeneous sample from that state.

This method of using state-level means and standard deviations to compare performance and reflect heterogeneity in performance imposes the additional assumption that statewide variance in achievement would be similar within each grade and state if the same vertically scaled test were used for each grade and state. This is because the statewide means and standard deviations

[25] An additional reason researchers might chose to standardize by the state mean and standard deviation is if control group samples in each state are too small to reliably estimate within-sample means and standard deviations (Hedges 1981).

are used to rescale the test scores relative to the statewide distribution of achievement. For example, if the study samples from each grade and state were representative of the statewide populations for each grade and state, then the resulting rescaled scores would have a mean of zero and a standard deviation of one for every grade and state. This result implies that any differences in the means or standard deviations across grades and states are artifacts of the tests, and can be suitably removed by rescaling.

Although it is impossible to fully evaluate this assumption, results from the Florida Comprehensive Assessment Test (FCAT) vertical developmental scale suggest that variance in math and reading scores is fairly consistent across narrow grade spans (for example, no more than three grade levels), and that variance in math and reading scores may decrease substantially over larger grade spans (for example, the variance in tenth-grade math scores is less the half the variance of third-grade math scores) (Coxe 2002). Similar patterns can be seen in the vertical scale of the Stanford-9 reading and math tests (Harcourt 1997).

Regarding variation for a single grade across multiple states, data from the National Assessment of Educational Progress (NAEP) suggest that variation in student achievement for several subjects in fourth or eighth grade is fairly consistent across groups or clusters of states (National Center for Education Statistics 2008). Although there are small but significant differences in within-state achievement variation, the variance estimate for any one state is typically not significantly different from the variance estimates for more than half of the states in the nation. This consistency in variation across grades and states suggests that combining impact estimates across grades and states in RCTs might be reasonable, so long as the grade span is not

wide (for example, no more than three or perhaps four grades) and so long as the states included in the study have similar within-state variability on the NAEP tests.[26]

In addition to assumptions about consistency in means and variances of student performance, a second set of assumptions requires that the *shape* of the distributions of achievement scores be similar across grades and states. The plausibility of this second assumption can be evaluated by comparing the shapes of the distributions of pretest scores from each grade and state through graphical displays (for example, boxplots, histograms, normal-quantile plots). If the distributions of pretest scores appear similar, then the simple linear transformation to z-scores might be sufficient.[27]

If the second assumption is violated, and differences in the distributions cannot be attributed to differences in the samples of students (that is, the target population is similarly represented in each grade and state), then a nonlinear transformation of test scores might be a more appropriate option. The most common nonlinear transformation used to link test scores is called equipercentile equating or linking (see Kolen and Brennan 2004).

In its most basic form, the equipercentile equating approach involves first converting test scores to percentile ranks in each grade and state. The percentile ranks are then converted to z-scores by substituting the value of a z-score from the standard normal distribution associated with each percentile rank. As with the linear transformation, the equipercentile approach assumes that differences in content tested are negligible (that is, for impacts on specific skills) or attributable to intentional variation in state standards (that is, for impacts on standards proficiency). Unlike the linear transformation, the equipercentile approach is able to remove

[26] State-level means and standard deviations on the NAEP in numerous subjects and grades since 1990 can be accessed through the online NAEP Data Explorer at http://nces.ed.gov/nationsreportcard/naepdata/

[27] For an example in which these methods were used to link different assessments across grades, see May and Supovitz (2006).

differences in the shapes of the distributions of test scores from different states. Implied in this process is the assumption that distribution differences are due to differences in the concentrations of easy and hard items on each state's test, and not due to actual differences in the distributions of student achievement across states.

The Importance of a Consistent Reference Population. The objective of any rescaling is to place the test scores on a common metric and *ensure that the interpretation of impact estimates is comparable across grades and states*. This requires that effect estimates reflect treatment-control differences not only in a common scale but also *for a common reference population* (Dong, Maynard, and Perez-Johnson 2008; Lipsey and Wilson 2001; Cooper 1998; Hedges and Olkin 1985). For example, a Cohen's *d* standardized mean difference (Lipsey 1990; Cooper 1998; Cooper and Hedges 1994; Cohen 1988), the most common standardized effect size, is highly sensitive to the reference population whose standard deviation provides the scale for this statistic. In a study in which the target population is consistently represented across grades and states, converting to z-scores within grades and states sets the sample standard deviation to 1.0 in each grade and state, yielding impact estimates in units equal to the estimate of the within-grade standard deviation of the target population.

If the representation of the study sample varies across grades or states (for example, study participants might be relatively disadvantaged in one state and more broadly representative of the overall student population in another state), and statewide means and standard deviations are used to rescale individual scores to a comparable metric, the resultant differences in the standard deviations of scores in each grade or state reflect differences in representation of the target population. When calculating an overall standardized effect size, the standard deviation of the

sample that best represents the target population[28] might be used as a divisor to convert the unstandardized effect size estimate into a rescaled effect size for the specific population of students targeted by the intervention. This would produce an effect size in units equal to the standard deviation of the target population. In any case, without effective rescaling of the individual test scores or calculation of comparable standardized effect estimates separately for each grade and state, combining results across grades and states might produce misleading results.

It is therefore important to determine the extent to which individual state samples represent the overall population targeted by the intervention, and when justified, to rescale scores so that a consistent estimate of outcome variability is used to standardize the impact estimate in different states. To rescale scores from distinct assessments and make them directly comparable, evaluation researchers have two main alternatives from which to choose—using the mean and standard deviation of the sample control group versus using the mean and standard deviation of the population of students in the state. In cases in which there is not sufficient comparability of the study sample across states, we recommend using the state distribution. This rescales each student's score to represent his or her performance relative to other students statewide. Because most RCTs involve subpopulations of students instead of a statewide target population, the standard deviation of the rescaled scores will be less than one. Thus, the treatment-control mean difference must still be divided by the standard deviation of the control group (or another

[28] In general, the choice of estimate should be justified such that it provides a reasonable approximation for the standard deviation of the *population* of students targeted by the intervention. The standard deviation for the target population might be estimated using data from a single grade and state or as a pooled estimate across grades and states. If tests with different score scales are used across multiple grades or states, then the standard deviation for each grade and state would first need to be expressed as the square-root of the ratio of sample variance to statewide variance. For example, a rescaled standard deviation of 0.5 denotes a sample variance (0.25) that is one-quarter of the statewide variance.

estimate of the standard deviation of achievement in the target population) in order to produce a traditional standardized effect size for the target population.

Combining Impact Estimates Using Meta-Analysis. Although any method for combining results across multiple grades or states can be thought of as a variant of meta-analytic methods, this section focuses on traditional meta-analytic models used to combine separate impact estimates (Glass, McGraw, and Smith 1981; Cooper and Hedges 1994). Because effect estimates in meta-analytic models must be on the same scale (for example, standardized mean differences) it is very likely that the researchers will need to use a linear or nonlinear linking technique described above to rescale the test scores or the impact estimates prior to analyses. If linear or nonlinear transformations of individual test scores cannot be implemented due to differences in the study samples across grades or states, separate impact analyses should be conducted for each grade and state, with combined effect size estimates produced only when scores from different states can be rescaled relative to state-level means and variances as described above.

Although similar in some respects to pooling individual test scores after rescaling to z-scores or another common metric, a meta-analysis is theoretically different in that data from different state tests are not treated as though they produce equated scores that can be pooled in a traditional analysis. Instead, a meta-analysis provides an estimate of the distribution of treatment effects from different studies. Variation in treatment effects is expected across grades, states, or both, and this variation may be explained by contextual variables that reflect differences in the tests or in study contexts. When the study design and analysis results support the notion that the variation in effect sizes is (a) due to random sampling variation, (b) adequately explained by contextual measures, or (c) ignorable based on the need for an impact estimate that is pooled across different sets of state standards, then an average treatment effect may be produced along with the standard error of the estimate. Otherwise, if the separate impact estimates cannot be

pooled, then the distribution of effects may be presented without an average treatment effect. It is because of this ability to explicitly evaluate assumptions that we recommend a meta-analytic approach as opposed to indiscriminately pooling rescaled scores whenever cross-state or cross-grade average impact estimates are sought.

In the classical meta-analytic approach, separate effect estimates are produced for each grade and state and then combined using weighted average effect estimates or meta-regression models (including HLM meta-analytic models).[29] In the weighted average estimate approach, each grade or state is treated as a separate substudy with a separate analysis and associated impact estimates. A second stage of analysis combines the grade-specific or state-specific estimates by averaging. Typically, the impact estimates from each grade and state are weighted by their associated sample size or by the inverse of their standard error of estimate.

Alternatively, a meta-regression model could be used to produce a combined effect size. These meta-regression models can be categorized into (a) those that use two stages of analysis in which grade- and state-specific impact estimates are produced in the first stage and then used as the dependent variable in the second stage; or (b) those that rely on a single statistical model in which student-level data are clustered by grade and state via random coefficients (for example, random slopes in HLM models) or fixed coefficients (for example, interactions in a regression model).

[29] Note that a pooled multistate HLM analysis of rescaled scores should produce standardized effect sizes that are similar to weighted averages of standardized effect sizes calculated separately for each grade and/or state (Raudenbush & Bryk, 2002). However, a key benefit of analyzing pooled individual-level data is the ability to produce more efficient estimates through multilevel analyses if the assumptions about the structure of the error term underlying the multistate HLM model hold (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). If the assumptions do not hold, results of this model will be biased. In that case, other models using robust standard errors via generalized estimating equations or Taylor-series estimation might be more appropriate (Liang & Zeger, 1986; Goldstein, 2003).

Method (a), the meta-regression using two stages, is useful only when there is a large number of impact estimates to be combined (at least 10 but preferably 30 or more) and moderator analyses are required (see footnote 24). Not surprisingly, although this method is often employed in meta-analytic and systematic reviews of literature comprised of many separate studies, it is unlikely to be appropriate for IES-funded RCTs given the relatively small number of grades and states in a single study.

Method (b), which employs a single regression model and uses fixed or random coefficients to distinguish impact estimates across grades or states, might be applicable to most multigrade or multistate RCTs. In this case, it is essential that the test scores from each grade and state be rescaled (as necessary) to a common scale (for example, grade-specific z-scores or adjusted z-scores based on statewide means and standard deviations) so that impact estimates from each grade and state are on the same scale.

In addition, whether to utilize fixed or random coefficients/effects in a meta-analysis is a key consideration. Fixed effects meta-analytic models impose the assumption that a single true impact estimate applies to every grade and state, and that differences in estimated impacts across grades and states are due purely to sampling variation and are not indicative of systematic differences in impacts across states. This approach may employ interaction terms involving the grade, state, and treatment indicators to explicitly test whether treatment effects are different across grades and/or states. Alternatively, random effects meta-analytic models assume that treatment effects will vary across grades and states, usually with the specific assumption that the impact estimates are reflective of a sample of impact estimates drawn from the normally distributed population of impact estimates for the population of contexts. Random effects meta-analytic models also allow one to test and model variance in impact estimates using moderator variables (see footnote 24).

Another distinction between fixed and random effects models is that the former produces results that cannot be generalized beyond the sample of grades and states in the study, whereas the latter considers the grades and states in the study to be a sample from a larger population of grades and states to which the results might be generalized. In studies that include multiple grades and multiple states, it is possible to use fixed effects for grades (because generalization beyond the grades included in the study might not be a goal) and random effects for states (permitting generalization to states not included in the study).

It is also important to recognize that, although the assumptions underlying a random effects analysis may be plausible in some studies, a limitation of the random effects analysis is that it requires a relatively large number of grades or states to produce stable estimates. While a fixed effects analysis can be run with as few as two grades or states, maximum likelihood estimation of random effects models can become unstable when the number of states is small (for example, fewer than 10).[30] Given the routinely small number of grades and states in IES-funded RCTs, it is likely that fixed effects methods are most applicable for typical study designs; however, it is important to recognize that use of fixed effects methods implies that the results will not be generalized beyond the sample of grades and states in the study.

Combining Effect Estimates Using Proficiency Scores. The analysis of proficiency category scores is even more complicated than the analysis of scaled scores in multistate or multigrade studies. This is because state assessments have wide variation in cut points and corresponding levels of difficulty for defining proficiency (NCES 2007). This suggests that the impact estimates for an intervention might depend on the difficulty of the state test. On the other

[30] If a random effects model must be run with a small number of states, it might be possible to produce unbiased and consistent estimates using Bayesian estimation. The drawback of the Bayesian approach is that the analyst might be forced to make strong assumptions about the variance in the distribution of treatment effects when the number of grades or states in a study is small.

hand, one might invoke the same argument posed at several points in this report—if the research questions focus on the impacts of the intervention on students' proficiency rates, then differences in the difficulty of the assessments could be ignored because definitions of proficiency are set by state policy and are indicative of the natural variation in what it means to be proficient. A less-extreme position would involve a "middle-ground" approach in which state-specific impact estimates are produced, and variation in impacts across states is modeled using random effects. In studies involving several states, this variation may be analyzed to determine how differences in the difficulty of the tests might moderate program impacts on proficiency rates.

IV. CONCLUSIONS AND RECOMMENDATIONS

The relative appeal and use of statewide proficiency assessments as sources of outcome measures in education randomized control trials (RCTs) has grown in recent years. This discussion paper examined the issues, methods, and assumptions associated with the use of state test data in education experiments. An important theme emerging from this work is that there are numerous important factors that researchers should carefully consider when deciding whether and how to use state test data in RCT evaluations. Such decisions typically have serious implications for the validity and precision of RCT results.

A number of recommendations pertaining to the design and conduct of RCTs flow from our discussions and should help guide researchers considering using state assessments as a source of outcome measures in their studies. They include the following:

1. Gauge the alignment of specific assessments with the outcome objectives of, and research questions about, the intervention of interest.

Arguably the most important first step in assessing the suitability of state tests is identifying the outcomes that the intervention is intended to affect. After defining the outcomes of interest, research teams should gauge the degree to which specific assessments are aligned with those objectives, focusing on the central research questions. If the intervention is expected to affect a relatively narrow and specific set of skills, it is important to gauge whether those skills are captured sufficiently well by the assessment, whether the scores reported include those that pertain to these specific skills, and whether the information across grades and/or states is consistent. If the intervention seeks to improve students' proficiency on state standards, then variation in test content appropriately represents the variation in proficiency goals and standards across grades and states. In other words, in this latter case, state assessments are aligned by

definition and the decision to use state tests as a source of outcome measures and/or to combine results across grades and states seems easier to justify.

In evaluating program impacts, it is also important to identify not only the ultimate outcome objectives but also the intermediate outcomes and mechanisms through which the intervention might achieve such objectives. Elaborating this theory of action can help identify additional outcomes and/or processes that the study should measure. By estimating program impacts across a set of relevant outcomes, an RCT might provide information about potential variation in effects across different outcomes. It is nevertheless important to explicitly connect the outcomes examined to the intervention's theory of action, and to focus mainly on the primary outcomes the intervention intends to influence. When multiple outcome measures are examined, researchers also need to adjust significance levels to account for multiple comparisons.

2. Ensure that the assessment is reliable and appropriate for the study target population.

The power of an RCT to detect program effects is directly related to the reliability of the outcome measure used. It is essential that researchers select instruments that have demonstrated high reliability, producing test scores that are relatively free of random measurement error. It is important to note that a state test that has been shown to produce reliable scores for a statewide or national population might produce unreliable scores if that test is used with a sample of students who exhibit performance that is substantially above or below average. A test that is too easy (or too hard) for the study sample might produce many (near) perfect (or zero) scores, making it impossible to distinguish between the performance of many students and potentially masking program impacts. Other important considerations include content coverage, test format, high versus low stakes, participation rates, and testing accommodations or exceptions.

3. Whenever possible, collect and use baseline measures.

When outcome measures include state assessments, the added effort and/or cost to obtain data from prior years is usually well justified by the associated increases in power and other benefits of having baseline data available. The typically high correlations between waves of annual achievement test scores yield dramatic increases in the statistical power of an RCT study when prior outcome scores are included as covariates in statistical models of program impacts. Even if individual-level data are unavailable, aggregate school-level data might be easily obtained from school accountability reports for use as a covariate in studies in which schools represent the unit of random assignment. Baseline data are also useful in confirming the equivalence of treatment and comparison samples, and in examining the potential for nonresponse bias for impact estimates.

4. Carefully consider whether and how to combine results based on distinct assessments.

To combine results based on distinct assessments across grades or states, researchers must demonstrate that several important conditions are met. First and foremost, differences in the assessments must be viewed as ignorable and reflecting expected variation in the definition of outcomes targeted by the intervention (for example, students' ability to meet their state's proficiency standards). If the goal of study is to demonstrate that an intervention has an effect on specific skills or knowledge, then combining results across grades or states might be inappropriate unless each test provides valid, reliable, and comparable data on the targeted outcomes.

If this first condition is met, combining results across states or grades may be appropriate and researchers can choose among several analytic approaches to combine results. The simplest and most powerful approach to produce combined impact estimates involves running multigrade or multistate analyses of pooled individual test scores, once these have been rescaled to z-scores

or some other common metric. This approach, however, imposes strong assumptions on the data. First, the study sample from each grade and state must represent a similar cross-section of the overall population of students targeted by the intervention. Second, the distributions of scores from each test must be identical, except for differences in the means and standard deviations of the scale scores.

As discussed in Chapter II, when these further conditions are met, results across grades or states may be combined using relatively straightforward analytic approaches. Because these determinations are subjective, however, researchers are responsible for explicitly making the case that variation in study samples does not preclude combining results.

An alternative to pooling rescaled student-level test scores—and our recommended approach to produce cross-grade or cross-state impact estimates—is to employ meta-analytic strategies. These strategies treat the estimates based on distinct assessments as estimates from separate, independent studies that jointly describe a distribution of treatment effects for the intervention of interest. While computationally more intensive, the strength of these approaches is that they can explicitly test whether the assumptions necessary to generate estimates of average treatment effects are met. Under the meta-analytic framework, the estimates for different grades or states may be combined using a weighted average or analyzed using a meta-regression model. Either approach seems appropriate for those RCTs that are conducted across states or grades and make use of state assessment data.

REFERENCES

Achieve, Inc. Three paths, one destination: Standards-based reform in Maryland, Massachusetts, and Texas. Washington, DC: Author, 2002.

American Educational Research Association. "Standards and tests: Keeping them aligned." Research Points, vol. 1, no. 1, 2003. Washington, DC: Author.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, D.C.: Author, 1999.

Allen, M.J., and W.M. Yen. Introduction to Measurement Theory. Monterey, CA: Brooks/Cole, 1979.

Allison, D.B., R.L. Allison, M.S. Faith, F. Paultre, and F.X. Pi-Sunyer. "Power and money: designing statistically powerful studies while minimizing financial costs." Psychological Methods, vol. 2, 1997, pp. 20–33.

Amrein, A.L., and D.C. Berliner. An Analysis of Some Unintended Consequences of High-Stakes Testing. Tempe, AZ: The Great Lakes Center for Education Research & Practice, Arizona State University, 2002.

Baron, R. M., & Kenny, D. A. (1986). "The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." Journal of Personality and Social Psychology, vol. 51, 1986, pp. 1173-1182.

Bloom, H., L. Richburg-Hayes, and A.R. Black. "Using covariates to improve precision for studies that randomize schools to evaluate educational interventions." Educational Evaluation and Policy Analysis, vol. 29, no. 1, 2007, pp. 30-59.

Bracey, G.W. Put to the Test: An Educator's and Consumer's Guide to Standardized Testing (second edition). Bloomington, IN: Phi Delta Kappa International, 2002.

Goertz, Margaret, Mark Duffy, and Kerstin Carlson Le Floch. "Assessment and Accountability Systems in the 50 States: 1999-2000." Philadelphia, PA: Consortium for Policy Research in Education, Research Report No. 46, March 2001.

Cronin, J. A Study of the Ongoing Alignment of the NWEA RIT Scale with Assessments from the Montana Comprehensive Assessment System (MontCAS). Lake Oswego, OR: Northwest Evaluation Association, 2005. (ERIC Document Reproduction Service No. ED491218).

Cohen, J. Statistical Power for the Behavioral Sciences (Second Edition). Hillsdale, NJ: Erlbaum, 1988.

Council of Chief State School Officers (CCSSO). "Key State Education Policies on PK-12 Education: 2004." Washington, DC: CCSSO, 2005.

Council of Chief State School Officers (CCSSO). "Statewide Student Assessment 2007-08 SY: Math, ELA, and Science." Washington, D.C.: CCSSO, 2008. Retrieved from http://www.ccsso.org/projects/Accountability_Systems/State_Profiles/ on October 24, 2008.

Cooper, H. Synthesizing Research (3rd ed.): A Guide for Literature Reviews. Applied Social Research Methods Series, Volume 2. Thousand Oaks, CA: Sage, 1998

Cooper, H., and L.V. Hedges. The Handbook of Research Synthesis. New York: Russell Sage, 1994.

Coxe, B. "FCAT Developmental Score Scale." Unpublished memorandum. August 14, 2002. Retrieved from http://info.fldoe.org/docushare/dsweb/Get/Document-473/DPSM03-015.pdf on December 31, 2008.

Cronbach, L.J., and L. Furby. "How should we measure 'change'—or should we?" *Psychological Bulletin*, vol. 74, 1970, pp. 68-80.

Darling-Hammond, L. "Testimony Before the House Education and Labor Committee on the Re-Authorization of No Child Left Behind." Washington, DC, September 10, 2007.

Dong, N., R.A. Maynard, and I. Perez-Johnson, I. "Averaging Effect Sizes Within and Across Studies of Interventions Aimed at Improving Child Outcomes." *Child Development Perspectives*, vol. 2, no. 3, 2008, pp. 187-197.

Feuer, M.J., P.W. Holland, B.F. Green, M.W. Bertenthal, and F.C. Hemphill. Uncommon Measures: Equivalence and Linkage Among Educational Tests. Washington, DC: National Academy Press, 1999.

Finn, C.E., M.J. Petrilli, and L. Julian. The State of State Standards 2006. Washington, D.C.: Thomas B. Fordham Institute, 2006.

General Accounting Office (GAO). "No Child Left Behind Act. Most Students with Disabilities Participated in Statewide Assessments, but Inclusion Options Could Be Improved." Washington, D.C.: GAO, July 2005.

Glass, G.V., B. McGraw, and M.L. Smith. Meta-Analysis in Social Research. Beverly Hills, CA: SAGE, 1981.

Glazerman, S., D.M. Levy, and D. Myers. "Nonexperimental Replications of Social Experiments: A Systematic Review." Princeton, NJ: Mathematica Policy Research, Inc., 2002.

Goldstein, H. Multilevel Statistical Models. Third edition. London: Edward Arnold, 2003.

Hambleton, R.K., and H. Swaminathan. Item Response Theory: Principles and Applications. Hingham, MA: Kluwer, 1984.

Hambleton, R.K., H. Swaminathan, and H.J. Rogers. Fundamentals of Item Response Theory. Newbury Park, CA: Sage Press, 1991.

Harcourt. Stanford Achievement Test Series, ninth edition, spring norms book. San Antonio, TX: Author, 1997.

Hedges, L.V. and I. Olkin. Statistical methods for meta-analysis. Orlando, FL: Academic Press, 1985.

Holland, P.W., and N.J. Dorans. "Linking and Equating." In Educational Measurement (fourth edition), edited by R.L. Brennan. Westport, CT: American Council on Education/Praeger, 2006.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement (pp. 3-25). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kohn, A. The Case Against Standardized Testing: Raising the Scores, Ruining the Schools. Portsmouth, NH: Heinemann, 2000.

Kolen, M.J., and R.L. Brennan. Test Equating, Scaling, and Linking: Methods and Practices (second edition). New York: Springer, 2004.

Liang, K. Y. and Zeger, S. L. "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, vol. 73, 1986, pp. 13–22.

Linn, R.L. "Linking results of distinct assessments." *Applied Measurement in Education*, vol. 6, no. 1, 1993, pp. 83-102.

Lipsey, M.W. and D.B. Wilson. Practical Meta-Analysis. Applied Social Research Methods Series, Volume 49. Thousand Oaks, CA: Sage, 2001.

Littell, R., Milliken, G., Stroup, W., Wolfinger, R., and Schabenberger, O. SAS for Mixed Models (second edition): Cary, NC: SAS Press, 2006.

Lord, F.M. and M.R. Novick. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Welsley Publishing Company, 1968.

Martineau, J. A. "Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability." *Journal of Educational and Behavioral Statistics*, vol. 31, no. 1, 2006, pp. 35-62.

May, H. "The Reality of Designing Field Experiments in Education: Using Monte Carlo Methods for Power Analysis and Design Decisions." Paper presented at the meeting of the American Education Research Association, Montreal, Canada, 2005.

May, H., and M.A. Robinson. A Randomized Evaluation of Ohio's Personalized Assessment Reporting System (PARS). Philadelphia, PA: Consortium for Policy Research in Education, 2007.

May, H., and J.A. Supovitz. "Capturing the cumulative effects of school reform: An 11-year study of the impacts of America's Choice on Student Achievement." *Educational Evaluation and Policy Analysis*, vol. 28, no. 3, 2006, pp. 231-257.

Mislevy, R.J. Linking Educational Assessments: Concepts, Issues, Methods, and Prospects. Princeton, NJ: Educational Testing Service, 1992.

Muthén, L.K., and B.O. Muthén. "How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power." *Structural Equation Modeling*, vol. 9, no. 4, 2002, pp. 599-620.

National Center for Educational Statistics (NCES). "Mapping 2005 State Proficiency Standards onto the NAEP Scales (NCES 2007-482)." Washington, DC: U.S. Department of Education, 2007.

National Center for Educational Statistics. "National Assessment of Educational Progress [online data]." Washington, DC: U.S. Department of Education, 2008. Retrieved from http://nces.ed.gov/nationsreportcard/nde/ on October 29, 2008.

National Research Council. "Common Standards for K-12 Education? Considering the Evidence: Summary of a Workshop Series." Washington, DC: National Academy of Sciences, 2008.

Neter, J., M.H. Kuter, C.J. Nachtscheim, and W. Wasserman. Applied Linear Statistical Models (fourth edition). Chicago: Irwin, 1996.

Petrilli, M. "The Proficiency Illusion." Presentation to the National Research Council Workshop on Assessing the Role of K-12 Academic Standards in States, April 2008. Retrieved from http://www7.nationalacademies.org/cfe/Petrilli%20Presentation.pdf on October 29, 2008.

Porter, A., M. Polikoff, and J. Smithson. "Is there a de facto national curriculum? Evidence from state standards." Paper prepared for the National Research Council Workshop on Assessing the Role of K-12 Academic Standards in States, January 2008. Retrieved from http://www7.nationalacademies.org/cfe/Porter_Smithson%20State%20Standards%20Paper_Tables.pdf on May 15, 2008.

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

Riddle, W. Adequate Yearly Progress (AYP): Implementation of the No Child Left Behind Act. Washington DC: Congressional Research Service, 2005.

Rogosa, D.R., and J.B. Willett. "Demonstrating the reliability of the difference score in the measurement of change." *Journal of Educational Measurement*, vol. 20, 1983, pp. 335-343.

Rothman, R., J.B. Slattery, J.L. Vranek, and L.B. Resnick. Benchmarking and Alignment of Standards and Testing. (CSE Technical Report 566.) Los Angeles: University of California-Los Angeles, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, 2002.

Sanders, W. "Comparisons Among Various Educational Assessment Value-Added Models." Paper presented at the National Conference on Value-Added, Columbus, OH, October 2006.

Schmidt, W.H., H.C. Wang, and C.C. McKnight. "Curriculum coherence: An examination of U.S. mathematics and science content standards from an international perspective." *Journal of Curriculum Studies*, vol. 37, no. 5, 2005, pp. 525-559.

Segal, C. "Motivation, Test Scores, and Economic Success." Unpublished manuscript, 2006. Retrieved from http://www.people.hbs.edu/csegal/motivation_test_scores.pdf on December 3, 2008.

Shadish, W., T. Cook, and D. Campbell. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin, 2002.

Singer, J.D., and J.B. Willett. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. New York: Oxford University Press, 2003.

Thurlow, Martha, Christopher Johnson, and Ruth Ryder. "Accountability for Performance in Assessment." Presentation at the National Accountability Conference, New Orleans, LA, October 4-5, 2004.

U.S. Department of Education. "Family Educational Rights and Privacy, Final Rule." 34 CFR Part 99. Federal *Register*, vol. 73, no. 237, December 2008, pp. 74806–74855. Retrieved from http://www.ed.gov/legislation/FedRegister/finrule/2008-4/120908a.pdf on October 6, 2009.

U.S. Department of Education. "State and Local Implementation of the No Child Left Behind Act." Washington, DC: ED and RAND, 2007. Retrieved from http://www.rand.org/pubs/reprints/2007/RAND_RP1303.pdf on October 22, 2008.

U.S. Department of Education. "Assistance to States for the Education of Children with Disabilities and Preschool Grants for Children with Disabilities; Final Rule." 34 CFR Parts 300 and 301. *Federal Register*, vol. 71, no. 156, August 2006, pp. 46540–46845. Retrieved from http://idea.ed.gov/download/finalregulations.pdf on October 6, 2009.

Webb, N.L. "Issues Related to Judging the Alignment of Curriculum Standards and Assessments." *Applied Measurement in Education*, vol. 20, no. 1, 2007, pp. 7-25.

Wildt, A.R., and O. Ahtola. Analysis of Covariance. Thousand Oaks, CA: Sage Publications, 1978.

Wise, S.L., and C.E. DeMars. "Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions." *Educational Assessment*, vol. 10, no. 1, 2005, pp. 1-17.

Zimmerman, D.W., and R.H. Williams. "Note on the reliability of experimental measures and the power of significance tests." *Psychological Bulletin*, vol. 100, no. 1, 1986, pp.123-124.

APPENDIX A

STATE TESTING PROGRAMS UNDER NCLB

This appendix aims to provide some context on state assessment programs for researchers who might be unfamiliar or have limited knowledge about the current landscape and recent evolution of such programs in the United States. Although testing programs are continuously changing as states work to meet federal requirements and improve performance, this section aims to provide a useful snapshot for researchers working to identify factors that could affect the design of their studies.

In this appendix, we describe key provisions of the No Child Left Behind Act of 2001 (NCLB) that have influenced state testing policies and the overall availability of student assessment data, trends in state testing since NCLB was introduced, and issues related to the alignment of state tests to academic content and performance standards. Our discussion of these topics is based on reviews of information on state assessment policies from the Council of Chief State School Officers (CCSSO) and key reports on student testing.

1. Key NCLB Provisions That Influence State Testing Policies

The No Child Left Behind Act of 2001 was signed into law by President George W. Bush on January 8, 2002.[31] It is the reauthorization of the Elementary and Secondary Education Act, which governs the distribution and use of Title I funds, the federal government's principal aid program for the education of disadvantaged students.

At the core of NCLB are a number of provisions requiring, as a condition for receipt of Title I funds, that states implement comprehensive student testing programs. By the 2005-2006 school year, states had to test students annually in mathematics and English Language Arts (ELA) in grades three through eight and once in high school. Starting in 2007-2008, states also had to test

[31] As of October 6, 2009, the complete text of the NCLB Act could be obtained from the U.S. Department of Education website (http://www.ed.gov/policy/elsec/leg/esea02/index.html), along with a variety of other useful summary and overview materials (see http://www.ed.gov/nclb/overview/intro/execsumm.pdf).

students in science at least once in each of the following grade periods: 3 through 5, 6 through 9, and 10 through 12. In addition, NCLB requires that assessments be aligned with states' academic content standards, which states can accomplish either by developing assessments specifically designed to reflect those standards or by modifying commercially available "off-the-shelf" tests.

Other key provisions of NCLB that influence state testing policies and the overall availability of test scores for individual students include the following:

Adequate Yearly Progress Toward Proficiency. NCLB requires that *all* students reach proficiency in the *state-defined* standards by the spring of the 2013-2014 academic year, as measured by performance on state tests. Adequate yearly progress (AYP) is the measure by which schools, districts, and states are held accountable for student progress toward this 100 percent proficiency goal. Based on 2001-2002 test data, states set their baseline proficiency rates. States were then required to specify yearly benchmarks for how students would progress to meet the goal of 100 percent proficiency by 2014. To achieve AYP, 95 percent of students in a school as a whole must meet or exceed the "annual measurable objectives" set by the state for a given academic year. Schools or districts that fail to make AYP for two consecutive years are identified as "in need of improvement."

Statewide Accountability Systems. NCLB requires states to develop a single accountability system to determine whether all students and key subgroups of students are meeting AYP. All students must be assessed using the same state assessment (with limited exceptions, described below) and AYP definitions must apply to all public schools and districts in the state, Title I and non-Title I.

Student Participation Requirements. An additional condition to achieve AYP is that at least 95 percent of the students enrolled in a school or local education agency (LEA) must take the state tests. The participation rate must also reach 95 percent for "numerically significant"

student subgroups, which include various racial/ethnic subgroups, socioeconomically disadvantaged students, English language learners (ELLs), and students with disabilities.

Testing Accommodations and Exemptions. Because the assessments play a major role in states' accountability systems, NCLB provisions allow some modifications to the typical assessment scenario to improve fairness. For example, NCLB allows ELL students to be exempted from state testing in their first year in school. ELL students must participate in the state testing program thereafter, but may take the state test in their native language. Another common accommodation involves a testing proctor reading aloud portions of the math assessment to ELL students. Similarly, special education students may receive accommodations (for example, extended time), an alternate version of the test (for example, large-print or Braille versions), or be administered an entirely different assessment (for example, a portfolio assessment) that reflects academic standards and goals that apply specifically to them (that is, those specified in an Individualized Education Program or IEP) and are different from those that apply to the general student population (U.S. Department of Education 2006).

2. Characteristics of and Recent Trends in State Testing Programs

Statewide assessment programs were already prevalent before NCLB was enacted. A 2001 study by the Consortium for Policy Research in Education (CPRE) at the University of Pennsylvania found that in the years before NCLB was enacted, 48 states already had statewide student assessment programs (Goertz et al. 2001). (The two remaining states—Iowa and Nebraska—allowed districts to choose whether and how to assess students.) The same study nevertheless found wide cross-state variation in how often students were tested (for example, how many and which grades) and in the types of tests administered to students (for example, nationally-normed versus state-developed criterion-referenced tests).

Statewide assessment programs have nevertheless become more uniform, closely reflecting NCLB requirements. Tables A.1, A.2, and A.3 (presented at the end of this appendix) reflect data from the CCSSO on the assessments used and grades tested in ELA, mathematics, and science, respectively, during two time periods—2003-2004 and 2007-2008 for ELA and mathematics, and 2004-2005 and 2007-2008 for science—for the 50 states and the District of Columbia (CCSSO 2005, 2008). As the tables show, in 2003-2004, there was still notable variation in state assessment programs along the dimensions examined. By 2007-2008, however, all states complied with NCLB's requirements to test students yearly in grades 3-8 and at least once in grades 10-12 in mathematics and ELA. Thirty-four states (67 percent) tested students solely in the grades required by NCLB.

States test high school students in one or more of grades 10 through 12. As of the 2007-2008 academic year, the majority of states tested students in grade 10 (for example, 53 percent for mathematics, see Table A.1) and a few states (for example, Iowa and South Dakota) test students in grade 12. Some states (for example, Nevada and New Hampshire) tested students in multiple grades in high school. However, states rarely tested ninth-grade students (which is not mandated by NCLB). Some states (for example, Maryland and North Carolina) administered end-of-course exams instead of testing high school students in specific grades.

Only a handful of states test very young students. Examples of states that, as of 2007-2008, tested students below grade three include California and Delaware. This pattern likely reflects both the lack of an NCLB testing mandate for these grades as well as the difficulties in assessing young children economically and reliably. As of 2007-2008, only seven and eight states tested students at least once in grades Kindergarten through two in mathematics and ELA, respectively (see Tables A.1 and A.2).

As NCLB science testing requirements have come into effect, many states have added

science to their lists of subjects tested. The exact grades tested vary across states, however. In 2007-2008, 46 states tested students in science at least once in the required grade blocks (as compared to 35 states in 2004-2005). As of 2007-2008, Maine, Maryland, and Nevada did not test students in science in grades 10 through 12, while Arkansas and the District of Columbia were still developing their science assessment programs.

Use of nationally normed assessments is now rare. Goertz et al. (2001) note that 31 states used nationally-normed tests in their state assessment programs in 1999-2000. Since passage of NCLB, most states have nevertheless opted to develop state-specific assessments to test students in all three NCLB-mandated subjects. By 2007-2008, the number of states using nationally normed assessments had decreased to seven for mathematics (Table A.1) and four for ELA (Table A.2). As of 2007-2008, only one state (Alabama) used a nationally normed assessment to test students in science (Table A.3). Notably, those states that in 2007-2008 still used nationally normed assessments in mathematics and ELA administered them *in addition* to state-specific tests in the same grades, or administered them only in high school.

However, some states contract with commercial testing companies to develop customized assessments. Although the use of "off-the-shelf" nationally normed assessments has become less common, many states have contracted with commercial test developers such as CTB/McGraw-Hill, Educational Testing Service (ETS), Pearson Assessment, and Riverside Publishing.[32] We were unable to locate information on the degree to which, in such instances, state tests draw upon or are derived from the item banks for the nationally normed assessments sold by these same

[32] For example, CTB/McGraw Hill offers "state specific" assessment products that reportedly are aligned with the content standards of 15 states, including California, New York, Florida, New Jersey, Pennsylvania, and Ohio (http://www.ctb.com/products/category_home.jsp?FOLDER%3C%3Efolder_id=2534374302134883&bmUID=1220106041853; accessed on October 6, 2009). Riverside Publishing claims to have "collaborated with over half of U.S. states to provide assessment programs designed to meet their state-specific, large scale testing needs" (http://www.riverpub.com/large-scaleprograms/; accessed on October 6, 2009).

testing vendors. However, this contracting practice suggests that the format and content of some state tests may be closely related to the format and content of some nationally normed student assessments.

State tests rely primarily on multiple-choice items to measure student performance. *Quality Counts 2008* indicates that, in 2007-2008, the assessment programs of 49 states (all except Nebraska) and the District of Columbia included multiple-choice test items (*Education Week* 2008). Most state tests rely on about 40 to 50 multiple-choice items per subject tested (Webb 2007), which translates to only one or two items per standards-based objective assessed. Thus, many academic objectives are typically left unassessed in any given year. Multiple-choice items produce very reliable test scores, but some educators and psychologists argue that they do a poor job of measuring higher-order skills (Darling-Hammond 2007; Bracey 2002; Kohn 2000). The typical design of state tests is not surprising given their intended use: determining a student's level of proficiency relative to state standards. Such use requires highly reliable scores—justifying the use of multiple-choice items—that represent a student's proficiency across the entire set of standards for a particular grade—justifying a broad sampling of items across many standards.

There are important differences in test content and performance standards across states. Studies that have examined content standards and proficiency levels across states (separately) conclude that both vary widely. For example, Porter, Polikoff, and Smithson (2008) examined the state assessment programs of 31 states for grades three through eight in mathematics, ELA, and science; they found more overlap in standards across grades in a given state than for a given grade across states. Similarly, an NCES (2007) study used traditional psychometric equating techniques to link assessments from all 50 states to the National Assessment of Educational Progress (NAEP). This study found that the NAEP test scores corresponding to states'

proficiency cutoffs for state tests in ELA and mathematics for grades four and eight ranged from a high of 12 points *above* the NAEP cut score for "proficient" performance to a low of 45 points *below* the NAEP cut score for "basic" performance. Petrilli (2008) examined proficiency cut scores in 26 states and concluded that these varied tremendously in the difficulty level represented.

Schools and LEAs also vary in the participation rates they are able to achieve. As noted, NCLB provisions set a national standard of 95 percent for the participation of students and subgroups in state assessment programs. According to a 2007 study commissioned by ED, among the 25 percent of U.S. schools that did not make AYP in 2003-2004, 6 percent failed solely because of their test participation rates (U.S. Department of Education 2007). In other words, a little more than one percent of schools in the United States did not make AYP solely because of their test participation rates.

Special education students might have lower participation rates. A 2004 study sponsored by the National Center for Education Outcomes found that the participation rates for students with an IEP could differ within states by as much as 40 to 50 percentage points (Thurlow 2004). However, only eight states had differences greater than 25 percentage points in the participation rates of IEP and non-IEP students. According to the Government Accounting Office (GAO 2005), in 2003-2004, eight states—Alabama, Arkansas, the District of Columbia, Georgia, New Mexico, New York, Pennsylvania, and Texas—had participation rates in ELA exams below 95 percent for students with disabilities, as compared to four states—Alabama, the District of Columbia, Georgia, and Texas—with participation rates below 95 percent for all students. The GAO nevertheless concluded that, for the United States as a whole, the participation rates for special education students were generally similar to those for all students.

State testing policies also influence the completeness of student test data. Under NCLB, states independently determine a testing window within which students must take or make up the state assessment. Longer testing windows allow time for more students to be tested.[33] Some states (for example, California, Colorado, and Washington) allow parents to opt out of testing for personal or religious reasons, excluding their children from having to take the state assessment.

3. The Future of State Assessment Systems

The No Child Left Behind Act of 2001 expired in 2007 and various proposals for changes to the law have been offered as part of reauthorization efforts. Changes in regulations or priorities at the Federal, state, or other levels are likely to prompt important changes in state testing policies, which in turn would prompt changes in the types of data potentially available for research purposes. The diversity and ever-changing nature of state assessment systems heightens the importance that researchers be mindful of the issues and assumptions when using state tests for education evaluations.

[33] For example, in 2008-2009, New Jersey had a four-week testing window for grades three through eight, including the designated weeks for make-up testing (http://www.state.nj.us/education/assessment/schedule.shtml). In contrast, Texas requires that students take make-up exams within five days of the original testing date (http://ritter.tea.state.tx.us/student.assessment/admin/calendar/2007_2008_revised_01_17_08.pdf).

TABLE A.1

ASSESSMENTS AND GRADES TESTED—MATHEMATICS

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008		Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
	Test	Grades	Test	Grades	Test	Grades	Test	Grades			
Alabama	Alabama High School Graduation Exam	12	Alabama High School Graduation Exam	11	Stanford Achievement Test, 10th edition	3-8	Stanford Achievement Test, 10th edition	3-8	3-8,12	3-8,11	11
			Alabama Mathematics Test	3-8							
Alaska	Alaska Benchmark Exams	3,6,8	Standards-Based Assessments	3-10			TerraNova CAT/6	5,7	3,6,8,10	3-10	4,5,7,9
	High School Graduation Qualifying Exam (HSGQE)	10	High School Graduation Qualifying Exam (HSGQE)	10							
Arizona	Arizona Instrument to Measure Standards (AIMS)	3,5,8,10	Arizona Instrument to Measure Standards (AIMS)	3-8,10					3,5,8,10	3-8,10	4,6,7
Arkansas	Benchmark Exams	3-8	Benchmark Exams	3-8						3-8,HS	
	Algebra 1 End-of-Course Exam	HS EOC	End-of-Course Exams	HS EOC					3-8,HS		
	Geometry End-of-Course Exam	HS EOC									
California	California Standard Tests (CSTs)	2-11	California Standard Tests (CSTs)	2-7, +8-11 EOC					2-11	2-11	
	Achievement Test (CAT/6)	3,7	California Achievement Test (CAT/6)	3,7							
	California High School Exit Exam (CAHSEE)	10,11	California High School Exit Exam (CAHSEE)	10							
California	California High School Proficiency Exam (CHSPE)	10									
Colorado	Colorado Student Assessment Program	5-10	Colorado Student Assessment Program	3-10					5-10	3-10	3,4
Connecticut	Connecticut Mastery Test (CMT)	4,6,8	Connecticut Mastery Test (CMT)	3-8					4,6,8,10	3-8,10	3,5,7

TABLE A.1 *(continued)*

	State Assessments				National Assessments				Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
	2003-2004		2007-2008		2003-2004		2007-2008				
	Test	Grades	Test	Grades	Test	Grades	Test	Grades			
	Academic Performance Test (CAPT)	10	Connecticut Academic Performance Test (CAPT)	10							
		Grades									
Delaware	Delaware Student Testing Program	2-10	Delaware Student Testing Program	2-10					2-10	2-10	
Connecticut											
District of Columbia			District of Columbia Comprehensive Assessment System	3-8,10	Stanford Achievement Tests, Ninth Edition (SAT-9)	1-11			1-11	3-8,10	
Florida	Florida Comprehensive Assessment Test	3-10	Florida Comprehensive Mathematics Assessment Test	3-10					3-10	3-10	
Georgia	Criterion-Referenced Competency Tests (CRCT)	4,6,8	Criterion- Referenced Competency Tests (CRCT)	1-8	Iowa Tests of Basic Skills (ITBS/A)	3,5,8			3-6, 8-12	1-8, 11, EOC	1,2,7
	Georgia High School Graduation Tests (GHSGT)	11-12	Georgia High School Graduation Tests (GHSGT)	11							
	End of Course Test (EOCT)	9-12	End of Course Test (EOCT)	EOC							
Hawaii	Hawaii Content and Performance Standards (HCPS) II (SAT-9 based)	3,5,8,10	Hawaii Content and Performance Standards (HCPS) II (SAT-9 based)	3-8,10					3,5,8,10	3-8,10	4,6,7
Idaho	Idaho State Achievement Test (ISAT)	2-10	Idaho State Achievement Test (ISAT)	2-10					2-10	2-10	
Illinois	Illinois Standards Achievement Test (ISAT)	3,5,8	Illinois Standards Achievement Test (ISAT)	3-8					3,5,8,11	3-8,11	4,6,7
	Prairie State Achievement Examination (PSAE)	11	Prairie State Achievement Examination (PSAE)	11							

TABLE A.1 *(continued)*

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008		Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
	Test		Test	Grades	Test	Grades	Test	Grades			
Indiana	Indiana Statewide Testing for Educational Progress Plus (ISTEP+)	3,6,8,10 Grades	Indiana Statewide Testing for Educational Progress Plus (ISTEP+)	3-10					3,6,8,10	3-10	4,5,7,9
			Graduation Qualifying Exam	10							
Iowa	Iowa Tests of Basic Skills (ITBS)	4,8	Iowa Tests of Basic Skills (ITBS)	K-12					4,8,11	K-12	K,1,2,3,5,6, 7,9,10,12
	Iowa Tests of Educational Development (ITED)	11									
Kansas	Kansas Computerized Assessments (KCA)	4,7,10	Kansas State Assessment	3-8,11					4,7,10	3-8,11	3,5,6,8,11
Kentucky	Kentucky Core Content Test	5,8,11	Kentucky Core Content Test	3-8,11	CTBS/5 Survey Edition	3,6,9			3,5,6,8,9,11	3-8,11	4,7
Louisiana	Louisiana Educational Assessment Program (LEAP 21)	4,8	Louisiana Educational Assessment Program	4-8					4,8,10,11	3-11	3,5,6,7,9
	Graduation Exit Examination (GEE 21)	10,11	Graduation Exit Examination (GEE)	10,11							
			Integrated Louisiana Educational Assessment Program (iLEAP)	3,5,6,7,9							
Maine	Maine Educational Assessment (MEA)	4,8,11	Maine Educational Assessment (MEA)	3-8			Scholastic Assessment Test (SAT)	HS (11?)	4,8,11	3-8,HS	3,5,6,7,9,10 ,12
Maryland	Maryland School Assessments (MSA)	3-8, EOC	Maryland School Assessments (MSA)	3-8					3-8,EOC	3-8,EOC	
	Maryland High School Assessment	EOC	Maryland High School Assessment	EOC							

TABLE A.1 *(continued)*

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008		Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
	Test	Grades	Test	Grades	Test	Grades	Test	Grades			
Massachusetts	Massachusetts Comprehensive Assessment System (MCAS)	4,6,8,10 Grades	Massachusetts Comprehensive Assessment System (MCAS)	3-8,10					4,6,8,10	3-8,10	3,5,7
Michigan	Michigan Educational Assessment Program (MEAP)	4,8,11	Michigan Educational Assessment Program (MEAP)	3-8					4,8,11	3-8,11	3,5,6,7
			Michigan Merit Examination	11							
Minnesota	Minnesota Comprehensive Assessments (MCAs)	3-8,11	Minnesota Comprehensive Assessments-Series II (MCA-II)	3-8,11					3-8,11	3-8,11	
Mississippi	Mississippi Curriculum Test	2-8	Mississippi Curriculum Test	2-8	TerraNova CTBS/5	6			2-8,HS EOC	2-8,HS EOC	
	Subject Area Testing Program (SATP)	HS EOC	Subject Area Testing Program (SATP)	HS EOC							
Missouri	Missouri Assessment Program (MAP)	4,8,10	Missouri Assessment Program (MAP)	3-8, EOC Algebra					4,8,10	3-8, EOC Algebra	3,5,6,7,EOC Algebra
Montana	Montana Comprehensive Assessment System (MontCAS)	4,8,10	Montana Comprehensive Assessment System (MontCAS)	3-8,10					4,8,10	3-8,10	3,5,6,7
Nebraska	School-Based Teacher-Led Assessment and Reporting System (STARS)	4,8,11	School-Based Teacher-Led Assessment and Reporting System (STARS)	3-8,11					4,8,11	3-8,11	3,5,6,7
Nevada	Criterion-referenced tests	3-8	Criterion-referenced tests	3-8			Iowa Tests of Basic Skills (ITBS)	4-8	3-8,10-12	3-8,10-12	
	High School Proficiency Examination	10-12	High School Proficiency Examination	10-12			Iowa Tests of Educational Development (ITED)	10			

TABLE A.1 *(continued)*

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008		Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
	Test	Grades	Test	Grades	Test	Grades	Test	Grades			
New Hampshire	New Hampshire Educational Improvement Assessment Program (NHEIAP)	3,6,10	New Hampshire Educational Improvement Assessment Program (NHEIAP)	10					3,6,10	3-8,10,11	4,5,7,11
			New England Common Assessment Program	3-8, Pilot:11							
New Jersey	New Jersey Skills & Knowledge Assessment (NJ ASK)	3,4	New Jersey Skills & Knowledge Assessment (NJ ASK)	3-7					3,4,8,11	3-8,11	5,6,7
	Grade Eight Proficiency Assessment (GEPA)	8	Grade Eight Proficiency Assessment (GEPA)	8							
	High School Proficiency Assessment (HSPA)	11	High School Proficiency Assessment (HSPA)	11							
New Mexico	New Mexico Achievement Assessment Program	3-9	New Mexico Student Assessment Program	3-9					3-10	3-10	
	New Mexico High School Competency Examination	10	New Mexico High School Competency Examination	10							
New York	Grade 4 and 8 Mathematics Assessment	4,8	Mathematics Assessment Tests	3-8					4,8,9-12	3-8,EOC	3,5,6,7
	Comprehensive Examination in Mathematics	9,10,11, 12	High School Regents Examination	EOC							
North Carolina	End-of-Grade Mathematics	3-8	End-of-Grade Mathematics	3-8					3-8,10,EOC	3-9,EOC	9
Regents	End of Course	EOC	End of Course	EOC							
	North Carolina Competency Test	10	North Carolina Competency Test	9							
			North Carolina HS Comprehensive Test	10							

TABLE A.1 *(continued)*

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008		Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
	Test		Test	Grades	Test	Grades	Test	Grades			
North Dakota	North Dakota State Assessment	4,8,12 Grades	North Dakota State Assessment	3-8,11					4,8,12	3-8,11	3,5,6,7,11
Ohio	Ohio Proficiency Test	4,6,9	Ohio Proficiency Test	9					4,6,9	3-10	3,5,7,8,10
			Ohio Achievement Test	3-8							
			Ohio Graduation Test	10							
Oklahoma	Oklahoma Core Curriculum Tests	3-8, HS	Oklahoma Core Curriculum Tests	3-8,HS EOC	Stanford 9 Achievement Test	3	Stanford 9 Achievement Test	3	3-8, HS	3-8,HS EOC	
Oregon	TESA Knowledge and Skills Tests	3-8, CIM	TESA Knowledge and Skills Tests	3-8, CIM					3-8, CIM	3-8, CIM	
Pennsylvania	Pennsylvania System of School Assessment (PSSA)	3,5,8,11	Pennsylvania System of School Assessment (PSSA)	3-8,11					3,5,8,11	3-8,11	4,6,7
Rhode Island			New England Common Assessment Program	3-8	New Standards Reference Exams	4,8,10	New Standards Reference Exams	11	4,8,10	3-8,11	3,5,6,7,11
South Carolina	Palmetto Achievement Challenge Test (PACT)	3-8	Palmetto Achievement Challenge Test (PACT)	3-8					3-8,10, EOC	3-8,10, EOC	
	Basic Skills Assessment Program	10	High School Assessment Program	10							
	End of Course Examination Program (EOCEP)	EOC	End of Course Examination Program (EOCEP)	EOC							
South Dakota	Dakota STEP	3-8, 11	Dakota STEP	3-8,11						2-12	2,9,12
	Dakota Assessment of Content Standards (DACS)	3,6,10	Dakota Assessment of Content Standards (DACS)	2-12					3-8,10,11		
			Achievement Series Assessments	2-12							

TABLE A.1 *(continued)*

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008		Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
	Test	Grades	Test	Grades	Test	Grades	Test	Grades			
Tennessee	Tennessee Comprehensive Assessment Program (TCAP)	3-8	Tennessee Comprehensive Assessment Program (TCAP) Achievement Tests	3-8					3-8,EOC	3-8,EOC	
	Gateway Tests	EOC	Tennessee Comprehensive Assessment Program (TCAP) Gateway Tests	EOC							
Texas	Texas Assessment of Knowledge and Skills (TAKS)	3-11	Texas Assessment of Knowledge and Skills (TAKS)	3-11					3-11	3-11	
Utah	Utah Core Curriculum	3-7, EOC	Utah Core Curriculum	3-11					3-7,10,EOC	3-11	8,9,11
	Utah Basic Skills Competency Tests	10	Utah Basic Skills Competency Tests	10							
Vermont			New England Common Assessment Program	3-8	New Standards Reference Examinations (NSRE)	4,8,10	New Standards Reference Examinations (NSRE)	10	4,8,10	3-8,10	3,5,6,7
Virginia	Standards of Learning (SOL) Assessments	3,5,8, EOC	Standards of Learning (SOL) Assessments	3-8,EOC					3,5,8 EOC	3-8,EOC	4,6,7
Washington	Washington Assessment of Student Learning (WASL)	4,7,10	Washington Assessment of Student Learning (WASL)	3-8,10	Iowa Tests of Basic Skills (ITBS)	3,6			3,4,6,7,9,10	3-8,10	5,8
					Iowa Tests of Educational Development (ITED)	9					
West Virginia	West Virginia Educational Standards Tests (WESTEST)	3-8,10	West Virginia Educational Standards Tests (WESTEST)	3-8,10					3-8,10	3-8,10	
Wisconsin	Wisconsin Knowledge & Concepts Examinations (WKCE)	4,8,10	Wisconsin Knowledge & Concepts Examinations (WKCE)	3-8,10					4,8,10	3-8,10	3,5,6,7

TABLE A.1 *(continued)*

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008				
	Test		Test	Grades	Test	Grades	Test	Grades	Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
Wyoming	Wyoming Comprehensive Assessment System (WyCAS)	4,8,11 Grades	Proficiency Assessments for Wyoming Schools (PAWS)	3-8,11					4,8,11	3-8,11	3,5,6,7

Source: Council of Chief State School Officers and State Departments of Education, 2008.

EOC = End of Course; CIM = Certificate of Mastery.

TABLE A.2

ASSESSMENTS AND GRADES TESTED—ENGLISH LANGUAGE ARTS

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008		Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
State	Test	Grades	Test	Grades	Test	Grades	Test	Grades			
Alabama	Alabama High School Graduation Exam	12	Alabama High School Graduation Exam	11	Stanford Achievement Test, 10th edition	3-8			K1,K2,3-8,12	3-8,11	12
			Alabama Reading Test	3-8	Dynamic Indicator of Basic Early Literary Skills	K1-K2					
Alaska	Alaska Benchmark Exams	3,6,8	Standards-Based Assessments	3-10			TerraNova CAT/6	5,7	3,6,8,10	3-10	4,5,7,9
	High School Graduation Qualifying Exam (HSGQE)	10	High School Graduation Qualifying Exam (HSGQE)	10							
Arizona	Arizona Instrument to Measure Standards (AIMS)	3,5,8,10	Arizona Instrument to Measure Standards (AIMS)	3-8,10					3,5,8,10	3-8,10	4,6,7
Arkansas	Benchmark Exams	3-8	Benchmark Exams	3-8					3-8,11	3-8,11	
	Literacy Exam	11	Literacy Exam	11							
California	California Standard Tests (CSTs)	2-11	California Standard Tests (CSTs)	2-11					2-12	2-11	
	California Achievement Test (CAT/6)	3,7	California Achievement Test (CAT/6)	3,7							
	California High School Exit Exam (CAHSEE)	10,11	California High School Exit Exam (CAHSEE)	10							
	California English Language Development Test (CELDT)	2-12									
	California High School Proficiency Exam (CHSPE)	10									
Colorado	Colorado Student Assessment Program	3-10	Colorado Student Assessment Program	3-10					3-10	3-10	
Connecticut	Connecticut Mastery Test (CMT)	4,6,8	Connecticut Mastery Test (CMT)	3-8					4,6,8,10	3-8,10	3,5,7
	Connecticut Academic Performance Test (CAPT)	10	Connecticut Academic Performance Test (CAPT)	10							
Delaware	Delaware Student Testing Program	2-10	Delaware Student Testing Program (Reading)	2-10					2-10	2-10	
			Delaware Student Testing Program (Writing)	3-10							
District of Columbia			District of Columbia Comprehensive Assessment	3-8,10	Stanford Achievement Tests, Ninth Edition (SAT-9)	1-11			1-11	3-8,10	

TABLE A.2 *(continued)*

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008		Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
State	Test	Grades	Test	Grades	Test	Grades	Test	Grades			
			System								
Florida	Florida Comprehensive Assessment Test	3-10	Florida Comprehensive Reading Assessment Test	3-10					3-10	3-10	
	FCAT Reading SSS	4,8,10									
Georgia	Criterion-Referenced Competency Tests (CRCT)	4,6,8	Criterion-Referenced Competency Tests (CRCT)	1-8	Iowa Tests of Basic Skills (ITBS/A)	3,5,8			3-6,8-12	1-8,11,EOC	1,2,7
	Georgia High School Graduation Tests (GHSGT)	11-12	Georgia High School Graduation Tests (GHSGT)	11							
	End of Course Test (EOCT)	9-12	End of Course Test (EOCT)	EOC							
	Writing Assessment (GHSWT and MGWA)	3,5,8,11									
Hawaii	Hawaii Content and Performance Standards (HCPS) II (SAT-9 based)	3,5,8,10	Hawaii Content and Performance Standards (HCPS) II (SAT-9 based)	3-8,10					3,5,8,10	3-8,10	4,6,7
Idaho	Idaho State Achievement Test (ISAT)	-10	Idaho State Achievement Test (ISAT)	2-10					2-10	2-10	
Illinois	Illinois Standards Achievement Test (ISAT)	3,5,8	Illinois Standards Achievement Test (ISAT)	3-10					3,5,8,11	3-11	4,6,7,9,10
	Prairie State Achievement Examination (PSAE)	11	Prairie State Achievement Examination (PSAE)	11							
Indiana	Indiana Statewide Testing for Educational Progress Plus (ISTEP+)	3,6,8,10	Indiana Statewide Testing for Educational Progress Plus (ISTEP+)	3-10					3,6,8,10	3-10	4,5,7,9
			Graduation Qualifying Exam	10							
Iowa	Iowa Tests of Basic Skills (ITBS)	4,8	Iowa Tests of Basic Skills (ITBS)	K-12					4,8	K-12	K-3,5-7,9-12
	Iowa Tests of Educational Development (ITED)	11									
Kansas	Kansas Computerized Assessments (KCA)	4,7,10	Kansas State Assessment	3-8,HS					4,7,10	3-8,HS	3,5,6,8,9,11,12
Kentucky	Kentucky Core Content Test	4,7,10	Kentucky Core Content Test	3-8,10	CTBS/5 Survey Edition	3,6,9			3,4,6,7,9,10,12	3-8,10	5,8
	Writing Portfolio/Writing on Demand	4,7,12									
Louisiana	Louisiana Educational Assessment Program (LEAP 21)	4,8	Louisiana Educational Assessment Program	4-8					4,8,10,11	3-11	3,5,6,7,9
	Graduation Exit Examination (GEE 21)	10,11	Graduation Exit Examination (GEE)	10,11							
			Integrated Louisiana Educational Assessment	3,5,6,7,9							

TABLE A.2 *(continued)*

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008		Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
State	Test	Grades	Test	Grades	Test	Grades	Test	Grades			
			Program (iLEAP)	3-8							
Maine	Maine Educational Assessment (MEA)	4,8,11	Maine Educational Assessment (MEA)				Scholastic Assessment Test (SAT)	HS (10)	4,8,11	3-8,HS	3,5,6,7,9,10,12
Maryland	Maryland School Assessments (MSA)	3,5,8,10	Maryland School Assessments (MSA)	3-8					3,5,8,10, EOC	3-8,EOC	4,6,7
	Maryland High School Assessment	EOC	Maryland High School Assessment	EOC							
Massachusetts	Massachusetts Comprehensive Assessment System (MCAS)	3,4,7,10	Massachusetts Comprehensive Assessment System (MCAS)	3-8,10					3,4,7,10	3-8,10	5,6,8
Michigan	Michigan Educational Assessment Program (MEAP)	4,7,11	Michigan Educational Assessment Program (MEAP)	3-8					4,7,11	3-8,11	3,5,6,8
			Michigan Merit Examincation	11							
Minnesota	Minnesota Comprehensive Assessments (MCAs)	3-8,10	Minnesota Comprehensive Assessments-Series II (MCA-II)	3-8,10					3-8,10	3-8,10	
Mississippi	Mississippi Curriculum Test	2-8	Mississippi Curriculum Test	2-8	TerraNova CTBS/5	6			2-8,HS EOC	2-8,HS EOC	
	Subject Area Testing Program (SATP)	HS EOC	Subject Area Testing Program (SATP)	HS EOC							
Missouri	Missouri Assessment Program (MAP)	3,7,11	Missouri Assessment Program (MAP)	3-8, EOC English II					3,7,11	3-8, EOC English II	4,5,6,8,EOC English II
Montana	Montana Comprehensive Assessment System (MontCAS)	4,8,10	Montana Comprehensive Assessment System (MontCAS)	3-8,10					4,8,10	3-8,10	3,5,6,7
Nebraska	School-Based Teacher-Led Assessment and Reporting System (STARS)	4,8,11	School-Based Teacher-Led Assessment and Reporting System (STARS)	3-8,11					4,8,11	3-8,11	3,5,6,7
Nevada	Criterion-referenced tests High School Proficiency Examination	3-8 10-12	Criterion-referenced tests High School Proficiency Examination	3-8 10-12					3-8,10-12	3-8,10-12	
New Hampshire	New Hampshire Educational Improvement Assessment Program (NHEIAP)	3,6,10	New Hampshire Educational Improvement Assessment Program (NHEIAP)	10					3,6,10	3-8,10,11	4,5,7,11
			New England Common Assessment Program	3-8, Pilot:11							
New Jersey	New Jersey Skills &	4									

TABLE A.2 *(continued)*

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008		Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
State	Test	Grades	Test	Grades	Test	Grades	Test	Grades			
	Knowledge Assessment (NJ ASK)		New Jersey Skills & Knowledge Assessment (NJ ASK)	3-7					4,8,11	3-8,11	3,5,6,7
	Grade Eight Proficiency Assessment (GEPA)	8	Grade Eight Proficiency Assessment (GEPA)	8							
	High School Proficiency Assessment (HSPA)	11	High School Proficiency Assessment (HSPA)	11							
New Mexico	New Mexico Achievement Assessment Program	3-9	New Mexico Student Assessment Program	3-9					3-9,10	3-9,10	
	New Mexico High School Competency Examination	10	New Mexico High School Competency Examination	10							
New York	English Language Arts Test	4,8	English Language Arts Test	3-8					4,8,10-12	3-8,EOC	3,5,6,7, EOC
	Regents Comprehensive Examination in Mathematics	10,11,12	High School Regents Examination	EOC							
North Carolina	End-of-Grade Reading Comprehension	3-8	End-of-Grade Reading Comprehension	3-8					3,8,10, EOC	3-9,EOC	9
	End of Course	EOC	End of Course	EOC							
	North Carolina Competency Test	10	North Carolina Competency Test	9							
			North Carolina HS Comprehensive Test	10							
North Dakota	North Dakota State Assessment	4,8,12	North Dakota State Assessment	3-8,11					4,8,12	3-8,11	3,5,6,7,11
				9							
Ohio	Ohio Proficiency Test	6,9	Ohio Proficiency Test						6,9	3-10	3-5,7,8,10
			Ohio Achievement Test	3-8							
			Ohio Graduation Test	10							
Oklahoma	Oklahoma Core Curriculum Tests	3-8,HS EOC	Oklahoma Core Curriculum Tests	3-8,HS EOC	Stanford Achievement Tests, Ninth Edition (SAT-9)	3			3-8,HS EOC	3-8,HS EOC	
Oregon	TESA Knowledge and Skills Tests	3-8, CIM	TESA Knowledge and Skills Tests	3-8, CIM					3-8, CIM	3-8, CIM	
Pennsylvania	Pennsylvania System of School Assessment (PSSA)	3,5,8,11	Pennsylvania System of School Assessment (PSSA)	3-8,11					3,5,8,11	3-8,11	4,6,7
Rhode Island	Rhode Island Writing Assessment	3,7,10,11	New England Common Assessment Program	3-8	New Standards Reference Exams	4,8,10	New Standards Reference Exams	11	3,4,7,8,10, 11	3-8,11	5,6
South Carolina	Palmetto Achievement Challenge Test (PACT)	3-8	Palmetto Achievement Challenge Test (PACT)	3-8					3,8,10, EOC	3-8,10,EOC	
	Basic Skills Assessment	10	High School Assessment	10							

TABLE A.2 *(continued)*

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008		Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
State	Test	Grades	Test	Grades	Test	Grades	Test	Grades			
	Program End of Course Examination Program (EOCEP)	EOC	Program End of Course Examination Program (EOCEP)	EOC							
South Dakota	Dakota STEP Dakota Assessment of Content Standards (DACS)	3-8,11 3,6,10	Dakota STEP Dakota Assessment of Content Standards (DACS)	3-8,11 2-12					3-8,10,11	2-12	2,9,12
			Achievement Series Assessments	1-12							
Tennessee	Tennessee Comprehensive Assessment Program (TCAP)	3-8	Tennessee Comprehensive Assessment Program (TCAP) Achievement Tests	3-8					3-8,EOC	3-8,EOC	
	Gateway Tests	EOC	Tennessee Comprehensive Assessment Program (TCAP) Gateway Tests	EOC							
Texas	Texas Assessment of Knowledge and Skills (TAKS)	3-11	Texas Assessment of Knowledge and Skills (TAKS)	3-11					3-11	3-11	
Utah	Utah Core Curriculum Utah Basic Skills Competency Tests	3-11 10	Criterion-Referenced Tests Utah Basic Skills Competency Tests	3-11 10					3-11	3-11	
Vermont	Vermont Developmental Reading Assessment (VT-DRA)	2	Vermont Developmental Reading Assessment (VT-DRA)	2	New Standards Reference Examinations (NSRE)	4,8,10	New Standards Reference Examinations	10	2,4,8,10	2-8,10	3,5,6,7
			New England Common Assessment Program	3-8							
Virginia	Standards of Learning (SOL) Assessments	3,5,8, EOC	Standards of Learning (SOL) Assessments	3-8,EOC					3,5,8 EOC	3-8,EOC	4,6,7
Washington	Washington Assessment of Student Learning (WASL)	4,7,10	Washington Assessment of Student Learning (WASL)	3-8,10	Iowa Tests of Basic Skills (ITBS)	3,6			3,4,6,7,9,10	3-8,10	5,8
					Iowa Tests of Educational Development (ITED)	9					
West Virginia	West Virginia Educational Standards Tests (WESTEST)	3-8, 10	West Virginia Educational Standards Tests (WESTEST)	3-8,10					3-8,10	3-8,10	
			Writing Assessment	4,7,10							5,6,7
Wisconsin	Wisconsin Knowledge & Concepts Examinations (WKCE)	4,8,10	Wisconsin Knowledge & Concepts Examinations (WKCE)	3-8,10					3,4,8,10	3-8,10	
	Wisconsin Reading	3									

TABLE A.2 *(continued)*

	State Assessments				National Assessments						
	2003-2004		2007-2008		2003-2004		2007-2008				
State	Test	Grades	Test	Grades	Test	Grades	Test	Grades	Combined Grades Tested 2003-2004	Combined Grades Tested 2007-2008	Grades Added
	Comprehension (WRCT)										
Wyoming	Wyoming Comprehensive Assessment System (WyCAS)	4,8,11	Proficiency Assessments for Wyoming Schools (PAWS)	3-8,11					4,8,11	3-8,11	3,5,6,7

Source: Council of Chief State School Officers and State Departments of Education, 2008.
EOC = End of Course; CIM = Certificate of Mastery.

TABLE A.3

ASSESSMENTS AND GRADES TESTED—SCIENCE

	State Assessments				National Assessments						
	2004-2005		2007-2008		2004-2005		2007-2008		Combined Grades Tested 2004-2005	Combined Grades Tested 2007-2008	Grades Added
State	Test	Grades	Test	Grades	Test	Grades	Test	Grades			
Alabama	Alabama High School Graduation Exam	10	Alabama High School Graduation Exam	11			Stanford Achievement Test, 10th Edition	5,7	10	5,7,11	5,7,11
Alaska			Science Assessment (pilot testing)	4,8,10						4,8,10	4,8,10
Arizona	Arizona Instrument to Measure Standards (AIMS)	4,8, and Bio	Arizona Instrument to Measure Standards (AIMS)	4,8,10					4,8, and Bio	4,8,10	10
Arkansas										-	-
California	STAR California Standard Tests (CSTs)	5, 9-11	STAR California Standard Tests (CSTs)	5,8,10 and 9-11 EOC					5, 9-11	5,8,10 and 9-11 EOC	8
Colorado	Colorado Student Assessment Program	5,8,10	Colorado Student Assessment Program	5,8,10					5,8,10	5,8,10	
Connecticut	Connecticut Academic Performance Test (CAPT)	5,10	Connecticut Academic Performance Test (CAPT) Connecticut Mastery Test (CMT)	10 5,8					5,10	5,8,10	8
Delaware	Delaware Student Testing Program	4,6,8,11	Delaware Student Testing Program	4,6,8,11					4,6,8,11	4,6,8,11	
District of Columbia										-	
Florida	Florida Comprehensive Assessment Test Science	5,8,11	Florida Comprehensive Assessment Test	5,8,11					5,8,11	5,8,11	
Georgia	Criterion-Referenced Competency Tests (CRCT) End of Course Test (EOCT)	4,6,8 9-12	Criterion-Referenced Competency Tests (CRCT) End of Course Test (EOCT) Georgia High School Graduation Tests	3-8 EOC 11	Iowa Tests of Basic Skills (ITBS/A)	3,5,8			3-6,8-12	3-8,11,EOC	7
Hawaii	Hawaii Content and Performance Standards (HCPS) II (SAT-9 based)	5,7,11	Hawaii Content and Performance Standards (HCPS) II (SAT-9 based)	5,7,11					5,7,11	5,7,11	

TABLE A.3 *(continued)*

Idaho	Science Idaho State Achievement Test (ISAT)	5,7,10	Science Idaho State Achievement Test (ISAT)	5,7,10	5,7,10	5,7,10	
Illinois	Illinois Standards Achievement Test (ISAT)	4,7	Illinois Standards Achievement Test (ISAT)	3-8	4,7,11	3-8,11	3,5,6,8
	Prairie State Achievement Examination (PSAE) Science	11	Prairie State Achievement Examination (PSAE) Science	11			
Indiana	Indiana Statewide Testing for Educational Progress Plus (ISTEP+)	5	Indiana Statewide Testing for Educational Progress Plus (ISTEP+)	5,7, Biology EOC	5	5,7, Biology EOC	7,Biology EOC
Iowa	Iowa Tests of Basic Skills (ITBS)	8,11	Iowa Tests of Basic Skills (ITBS)	5,8,11	8,11	5,8,11	5
Kansas	Kansas Computerized Assessments (KCA)	4,7,10	Kansas State Assessment	4,7,10	4,7,10	4,7,10	
Kentucky	Kentucky Core Content Test	4,7,11	Kentucky Core Content Test	4,7,11	4,7,11	4,7,11	
Louisiana	Louisiana Educational Assessment Program (LEAP 21)	4,8	Louisiana Educational Assessment Program	4,8	4,8,10,11	3-11	5,6,7,9
	Graduation Exit Examination (GEE 21)	10,11	Graduation Exit Examination (GEE)	10,11			
			Integrated Louisiana Educational Assessment Program (iLEAP)	3,5,6,7,9			
Maine	Maine Educational Assessment (MEA)	4,8,11	Maine Educational Assessment (MEA)	4,8	4,8,11	4,8	
Maryland			Maryland School Assessment (MSA)	5,8		5,8	5,8
Massachusetts	Massachusetts Comprehensive Assessment System (MCAS)	5,8,9,10	Massachusetts Comprehensive Assessment System (MCAS)	5,8,9,10	5,8,9,10	5,8,9,10	
Michigan	Michigan Educational Assessment Program (MEAP)	5,8,11	Michigan Educational Assessment Program (MEAP)	5,8	5,8,11	5,8,11	
			Michigan Merit Examination	11			
Minnesota			Minnesota Comprehensive Assessments-Series II (MCA-II)	5,8,HS		5,8,HS	5,8,HS
Mississippi	Elementary/Middle Grades Science Assessments	5,8	Elementary/Middle Grades Science Assessments	5,8	5,8	5,8,EOC	EOC

TABLE A.3 *(continued)*

			Mississippi Area Subject Testing	EOC					
Missouri	Missouri Assessment Program (MAP)	3,7,10	Missouri Assessment Program (MAP)	3,7,Biology EOC			3,7,10	3,7,Biology EOC	Biology EOC
Montana			Montana Comprehensive Assessment System (MontCAS)	4,8,11	Iowa Tests and Basic Skills	4,8	4,8,11	4,8,10,11	10
			Montana's Criterion-Referenced Test	4,8,10	Iowa Tests of Educational Development	11			
Nebraska	School-Based Teacher-Led Assessment and Reporting System (STARS)	4,8,11	School-Based Teacher-Led Assessment and Reporting System (STARS)	4 or 5,8,11			4,8,11	4 or 5,8,11	5
Nevada	Criterion-referenced tests	5,8	Criterion-referenced tests	5,8	Iowa Tests and Basic Skills	4,7	4,5,7,8,10	5,8	
					Iowa Tests of Educational Development	10			
New Hampshire	New Hampshire Educational Improvement Assessment Program (NHEIAP)	6,10	New Hampshire Educational Improvement Assessment Program (NHEIAP)	6,10			6,10	4,6,8,10,11	4,8,11
			Tri-State Assessment End of Grade	4,8,11					
			New England Common Assessment Program (Tri-State) Science (pilot)	4,8,11					
New Jersey	New Jersey Skills & Knowledge Assessment (NJ ASK)	4	New Jersey Skills & Knowledge Assessment (NJ ASK)	4			4,8,11	4,8,11	
	Grade Eight Proficiency Assessment (GEPA)	8	Grade Eight Proficiency Assessment (GEPA)	8					
	High School Proficiency Assessment (HSPA)	11	High School Proficiency Assessment (HSPA)	11					
New Mexico			New Mexico Student Assessment Program	3-9,11				3-9,11	3-9,11
New York	Regents Comprehensive Examination in Science	4,8,HS	Science Examination, Regents	4,8,HS			4,8,HS	4,8,HS	
			End of Course Test	HS					
North Carolina	End of Grade	5,8	End-of-Grade Science	3-8			5,8,HS	3-8,EOC	3,4,6,7
	End of Grade: Biology	HS	End of Course	EOC					
North Dakota			North Dakota State Assessment	4,8,11				4,8,11	4,8,11
Ohio	Ohio Proficiency Test	4,6,11,12	Ohio Proficiency Test	9			4,6,10-12	5,8-10	5,8,9
	Ohio Graduation Test	10	Ohio Graduation Test	10					
			Ohio Achievement Test	5,8					

TABLE A.3 *(continued)*

Oklahoma	Oklahoma Core Curriculum Tests	5,8,EOC	Oklahoma Core Curriculum Tests	5,8,EOC	5,8,EOC	5,8,EOC	
Oregon	TESA Science Knowledge and Skills Tests	5,8,CIM	TESA Science Knowledge and Skills Tests	5,8,CIM	5,8,CIM	5,8,CIM	
Pennsylvania	Pennsylvania System of School Assessment (PSSA)	4,8,11	Pennsylvania System of School Assessment (PSSA)	4,8,11	4,8,11	4,8,11	-
Rhode Island			Tri-State Science Assessment	4,8,11		4,8,11	4,8,11
South Carolina	Palmetto Achievement Challenge Test (PACT) End of Course Examination Program (EOCEP)	3-8 3-8	Palmetto Achievement Challenge Test (PACT) End of Course Examination Program	3-8 EOC	3-8	3-8,EOC	EOC
South Dakota	Dakota STEP Dakota Assessment of Content Standards (DACS)	3-8,11 2-8	Dakota STEP Dakota Assessment of Content Standards (DACS) Achievement Series Assessments	3-8,11 2-10 1-12	3-8,11	1-12	1,2,9,10,12
Tennessee	Tennessee Comprehensive Assessment Program (TCAP) Gateway Tests	3-8,11 EOC	Tennessee Comprehensive Assessment Program Achievement Tests Tennessee Comprehensive Assessment Program Gateway Tests	3-8 EOC	3,8,11,EOC	3-8,EOC	
Texas	Texas Assessment of Knowledge and Skills (TAKS)	5,10,11	Texas Assessment of Knowledge and Skills (TAKS)	5,8,10,11	5,10,11	5,8,10,11	8
Utah	Utah Performance Assessment System for Students (U-PASS) - includes the Utah Core CRTs	4-11	Science Core Criterion-Referenced Tests	4-8,11	4-11	4-8,11	
Vermont	Vermont—PASS	5,9,11	Vermont—PASS	5,9,11	5,9,11	5,9,11	
Virginia	Standards of Learning (SOL) Assessments	3,5,8,HS	Standards of Learning Assessment	3,5,8,HS	3,5,8,HS	3,5,8,HS	
Washington	Washington Assessment of Student Learning (WASL)	5,8,10	Washington Assessment of Student Learning (WASL)	5,8,10	5,8,10	5,8,10	

TABLE A.3 *(continued)*

West Virginia	West Virginia Educational Standards Tests (WESTEST)	3-8,10	West Virginia Educational Standards Tests (WESTEST)	3-8,10		3-8,10	3-8,10
Wisconsin	Wisconsin Knowledge & Concepts Examinations (WKCE)	4,8,10	Wisconsin Knowledge & Concepts Examinations (WKCE)	4,8,10		4,8,10	4,8,10
Wyoming	Proficiency Assessments for Wyoming Students (PAWS)	4,8,11	Proficiency Assessments for Wyoming Students (PAWS)	4,8,11		4,8,11	4,8,11

Source: Council of Chief State School Officers and State Departments of Education, 2008.

EOC = End of Course, CIM = Certificate of Mastery; CRT = criterion-referenced test.

APPENDIX B

HOW NCEE-FUNDED EVALUATIONS USE STATE TEST DATA

As noted, the appeal and ease of using state assessment data for evaluation purposes has grown in recent years. To provide a richer sense of the types of evaluations that use state assessment data and how rigorous evaluations may use such data, we gathered information about studies funded by the National Center for Education Evaluation and Regional Assistance (NCEE) that use state assessment data. We also examined the reasons why research teams viewed state assessments as an appropriate source of outcome data for their studies and the issues they encountered or anticipated in using such data.

1. Which NCEE-Funded Evaluations Use State Data?

To identify a set of rigorous studies that have used or plan to use state assessments, we reviewed study descriptions, unpublished design documents, and published reports (when available) for NCEE-funded evaluations begun or completed during the past five years. Our review included both studies sponsored by NCEE's evaluation division—34 in total[34]—as well as randomized control trials (RCTs) being conducted by the NCEE-funded Regional Educational Laboratories (RELs)—24 in total. Notably, our review did not include the investigator-initiated studies funded through Institute of Education Sciences (IES) research grants, which likely include additional examples of rigorous evaluations making use of state assessment data. Gathering information about such studies would have required obtaining unpublished information from dozens of principal investigators, so it was deemed beyond the scope of our review.

Among the 58 studies described above, we identified 21 that planned to use or have used state assessments as a source of outcome measures. These included 12 REL-initiated RCTs and 9

[34] See http://ies.ed.gov/ncee/projects/evaluation/index.asp for links to descriptions of ongoing NCEE research projects.

NCEE-sponsored evaluations. Table B.1 (provided at the end of this appendix) provides basic descriptions of these studies.

The studies identified evaluate a diverse set of interventions. These range from system-wide educational reforms—for example, Charter Schools (1)[35] and Success in Sight (8)—to broad-based or subject-specific professional development for teachers—for example, Teacher Induction (3), Pacific CHILD (5); and Early Reading PD (6); to subject-specific curricula or instructional programs—for example, Virtual Algebra (9) and AMSTI (14)—to supplementary instructional approaches—for example, After-School EAI (2), MAP Assessment (11), and CASL (13). An important common denominator across many of these evaluations is a focus on improving students' academic achievement broadly defined and/or students' general ability to meet academic standards.

Importantly, many of the aforementioned studies are ongoing and, hence, their designs and other details may change by the time the studies are completed and results are published. The main purpose of our review, however, was to help anchor the discussion presented in the main body of this report on issues related to the use of state tests to which research teams had called attention. The examples presented in this appendix are illustrative only. They do not describe the use of state tests in education experiments or the perspectives of education researchers about state test data in a representative way. Consistent with the purposes of our review, we therefore omit the identities of, and details about, individual studies from the discussion that follows.

2. What State Data Are Collected?

[35] We refer to studies by the shorthand name provided in the second column in Table B.1. The numbers provided in parentheses after each study correspond to those assigned in that same table.

Many studies use assessment data from multiple states. Nine of the identified studies were being conducted in, and therefore only collect test data from, a single state. An equal number of studies nevertheless collected or planned to collect data in multiple states, which suggests that considerations about whether it is appropriate to combine impact estimates based on distinct state assessments and how best to do this are relevant for many studies. One of the reviewed studies planned to collect test data from 16 states, and two other studies had collected or planned to collect test data from 13 states each. States in which NCEE-funded studies would be collecting test data included Alabama, Arizona, California, Colorado, Connecticut, Delaware, Florida, Georgia, Hawaii, Illinois, Indiana, Kansas, Kentucky, Louisiana, Massachusetts, Maryland, Maine, Michigan, Minnesota, Missouri, New Jersey, New Mexico, New York, North Carolina, Ohio, Oregon, Pennsylvania, Texas, Utah, Vermont, and Wisconsin.

Not surprisingly, studies examine intervention effects on grades commonly tested. Consistent with No Child Left Behind (NCLB) requirements and state assessment policies, most studies used state test data to examine intervention outcomes in those grades in which students must be tested yearly (that is, grades three to eight). Some studies included second grade students in their research samples but often excluded these students from analyses of achievement impacts, because state assessment data are unavailable. (These students and/or classrooms are nevertheless included in other study components—for example, analyses of intervention impacts on teacher practices.) Several studies examined the impacts of an intervention tested across multiple grades. Researchers must therefore also consider whether to aggregate impact estimates based on the distinct assessments administered to students in different grades, within a given state, or across multiple states.

Studies also tend to examine impacts on commonly tested subjects. Most studies planned to collect students' test scores in locally or state-administered mathematics and/or reading

assessments, reflecting the focus of the intervention being evaluated. Some studies would examine impacts on *both* reading and mathematics test scores; however, the analyses of these test scores are generally distinct. Reflecting the gradual expansion of state testing to additional subjects, one study planned to collect student scores on the state's science test, in addition to scores on the state tests in reading and mathematics.

Several studies collect state test scores in addition to administering their own assessments. This could make it possible to examine whether impact estimates based on state assessment data are consistent with results based on a common study-administered assessment.

3. How Do Studies Use State Assessment Data?

In reviewing how NCEE-funded studies have used or plan to use state assessment data, we focused on two main issues: (1) the types of scores examined; and (2) how impact estimates were computed and, if applicable, aggregated across grades and/or states.

Studies generally examine overall achievement scores in a given subject area. These can be reported in several different metrics including scale scores, normal curve equivalents, and percentile rankings. Some studies planned to estimate impacts *both* on scale scores (or other continuous measures of achievement) and on proficiency rates. One study estimated impacts on scale scores and on the percentage of students achieving above or below the districts' pre-intervention average reading test score. A few studies planned to estimate impacts on subscales of achievement.

Impacts are commonly estimated in effect size units. Because test scores across grades and/or states rarely share a common scale, most research teams planned to standardize test scores to have a common mean of zero and a standard deviation of one (that is, convert them to z-scores). Such standardization would enable researchers to describe impact estimates as effect sizes, which facilitates the comparison or aggregation of impact estimates based on distinct

assessments. The scale scores (or other continuous measures of achievement examined) for students in a given grade and state are converted to z-scores by subtracting the mean score for that grade/test and dividing this difference (or deviation from the mean) by the standard deviation of scores for that grade/test.

Studies use different standardization strategies. To convert the scale scores into z-scores, some studies used *sample-based* estimates of the means and standard deviations for students taking a given assessment in a given grade. Other studies used the *state-reported* means and standard deviations reported for the overall student population. Such decisions are important as they influence the precision of impact estimates (since sample-based parameter estimates are typically less precise because they are based on smaller sample sizes). They also influence the interpretation of impact estimates—relative to the distribution of achievement for students or schools similar to those included in the study or relative to the distribution of achievement for a broader, statewide student population.

Many studies aggregate effect size impact estimates across states and/or grades. One study conducted all treatment-control comparisons using z-scores *for students taking the same tests* (that is, within the same grade and district) to ensure that treatment status was not confounded with properties of the test(s). Impact estimates were then aggregated across districts and grades to generate overall estimates for the intervention. Two other studies planned to treat individual states as separate samples, estimate impacts within each state, and then combine the separate impact estimates across states for an estimate of the overall effect of the intervention. Notably, for studies still underway, study design and analysis documents did not always specify how effect size estimates would be combined (that is, as simple or weighted averages).

A few studies do not aggregate effect size estimates, reflecting unique design features. For instance, one study planned to conduct analyses separately for grades four and five because

participating schools were randomly assigned to test the intervention in one grade or the other. Another study examining state tests scores for students across several consecutive grades planned analysis focusing on each grade separately since students in these various grades could have different amounts of exposure to the intervention being evaluated.

Other studies examine impacts using vertically scaled assessments. The availability of vertically equated scores was expected to enable one study team to analyze together the scores of students in two consecutive grades and estimate the intervention's effects on academic performance for both grades combined. Another study planned to estimate impacts as (yearly) deviations from the average trajectory of learning for students across five grades.

4. Why Did Studies View State Tests as an Appropriate Source of Outcome Measures?

Although information was not uniformly available, study documents sometimes included statements about the reasons why research teams chose to use state tests in their studies. Research teams cited the following reasons.

Potential for Greater Policy Relevance. One study team noted that, while district-administered test scores may not cover every relevant domain of student achievement, they captured the content that schools deem most important or worthy of assessing. Documents for another study indicated that the study would be estimating impacts using state data because educational authorities care about student performance in these high-stakes tests. A third study noted that using state assessments would enable researchers to estimate the extent to which program implementation influences student achievement relative to NCLB goals. Researchers for a fourth study described the state assessment as a policy-relevant achievement measure.

Minimize Test Burden. One study team anticipated that relying on locally administered tests would help overcome likely resistance to additional student testing by participating entities, as well as reduce evaluation costs. Another study team viewed using state test data as facilitating

school recruitment, because no additional test administrations had to be required from participating schools or students.

Possible Greater Reliability of Achievement Measures. In addition to administering a brief, common assessment to all study participants, which was the main source of outcome measures, one study team also collected state test scores. This team anticipated that the state-administered tests were more likely to be "full battery" and, therefore, might measure achievement more reliably.

Alignment with the Intervention. One study team indicated that state standards mandate the teaching of concepts that represented the focus of the intervention being evaluated. The state tests, in turn, were expected to be aligned with the required curriculum content and, therefore, with the intervention being evaluated.

5. What Challenges Were Anticipated in Using State Test Data?

Study documents identified several important challenges related to using state tests. Although these fail to represent an exhaustive inventory, they provide a sense of the issues that researchers worry about when using state tests for evaluation purposes. Challenges mentioned explicitly in study documents included the following.

Estimation and Interpretation of Aggregated Impact Estimates. The fact that locally administered tests vary in their scales as well as in the subjects and content covered poses important evaluation challenges. For example, one study estimated treatment-control differences "within grade and district"—that is, test scores were standardized to describe student performance relative to other students in the same grade and district—in order to provide a common standard for treatment and control groups across all study sites. Another study team noted that using proficiency percentages and standardized scores would yield common outcomes but would not make the meaning of the measures uniform across states; hence, the pooled impact

of the intervention should be interpreted as the average impact of the intervention on the skills measured by the individual state assessments. A third study team noted that effect size impact estimates need to be interpreted in terms of the variance in scores on each state assessment.

Possible Floor Effects. Especially for interventions targeting English language learner (ELL) students, researchers sometimes expressed concern about the ability of standardized tests to capture improvements in student outcomes. One study team, for example, expressed concern that some study participants might "bottom out" on the state tests making it difficult to record meaningful learning gains. For this reason, the study planned to student performance over multiple years and according to students' starting achievement levels. Another study team expressed concerns that the state ELA test might be subject to floor effects in study locations where most students are not native speakers of English and/or classroom use of English is limited. For this reason, the study team planned to collect data on, and estimate impacts using, both state-administered ELA and English as a second language (ESL) assessments. Similar concerns applied to other studies focused on low-performing students.

Insufficient Alignment with the Intervention. One study team hypothesized that misalignment could help account for the absence of an impact on student achievement, although the study could not test this hypothesis directly since scores on an alternative, more closely aligned test were not available. Another study team ultimately determined that the state test was not as well aligned with the intervention as originally thought, because it was designed to cover the broad set of skills covered in the state's curriculum, while the intervention focuses on particular concepts. The limited coverage of these concepts in the state's standardized test would reduce the study's ability to detect impacts.

Missing Data. Another key concern in collecting and analyzing state test data is the potential for different participation rates among treatment and control group members, especially

if exposure to the intervention influences which students participate in standardized testing. Such concerns were raised by two study teams. Study documents suggested that investigators planned to look for evidence of differential participation rates as a possible source of bias in impact estimates, but did not specify how potential biases might be addressed.

TABLE B.1

NCEE-FUNDED STUDIES USING STATE ASSESSMENTS: STUDY BACKGROUND INFORMATION

Num	Study Name (*Short Form*)	Organization(s) Conducting the Study	Nature of Intervention	Study Design/Unit of Random Assignment	Target Population/Subjects	Findings Available
Studies Using Data from Multiple States						
1	Evaluation of the Impact of Charter School Strategies *(Charter Schools)*	Mathematica Policy Research, Inc. (Mathematica)	Charter schools are public schools that have been granted autonomy over their operations and freedom from state and district regulations that govern other public schools; however, they must meet accountability standards.	Individual students are randomly assigned by the charter schools through their admission lotteries. Control students are those who applied but were not admitted.	Middle School - entering 5th, 6th, or 7th grade Cohort 1 has 20 schools and Cohort 2 has 23 schools	Not yet reported
2	Impact Evaluation of Enhanced Academic Instruction (EAI) For After-School Programs *(After-School EAI)*[b]	Manpower Development Research Corporation (MDRC), Public/Private Ventures, Bloom Associates, Survey Research Management	Two models providing 45 minutes of formal academic instruction in reading (Success for All's Adventure Island) or math (Harcourt Athletics) in after-school programs.	Random assignment is at the student level. Students were randomly assigned by grade within each after-school center. Control students receive regular after-school services.	2nd to 5th grades Analysis for math program includes 1,961 students; analysis for reading program includes 1,828 students	Year 1—2008
3	An Impact Evaluation of Comprehensive Teacher Induction Programs *(Teacher Induction)*[c]	Mathematica; Center for Education Leadership, WestEd	Two high intensity induction programs were chosen—the Santa Cruz New Teacher Project by the New Teacher Center at the University of California-Santa Cruz, and the Pathwise Framework Induction Model by the Educational Testing Service in Princeton, NJ. A prominent feature of the models is the use of mentors who are trained extensively and released from teaching to devote an entire year to supporting new teachers.	Schools were randomly assigned to treatment group (received comprehensive teacher induction services) or control (took part in district's usual teacher induction program).	2nd to 6th grades Total sample is 210 treatment schools and 208 control schools	Year 1—2008

TABLE B.1 *(continued)*

Num	Study Name (*Short Form*)	Organization(s) Conducting the Study	Nature of Intervention	Study Design/Unit of Random Assignment	Target Population/Subjects	Findings Available
4	Improving Adolescent Literacy Across the Curriculum in High Schools: An Evaluation of the Strategic Instruction Model's Content Literacy Continuum *(CLC)*[d]	Midwest REL	CLC is a school-wide literacy-across-the-curriculum model that includes teaching practices to help teachers organize and present information and a literacy-focused curriculum that embeds literacy instruction in content-area instruction.	Random assignment was done at the school level, separately within each school district.	9th grade There are 33 schools from 8 districts	Not yet reported
5	Evaluation of Principles-Based Professional Development to Improve Reading Comprehension for English Language Learners *(Pacific CHILD)*[e]	Pacific REL	Pacific CHILD adapted to teachers of English language arts for 4th and 5th grade ELL. Two-year professional development intervention.	Random assignment was done at the school level, with blocking to ensure that treatment and control groups were balanced in terms of size and location.	4th and 5th grades 23 treatment schools and 23 control schools	Not yet reported
6	The Impact of Professional Development Models and Strategies on Teacher Practice and Student Achievement in Early Reading *(Early Reading PD)*[f]	American Institutes for Research (AIR), MDRC	Two professional development methods: a 5-day, content-focused summer institute with three days of follow-up through the school year; and the institute and follow-up days plus coaching by an in-school reading specialist trained in a particular coaching approach.	Random assignment was done at the school level, with equal numbers of schools assigned to a treatment A group, a treatment B group, and a control group that participated only in the district's usual PD services.	2nd grade There are 90 schools in six districts, with a total of 270 second grade teachers	Final—2008
7	Lessons in Character Study *(LIC)*[g]	West REL =	LIC is a comprehensive, schoolwide English language arts-based character education program that includes core reading and writing curricula and support materials that reinforce good character and language arts learning standards.	Random assignment is at the school level.	2nd to 5th grades 15,000 students and their teachers at 50 schools (25 of which are receiving the intervention)	Not yet reported

TABLE B.1 *(continued)*

Num	Study Name (*Short Form*)	Organization(s) Conducting the Study	Nature of Intervention	Study Design/Unit of Random Assignment	Target Population/Subjects	Findings Available
8	A Study of the Effectiveness of the Success in Sight School Improvement Intervention *(Success in Sight)*[h]	Central REL	Success in Sight is a whole-school reform model emphasizing data-based decision making, research-based strategies for school improvement, community building, and shared leadership.	Schools will be randomly assigned within districts.	3rd to 5th grades	Not yet reported
9	Eighth-Grade Access to Algebra I: A Study of Virtual Algebra *(Virtual Algebra)*[i]	Northeast REL	An online course in Algebra I taught by a certified algebra instructor.	Random assignment occurred at the school level, and schools were blocked based on state and size.	8th grade 70 schools with a total of 479 Algebra-ready students and 2,081 total 8th grade students	Not yet reported
10	Impact Evaluation of the U. S. Department of Education's Student Mentoring Program *(Mentoring)*[j]	Abt Associates; Branch Associates, Moore & Associates, Center for Resource Management (CRM)	The U.S. Department of Education's Student Mentoring Program is designed to assist local educational agencies and community-based organizations to promote mentoring programs for children with greatest need in the 4th through 8th grades.	The study uses random assignment at the student level.	4th to 8th grades	Final – 2009

TABLE B.1 *(continued)*

Studies Using Data from One State						
11	The Efficacy of the Measures of Academic Progress (MAP) and Its Associated Training on Differentiated Instruction and Student Achievement *(MAP Assessment)*[k]	Midwest REL	The MAP system includes a portfolio of assessment tests designed to be predictive of student performance on standardized tests and training in use of the tests to help teachers tailor instruction to individual student needs.	The study uses a cluster-randomized design with a delayed-treatment control group. Grade within school is the unit of assignment. Teachers in the control group will receive their schools' customary form of PD and will receive the treatment two years later.	4th and 5th grades 32 schools with a total of about 128 teachers	Not yet reported
12	The Effects of a Hybrid Secondary School Course in Algebra I on Teacher Practices, Classroom Quality and Adolescent Learning *(Hybrid Algebra)*[l]	Appalachia REL	Online Algebra 1 module customized to be combined with a traditional Algebra 1 course. Professional development and instructional supports provided by Kentucky Virtual High School for its hybrid teachers.	Random assignment is at the school level.	9th grade 41 schools	Not yet reported
13	A Study of the Effects of Using Classroom Assessment for Student Learning *(CASL)*[m]	Central REL	CASL is a content-neutral professional development program to help teachers improve their use of formative assessment to help students succeed in school.	Schools will be randomly assigned to an intervention group or a control group that will continue with regular PD.	4th and 5th grades 64 schools with three to six teachers in each school.	Not yet reported

TABLE B.1 *(continued)*

14	The Effectiveness of the Alabama Math, Science, and Technology Initiative *(AMSTI)*[n]	Southeast REL	AMSTI promotes student achievement in math and science through improving pedagogical practices of teachers, integration of technology in science and math curricula, and provision of technology to schools. Alabama Dept. of Ed. awards grants for two years of professional development, technology-oriented tools and resources, and ongoing technical support from local colleges.	Schools were paired on the basis of grade configuration, math scores, percentage of students qualifying for free or reduced-price lunch, and the percentage of minority students. A coin toss determined which school in each pair would be in the treatment group.	4th to 8th grades 40 schools (20 treatment/20 control); 174 treatment teachers, 150 control teachers	Not yet reported
15	*Understanding Science* and the Academic Literacy of English Learners *(Understanding Science)*[o]	West REL	Understanding Force and Motion, a professional development course aimed at improving teacher science content knowledge and pedagogy.	Randomization will occur at the teacher level, nested within district/region.	8th grade 120 total teachers, each teaching two physical science classes of 20-25 students, for a total sample of 4,800 students	Not yet reported
16	Quality Teaching for English Learners *(QTEL)*[p]	West REL	QTEL aims to improve the capacity of teachers to support the linguistic, conceptual, and academic development of ELLs by providing summer professional development sessions, individual coaching and classroom support, and collaborative planning sessions.	The study uses a cluster random assignment design with schools as the unit of randomization.	6th to 8th grades 52 schools (26 treatment); approximately 600 teachers and 55,000 students	Not yet reported
17	Assessment Accommodations for English Language Learners *(ELL Accommodations)*[q]	West REL	The development of a modified math test to better assess the skills of ELL students without affecting assessment for non-ELL students (linguistic modification).		7th and 8th grades	Not yet reported

TABLE B.1 *(continued)*

18	Closing the Reading Gap *(Striving Readers)*[r]	Florida State University (FSU), Corporation for the Advancement of Policy Evaluation	Four reading program interventions were selected: Corrective Reading, Failure Free Reading, Spell Read P.A.T., and Wilson Reading.	The study uses random assignment at two levels: schools were randomly assigned to one of the four interventions and within each school students were randomly assigned to the treatment or control groups.	3rd and 5th grades 779 students	Final—2007
Other Studies						
19	Impact Evaluation of a School-Based Violence Prevention Program *(Violence Prevention)*[s]	RTI International, Pacific Institute for Research (PIRE), Tanglewood Research, Inc.	A curriculum-based violence-prevention program, Responding in Peaceful and Positive Ways (RIPP), and a whole-school program, Best Behavior, are implemented together.		4th to 8th grades	Data collected for students entering the program in 2005-2006 and 2006-2007
20	The Impact of Professional Development Strategies on Teacher Practice and Student Achievement in Math *(Math PD)*[t]	AIR, MDRC	A math professional development model that focuses on student misconceptions in the areas of fractions, decimals, ratios, percentages, and proportions. The professional development consists of a three-day summer institute, five day-long seminars during the school year, and 10 days of additional coaching support		Not available	Not yet reported
21	Impact Evaluation of the Upward Bound's Increased Focus on Higher-Risk Students[u]	Abt Associates; Urban Institute; Berkeley Policy Associates				Study was cancelled in March 2008

NOTES: The sources of information reviewed on these studies include publicly available reports and summaries, as well as unpublished study design documents that were current as of November of 2008. Publicly available sources of information for each study are noted below.
NCEE = National Center for Education Evaluation and Regional Assistance.

TABLE B.1 *(continued)*

SOURCES:

a. U.S. Department of Education. "Evaluation of the Impact of Charter School Strategies." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/projects/evaluation/choice_charter.asp on October 6, 2009.

b. Black, Alison R., Fred Doolittle, Pei Zhu, and others (2008). *The Evaluation of Enhanced Academic Instruction After-School Programs: Findings After the First Year of Implementation* (NCEE 2008-4022). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

c. Glazerman, Steven, Sarah Dolfin, Martha Bleeker, and others. (2008). *Impacts of Comprehensive Teacher Induction: Results from the First Year of a Randomized Controlled Study* (NCEE 2009-4034). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Science, U.S. Department of Education.

d. U.S. Department of Education. "Improving Adolescent Literacy Across the Curriculum in High Schools (Content Literacy Continuum, CLC)." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=34 on October 6, 2009.

e. U.S. Department of Education. "Evaluation of Principles-Based Professional Development to Improve Reading Comprehension for English Language Learners." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=61 on October 6, 2009.

f. Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Kazuaki Uekawa, Audrey Falk, Howard Bloom, Fred Doolittle, Pei Zhu, and Laura Sztejnberg. *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement* (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

g. U.S. Department of Education. "Effects of the Lessons in Character English Language Arts Character Education Program on Behavior and Academic Outcomes." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/rct_91.asp on October 6, 2009.

h. U.S. Department of Education. "The Effects of Success in Sight as a School Improvement Intervention." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/rct_20.asp on October 6, 2009.

i. REL Northeast and Islands. "Eighth Grade Access to Algebra I: A Study of Virtual Algebra." Retrieved from http://www.virtualalgebrastudy.org/documents/Virtual_Algebra_Study_033108.pdf on October 6, 2009.

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

k. U.S. Department of Education. "Efficacy of Frequent Formative Assessment for Improving Instructional Practice and Student Performance, Given Variations in Training to Use Assessment Results." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=36 on October 6, 2009.

l. U.S. Department of Education. "The Effects of Hybrid Algebra I on Teaching Practices, Classroom Quality, and Adolescent Learning." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/rct_8.asp on October 6, 2009.

m. U.S. Department of Education. "The Effects of Classroom Assessment for Student Learning (CASL) on Student Achievement." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/rct_18.asp on October 6, 2009.

n. U.S. Department of Education. "The Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI)." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=69 on October 6, 2009.

TABLE B.1 *(continued)*

o. U.S. Department of Education. "Impact of the Understanding Science Professional Development Model on Science Achievement of English Language Learner Students." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/rct_87.asp on October 6, 2009.

p. U.S. Department of Education. "Quality Teaching for English Learners (QTEL)." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=88 on October 6, 2009.

q. U.S. Department of Education. "Effect of Linguistic Modification of Math Assessment Items on English Language Learner Students." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/rct_92.asp on October 6, 2009.

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

s. U.S. Department of Education. "Impact Evaluation of a School-Based Violence Prevention Program." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/projects/evaluation/other_violence.asp on October 6, 2009.

t. U.S. Department of Education. "The Impact of Professional Development Strategies on Teacher Practice and Student Achievement in Math." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/projects/evaluation/tq_mathematics.asp on October 6, 2009.

u. U.S. Department of Education. "Impact Evaluation of the Upward Bound's Increased Focus on Higher-Risk Students." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/projects/evaluation/upward.asp on October 6, 2009.