3-9-2004

Exploring Data and Methods to Assess and Understand the Performance of SSI States: Learning from the Cases of Kentucky and Maine

Jaekyung Lee
Co-Principal Investigator; University of Maine, Orono

Walter McIntire
Co-Principal Investigator; University of Maine, Orono

Theodore Coladarci
Co-Principal Investigator; University of Maine, Orono

Final Report for Period: 09/1999 - 08/2002
Submitted on: 03/09/2004
Award ID: 9970853
Principal Investigator: Lee, Jaekyung .
Organization: University of Maine
Submitted By:

Title:
Exploring Data and Methods to Assess and Understand the Performance of SSI States: Learning from the Cases of Kentucky and Maine

Project Participants

Senior Personnel

Name: Lee, Jaekyung
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Coladarci, Theodore
Worked for more than 160 Hours: Yes
Contribution to Project:
Dr. Theodore Coladarci replaced Dr. Walter McIntire as Co-PI since September 2000 because of Dr. McIntire's retirement.

Post-doc

Graduate Student

Undergraduate Student

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

We had three project consultants, Prof. Benjamin Wright from the University of Chicago, Prof. Ken Wong from the University of Chicago, and Ruey Yehle, Curriculum Coordinator from the Orono School District.

Activities and Findings

Research and Education Activities:
Our research team has collected and analyzed national, state and local student assessment data in Maine and Kentucky. We had meetings with state and local assessment specialists as well as nationally recognized consultants. We shared and interpreted our data analysis results and other relevant sources of information. We presented papers based on our research at the 2000 and 2001 AERA national conferences and at local superintendents meeting. We also educated our graduate students in educational measurement and statistics courses about our research. We are preparing manuscripts for publication in academic journals and reports.

Findings: (See PDF version submitted by PI at the end of the report)

Our study addressed two interrelated questions regarding the use of national and state assessment databases during the first year (9/99-8/00). First, do national and state assessments provide the same information on the performance of a system? Second, what are the factors that might affect the discrepancies between national and state assessment results? Kentucky and Maine were chosen for this case study in which three key aspects of educational system performance were examined: the performance level of students, the equality of student achievement and the progress of student achievement.

While there were close similarities between the four categories in the NAEP and the corresponding four categories in the state assessments, the percentage of students who perform at or above high proficiency levels in the Maine and Kentucky assessments (i.e., 'Advanced' on the MEA, 'Proficient' on the KIRIS) were not substantially different from the national assessment results (i.e., 'Proficient' on the NAEP). These similarities, relative to many other states, indicate that those two states? assessment standards are more consistent with national standards and that the MEA and KIRIS cutpoints for mathematics proficiency are as high as NAEP. However, the results were not entirely consistent across grades and years. This may be attributed to the fact that the definitions of performance standards and the methods of standard setting were different.

On the other hand, the national and state assessments were relatively consistent in their estimation of achievement gaps between students with different background characteristics (such as gender, race, parental education and academic readiness). However, there was also some tendency that the size of achievement gaps appeared smaller on states? own assessments than on the NAEP. This may be attributed primarily to the fact that the NAEP test items had more discrimination power on average than the state assessments with regard to differentiating students at different levels of achievement. At the same time, the discrepancy in the size of student achievement gaps might be also attributed to some external, state policy-related factors functioning as potential achievement equalizers.

Both states reported increased student achievement based on their statewide assessment results. Because the NAEP and state assessments employed different scales for test scores, a common metric in standard deviation units was established. The sizes of achievement gains from the states' own assessments (i.e., gain scores from 1992 through 1996) turned out to be greater than their counterparts from the NAEP. This may be attributed to the fact that the states' own assessments were high-stakes tests and thus have had greater impacts on curriculum and instruction than the national assessment. A further complicating factor is that changes in testing formats and the equating strategies employed created more tenuous linkages between the assessment results from remote years.

During the second year (9/00-08/01) of our project, we examined multimeasure and multilevel analysis methods for evaluating systemic school reform. First, we examined ways to cope with the challenges of considering measures from multiple sources of school system and combining multiple measures of student achievement data (measurement issue). Second, we examined ways to tackle the challenges of considering multiple levels of influences on student achievement and attributing achievement results to school effects (attribution issue). Finally, we discussed the utility and limitations of multi-level and multi-measure approaches to evaluation of systemic school reform.

Our results suggest that it is not necessary to weight each measure before forming an achievement composite to classify student performance. If intercorrelations vary in magnitude, however, then it may be advisable to weight each measure to reflect the measure’s association with the underlying principal component.

Our results also point to the possible hazards of classifying student achievement based on a single measure. single-measure classification tended to result in additional students identified as meeting the standard.

We have tested three different multilevel models of estimating school effects. Partially conditional model (with adjustment for student-level demographic differences) is regarded as fairer than fully unconditional model (without any adjustment) as it considers student background factors that schools cannot control. Fully conditional model may be fairer than partially conditional model as it further takes into account school-level compositional effects beyond individual student-level effects.

Our analysis of school effects also involved estimating student achievement gaps with regard to background characteristics (i.e., race and SES in our case). We found that while average achievement varies significantly among schools in both states, their racial and social gaps vary little among schools. This means that much of the observed variability in achievement gaps is sampling variance and, as a result, cannot be explained by school factors.

Training and Development:

This project has helped us deepen our understanding of large-scale student assessment data and sharpen our skills for educational measurement and evaluation.

We are developing methods to examing the consistencies and discrepancies among national, state and local student assessments.

Outreach Activities:

Journal Publications

Books or Other One-time Publications

Web/Internet Site

URL(s):

Description:

Other Specific Products

Contributions

Contributions within Discipline:

Our research findigns contribute to educational measurement/assessment and policy analysis fields by providing new information on the consistentcy of national and state assessments and updating our knowledge base on the issue of evaluating educational performance.

Contributions to Other Disciplines:

Contributions to Human Resource Development:

Contributions to Resources for Research and Education:

Our research has produced two papers presented at the American Edcuational Research Association. Those papers have been archived in Educational Resoruces Information Center (ERIC) database and become publicly accessible for research and education.

Contributions Beyond Science and Engineering:

Categories for which nothing is reported:

Organizational Partners

Activities and Findings: Any Outreach Activities

Any Journal

Any Book

Any Product

Contributions: To Any Other Disciplines

Contributions: To Any Human Resource Development

Contributions: To Any Beyond Science and Engineering

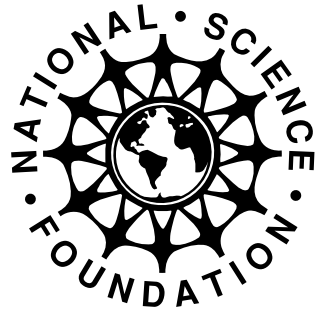
Research Report No. 1
Statewide Systemic Initiatives (SSI) Study

Using National and State Assessments to Evaluate the Performance of State Education Systems: Learning from the Cases of Kentucky and Maine

Jaekyung Lee, Ph.D.
Walter G. McIntire, Ph.D.
University of Maine

July 2002

Sponsored by the National Science Foundation



Research Report No. 1
Statewide Systemic Initiatives (SSI) Study

Using National and State Assessments to Evaluate the Performance of State Education Systems: Learning from the Cases of Kentucky and Maine

by

Jaekyung Lee, Ph.D.
Walter G. McIntire, Ph.D.
University of Maine

Prepared for
Division of Research, Evaluation and Communication
Directorate for Education and Human Resources
National Science Foundation
Arlington, VA

July 2002

NSF Grant No. REC-9970853

Prepared by the University of Maine, Orono, Maine,
for the Division of Research, Evaluation and Communication,
Directorate for Education and Human Resources,
Bernice Anderson, Program Officer

July 2002

The conduct of this study and preparation of this report was sponsored by the National Science Foundation, Directorate for Education and Human Resources, Division of Research, Evaluation and Communication, under Grant No. REC-9970853. Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily represent the views of the National Science Foundation. This report and other related publications are available on the World Wide Web: www.ume.maine.edu/naep/SSI

Table of Contents

Preface

Evaluation of systemic school reform requires a systemic approach to data collection and analysis. The National Science Foundation's Statewide Systemic Initiatives (SSI), comprehensive state policies aimed at broad student populations, consider the effects of change on the total system over a sufficient period of time, and thus are distinctive in terms of the scale and nature of programs. We need to identify and fill the gaps between currently available data and methods and desired ones in assessing and understanding the performance of SSI states. We selected two SSI states, Kentucky and Maine, to explore two research questions:

First, what information is available on the academic performance of state education systems? While there are several ways to measure academic performance, we chose to focus on student achievement in mathematics. We examined whether and how the current national and state assessments can be used, together, to inform us of statewide academic performance. We also examined national and state assessments to determine if they produce inconsistent results and to explore reasons. Second, what methodological challenges are posed by multiple measures such as national, state, and local assessments as we seek to evaluate student and school performance? We attempted to identify appropriate methods for analyzing multi-dimensional achievement data: multiple measures of achievement collected through multiple types of assessments in the multiple levels of school system at multiple time points.

Research Report No. 1 is the product of our first-year SSI research study project, "Exploring Data and Methods to Assess and Understand the Performance of SSI States." During our first project year, we have focused on the first research question and produced significant findings. This first-year study examined the consistency of the National Assessment of Educational Progress (NAEP) and state assessments as statewide educational performance measures. Two states, Kentucky and Maine, were chosen for the study, and their students' 4th and 8th grade mathematics achievement data during the 1992-96 period were examined. Similarities and discrepancies between the NAEP and state assessments were examined in terms of three major statewide performance indicators that they produce: students' performance level, achievement gap, and achievement gain.

All of the research in this report was conducted by Dr. Jaekyung Lee (PI) and Dr. Walter McIntire (Co-PI). We are very grateful to the National Science Foundation for its financial support and to the University of Maine College of Education and Human Development for its administrative support. We acknowledge that both Maine and Kentucky state education agencies provided essential help by sharing their states' student assessment data and reports. We emphasize that the views expressed herein are solely those of the authors. Our special thanks go to Dr. Bernice Anderson at the National Science Foundation, Dr. Benjamin Wright, and Dr. Kenneth Wong who provided guidance and feedback throughout our project. We also thank Yuhong Sun, Jacqueline Henderson, Mary Anne Royal, and Amy Cates at the University of Maine, who provided research assistance and/or editorial assistance.

Summary

This study examined two major questions. Do national and state assessments provide consistent information on the performance of state education systems? What accounts for discrepancies between national and state assessment results if they are found?

Data came from national and state assessments in grade 4 and grade 8 mathematics from 1992 to 1996 in Maine and Kentucky: National Assessment of Educational Progress (NAEP), Kentucky Instructional Results Information System (KIRIS), and Maine Educational Assessment (MEA). Here is a very brief summary of major research findings:

1. NAEP and state assessments reported inconsistent results on the performance level of students in Maine and Kentucky across grades and years. Both MEA and KIRIS appear to have more rigorous performance standards, which reduces the percentage of students identified as performing at Proficient/Advanced level. These discrepancies may be understood in light of the differences between the NAEP and state assessments in their definitions of performance standards and the methods of standard setting.

2. The size of achievement gaps between different groups of students appeared somewhat smaller on state assessments than on the NAEP. The discrepancies may be explained by examining the differences between NAEP and state assessments in the representation of different student groups in their testing samples, the distribution of item difficulties in their tests, and differential impact of state assessment on low-performing students/schools.

3. The sizes of achievement gains from the states' own assessments were considerably greater than that of NAEP's. At the same time, the amount of difference is not always consistent across grades. These gaps and inconsistencies might be related to differences between the national and state assessments in the stakes of testing for school systems and changes in test format that impact test equating.

The study findings raise cautions in using either national or state assessment results alone to evaluate the performance of particular state education systems. This report also provides some preliminary analyses of the sources of inconsistencies and discrepancies between national and state assessments. Although these findings may not be generalized to all states, they suggest that policymakers and educators become more aware of the unique features and limitations of current national and state assessments. While the NAEP assessment can be used to cross-check and validate the states' own assessment results, each state's unique assessment characteristics (both policy and technical aspects) need to be considered. The study gives us implications for comparing and/or combining the results from national and state assessments.

I. Research Objectives

Since 1991, the National Science Foundation (NSF) has signed cooperative agreements with 26 states to undertake ambitious and comprehensive initiatives to reform science, mathematics, and technology education. This effort to improve public education is known as the Statewide Systemic Initiatives (SSI). While one of the NSF's drivers for systemic reform required improvement in the achievement of all students, the SSI program also explicitly requested that participating states seek ways to ensure that their systemic initiatives addressed equity issues.

Given statewide systemic reform efforts for academic excellence and equity, we need to know what information is available on the performance of state education systems. While the National Assessment of Educational Progress (NAEP) and individual state student assessments have been used to inform us of state-level performance, problems exist. On one hand, states are having difficulty in realigning their student assessment systems and tracking student achievement (CPRE, 1995). Moreover, most states use their statewide assessments for several purposes, some of which are incompatible (Bond, Braskamp, & Roeber, 1996). On the other hand, the NAEP state assessments provide highly comparable information on student achievement across the states, but they are not specifically aligned with the policies and standards of any given state. Thus, we need to examine whether and how the current NAEP and states' own student assessments can be used to inform us of systemwide academic performance. We also need to examine if the national and state assessments produce consistent results on the proficiency levels of students, the achievement gaps among different groups of students and their academic progress.

Our study is based on the premise that one must use multiple measures if the measures are to be used for evaluation that will result in consequences for students and/or their school systems (see AERA, APA, & NCME, 1999). Using more than a single measure may enhance the validity and fairness of evaluation. Nevertheless, it is really challenging to compare and link the results from national and state assessments which share some common technical features as a large-scale student assessment tool, but remain different in many other ways (NRC, 1999). If we simply focus on assessment results and compare them without looking into the assessments themselves, we are likely to draw erroneous conclusions. Once we make sure that the assessments are appropriate and comparable, then we must determine how to analyze the results which might be similar in some aspects and different in the others. One may be tempted to combine the results from two assessments by simply averaging them. But this approach can yield biased evaluations without considering each assessment's unique features (i.e., goals, content, process, context, consequences) and technical qualities. We need to identify factors that produces discrepancies and make evaluation conditional upon those factors.

In light of these concerns, we conducted a systematic analysis of currently available statewide student assessment data ,using NAEP and state assessments. We addressed the consistency of these assessments for producing information on the performance of states. The objective of this study was to identify and explain the gaps between national and state assessments in light of three major educational system performance indicators: (1) students' performance level, (2) achievement gap, and (3) achievement gain. We also explored some of the factors that might explain any discrepancies in the NAEP and state assessment results.

II. Research Methods and Findings

We selected and examined two SSI states, Kentucky and Maine, which (1) put student assessment systems in place early enough to gather baseline data and monitor their progress, (2) made their assessments more in line with the goals of their education reform initiatives than other states, and (3) adopted similar performance standards to those in the NAEP. We utilized data collected from the states' student assessments, that is, Kentucky Instructional Results Information System (KIRIS) and Maine Educational Assessment (MEA) in mathematics at grade 4 and grade 8 from 1992 through 1996. We also used the NAEP state assessment data for cross-check and cross-state comparisons: the NAEP state mathematics assessments in both Maine and Kentucky were collected for 4th and 8th graders in 1992 and 1996. The NAEP state assessment was administered to a random sample of each state's fourth and eighth graders while both MEA and KIRIS were given to the virtually entire populations of Maine and Kentucky fourth and eighth graders. Our data do not include students which were exempt or absent from testing and whose test scores were not reported or missing for any reasons.

Several concerns have been raised about what data is required to adequately assess the performance of a system (Laguarda et al., 1994). Do the tests exist? If so, are they aligned with the curriculum content promoted by national and state education goals? Are the results available in a form compatible with national and state performance standards? Have the assessments been equated across the years and grade levels to track performance gains? Assessments in our study states meet the above-mentioned criteria, but it remains to be seen whether these state assessments produce the same information as the NAEP regarding the performance of the systems as a whole. We not only conducted analysis of the raw data but also reviewed information available from existing technical reports or manuals on the NAEP and state assessments. In the following sections, three major aspects of educational system performance are examined: the level of student achievement, the size of the student achievement gap, and the amount of achievement gain.

How Do Students Measure Up Against National and State Performance Standards?

Previous comparisons of national and state assessment results have shown that the percentages of students reaching the proficient level on NAEP are generally lower than on the state assessments. These results have been interpreted by educational policymakers as implying that for many states, NAEP proficiency levels are more challenging than the states' own and that state standards are still not high enough (see National Education Goals Panel, 1996). However, differences between NAEP and state assessments in the purpose of their performance standards were also noted and their comparability was questioned (Linn, 2000). The issue of comparability is much less problematic in the cases of Maine and Kentucky assessments, because they modeled their frameworks closely after NAEP and adopted very challenging performance standards.

The NAEP achievement levels, as authorized by the NAEP legislation and adopted by the National Assessment Governing Board (NAGB), are collective judgments, gathered from a broadly representative panel of teachers, education specialists, and members of the general public, about what students should know and be able to do relative to a body of content reflected in the NAEP assessment frameworks. For reporting purposes, the achievement level cut scores for each grade are placed on the traditional NAEP scale resulting in four ranges: Below Basic, Basic, Proficient, and Advanced.

Both Maine and Kentucky have achievement levels that are very similar to the NAEP levels. In Maine, proficiency levels were introduced into the MEAs in 1995, and students were identified as being in Novice, Basic, Advanced, or Distinguished levels of achievement. In Kentucky, four corresponding categories were established for the KIRIS in 1992: Novice, Apprentice, Proficient, and Distinguished. While Kentucky set its student performance goal at the Proficient level on the KIRIS as a result of statewide education reform (i.e., 100% of students proficient in 20 years), Maine did not specifically link their performance standards with the MEA proficiency levels. Despite the lack of a standards-assessment linkage, it was reasonable to say that Maine also set its performance expectation for all students to the level of being "Advanced" on the MEA. Category labels and brief generic definitions are shown in Table 1.

Table 1. Comparison of NAEP, KIRIS and MEA Definitions of Student Performance Levels

NAEP	KIRIS	MEA
Below Basic Students have little or no mastery of knowledge and skills necessary to perform work at each grade level.	Novice The student is beginning to show an understanding of new information or skills.	Novice Maine students display partial command of essential knowledge and skills.
Basic Students have partial mastery of knowledge and skills fundamental for proficient work.	Apprentice The student has gained more understanding, can do some important parts of the task.	Basic Maine students demonstrate a command of essential knowledge and skills with partial success on tasks involving higher-level concepts, including application of skills.
Proficient Students demonstrate competency over challenging subject matter and are well prepared for the next level of schooling.	Proficient The student understands the major concepts, can do almost all of the task, and can communicate concepts clearly.	Advanced Maine students successfully apply a wealth of knowledge and skills to independently develop new understanding and solutions to problems and tasks.
Advanced Student show superior performance beyond the proficient grade-level mastery.	Distinguished The student has deep understanding of the concept or process and can complete all important parts of the task. The student can communicate well, think concretely and abstractly, and analyze and interpret data.	Distinguished Maine students demonstrate in-depth understanding of information and concepts.

In order to see how students in Kentucky and Maine meet national and state performance standards, we compared NAEP and state math assessment results on student performance in 1992 and 1996 (1996 only for Maine because the MEA lacked performance standards in 1992). As shown in Table 2, the percentage of students at or above the NAEP Proficient level is smaller than at or above the MEA Advanced level. Specifically, the difference is remarkable at grade 8: 31% of Maine eighth grade students meet the NAEP's Proficient level in math as of 1996, whereas only 9% of the students meet the MEA's Advanced level. Thus, as Maine sticks more to the state's own performance goals, it ends up with a longer way to go. On the other hand, the definition of Basic performance level seems to be more convergent between the NAEP and MEA. Whether we base our judgment of Maine students' performance on the NAEP or MEA achievement levels, we come to the same conclusion that approximately one fourth of the student population in Maine does perform below the Basic level across grades and subjects examined.

Table 2. Percentages of Maine 4th and 8th Graders by Performance Level on 1996 NAEP and MEA Mathematics

NAEP		MEA	
Grade 4			
Advanced	3	Distinguished	8
Proficient	24	Advanced	15
Basic	48	Basic	55
Below Basic	25	Novice	22
Grade 8			
Advanced	6	Distinguished	1
Proficient	25	Advanced	8
Basic	46	Basic	62
Below Basic	23	Novice	29

On the other hand, comparison of NAEP and KIRIS assessment results reveal more inconsistent performance patterns. Table 3 shows the results of 1992 assessments in which the percentage of students below the NAEP Basic level is smaller than the KIRIS Novice level, whereas the percentage of students at or above the NAEP and KIRIS Proficient level is more congruent. However, the results of the 1996 assessments reversed the pattern: the percentage of students below the NAEP Basic level is greater than the KIRIS Novice level (see Table 4).

Table 3. Percentages of Kentucky 4th and 8th Graders by Performance Level on 1992 NAEP and KIRIS Mathematics

NAEP		KIRIS	
Grade 4			
Advanced	1	Distinguished	2
Proficient	12	Proficient	3
Basic	38	Apprentice	31
Below Basic	49	Novice	65
Grade 8			
Advanced	2	Distinguished	3
Proficient	12	Proficient	10
Basic	37	Apprentice	24
Below Basic	49	Novice	63

Table 4. Percentages of Kentucky 4th and 8th Graders by Performance Level on 1996 NAEP and KIRIS Mathematics

NAEP		KIRIS	
Grade 4			
Advanced	1	Distinguished	5
Proficient	15	Proficient	9
Basic	44	Apprentice	56
Below Basic	40	Novice	30
Grade 8			
Advanced	1	Distinguished	12
Proficient	15	Proficient	16
Basic	40	Apprentice	36
Below Basic	44	Novice	36

By and large, the performance standards for the KIRIS and MEA appear to have been set at comparable or even higher levels than the standards for NAEP: the percentage of students at or above the NAEP Proficient level is equal to or smaller than at or above the KIRIS Proficient level and MEA Advanced level. Nevertheless, the comparison of the NAEP, MEA and KIRIS assessment results identified inconsistent percentages of students in their corresponding performance categories. In the following sections, we explored potential factors that might explain those gaps or inconsistencies in standards-based performance results by examining how the definition of performance standards and standard-setting method differed between the national and state assessments.

Differences in the Definition of Performance Standards

As shown above, NAEP, Kentucky and Maine assessments all employed four performance standards or achievement levels. It appears that each tried to keep the standards to a reasonable number, avoiding potential problems with too few (no recognition of modest progress) or too many standards (inaccuracy of classification). Further, the KIRIS technical manual (1995) describes the difficulty that Kentucky faced in naming performance standards, particularly choosing the term "proficient" for the level of success:

> Its only drawback was that NAEP uses that term; since KIRIS will be linked to NAEP, and because NAEP's standard of "proficient" likely will be at least somewhat different from Kentucky's, there was concern about confusion between the two. However, all things considered, "Proficient" was judged to be the most appropriate term. (p. 65)

However, the real issue is operational definitions. The definition of standards affects the level of cut scores associated with the standards (Jaeger & Mills, 2001). Part of the differences between NAEP and state performance results can be explained by comparing performance level definitions by subject and grade. NAEP has both grade-specific and subject-specific definitions of performance levels, while the MEA has only subject-specific definitions and KIRIS lacks both subject-specific and grade-specific standards. Particularly the KIRIS performance standards were criticized for their vagueness (Hambleton et al., 1995). The presence or absence of clearly-stated and well-specified definitions of performance standards and achievement levels by grade and subject may help explain the differences in outcomes.

Table 5 provides definitions of MEA and NAEP math achievement levels; the 4th grade-specific definition is shown for NAEP while an across-grade definition is shown for the MEA. It is obvious that the NAEP has more clear and specific definitions with performance indicators than does the MEA. Definitions of "Basic" look very similar in that both assessments require demonstrations of student ability to solve some simple, routine problems with limited reasoning and communication. In contrast, the MEA definition of "Advanced" appears somewhat more rigorous than the NAEP definition of "Proficient": the former requires the student to solve both routine and non-routine (many) problems with effective reasoning and communication, whereas the latter requires the student to consistently solve routine problems (as distinct from complex, nonroutine problems) with successful reasoning and communication. However, both the complexity and non-routineness of any math problem is a matter of degree and subject to personal judgement. Consequently, without careful elaboration of standards by subject and grade, it is very unlikely that we will find congruence between national and state assessments in the percentages of students at the proficiency levels even with similar generic definitions and labels.

Table 5. Comparison of NAEP and MEA Definition of Mathematics Performance Levels

NAEP (Grade 4-Specific)	MEA (Grade-Free)
Below Basic	Novice. Maine students demonstrate some success with computational skills, but have great difficulty applying those skills to problem-solving situations. Mathematical reasoning and communication skills are minimal.
Basic. Fourth-grade students should show some evidence of understanding the mathematical concepts and procedures in the five NAEP content strands. Estimate and use basic facts to perform simple computations with whole numbers; show some understanding of fractions and decimals; and solve some simple real-world problems; use four-function calculators, rulers, and geometric shapes (though not always accurately). Their written responses are often minimal and presented without supporting information.	Basic. Maine students can solve routine problems, but are challenged to develop appropriate strategies for non-routine problems. Solutions sometimes lack accuracy; reasoning and communications are sometimes limited.
Proficient. Fourth-grade students should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content strands. Use whole numbers to estimate, compute, and determine whether results are reasonable; have a conceptual understanding of fractions and decimals; solve real-world problems; use four-function calculators, rulers, and geometric shapes appropriately; employ problem-solving strategies such as identifying and using appropriate information. Their written solutions are organized and presented both with supporting information and explanations of how they were achieved.	Advanced. Maine students solve routine and many non-routine problems and determine the reasonableness of the solutions using estimation, patterns and relationships, connections among mathematical concepts, and effective organization of data. These students make important connections of mathematics to real-world situations, do accurate work, and communicate mathematical strategies effectively.
Advanced. Fourth-grade students should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problems in the five NAEP content strands. Solve complex and non-routine real-world problems; display mastery in the use of four-function calculators, rulers, and geometric shapes; draw logical conclusions and justify answers and solution process; go beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely.	Distinguished. Maine students demonstrate an in-depth understanding of mathematics by applying sound reasoning to solve non-routine problems using efficient and sometimes innovative strategies. These students make connections among mathematical concepts and extend their understanding of specific problems to more global or parallel situations. They can communicate mathematically with effectiveness and sophistication

Source. Figure 3.1 in Reese et al. (1997). *NAEP 1996 Math Report Card for the Nation and the States;* Maine Department of Education (1996). *MEA Performance Level Guide: Grade 4.*

Differences in Standard-Setting (Identification of Cut Scores) Method

The NAEP math achievement levels were set following the 1990 assessment and further refined following the 1992 assessment. In developing the threshold values (cut scores) for the levels, a panel of judges rated a grade-specific item pool using the policy definitions of the NAGB. The NAEP performance standard-setting process employed a variant of Angoff method (NCES, 1997). The judges (24 at grade 4 and 22 at grade 8) rated the questions in terms of the expected probability that a student at a borderline achievement level would answer the questions correctly (for multiple-choice and short constructed-response items) or receive scores of 1, 2, 3, and 4 for the extended constructed-response items. The results from the first round of approximation were adjusted by going through subsequent rounds of review/revision processes.

The 1992 math achievement levels were evaluated by several groups including the National Academy of Education. They raised serious concerns about the reliability and validity of the current achievement levels, concluding that the Angoff judgement method was not reasonable and could yield misleading interpretations (see Shepard et al., 1993; U.S. General Accounting Office, 1993). The MEA Performance Level Guide (1994-95) from Maine Department of Education also criticizes the NAEP standard-setting process as unrealistic and unreliable. It emphasizes the need for a different approach for the MEA in that the MEA employs a totally open-response format (scored on a 0-4 scale). Thus, the MEA standard-setting process utilized a totally different method which involved judges matching actual student work to the pre-determined definitions. By matching student work to the performance level definitions, ranges of the scale where cut-points are likely to be found were identified. Once the ranges were identified, judges examined large volumes of student work within the range and the cut points were identified based on the ratings of all judges.

The Kentucky standard-setting process shares some common features with Maine. First, Kentucky's standard setting was done on open-response items only; no multiple-choice items were included in the process. Second, standard setting was done by examining actual student work rather than by investigating test items. Third, standard setting was initiated as a result of standards-based statewide education reform and designed for monitoring systemwide progress toward the goal.

Studies show that different standard setting methods yield inconsistent results (Jaeger, 1989). In our case, it is not clear how the use of different standard-setting methods affected the cut scores and resulting estimation of the percentage of students at multiple achievement levels. The lack of comparability across different standard setting methods is further complicated by the use of different performance level definitions by NAEP and state assessments. Any effort to directly compare and/or combine NAEP and state assessments' performance level results may be misleading without considering these differences and their potential influences.

How Do Student Achievement Gaps Appear on National and State Assessments?

When the performance of a school system is evaluated from an equity perspective, the size of student achievement gap becomes an important indicator of the system performance. We examined whether the sizes of achievement gaps between different groups of students are consistent between the states' own assessments and the NAEP. We selected four major student background variables (i.e., gender, race, parental education, and Title I program participation) that are available both in the national and state assessments and computed standardized gap estimates (see Table 6 and Table 7). As the student achievement gaps reported in standard deviation units incorporate differences in test score distribution as scaling

artifacts, any discrepancies between the national and state assessments in the size of achievement gaps among the same student groups requires explanation.

By and large, the standardized gap estimates in standard deviation units turned out to be smaller on the state's own assessments than on the NAEP although their discrepancies were very modest. The only exception to this pattern was a gender gap in Kentucky 8th grade math where the gap appeared larger on NAEP than on KIRIS. Regardless of the type of assessment in both states, however, it needs to be noted that the score differences between male and female students are relatively very small (hardly different from zero) in comparison with racial, social or academic gaps. In Maine, the gap between students whose parents had a high school education or more and students whose parents had less than a high school education was as large as the gap between Title I students and non-Title I students. In Kentucky, the gap between white and minority students was also as large as the gap between Title I students and non-Title I students.

Table 6. Maine 8th Grade Math Achievement Gaps on 1996 MEA and NAEP by Gender, Parental Education, and Title I Participation

	Standardized Gap		
Assessment	Gender	Parental Education	Title I
MEA	0.01	0.74*	0.80*
NAEP	0.06	0.86*	0.92*

Note: Parental education gap is between students who reported having parents with high school or more education vs. less than high school. Standardized gap is obtained by dividing the scale score gap between two concerned groups by their pooled standard deviation. Asterisk indicates that the gap is statistically significant at the .05 level.

Table 7. Kentucky 8th Grade Math Achievement Gaps on 1996 KIRIS and NAEP by Gender, Race, and Title I Participation

	Standardized Gap		
Assessment	Gender	Race	Title I
KIRIS	0.09*	0.53*	0.53*
NAEP	0.01	0.60*	0.85*

Note: Race gap is between white students and minority students. Standardized gap is obtained by dividing the scale score gap between two concerned groups by their pooled standard deviation. Asterisk indicates that the gap is statistically significant at the .05 level.

Differences in Testing Sample

Why do the gaps among different groups of students appear slightly larger on the NAEP than on the state assessments? One factor to consider is whether the NAEP testing sample is equivalent to the state assessment testing sample. Because NAEP employed a multistage stratified random sampling method, its sample was designed to properly represent major racial/ethnic and socioeconomic groups of students in each participating state (with an expectation of relatively small-size groups like Asian-Americans). In contrast, the state assessments do not involve any kind of sampling to select examinees, and their available testing samples are supposed to fully represent all student groups across the state. The exceptions include students with learning disabilities and limited English proficiency for whom the national and state assessments did not use exactly the same inclusion criteria for their testing and reporting.

To determine if the student groups compared in the previous section have equal representations in NAEP vs. state testing samples, we compared the percentage of students broken down by gender, race, parental education, and Title I. For gender and parental education, both NAEP and state assessments show exactly the same distributions. For race, there is 2% difference (11% minority for KIRIS vs. 13% minority for NAEP) in Kentucky but they are virtually identical considering the NAEP percent estimate's standard error of 1.03. For Title I participation, we found a significant difference: there is a 7% difference in Maine and 10% difference in Kentucky (see Figure 1). While Maine data shows slightly higher percentage of Title I students in the MEA than in the NAEP sample, Kentucky data shows the opposite pattern. We don't know the reason for these differences in both states but it might be due to misidentification or sampling error with regard to Title I group. Any overrepresentation or underrepresentation of Title I students in the samples who are mostly low-performing might be related to the difference between the NAEP and state assessments in their estimation of the Title I achievement gap.

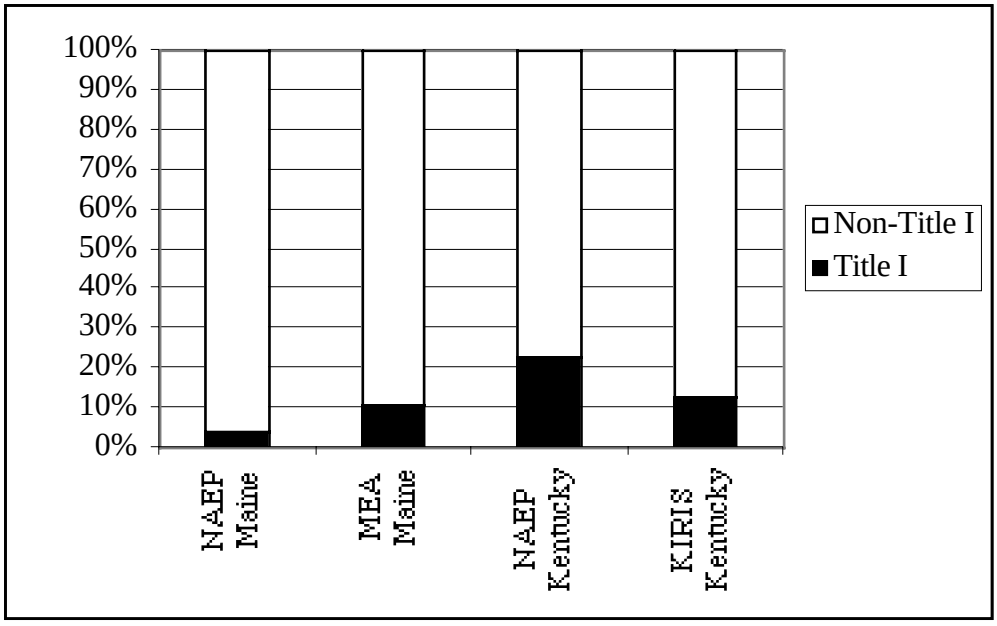


Figure 1. The Percentage of Title I vs. No-Title I Students in the 1996 NAEP, MEA, KIRIS 8th Grade Math Assessment Samples

Differences in Test Difficulty

Another potential factor that might influence the achievement gap estimates is test item difficulty. If some of the test items are more difficult for one group than for another group at the same level of proficiency, then it will affect the estimation of achievement gap. This can happen when the test items have an inherent bias or involve significant unequal opportunity to learn among different groups. Both the national and state assessments went through procedures to check against potential test bias and to conduct differential item functioning (DIF) analysis.

Assuming that all of the test items are equally difficult across different gender, race/ethnicity, and social groups , we need to consider how well those different assessments provide information on student achievement at different levels of proficiency. Although the assessments using more focused, challenging performance-type exams may provide richer information on the process of student learning (Neil et al., 1995), they may not serve all students equally well. Comparison of NAEP grade 8 mathematics test item information showed that the extended-response tasks provide much more information than both multiple-choice and short constructed-response items at the upper end of proficiency scale but less information at the lower end of the scale (see Dossey, Mullis, and Jones, 1993).

The NAEP employs more test items with a combination of multiple-choice and constructed-response items which produce wider range of item difficulties, whereas the state assessments with relatively limited number of only constructed-response items tend to have very narrow distributions of item difficulties (see Table 8 and Table 9). Lower item scores indicate greater difficulty, and both MEA and KIRIS may have been more difficult for low-performing students than the NAEP; most of the state test item scores are below .5. The MEA and KIRIS were likely to produce greater achievement gaps as they lacked test items that could measure student achievement at the lower end. Although our use of standardized gap measure takes into account potential difference in the score distributions, further investigation is needed.

Table 8. Maine Test Item (Easiness) Scores in 1996 MEA and NAEP 8th Grade Math

	Item Scores										
	.00-.10	.11-.20	.21-.30	.31-.40	.41-.50	.51-.60	.61-.70	.71-.80	.81-.90	.91-1.00	Total N
MEA	0	1	2	2	2	1	0	0	0	0	8
NAEP	0 (0)	4 (2)	20 (7)	18 (3)	14 (2)	18 (1)	17 (3)	21 (5)	25 (0)	10 (1)	147 (24)

Note. Only common items across test forms are available for the MEA. The number of entire MEA test items is 30 and all are polytomously-scored constructed-response items. Numbers in parenthesis indicate the number of polytomously-scored constructed-response items among all NAEP test items; the remainder includes multiple-choice items and dichotomously-scored constructed-response items. For dichotomoulsy-scored items (0, 1 scoring), the item score is the proportion of students who correctly answered each item. For polytomously-scored items, the item score is adjusted by dividing its mean by the maximum number of points possible.

Table 9. Kentucky Test Item (Easiness) Scores in 1996 KIRIS and NAEP 8th Grade Math

	Item Scores										
	.00-.10	.11-.20	.21-.30	.31-.40	.41-.50	.51-.60	.61-.70	.71-.80	.81-.90	.91-1.00	Total N
KIRIS	0	0	13	8	8	1	0	0	0	0	30
NAEP	1 (1)	15 (4)	21 (6)	24 (3)	17 (3)	19 (2)	17 (4)	13 (0)	15 (1)	5 (0)	147 (24)

Note. All of the above KIRIS items are polytomously-scored constructed-response items. Numbers in parenthesis indicate the number of polytomously-scored constructed-response items among all NAEP test items; the remainder includes multiple-choice items and dichotomously-scored constructed-response items. For dichotomoulsy-scored items (0, 1 scoring), the item score is the proportion of students who correctly answered each item. For polytomously-scored items, the item score is adjusted by dividing its mean by the maximum number of points possible.

The fact that state assessments in Maine and Kentucky were more challenging and difficult than their NAEP counterpart may reflect the two states' exceptionally high content and performance standards for all students. While the assessment by itself may be partly responsible for the discrepancy in the estimated size of student achievement gaps, we can think of the effect of broader assessment-driven state education policies and practices that might have functioned as achievement equalizers. Suppose that state assessment has a greater impact on lower-performing students and their schools which may pay more attention to the state test as an accountability measure and teach to the test. Student achievement scores on the state assessments may turn out to appear more equitable than on the NAEP. It remains to be investigated whether both states' assessment-driven school reform policies could have made any differential impact on schools at different performance levels and whether this could have made student achievement gaps appear smaller on the state assessments than on the NAEP.

How Much Has Student Performance Improved on National and State Assessments?

In the midst of standards-based school reform movement, every school system is expected to make continuous academic progress. The central question is whether the current NAEP and state assessments allow us to consistently keep track of system performance. To examine this issue, we first looked at changes in MEA and KIRIS student performance. Table 10 shows that the overall Maine performance trends in mathematics are highly positive across grade levels over the 1990-1997 period. Table 11 also shows that the overall Kentucky performance trends in mathematics are highly positive across grade levels over the 1992-1998 period. This successive cohort comparison method requires that the same grades of students are tested successively over time and their test scores are compared. The validity of this method for evaluating a school system's academic progress may be challenged if there are significant demographic changes in its student population over time and high level of student mobility during the school years. But we assumed that this potential problem is highly minimal at the aggregate state level.

Table 10. 1990-1997 MEA State Average Scale Score Trends in Mathematics

	1990	1991	1992	1993	1994	1995	1996	1997
Grade 4	255	265	270	270	285	285	330	320
Grade 8	300	305	305	315	325	325	350	360

Note. Scores were held constant in 1995 because of the change in test format.

Table 11. 1992-1998 KIRIS State Accountability Index Score Trends in Mathematics

	1992	1993	1994	1995	1996	1997	1998
Grade 4/5	17.8	22.3	34.2	41.8	38.9	44.8	44.4
Grade 7/8	23.8	22.8	31.4	48.9	47.3	53.8	51.4

Note. Math index is based upon the combination of on-demand and portfolio scores for 1993 and 1994 and on-demand scores only for 1995-1998.

Despite such positive performance trends based on the state assessment results, it is worthy to examine whether both Maine and Kentucky students made comparable amount of progress on the National Assessment of Educational Progress in mathematics. Earlier comparison of the KIRIS and NAEP achievement gains showed discrepancies (Hambleton et al., 1995). Using the NAEP and state assessment 4th and 8th grade math results in 1992 and 1996, we compared achievement gains from 1992 to 1996.

Tables 12 and 13 compare Maine student performance improvement levels based on the NAEP and MEA assessment results. Because NAEP and MEA scores employ different scales, a common metric in standard deviation units was established. Specifically, student standard deviations as obtained from the MEA 1996 mathematics assessment results were used to compute MEA standardized gain, while Maine's standard deviations from the 1996 NAEP state assessment results were used to compute NAEP standardized gain.

Table 12. Maine 4th Grade Math Score Gains on MEA and NAEP from 1992 to 1996

Assessment	1992	1996	Raw Gain	Standardized Gain
MEA	270	330	60*	0.39
NAEP	231	232	1	0.03

Note. Asterisk indicates that the gain is statistically significant at the .05 level.

Table 13. Maine 8th Grade Math Score Gains on MEA and NAEP from 1992 to 1996

Assessment	1992	1996	Raw Gain	Standardized Gain
MEA	305	350	45*	0.34
NAEP	279	284	5*	0.16

Note. Asterisk indicates that the gain is statistically significant at the .05 level.

Tables 14 and 15 compare Kentucky student performance improvement levels based on the NAEP and KIRIS assessment results. Because NAEP and KIRIS report gains in the percent of students meeting their own performance standards, a common metric in Cohen's h units was established. Specifically, percents of students at or above Proficient level as obtained from the KIRIS 1992 and 1996 assessment results were used to compute KIRIS standardized gain, while their counterparts from the 1992 and 1996 NAEP state assessment results were used to compute NAEP standardized gain.

Table 14. Kentucky 4th Grade Math Percent Proficient Gains on KIRIS and NAEP from 1992 to 1996

Assessment	1992	1996	Raw Gain	Standardized Gain
KIRIS	5	14	9*	0.32
NAEP	13	16	3	0.08

Note. Asterisk indicates that the gain is statistically significant at the .05 level.

Table 15. Kentucky 8th Grade Math Percent Proficient Gains on KIRIS and NAEP from 1992 to 1996

Assessment	1992	1996	Raw Gain	Standardized Gain
KIRIS	13	28	15*	0.38
NAEP	14	16	2	0.06

Note. Asterisk indicates that the gain is statistically significant at the .05 level.

As shown in Tables 12, 13, 14 and 15, we find overall statewide academic improvement in Maine and Kentucky since the early 1990s as measured by the MEA and KIRIS. However, the sizes of state math score gains tend to be somewhat greater than are observed in national assessment results (NAEP): ap-

proximately 13 times larger for grade 4 math, and twice as large for grade 8 math in the case of Maine; approximately 4 times larger for grade 4 math, and 6 times larger for grade 8 math in the case of Kentucky.

Both NAEP and state assessments face simultaneous goals of measuring trends in educational performance and providing information about student achievement on progressive curricular goals. NAEP uses several procedures to maintain the stability required for measuring trends, while still introducing innovations (Mullis et al., 1991). To keep pace with developments in assessment methodology and research about learning in each subject area, NAEP updates substantial proportions of the assessments with each successive administration. However, in some subject areas, NAEP conducts parallel assessments to provide separately for links to the past and the future. In the MEA and KIRIS, equating tests across years has been done by comparing any two adjacent years' test difficulties based on the items common to the tests both years. Nevertheless, drastic changes in the test content and format of tests raise doubts about whether their test equating is reliable and acceptable. In the following sections, we describe changes in the content and format of national and state assessments between 1992 and 1996, and explore how those changes might have affected results on test equating and performance gains.

Differences in Test Changes and Equating

Test specifications provide information on the content and format of national and state assessments. Table 16 shows the percentages of questions in 1992 and 1996 NAEP grade 4 and grade 8 math assessments. Questions could be classified under more than one content strand. It appears that changes were made in two content areas, "number sense, properties and operations" (fewer questions) and "algebra and functions" (more questions), which reportedly reflect the refinement of the NAEP math assessment to conform with recommendations from the NCTM standards (Reese at al., 1997).

Table 16. Percentage Distribution of NAEP Math Test Items by Content Strand and Grade

	Grade 4		Grade 8	
	1992	1996	1992	1996
Content Area				
Number Sense, Properties & Operation	45	40	30	25
Measurement	20	20	15	15
Geometry and Spatial Sense	15	15	20	20
Data Analysis, Statistics and Probability	10	10	15	15
Algebra & Functions	10	15	20	25
Total Percentage	100	100	100	100

Table 17. Percentage Distribution of KIRIS Math Test Items by Content Strand and Grade

	Grade 4		Grade 8	
	1992	1996	1992	1996
Content Area				
Number	13	14	20	16
Procedures	20	17	13	22
Space/Dimension	13	14	13	11
Measurement	13	14	20	16
Change	13	10	7	16
Structure	8	10	7	5
Data	20	21	20	14
Total Percentage	100	100	100	100

Source. Kentucky Department of Education (1995). *KIRIS Accountability Cycle 1 Technical Manual;* Kentucky Department of Education (1997). *KIRIS Accountability Cycle 2 Technical Manual.*

Reportedly, the curriculum and assessment frameworks for both the KIRIS and the MEA were based on those employed in creating NAEP tests. Table 17 shows the distribution of open-response KIRIS math items by year and grade across content areas. The entire KIRIS framework was consistent with the NAEP framework for mathematics. It appears that there were relatively large changes between 1992 and 1996 in KIRIS. Like NAEP, a single item in KIRIS often addresses more than one content area, which may have made the distribution of items less stable over time. The same can be said of the MEA.

While changes in test content tend to be minimal for both national and state math assessments, changes in test format and scoring standards also affect the stability of scores. The KIRIS, which started with a mix of performance exam items (i.e., writing portfolios, performance events, an on-demand essay, and open-response items) and multiple-choice items in 1992, later dropped multiple choice items. Liwewise, the MEA, which began as a combination of both multiple-choice and constructed-response questions, shifted to entirely constructed-response questions in 1995. The MEA 1994-1995 guide explains the rational for this change as follows:

> The findings of research studies are conclusive: heavy reliance on the multiple-choice format in high-stakes testing can have a negative effect on curriculum and instruction. On the other hand, the positive effect on curriculum and instruction associated with alternative modes of testing is widely recognized... MEA's use of "alternative" types of items is limited at this point to open-response items. Techniques for improving the data quality from portfolios and performance events for purpose of large-scale assessment are currently being investigated and refined. But the data quality from results of on-demand open-response testing, as used in Maine, is technically very sound. (p. 3)

Less dramatic but notable changes have been also made in the NAEP assessments. As a consequence of major revisions in the NAEP content framework in response to national standards, the 1990 NAEP assessment included a broad range of questions that required students to solve problems in both constructed-response and multiple-choice formats. For 1992, to increase NAEP's responsiveness to the then-published standards, the math assessment was nearly doubled in scope to provide greater emphasis on constructed-response questions and innovative problem-solving situations (Dossey, Mullis, and Jones, 1993). In 1996 NAEP testing, more than 50% of student assessment time was devoted to constructed-response questions.

Figure 2 illustrates these changes. While both national and state assessments shifted from multiple-choice items (MC) to more constructed-response questions (CR) including extended constructed-response questions that required students to provide an answer and a corresponding explanation, the extent of changes was greater in state assessments than NAEP between 1992 and 1996. If test score tends to drop right after introduction of a new test form (Linn et al., 1990), we might expect relatively smaller achievement gains on state assessment that changed its test format more substantially. But the pattern of actual achievement gains on NAEP, MEA, and KIRIS does not meet this expectation and asks for further examination of other factors that might have overridden the effect of test changes.

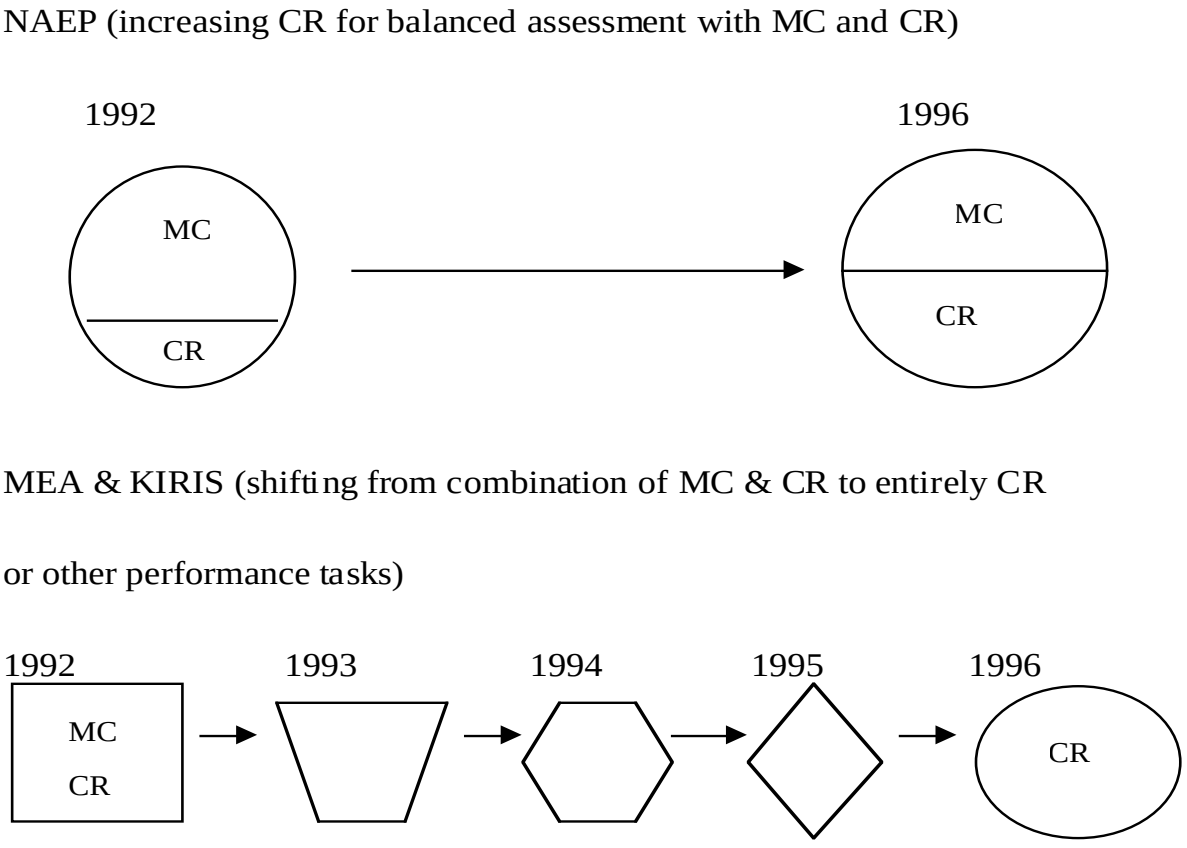


Figure 2. Changes in the Format of NAEP, MEA, and KIRIS from 1992 to 1996

Reliable estimation of achievement gains depends on robust test equating. NAEP, Kentucky and Maine assessments all used equivalent scaling and equating methods based on Item Response Theory. Nevertheless, there are differences between the NAEP and state assessments in their test equating frequency. NAEP equating was done directly between 1992 and 1996. MEA and KIRIS, which administer assessments every year, equating was done successively, that is, equating the 1993 assessment with its 1992 counterpart, the 1994 assessment with its 1993 counterpart and so on. Arrows in Figure 2 illustrate the difference in test equating process. This affects the reliability of equating: the equating of 1992 and 1996 test results is likely to be more reliable in NAEP than in the state assessments. In both the KIRIS and MEA relatively smaller percentages of items were used for equating, and this also might have increased the error of equating.

KIRIS proficiency level cut points for Accountability Cycle II (92/93 – 95/96) were linked to corresponding points for Cycle I (91/92 – 93/94). The method of linking was to determine the relationship between the original and revised 1992-93 scales using a linear transformation method (conversion of cut points based on changes in the mean and standard deviation of scale scores), and adjusting the proficiency level cutpoints accordingly. The accuracy of this adjustment also could have affected the gain in percent of students at the Advanced level from 1992 to 1996.

If equating happens regularly between successive years, the comparison of test results from remote years becomes less reliable because of the accumulation of equating errors. In other words, the link between 1992 and 1996 state assessment results should become more tenuous as a result of more drastic changes in the format of test as well as more frequent test administration and equating. To test this hypothesis, we attempted to check the stability of the linkage between the 1992 and 1996 state assessments by equating the two tests directly and comparing the results with the original gain scores that were obtained through the “chain-link” equating strategy. However, we found that there were no common items in the 1992 and 1996 MEA math assessments, which makes it impossible to equate them directly.

Differences in Test Stakes

In addition to the potential impact of changes in test format and related equating problems, one of the reasons for the greater achievement gains in Kentucky and Maine based on their state assessments might be the impact of the state assessments on school curriculum and instructional practices due to the stakes attached to the state test results. While there may be many other reasons for overstated or understated achievement gains (Wise & Hoffman, 2002), we here focus on the impact of high-stakes vs. low-stakes testing.

It is difficult to quantify how high the stakes of testing were and how much influence it might have had on actual test results. But when we simply compare the stakes of three assessments in terms of the consequences of testing for schools and school systems, it becomes obvious that the KIRIS has higher stakes than the MEA, which in turn has higher stakes than the NAEP (see Figure 3). In Kentucky, scores were used to measure school improvement and to give schools rewards or sanctions based on the adequacy of year-to-year progress. Not as high-stakes a test as the KIRIS, the MEA was designed primarily to provide information to schools to assist in making decisions about curricula and instruction. Reporting school performance to the public was also likely to produce moderate pressure on schools. This comparison of test stakes at the school or school district level, however, does not apply to the student level where neither state gave individual students substantial incentive to perform well on the state tests.

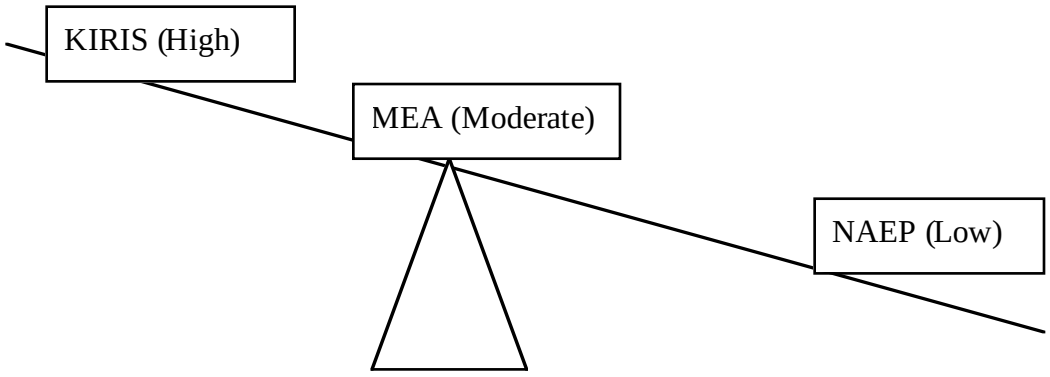


Figure 3. Contrast of NAEP, MEA, and KIRIS in the Level of Test Stakes

Given such moderate to high stakes attached to the KIRIS and the MEA for schools, it is likely that state assessment results show much greater improvement than national test results reveal. Linn (2000) explains the problem as follows:

> Divergence of trends (between a state's own assessment and NAEP) does not prove that NAEP is right and the state assessment is misleading, but it does raise important questions about the generalizability of gains reported on a state's own assessment, and hence about the validity of claims regarding student achievement. (p. 14)

The KIRIS technical manual noted that Kentucky students achieved gains on both NAEP and KIRIS but disregards the difference in the size of gains by saying that "As long as each measure provides an indication of whether changes over time are statistically significant, it is possibly to compare trends broadly. Comparing the magnitude of changes on one measure with magnitude of changes on another is more complicated, especially when multiple sets of scores are available for one or the other of the measures (such as scale score and standards-based percentage estimates) (KDE, 1997). But at the same time the manual raises the caution that some improvement in KIRIS scores is likely to occur as a result of directing school curricula toward the high-stakes test and preparing students for the test.

Our finding of the greater achievement gains in both Maine and Kentucky based on their own state assessments is consistent with the hypothesis that state assessments with serious consequences for schools would result in greater gains than NAEP without any stakes. However, our comparison of the two states in the amount of differences between NAEP and state assessment gains does not consistently support the expectation that Kentucky with relatively higher stakes would show greater differences than Maine; Maine reported greater gain than Kentucky at grade 4 while the pattern is reversed at the 8th grade level.

III. Discussion

Evaluation of systemic school reform requires us to investigate the adequacy and utility of the currently available data for assessing and understanding the performance of education systems. This study addressed two interrelated questions regarding the use of national and state assessment data. First, do national and state assessments provide the same information on the performance of states? Second, what are the factors that might explain the discrepancies between national and state assessment results? Kentucky and Maine were chosen for this study in which three key aspects of educational system performance were examined: achievement level, achievement gap, and achievement gain.

One might simply argue for using state assessment alone for evaluation of systemic school reform because it should be better able to capture the impact of state education reform policies than NAEP. It might be true that a state assessment better reflects state-specific reform goals because of stronger alignment with state curriculum standards, but it is also true that national assessment is more relevant to evaluating systemic reform that often goes beyond the boundary of a particular state given the influences of national standards and interstate benchmarking or comparisons. Table 19 provides a summary of consistent and inconsistent results in the national and state assessments as well as the factors that may account for the differences and should be considered in comparing and combining NAEP and state assessment results.

While there were seemingly close similarities between the four categories in NAEP and the corresponding four categories in state assessments, the percentage of students who perform at or above high proficiency levels in the Maine and Kentucky assessments (i.e., 'Advanced' on the MEA, 'Proficient' on the KIRIS) were not totally consistent with the national assessment results (i.e., 'Proficient' on the NAEP). Many other states also reported different results, but they tended to show the opposite patterns, i.e., greater percentage of students meeting the standard on the state's own assessment than on the NAEP. This indicates that these two states' assessment standards were uniquely higher than NAEP. However, the results were not entirely consistent across grades and years. This inconsistency might be due to differences between NAEP and state assessments in the definitions of performance standards and the methods of standards-setting. Therefore, extra caution is needed when comparing and/or combining the results on performance levels from NAEP and state assessments.

The national and state assessments were relatively consistent in their estimation of achievement gaps between students with different background characteristics. However, the size of achievement gaps were slightly smaller on the state assessments than on NAEP. Differences in the testing sample and the test itself may have influenced the results. While there was no significant difference between NAEP and state assessment data in the representation of major groups related to gender, race, and parental education, Title I students were not equally represented in the two assessments. On the other hand, NAEP had a wider range of item difficulty than the state assessments, and thus was better able to differentiate students performing at different achievement levels. These differences make it difficult to compare the size of the student achievement gaps between NAEP and state assessments. A further complicating factor is the possibility that state assessment had a greater impact on lower-performing students and their schools when they paid more attention to the state test as an accountability measure and teach to the test.

Both states reported increased student achievement based on their statewide assessment results. Because the NAEP and state assessments employed different scales for test scores, a common metric in standard

deviation units was established. The sizes of achievement gains from state assessments (i.e., gain scores from 1992 through 1996) turned out to be greater than their counterparts from NAEP. The state assessments went through more drastic changes in test format and more frequent test equating, which might have influenced the reliability of achievement gain estimates. Also, it is possible that student achievement gains were inflated by states' own assessments that were high-stakes tests and thus have had greater impacts on curriculum and instruction than NAEP.

This study explored a limited number of factors which might explain the discrepancies between national and state assessment results on school system performance. Further studies are needed to test not only the hypotheses presented in this report but also other alternative hypotheses. The findings from the two selected states may not be generalized to all states. With these caveats in mind, the study pinpoints the areas of consistency and inconsistency in the NAEP and state assessment results. It suggests that educational policymakers and practitioners become more aware of differences between current national and state assessments and potential biases and limitations in using only one of the two assessments to evaluate statewide educational system performance.

Table 19. Evaluation of the National (NAEP) and State (MEA/KIRIS) 4th and 8th Grade Math Assessment Results on Maine and Kentucky Education System Performance

	What the national and state assessments commonly say about	What the national and state assessments say differently about	What may account for the differences and should be considered for evaluation
Performance level	Majority of students found to perform below the Proficient/Advanced achievement level.	Percentage of students performing at or above Proficient level was smaller on state assessments than on NAEP. The size of this difference was also inconsistent across grade and year.	1. NAEP was more specific than KIRIS in defining its performance standards. 2. MEA standards were more rigorous than NAEP. 3. NAEP used test-centered standards-setting methods, whereas MEA and KIRIS used examinee-centered methods.
Achievement gap	The achievement gaps among different racial and socioeconomic groups of students were significant.	The achievement gaps were slightly smaller on state assessments than on NAEP. The size of this difference varied among the type of groups compared.	1. Percentage of Title I students in NAEP differed from its counterpart in MEA and KIRIS. 2. NAEP used test items with wider range of item difficulty than MEA and KIRIS.
Achievement gain	Statewide achievement gains (measured by increases in scale score or percent proficient) from 1992 to 1996 were positive.	The achievement gains were substantially smaller on state assessments than on NAEP.	1. MEA and KIRIS went through greater changes in test format and more frequent test equating than NAEP. 2. MEA and KIRIS had higher test stakes than NAEP.

References

American Educational Research Association, American Psychological Association, & National Council on Educational Measurement (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Bond, L. A., Braskamp, D., & Roeber, E. R. (1996). *The status of state student assessment programs in the United States: Annual Report.* Oakbrook, IL: NCREL.

Consortium for Policy Research in Education (1995). *CPRE policy briefs. Tracking student achievement in science and math: The promise of state assessment systems*. New Brunswick, NJ: Rutguers University.

Dossey, J. A., Mullis, I. V. S., & Jones, C. O. (1993). *Can students do mathematical problem solving?: Results from constructed response questions in NAEP's 1992 mathematics assessment.* Washington, DC: OERI, U.S. Department of Education.

Hambleton, R.K., Jaeger, R.M., Koretz, D., Linn, R., Millman, J., & Phillips, S.E. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System, 1991-1994.* A report prepared for the Office of Educational Accountability, Kentucky General Assembly.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (pp. 485-514). New York: Macmillan.

Jaeger, R. M., & Mills, C, N. (2001). An integrated judgement procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 313-338). Mahwah, NJ: Lawrence Erlbaum Associates.

Kentucky Department of Education (1995). *KIRIS Accountability Cycle 1 Technical Manual.* Kentucky: Author.

Kentucky Department of Education (1997). *KIRIS Accountability Cycle 2 Technical Manual.* Kentucky: Author.

Laguarda, K. G. et al. (1994). *Assessment programs in the statewide systemic initiatives (SSI) states: Using student achievement data to evaluate the SSI*. Washington, DC: Policy Studies Associates.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher.* 2(29), 4-16.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that "everyone is above average." *Educational Measurement: Issues and Practice*. 9(3), 5-14.

Maine Department of Education (1996). *1994-95 MEA Performance Level Guide: Grade 4.* Maine: Author.

Maine Department of Education (1996). *1994-95 MEA Performance Level Guide: Grade 8.* Maine: Author.

National Center for Education Statistics (1997). *Technical report of the NAEP 1996 state assessment program in mathematics.* Washington, DC: OERI.

National Education Goals Panel (1996). *Profile of 1994-95 state assessment systems and reported results.* Washington, DC: Author.

National Research Council (1999). *Uncommon measures,* M. Feuer, P. W. Holland, B. F. Greeen, M. W. Bertenthal, & C. Hemphill (Eds.)., Committee on Equivalency and Linkage of Educational Tests. Washington, DC: National Academy Press.

Neil, M., Bursh, P., Schaeffer, B., Thall, C., Yohe, M., & Zappardino, P. (1995). *Implementing performance assessments: A guide to classroom, school, and system reform.* Cambridge, MA: FairTest.

Reese, C. M., Miller, K. E., Mazzeo, J., & Dossey, J. A. (1997). *NAEP 1996 mathematics report card for the nation and the states*. Washington, DC: OERI.

Shepard, L. A., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement: An evaluation of the 1992 achievement levels* (A report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment). Stanford, CA: National Academy of Education.

U.S. General Accounting Office (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations* [GAO/PEMD-93-12]. Washington, DC: Author.

Wise, L. L., & Hoffman, R. G. (2002). How will assessment data be used to document the impact of educational reform? In R. W. Lissitz & W. D. Schafer (Eds.) *Assessment in education reform: Both means and ends* (pp. 146-161). Boston: Allyn and Bacon.

Exploring Data and Methods to Assess and Understand the Performance of SSI States: Learning from the Cases of Kentucky and Maine

Interim Report Submitted to the National Science Foundation

Principal Investigator:

Jaekyung Lee, Ph.D., Assistant Research Professor

College of Education and Human Development

University of Maine. (Phone: 207-581-2475)

Co-Principal Investigator:

Theodore Coladarci, Ph.D., Professor

College of Education and Human Development

University of Maine. (Phone: 207-581-2474)

September 2001

Project Summary

Evaluation of systemic school reform requires a systemic approach to data collection and analysis. The Statewide Systemic Initiatives (SSI), comprehensive state policies aimed at broad student populations, consider the effects of change on the total system over a sufficient period of time, and thus are distinctive in terms of the scale and nature of programs. We need to identify and fill the gaps between currently available data and methods and desired ones in assessing and understanding the performance of SSI states. We select two SSI states, Kentucky and Maine to explore this research question: What methodological challenges are posed by such multi-level, multi-dimensional time-series data as we seek to understand factors affecting system performance?

During our second project year (September 00 – August 01), we examined multimeasure and multilevel analysis methods for evaluating systemic school reform. First, we examined ways to cope with the challenges of considering measures from multiple sources of school system and combining multiple measures of student achievement data (measurement issue). Second, we examined ways to tackle the challenges of considering multiple levels of influences on student achievement and attributing achievement results to school effects (attribution issue). Finally, we discussed the utility and limitations of multi-level and multi-measure approaches to evaluation of systemic school reform.

During the next year, we will further examine the stability of school-level annual achievement gains from year to year and explore new methods to evaluate schools' academic progress over time. We will distribute our final report and other products widely to the public as well as to educational research and policy communities.

1. Research Objectives

During the last decade, many states have initiated systemic school reform. Systemic school reform is aimed at improving academic excellence for all students at all levels of the school system simultaneously (Smith & O'Day, 1991). Evaluation of systemic school reform calls for coordinated collection of information on student achievement at the different levels of school system (Roeber, 1995). At the same time, accountability piece of systemic school reform requires value-added school performance indicators. These policy imperatives lead us to investigate the adequacy and utility of methods for assessing and understanding the performance of a school system involved in systemic school reform.

In light of these concerns, we conduct a systematic analysis of student assessment data from Maine and Kentucky—the National Assessment of Educational Progress (NAEP) and state and local assessments—to address the issues of measurement and attribution involved in evaluating systemic school reform. This paper consists of three major sections. First, we examine ways to cope with the challenges of considering measures from multiple sources of school system and combining multiple measures of student achievement data (measurement issue). For this analysis, we use state and local assessment data collected in Maine. Second, we examine ways to tackle the challenges of considering multiple levels of influences on student achievement and attributing achievement results to school effects (attribution issue). For this analysis, we use NAEP data collected in Maine and Kentucky. Third, we discuss the utility and limitations of multi-level and multi-measure approaches to evaluation of systemic school reform.

2. Combining Multiple Measures of Achievement

A number of state and federal agencies now recommend or require multiple measures to assess student achievement (Ardovino, Hollingsworth, & Ybarra, 2000). However, no criteria about reliability, validity, and weighting in using multiple measures have been set by states like California (Jang, 1998). Currently available measures of student achievement are often inadequate for evaluation of systemic school reform, particularly when they rely on norm-referenced standardized tests and use percentile ranks as grade level standards. While local assessments are a potentially valuable source of additional measures, there is often insufficient consistency of the measures across sites. Despite these problems and challenges, districts have devised their own ways to combine multiple measures of achievement, which produces a great deal of variation from district to district (see Jang, 1998; Kalls, 1998; Law, 1998; Novak, Winters, & Flores, 2000).

In the present climate of standards-based education, school leaders in Maine also are being asked to think about assessment in new ways. Student achievement of the state standards, the *Learning Results*, must be measured by a combination of state and local assessments. Based on these assessments, local educators soon will be expected to "certify" a student's attainment of the *Learning Results* in order for the student to receive a high school diploma.

How should we approach the challenge of combining multiple measures of achievement for arriving at a single judgment of, say, "proficiency," or "meeting the

standard"? Specifically, what is an efficient and defensible method for combining multiple measures of achievement? This is the general question that we address in this section.

Data collection and analysis

We collected data from two sites in Maine, which were chosen because of their similarity in community size and proximity to the University of Maine. In both sites, we obtained the following achievement information for each student: (a) the mathematics subscale score on the 8th grade Maine Educational Assessment (MEA-M), (b) the mathematics subscale score on the locally administered standardized achievement test (ITBS in Site A and TerraNova in Site B), and (c) the course grade achieved in mathematics. In Site A ($n = 94$), all information was taken in the student's 8th grade year; in Site B ($n = 65$), the standardized achievement test and mathematics grades were obtained in the 9th grade (see Table 1). The MEA-M scores provide the only truly meaningful achievement information for comparing the two sites. From Table 2, one sees that the MEA-M mean for Site B was 17.76 points higher than that for Site A. With a pooled within-group standard deviation of 15.77, this mean difference corresponds to an effect size of $d = 17.76 \div 15.77 = +1.13$.

Creating a Common Scale for Mathematics Course Grade

As can be seen from Table 1, students in each site did not all enroll in the same level of mathematics. Our first task, then, was to create a single variable for "mathematics grade," even though it would comprise grades from different classes. Although we followed the same procedure in both sites, we will illustrate this procedure using data from Site A.

Site A students received a grade, on a 100-point scale, for either general mathematics (n = 59), algebra 1 (n = 29), or geometry (n = 6) (see Table 3). Because we believe that it makes little sense to regard a final grade in general mathematics as being comparable to the same grade in a higher level class, we weighted algebra 1 and geometry grades according to how these two groups of students performed on the MEA-M relative to the general mathematics students (see Table 4). Each of the two mean differences was converted to an effect size:

$$d_{21} = \frac{531.72 - 514.64}{9.46} = +1.81$$

$$d_{31} = \frac{555.00 - 514.64}{9.46} = +4.27$$

where d_{21} represents the difference in MEA-M scores between student enrolled in algebra 1and those taking general mathematics, and d_{31} the difference in MEA-M scores between geometry students and those taking general mathematics. Each effect size was then used to adjust upwards the mathematics grades for students enrolled in either algebra 1 or geometry. We did this by multiplying the pooled within-group standard deviation for mathematics grades (8.31) by either d_{21} or d_{31}, and then adding the product to the student's math grade. This resulted in an adjustment of +15.04 for each of the 29 algebra 1 students and +35.49 for the 6 geometry students. The resulting scale, which pools the three mathematics classes, is $\overline{X}$ = 89.24 and SD = 17.65.

Analyses and Results

Correlational Analyses

To examine the relationships among the results of state and local assessments, we obtained student-level within-site correlations among the three measures of student

achievement: (a) MEA-M, (b) the mathematics subscale score on the locally administered standardized achievement test (which we refer to as "ITBS/TN"), and (c) the weighted course grade achieved in mathematics ("COURSE").

As Table 5 shows, the three measures of mathematics achievement correlate substantially. Although these correlations are uniformly high, there is some variation in magnitude. Interestingly, COURSE correlates more highly with MEA-M than with ITBS/TN. This is not surprising, insofar as one would expect classroom assessments and the MEA to align with the *Learning Results* more than would be expected of a commercially available standardized test.

Classification Analyses

To explore an efficient and defensible method for combining multiple measures of achievement, we combined the three measures two different ways and compared the results by conducting classification analyses. As with the correlational analyses, these analyses were conducted within site.

Because of the standard setting process that was employed in the development of the Maine Educational Assessment, MEA-M scores can be stated in terms of performance levels that are tied to state standards:

exceeds the standard:	561
meets the standard:	541
partially meets the standard:	521
does not meet the standard:	<521

The critical score here is 541 (on a scale of 501-580), which is the cutscore that distinguishes between meeting the standard and not.

Although Maine school leaders soon will be expected to engage in standard

setting for their local assessments, the two sites in the present study, like most Maine school districts, have yet to implement standard setting. Consequently, neither COURSE nor ITBS/TN can be directly expressed as a performance level within the context of the *Learning Results*. However, because MEA-M correlates highly with both ITBS/TN and COURSE (Table 5), we can estimate, using simple regression, the critical cutscore for each of the latter two measures. We began by regressing ITBS/TN on MEA-M and, given the resulting equation, determined the predicted value of ITBS/TN for MEA-M = 541 (i.e., the designated cutscore for "meets the standard"). In Site A, for example, this regression equation is:

$$\text{ITBS/TN} = -676.487 + 1.4(\text{MEA-M})$$

which, for MEA-M = 541, yields an estimated cutscore of 80.91 (in percentile rank) for ITBS/TN. The analogous procedure was followed for COURSE. Again, for Site A this equation is:

$$\text{COURSE} = -443.307 + 1.019(\text{MEA-M})$$

which yields an estimated cutscore of 107.97 (in weighted grade) for COURSE. Thus, we identified in each site the score for ITBS/TN and for COURSE that corresponds to the MEA-M threshold for meeting the state standard.

We then transformed MEA-M, ITBS/TN, and COURSE to *z*-scores using the standard formula, but with one modification: We replaced the mean with 541 in the transformed MEA-M variable and the estimated cutscore (as described above) in the transformed COURSE and ITBS/TN variables. With this substitution, the sign of a *z*-score now indicates the student's performance relative to the MEA-M standard (rather than to the parent variable's mean).

Next, we formed an *unweighted* composite by taking the simple mean of the three transformed variables. A negative value on this composite went to the student who, on average, fell below the "standard" on all three measures. We also formed a *weighted* composite by (a) subjecting the three measures to a principal components analysis and (b) using the resulting component score coefficients to weight each measure in the formation of the composite. Each composite was dichotomized at 0, as were the transformed MEA-M, COURSE, and ITBS/TN variables. We then examined classification similarity by constructing a series of 2 x 2 tables.

The fundamental question is whether the unweighted and weighted composites classified students similarly. That is, when forming an achievement composite, is anything gained by weighting the measures that enter into the composite? As Table 6 shows, there was perfect agreement between the two sets of classifications. This no doubt reflects the relatively uniform correlations among MEA-M, ITBS/TN, and COURSE (Table 5) and, in turn, the relatively uniform component score coefficients that we obtained from the principal components analysis (see Table 7). In short, the results of this analysis indicate that weighting each measure is unnecessary. Thus, if the choice is between weighting or not weighting, the most efficient strategy for combining multiple measures would appear to be the latter. This assumes that correlations among measures are similar (which should be examined empirically) and that the measures are of equal importance. If either assumption does not hold, then weighting would be defensible.

A secondary question concerns the level of agreement between the classification based on the unweighted composite and that based on a single measure (see Tables 8-10). Except for the perfect agreement in Site A involving MEA-M, the levels of agreement are

fairly consistent, ranging from 89% to 92%. In these later cases, single-measure classification resulted in more students meeting the standard than when classification was based on the composite.

3. Identifying School Effects on Achievement

Student achievement is critically affected by variables at different levels of school organization. If academic achievement depends on the characteristics of students and teachers and/or the organizational context in which teaching and learning occurs, one cannot meaningfully assess school effects without considering these multi-level sources of influences (Keeves & Sellin, 1988). Previous studies of school effects in Maine and Kentucky analyzed aggregate school data to examine variation among schools in their performance status and gain, and found that poverty was the strongest and most consistent predictor of school performance in both states (Lee, 1998; Roeder, 2000). The past school performance indicators tend to focus on average test scores, which possibly conceal achievement differences among groups of students within each school. Consequently, these analyses are not sensitive to equity-related issues. Even when the effects of student-level background characteristics on achievement were considered to estimate value-added school performance, the effects are often assumed to be uniform across schools.

Multilevel analysis methods not only provide a means for formulating student-level and school-level regression models simultaneously, but they also provide more precise estimates of the relationships between predictors and outcomes at each level

(Bryk & Raudenbush, 1992). In particular, hierarchical linear modeling (HLM) is popular among educational researchers and evaluators for estimating school effects (see Phillips & Adcock, 1997; Weerasinghe, Orsak, & Mendro, 1997; Yen, Schafer, & Rahman, 1999). Because public schools do not randomly assign students and teachers across schools, multilevel methods that account for student and school context variables are regarded as the most rigorous means for estimating school effects (Phillips & Eugene, 1997). In fact, HLM has been found to produce more stable school effect estimates than ordinary least squares (OLS) or weighted least squares (WLS) methods (Yen et. al., 1999). This is true particularly when schools have few students and, thus, OLS estimates of the within-school regression parameter have low reliability.

Raudenbush and Willms (1995) discuss two different types of school effects: Type A and Type B effects. Type A effect is the difference between a child's actual performance and the expected performance had that child attended a typical school. This effect doesn't concern whether that effectiveness derives from school inputs (e.g., class size, teacher quality) or from factors related to school context (e.g., community affluence, parental support). By contrast, a Type B effect isolates the effect of a school's input from any attending effects of school context. The two indicators are appropriate for purposes of school choice and school accountability, respectively (Meyer, 1997). When HLM methods have been used to obtain school effect indices, researchers often did control for the influences of student background variables. However, the corresponding school-level compositional effects of these variables were not taken fully into account (see Weerasinghe, Orsak, & Mendro, 1997; Yen, Schafer, & Rahman, 1999). Raudenbush and Willms (1995) also suggest considering the possibility that a school will influence

different students differently. Yet there has been little research that systematically examines the achievement gaps among different groups of students as school effect indices.

How should we approach the challenge of identifying value-added contribution of schools to academic achievement for arriving at a judgment of, say, "effective"? Specifically, what is an efficient and defensible method for determining school effectiveness? This is the general question that we address in this section.

Data and Methods

In the present study, we use the data collected under 1996 NAEP 8th grade state math assessments for Kentucky and Maine. This allows us to compare the two states in terms of their school effects. The NAEP data are hierarchical in nature because students are nested within schools. HLM addresses the problem of students nested within schools. Further, the use of HLM on NAEP data copes with the problem of sampling error resulting from the multi-stage sampling in NAEP (see Arnold, 1993). Using HLM, we examine the effects of race and socioeconomic status on achievement at the student and school levels to estimate (a) adjusted school average achievement and (b) within-school racial and social gaps in achievement. We also examine relationships among the school performance indices obtained from HLM separately in each state. Finally, we compare schools in Maine and Kentucky from pooled HLM analyses and discuss implications of their differences for school effectiveness research.

Taking a multi-level organizational perspective and drawing on the relevant literature, we test three models of school effects separately for Maine and Kentucky: Model 1 (no predictors at the student and school levels), Model 2 (predictors at the

student level only, with grand-mean centering), and Model 3 (predictors both at the student and school levels, with grand-mean centering). Type A effect is estimated through Model 2 by removing the effect of student background variables. Type B effect is estimated through Model 3 by removing the effects of variables beyond a school's control (e.g., demographic composition). In this study, we consider only race and SES (socioeconomic status) factors. We believe that students' prior achievement (readiness for learning measured at the time of entry into current school) and mobility (length of stay in current school) factors must be considered to estimate authentic school effects but these data are not available in the NAEP.

All analyses were conducted using the HLM 5 program. Table 11 presents descriptive statistics for all variables used in these analyses. MRPCM1 through MRPCM5 are the five plausible values that make up the composite mathematics achievement outcome variable. WHITE is a dummy variable (1 = white, 0 = minority), and SES is a composite factor of parental education level, availability of reading materials at home, and school median income (standardized to have a mean of 0 and a standard deviation of 1 across states).

Model 1

Model 1, which includes no predictors at the student and school levels, partitions the total variance in mathematics achievement into its within- and between-school components. The school-level residual value from this model is used as an indicator of unadjusted school average performance.

Model 2

Model 2 adds student-level predictors by regressing mathematics achievement for student i within school j on race (WHITE) and socioeconomic status (SES). The Level 1 model (student level) is

$$(MRPCM)ij = \beta_{0j} + \beta_{1j}(WHITE)_{ij} + \beta_{2j}(SES)_{ij} + e_{ij}$$

where (MRPCM)ij is the composite mathematics achievement of student i in school j; $(WHITE)_{1ij}$ is the indicator of student i's race in school j; (SES)ij is the indicator of student i's socioeconomic status in school j; and e_{ij} is a Level 1 random effect representing the deviation of student $_{ij}$'s score from the predicted score based on the student-level model. Level 1 predictors are grand-mean centered so that the intercept, β_{0j}, can be interpreted as adjusted mean achievement for school j. This adjustment is chosen to sort out the unique effects of school on achievement after controlling for the influences of student/family characteristics.

The next step in HLM involves fitting an unconditional, or random, regression model at the school level (Level 2). Notice that all Level 1 regression coefficients are regarded as randomly varying across schools, and γ_{00} is the mean value of the school-level achievement outcome beyond the influences of student/family characteristics. r_{0j}, the school-level residual value from this regression, is used as an indicator of school average performance adjusted for racial and SES mixes of students. Likewise, r1j and r2j are used as indicators of racial and social achievement gaps respectively. The Level 2 (school level) model is

$$\beta_{0j} = \gamma_{00} + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + r_{1j}$$

$$\beta_{2j} = \gamma_{20} + r_{2j}$$

where β_{0j} represents school j's average mathematics achievement adjusted for its composition of students' racial and SES backgrounds; β_{1j} represents school j's racial gap (i.e., the achievement score gap between white and minority students); and β_{2j} represents school j's social gap (i.e., the extent to which students' SES differentiates their achievement).

Model 3

Model 3 adds two school-level predictors, or, school aggregate values of student-level predictors. Percent white (PWHITE) and average SES (AVSES) are added to explain between-school variation. r_{0j}, the school-level residual value from this regression, is used as an indicator of school average performance adjusted for racial and social composition effects. Model 3 is

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(PWHITE)_j + \gamma_{02}(AVSES)_j + r_{0j}$$

where $(PWHITE)_j$ is the proportion of white students (i.e., the mean of WHITE) in school j; and $(AVSES)_j$ is the mean SES of school j.

Results

Model 1 (fully unconditional model)

Decomposition of variance in the outcome variable shows that the two states have similar distributions of mathematics achievement between the school and student levels. In Maine,18% of variance exists at the school level and 82% at the student level; the

figures are 17% and 83%, respectively, in Kentucky. Residual school means from this model are called Model 1 average. The reliability estimate of these unadjusted school achievement averages is .80 in Maine and .79 in Kentucky, indicating that the sample means tend to be quite reliable as indicators of the true school means.

Model 2 (level-1 predictors only with grand-mean centering)

By using race and SES variables as predictors of math achievement at the student level (with grand-mean centering), we obtain adjusted school average achievement that takes into account differences among schools in their students' racial and social mixes. A residual school mean that is obtained after controlling for the effects of student-level predictors, as an indicator of value-added school performance, is called Model 2 average. The reliability of conditional school means (conditional reliability) becomes lower: .67 in Maine and .62 in Kentucky. As shown in Table 12, Model 2 average is correlated very highly with Model 1 average (r_{me}=.92 and r_{ky}=.87).

The effects of race and SES on achievement are used as indicators of academic inequity, as well as providing the basis for adjusting estimates of school effects. This assumes heterogeneity of regressions among schools and models the effects of student's race and SES on achievement as randomly varying at the school level. The within-school racial gap—the estimated average achievement gap between white and minority students within schools—is 12.1 (.41 standard deviations) in Maine and 16.8 (.57 *SD*) in Kentucky (see Table 13). The within-school social gap—the estimated effect of SES on achievement within schools—is 10.8 (.38 *SD*) in Maine and 10.6 (.36 *SD*) in Kentucky (see Table 13). In both states, these gaps are highly significant.

Maine and Kentucky show different patterns of relationships between achievement average and gap estimates (Table 12). In Maine, Model 2 average correlates positively with racial gap (.72) but negatively with social gap (-.63). Conversely, in Kentucky, Model 2 average correlates negatively with racial gap (-.28) but positively with social gap (.57). Higher performing schools in both states tend to have smaller gaps with regard to one background variable but larger gaps with regard to the other. This indicates that schools are not very effective in addressing both racial and social achievement gaps.

We should note that the reliability estimates of racial and social gaps are low: .13 and .21 in Maine, and .30 and .28 in Kentucky. Considering these reliabilities, it appears that both Maine and Kentucky schools vary little in their racial and social gaps. This is attributed to the fact that both states are highly homogeneous in racial composition. However, sufficient variability across schools on racial gap estimates does exist as the homogeneity of variance tests demonstrate significant variation (see the variance component chart in Table 13).

Model 3 (both level-1 and level-2 predictors with grand-mean centering)

School-level predictors of racial and social composition were used to make further adjustment for differences among schools in their average achievement due to composition effects. In Maine, both racial and social composition effects are not significant. This indicates that such school-level adjustment of performance for race and SES factors, in addition to the corresponding student-level adjustment, is not necessary (see Table 13). In Kentucky, only the social composition effect is significant, adding about 7 points to the within-school social gap estimate (see Table 13). Model 3 average—

residual school means after controlling for both student and school-level effects of race and SES—correlates .70 with Model 1 average and .94 with Model 2 average (see Table 12).

Pooled HLM analysis

In order to test differences in school performance between Maine and Kentucky, we pooled data from the two states and applied the same three models. However, we added a school-level dummy variable (MAINE) to indicate where a school's location (Maine = 1, Kentucky =0).

The results of the pooled HLM analyses are summarized in Table 4. First, the comparison of Maine and Kentucky schools without any control for background variables show that Maine schools perform significantly better than Kentucky schools: a gap of 17.18 (Model 1), or roughly 1.2 *SD*. The gap between Maine schools and their Kentucky counterparts in terms of average 8th grade mathematics achievement decreases about 40% when we control for their differences in students' racial and social background variables (gap = 9.97, Model 2). When we further control for school composition effects, the Maine-Kentucky school achievement gap becomes slightly smaller but remains statistically significant (gap = 6.18, Model 3). As Maine schools turn out to perform significantly better than Kentucky schools based on both Type A and Type B effect estimates, their effectiveness gap seems to come from sources related to schooling; students' prior achievement and mobility factors become less important when we compare schools across states (vs. within state). Despite the average school performance gap, it turned out that there are no significant differences between the two states' schools in terms of their racial and social gap estimates.

4. Discussion

Evaluation of systemic school reform requires that we evaluate school performance with multiple measures at mutliple levels of school system. This policy imperative makes data collection and analysis very challenging and complex. Despite the imperative, there is a lot of room for us to make technical choices that must be informed by scientific research. Although our results may not generalize to all states, they are expected to inform us about desired data and methods for a more systematic evaluation of systemic school reform. We caution that analytical methods themselves cannot cope with inherent measurement and attribution problems. We discuss implications of our research findings below.

Multi-measure Analysis of Student Achievement

Our results suggest that it is not necessary to weight each measure before forming an achievement composite to classify student performance. This is particularly true where measures are highly intercorrelated, as was the case here. If intercorrelations vary in magnitude, however, then it may be advisable to weight each measure to reflect the measure's association with the underlying principal component. Subsequent research would throw clarifying light on the merits of this recommendation, especially if the research involves multiple sites that differ with respect to the relatedness of the achievement measures they employ.

Having said this, we should acknowledge that high intercorrelations among measures are not sufficient for deciding in favor of an unweighted composite. That is,

one also should take into account the announced importance of each measure. For example, if a school district attaches greater importance to a district-wide assessment compared to, say, the standardized test that is annually administered, then the former should receive greater weight—even in the face of a high correlation between the two. Although there are various reasons why local achievement measures may differ in importance, a primary reason is the degree to which a measure aligns—in various respects (e.g., see Webb, 1997)—with the adopted standards. The reliability of assessment measures also need to be considered in developing weights.

Our results also point to the possible hazards of classifying student achievement based on a single measure. As Tables 8-10 illustrate, single-measure classification tended to result in additional students identified as meeting the standard. Are these students false positives? Because of two limitations of the present study, we unfortunately do not know. First, unlike MEA-M, which was designed to align with the *Learning Results*, neither ITBS/TN nor COURSE was constructed explicitly to reflect student attainment of these standards. This clearly is true for ITBS/TN, for no commercially available standardized achievement test is tailored to the standards of a particular state. And although teacher-constructed mathematics assessments (COURSE) in Maine arguably are more responsive to the *Learning Results*, the task of formally designing classroom assessments to demonstrably align with these standards still looms on the horizon for most Maine school districts. Clearly, in a standards-based climate, the integrity of an achievement composite depends, in part, on the extent to which the component measures are drawing on the same universe of standards. Without this assurance, we must interpret with caution the tendency of the single-measure

classifications to putatively overidentify students who meet the standard. Here, too, subsequent research could be illuminating, particularly if the research involves multiple sites that vary with respect to the degree to which each measure is of demonstrable alignment with the announced standards.

A second, and related, limitation of the present study is that neither site had engaged in formal standard setting for either ITBS/TN or COURSE—hence our decision to obtain regression estimates of ITBS/TN and COURSE cutscores, given the relationship between each measure and the MEA-M (for which the minimum score for "meets the standard" is known).

Multilevel Analysis of School Effects

We have tested three different models of estimating school effects. Model 2 is regarded as fairer than Model 1 as it considers student background factors that schools cannot control. Model 3 also may be fairer than Model 2 as it further takes into account school-level compositional effects beyond individual student-level effects and implies comparing "like with like." However, this position can be challenged in a situation where there is systematic covariation between school context and school practice variables. Raudenbush and Willms (1995, p. 332) point out the problem of causal inference:

> "Causal inference is much more problematic in the case of Type B effects because the treatment—school practice—is typically undefined so that the correlation between school context and school practice cannot be computed. Thus, even if the assignment of students to schools were strongly ignorable, the assignment of schools to treatments could not be."

Bryk and Raudenbush (1992, p.128) illustrate the problem where there exists differences in school staff quality that might confound the effects of school staff with the effects of student composition:

> "Suppose that [high SES] schools have more effective staff and that staff quality, not student composition, causes the elevated test scores. The results could occur, for example, if the school district assigned its best principals and teachers to the more affluent schools. If so, [Model 3] would give no credit to these leaders for their effective practices."

Conversely, one might argue that the differences among schools in school resources (including class size, teacher/administrator quality and instructional resources), possibly due to their different student demographic composition, are precisely what we need to remove for evaluating schools in fair ways. If high SES schools do a better job simply because they draw better staff, more resources, and better students, then this advantage should not be considered authentic "school" effects—i.e., differences among schools due to educational efforts and practices. Then, the task becomes to distinguish school inputs that are determined outside the school and sort out their effects as external school-level characteristics (Meyer, 1997). But this strategy can be more problematic when the school input variables are more highly correlated with school practice variables.

Thus, the fundamental issue is not simply a technical choice of estimation methods given the available data. Rather, the estimation of school effects requires that we define "school effects" and formulate an explicit model of these effects. In other words, this approach requires that the model be fully specified: all variables representing school input, practice, context, and student background would have to be measured and

included in the model in order to guarantee that the effects of school practice were unbiased. Nevertheless, school quality variables are generally more difficult to define and measure and the relevant data are expensive to collect (Raudenbush & Willms, 1995).

Our analysis of school effects also involved estimating student achievement gaps with regard to background characteristics (i.e., race and SES in our case). We found that while average achievement varies significantly among schools in both states, their racial and social gaps vary little among schools. This means that much of the observed variability in achievement gaps is sampling variance and, as a result, cannot be explained by school factors. Thus, at least in our data, it is not sensible to use student achievement gaps as school effect indices. It remains to be seen whether combination of state and local assessment measures would produce different results than those based on the NAEP.

References

Ardovino, J., Hollingsworth, J., & Ybarra, S. (2000). Multiple measures: Accurate ways to assess student achievement. Thousand Oaks, CA: Corwin Press.

Arnold, C. A. (1993). Using HLM with NAEP. Unpublished Paper Presented at the Advanced Studies Seminar on the Use of NAEP Data for Research and Policy Discussion, Washington, D.C.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models. Newbury Park: Sage Publication.

Jang, Y. (1998). Implementing standards-based multiple measures for IASA, Title I accountability using Terra Nova multiple assessment. Paper presented at the annual meeting of the AERA. (ED 426 084).

Keeves, J. P., & Sellin, N. (1988). Multilevel analysis. In J. P. Keeves (Ed.) Educational research, methodology, and measurement: An international handbook. New York: Pergamon Press.

Kolls, M. R. (1998). Standards-based multiple measures for IASA, Title I program improvement accountability: A vital link with district core values. Rowland unified school district. Paper presented at the annual meeting of the AERA. (ED 420 681).

Law, N. (1998). Implementing standards-based multiple measures for IASA, Title I accountability using Sacramento achievement levels. Paper presented at the annual meeting of the AERA. (ED 421 497).

Lee, J. (1998). Assessing the performance of public education in Maine: A national comparison. Orono, ME: University of Maine Center for Research and Evaluation.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. Economics of Education Review, 16(3), 283-301.

Novak, J. R., Winters, L., Flores, E. (2000). Using multiple measures for accountability purposes: one district's experience. Paper presented at the annual meeting of the AERA. (ED 443 846).

Phillips, G. W., & Adcock, E. P. (1997). Measuring school effects with HLM: data handling and modeling issues. Paper presented at the annual meeting of the AERA. (ED 409 330).

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. Journal of Educational and Behavioral Statistics. 20(4), 307-335.

Roeber, E. (1995). Emerging student assessment system for school reform. ERIC Digest (ED 389 959).

Roeder, P. W. (2000). Education reform and equitable excellence: The Kentucky experiment. Unpublished research paper.

Weerasinghe, D., Orsak, T., Mendro, R. (1997). Value added productivity indicators: A statistical comparison of the pre-test/post-test model and gain model. Paper presented at the annual meeting of the Southwest Educational Research Association. (ED 411 245)

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Research Monograph No. 6. National

Institute for Science Education (NISE), University of Wisconsin—Madison. Washington, DC: NISE.

Yen, S., Schafer, W. D., & Rahman, T. (1999). School effect indices: stability of one- and two-level formulations. Paper presented at the annual meeting of the AERA. (ED 430 029).

Table 1.
When achievement information was collected, by site.

achievement information ↓	*Site A* (*n* = 94)	*Site B* (*n* = 65)
Maine Educational Assessment (mathematics score)	8th grade	8th grade
Standardized achievement test, mathematics	8th grade (Iowa Test of Basic Skills; percentile ranks)	9th grade (Terra Nova; scaled scores)
course grade, mathematics	8th grade (course grade in general math, algebra 1, or geometry)	9th grade (course grade in applied math 1,integrated math, practical math 1, algebra 1, or geometry)

Table 2.
Distribution of MEA-M mathematics scores in each site.

	MEA-M performance	
course	*M*	*SD*
Site A (*n* = 94)	522.49	14.88
Site B (*n* = 65)	540.25	16.97
		SD_{pooled} = 15.77

Table 3.
Distribution of unweighted mathematics grades for each of three courses (Site A).

course	*M*	*SD*
general mathematics (*n* = 59)	78.24	9.26
algebra 1 (*n* = 29)	88.17	6.58
geometry (*n* = 6)	94.33	4.50
		SD_{pooled} = 8.31

Table 4.
Distribution of MEA-M mathematics scores for students in each of three mathematics courses (Site A).

	MEA-M performance	
course	*M*	*SD*
general mathematics (*n* = 59)	514.64	9.02
algebra 1 (*n* = 29)	531.72	10.82
geometry (*n* = 6)	555.00	5.33
		SD_{pooled} = 9.46

Table 5.
Correlations among measures of student achievement in mathematics.

	Site A	
	MEA-M	ITBS/TN
ITBS/TN	.81	
COURSE	.86	.72
	Site B	
ITBS/TN	.85	
COURSE	.84	.77

Table 6.
Classification similarity: unweighted and weighted composites.

Site A

		weighted composite	
		below standard	meets standard
unweighted composite	below standard	82	
	meets standard		12

Site B

		weighted composite	
		below standard	meets standard
unweighted composite	below standard	33	
	meets standard		32

Table 7.
Component score coefficients.

	Site A	Site B
MEA-M	.389	.389
ITBS/TN	.368	.376
COURSE	.354	.346

Table 8.
Classification similarity: Unweighted composite and MEA-M.

Site A
(100% agreement)

		MEA-M		
		below standard	meets standard	row total
unweighted composite	below standard	82		82
	meets standard		12	12
	column total	82	12	94

Site B
(92% agreement)

		MEA-M		
		below standard	meets standard	row total
unweighted composite	below standard	29	4	33
	meets standard	1	31	32
	column total	30	35	65

Table 9.
Classification similarity: Unweighted composite and ITBS/TN.

Site A
(91% agreement)

		ITBS/TN		
		below standard	meets standard	row total
unweighted composite	below standard	75	7	82
	meets standard	1	11	12
	column total	76	18	94

Site B
(91% agreement)

		ITBS/TN		
		below standard	meets standard	row total
unweighted composite	below standard	29	4	33
	meets standard	2	30	32
	column total	31	34	65

Table 10.
Classification similarity: Unweighted composite and COURSE.

Site A
(90% agreement)

		COURSE		
		below standard	meets standard	row total
unweighted composite	below standard	75	7	82
	meets standard	2	10	12
	column total	77	17	94

Site B
(89% agreement)

		COURSE		
		below standard	meets standard	row total
unweighted composite	below standard	28	5	33
	meets standard	2	30	32
	column total	30	35	65

Table 11.

Descriptive statistics of predictors and outcome variables for HLM analyses of Kentucky and Maine 1996 NAEP 8th grade math data

	Kentucky			Maine		
	n	*M*	*SD*	*n*	*M*	*SD*
			Student-level			
MRPCM1	2461	267.29	30.88	2258	285.22	30.51
MRPCM2	2461	267.14	31.00	2258	285.89	30.19
MRPCM3	2461	266.85	30.99	2258	284.95	30.17
MRPCM4	2461	267.01	30.87	2258	284.73	30.04
MRPCM5	2461	267.25	30.78	2258	285.11	30.32
WHITE	2535	0.87	0.33	2309	0.95	0.22
SES	2230	-0.40	0.94	2103	0.17	0.83
			School-level			
PWHITE	101	0.87	0.16	93	0.95	0.06
AVSES	101	-0.42	0.52	93	0.14	0.45

Table 12.

Correlations among school performance indicators

	Model 1 average	Model 2 average	Model 3 average	Racial gap
Model 2 average	0.87			
	0.92			
Model 3 average	0.70	0.94		
	0.82	0.97		
Racial gap	-0.24	-0.28	-0.23	
	0.61	0.72	0.77	
Social gap	0.34	0.57	0.53	-0.50
	-0.52	-0.64	-0.68	-0.96

Note. Upper values are for Kentucky and lower values are for Maine.

Table 13.

Summary of HLM Results

	Kentucky		Maine	
	Model 2	Model 3	Model 2	Model 3
	Estimation of Regression Coefficients (Fixed Effects)			
School-level Effects				
Adjusted Mean Outcome	266.58***	267.29***	283.92***	283.74***
PWHITE		-.39		38.01
AVSES		7.15**		3.27
Student-level Effects				
WHITE	16.79***	16.79***	12.11***	12.11***
SES	10.58***	10.58***	10.78***	10.78***
	Estimation of Variance Components (Random Effects)			
Adjusted Mean Outcome	90.39***	81.57***	91.86***	81.90***
WHITE	141.66***	141.66***	72.60**	72.60**
SES	21.42	21.42	16.50	16.50
	Percent of Outcome Variance Explained			
school-level	38.4	44.0	37.7	44.5
student-level	15.5	15.5	9.2	9.2

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 14.

Summary of Pooled HLM Results

	Model 1	Model 2	Model 3
	Estimation of Regression Coefficients		
School-level Effects			
Adjusted Mean Outcome	266.19***	270.29***	283.92***
MAINE	17.18***	9.97***	6.18**
PWHITE			4.41
AVSES			6.72***
Student-level Effects			
WHITE		16.77***	17.01***
SES		10.52***	10.02***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$