DOCUMENT RESUME

ED 442 871 TM 031 297

AUTHOR Lee, Jaekyung
TITLE Using National and State Assessments To Inform the Performance of Education Systems.
SPONS AGENCY National Science Foundation, Arlington, VA.
PUB DATE 2000-04-28
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Case Studies; Comparative Analysis; Educational Change; Elementary Secondary Education; National Competency Tests; *Performance Factors; *State Programs; Tables (Data); *Test Results; *Testing Programs
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

This study considered two questions about the use of national and state assessment databases: (1) Do state and national assessments provide the same information on the performance of an educational system? and (2) What are the factors that might affect the discrepancies between national and state assessment results? Kentucky and Maine were chosen for a case study. Four categories in the assessments of these states were compared with the same four categories of the National Assessment of Educational Progress (NAEP). While there were close similarities between the corresponding categories, it was risky to make direct comparisons without understanding how the NAEP and state assessments defined performance standards and how each state arrived at its own proficiency category labels. The percentage of students performing at or above high proficiency levels in the Maine and Kentucky assessments were not substantially different from the national assessment results. However, results were not entirely consistent across grades and years, a finding attributed to the fact that the definitions of performance standards and the methods of standard setting were different. The sizes of achievement gains from the state's own computations were greater than counterpart gains from the NAEP, something attributed to the high-stakes nature of the state assessments. These findings suggest that policymakers and educators need to become more aware of the uses and limitations of current national and state assessments as education information databases. (Contains 2 figures, 13 tables, and 12 references.) (SLD)

Using National and State Assessments to Inform the Performance of Education Systems

Jaekyung Lee

College of Education and Human Development

University of Maine

Paper presented at the Annual Meeting of the AERA (New Orleans, April 28, 2000)

This research was supported by a research grant from the National Science Foundation.

The views expressed herein are solely those of the author.

1. Research Objectives

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 and their academic growth.

In light of these concerns, I conducted a systematic analysis of currently available systemwide student assessments, that is, the NAEP and states' own assessments, and addressed the issue of the quality of data available for assessing and understanding the performance of states. The objective of this study is to identify and fill the gaps between currently available data and more desirable data in light of statewide systemic school reforms.

2. Research Methods and Findings

To explore the above questions, I examined two 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. I

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. I also used national assessment data for cross-check and cross-state comparisons: the NAEP state mathematics assessments were collected for 4th and 8th graders in 1992 and 1996. The NAEP state mathematics assessment was administered to a random sample of each state's fourth and eighth graders while both MEA and KIRIS were given to the entire populations of Maine and Kentucky fourth and eighth graders.

Several concerns have been raised about what data is required for adequately assessing 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? By and large, assessments in my study states, that is, Kentucky and Maine, 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.

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. Both Figure 1 and Figure 2 show the differences. 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 U.S. Department of Education Secretary Riley's House testimony at www.ed.gov/Speeches/04-1997/970429.html; Southern Regional Education Board President Musick's report at www.sreb.org/main/latestreports/MiscReports/set_stand.html). 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, where they modeled their frameworks closely after NAEP and adopted very challenging performance standards.

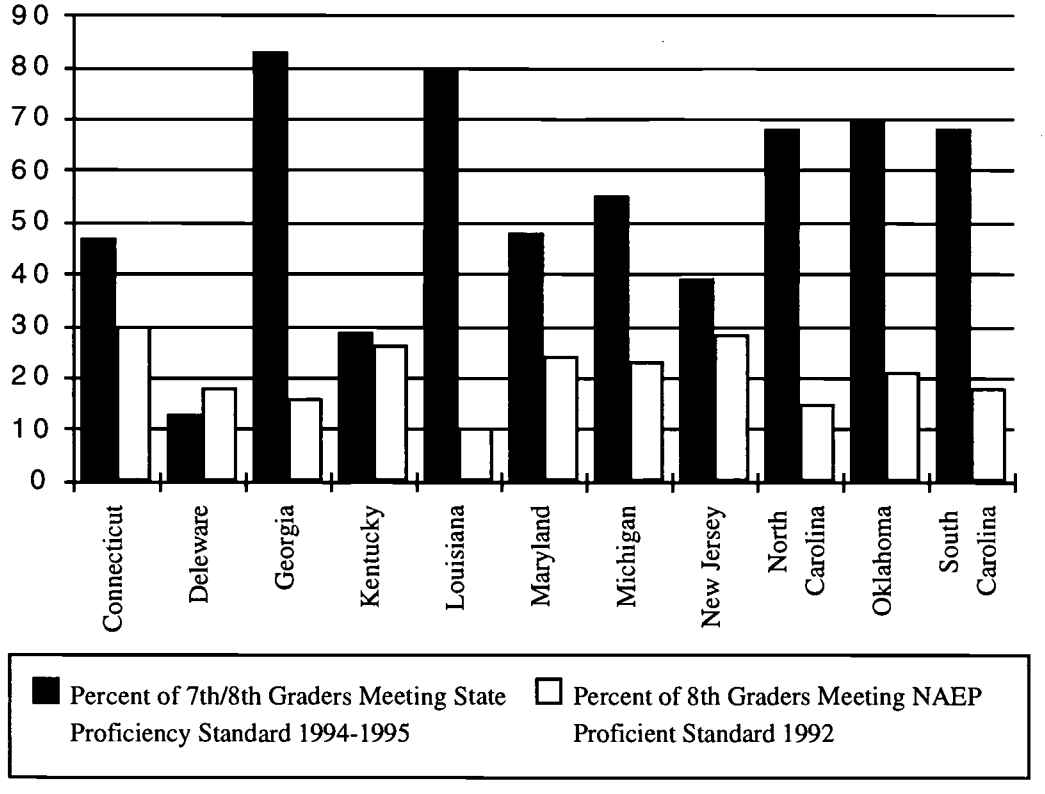

Figure 1. Comparison of National and State Assessment Results in 8th Grade Math

Note. Louisiana and Michigan state assessments are for 7th graders.
Source. Table 1 in Musick, M. D. Setting Education Standards High Enough (www.sreb.org/main/latestreports/MiscReports/set_stand.html).

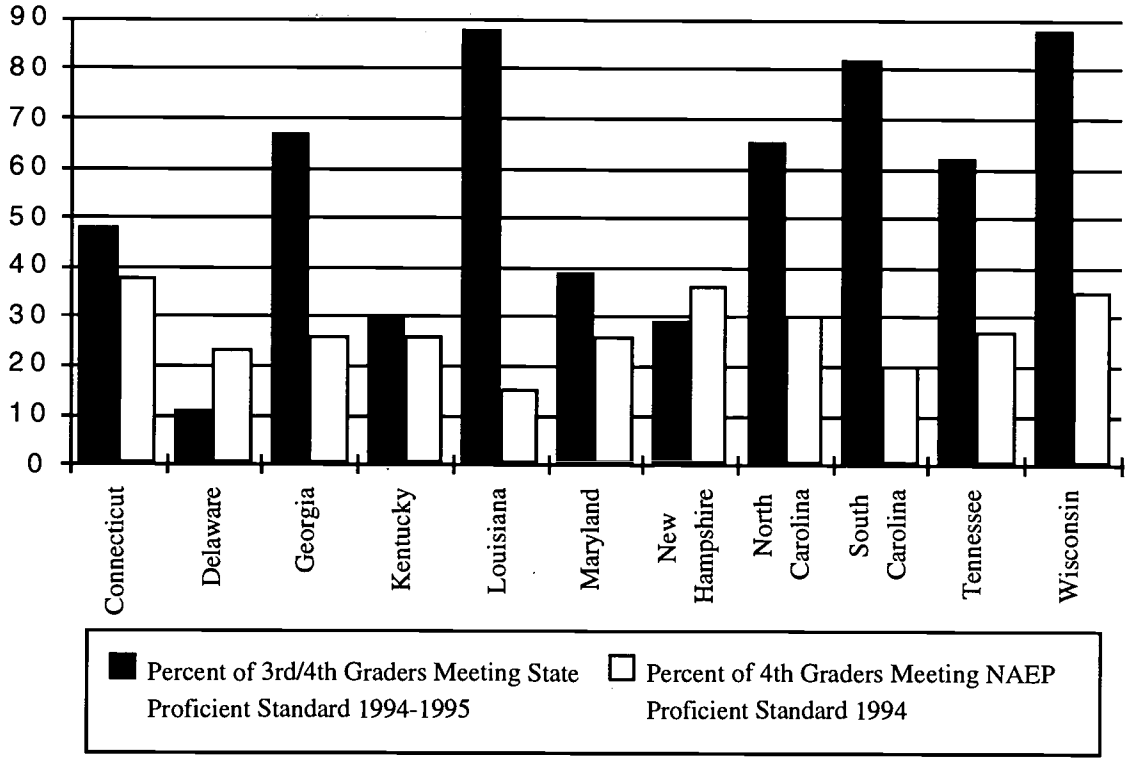

Figure 2. Comparison of National and State Assessment Results in 4th Grade Reading

Note. State assessment results in Delaware, Georgia, Louisiana, Maryland, New Hampshire, South Carolina and Wisconsin are for 3rd graders.
Source. Table 2 in Musick, M. D. Setting Education Standards High Enough (www.sreb.org/main/latestreports/MiscReports/set_stand.html).

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 level of Proficient on the KIRIS as a result of statewide education reform (i.e., 100% students proficient in 20 years), Maine did not specifically link their performance standards with the MEA proficiency levels. Despite the lack of standards-assessment linkage, it was reasonable to say that Maine also set their 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, I 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 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 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 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, I explored potential factors that might explain those gaps or inconsistencies in standards-based performance results by examining how performance standards were set for national and state assessments.

Differences in the Clarity and Specificity 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 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)

Indeed, the real issue is not so much with the name as with its operational definition. Part of the differences between NAEP and state performance results can be explained by comparing

their 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. The presence or absence of clearly-stated and well-specified definitions of performance standards and achievement levels by grade and subject was likely to cause 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 student ability to solve both routine and non-routine (many) problems with effective reasoning and communication, whereas the latter requires student ability 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 even at the proficiency levels with similar generic definitions and labels.

Table 5

Comparison of NAEP and MEA Definition of Math 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

(Table 5 Continued)

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 Performance Standard-Setting (Identification of Cut Scores) Processes

The NAEP math achievement levels were set following the 1990 assessment and further refined following the 1992 assessment. In developing the threshold values 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 an Angoff method. 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 the second and third rounds of review/revision processes.

The 1992 math achievement levels were evaluated by the National Academy of Education, which concluded that the current achievement levels raised serious concerns about their reliability and validity, were not reasonable (i.e., were set too high), and in the final analysis, should be abandoned by the end of the century. However, because NAGB did not agree with the results and believed in the value of standards-based reporting for the public, it decided to maintain the 92 math achievement levels (NCES, 1997).

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.

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

In the midst of this standards-based school accountability movement, the central question is whether the current NAEP and state assessments allow us to keep track of system performance. To examine this issue, I looked at time-series changes in MEA and KIRIS student performance. Table 6 shows that the overall Maine performance trends in mathematics are highly positive across grade levels over the 1990-1997 period. Table 7 also shows that the overall Kentucky performance trends in mathematics are highly positive across grade levels over the 1992-1998 period.

Table 6
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 7

1992-1998 KIRIS 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 states' own 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. Previous comparisons of the Kentucky and Maine assessment results with the NAEP in reading indicated some inflation of statewide gain scores (see Hambleton et al., 1995; Lee, 1998).

Tables 8 and 9 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 8

MEA and Maine NAEP Fourth Grade Average Math Scores, 1992 and 1996

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

Table 9

MEA and Maine NAEP Eighth Grade Average Math Scores, 1992 and 1996

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

Tables 10 and 11 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 10

Percent Kentucky 4th Graders at or above Proficient on KIRIS and NAEP Math, 1992 and 1996

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

Table 11

Percent Kentucky 8th Graders at or above Proficient on KIRIS and NAEP Math, 1992 and 1996

Assessment	1992	1996	Percent Gain	Standardized Gain
KIRIS	13	28	15	0.38
NAEP	14	16	2	0. 06

As shown in Tables 8, 9, 10 and 11, 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): approximately 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, any 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, I 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.

Changes in Test Content and Format

Test specifications provide information on the content and format of national and state assessments. Table 12 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 12. 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

Source: Table 1.1 in Reese et al. (1997). NAEP 1996 Mathematics Report Card for the Nation and the States.

Reportedly, the curriculum and assessment frameworks for both the KIRIS and the MEA were based on those employed in creating NAEP tests. Table 13 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 adddresses 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.

Table 13. 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.

While the 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 change from 1992 to 1996 in the NAEP math assessment was some shift toward more constructed-response questions, including extended constructed-response questions that required students to provide an answer and a corresponding explanation. In 1996, more than 50 percent of student assessment time was devoted to constructed-response questions. The MEA changed dramatically in 1995. The MEA began as a combination of both multiple-choice and constructed-response questions, but shifted to entirely constructed-response questions in 1995. The MEA 1994-95 guide explains the rationale 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)

Differences in Test Equating Strategies

NAEP, Kentucky and Maine assessments all used scaling and equating methods based on the Item Response Theory. With NAEP, equating was done directly between 1992 and 1996. With the MEA and KIRIS, which administer assessments every year, equating was done successively, that is, equating 1993 assessment with 1992 counterpart, 1994 assessment with 1993 counterpart

and so on. This will affect the size of equating error: the error of equating 92 and 96 test results is likely to be smaller in NAEP than in the state assessments. In both the KIRIS and MEA, relatively smaller percentage of items were used for equating, which also increases the error of equating.

In the 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 affects the gain in percent of students at the Advanced level from 1992 to 1996.

If error of equating happens regularly between successive years, the comparison of test results from remote years becomes less reliable by the accumulation of errors. In other words, the link between 1992 and 1996 state assessment results should become more tenuous as a result of more frequent changes in the content and format of tests as well as more repeated equating procedures.

Differences in Test Stakes

In addition to the impact of changes in test format and equating error, one of the reasons for inflated gains in Kentucky and Maine might be the impact of the state assessments on school curriculum and instructional practices due to stakes attached to the state tests. In Kentucky, scores are used to measure school improvement and 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 for schools to make decisions about curricula and instruction. But reporting school performance to the public was likely to promote teaching to the test. Given such moderate to high stakes attached to the KIRIS and the MEA, it is likely that states' own 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)

3. 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 state education systems. This study raised two interrelated questions regarding the use of national and state assessment databases. 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. While there were close similarities between the four categories in the NAEP and the corresponding four categories in the state assessments, it was risky to make direct comparisons without understanding how the NAEP and state assessments defined performance standards and how each state arrived at its own proficiency category labels and definitions. Thus, I investigated whether the measures and the methods of standard setting were substantially different between the NAEP and state assessments, and how those differences might have affected their performance-based student assessment results.

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 is attributed to the fact that the definitions of performance standards and the methods of standard setting were different.

Both states reported increased student achievement based on their statewide assessment results. I tried to compare such changes in the states' student performance based on the NAEP and each state's own assessment. 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 is 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.

Although these findings may not be generalized to all states, they inform us about the requirements of data collection for a more systemic evaluation of state education systems. It suggests that policy-makers and educators need to become more aware of the utilities and limitations of current national and state assessments as educational information databases.

References

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

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.

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

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

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.

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

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

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

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

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

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

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").