Do National and State Assessments Converge for Educational Accountability?
A Meta-Analytic Synthesis of Multiple Measures in Maine and Kentucky

Jaekyung Lee

* Graduate School of Education, State University of New York at Buffalo,

Online Publication Date: 13 June 2007


To link to this Article: DOI: 10.1080/08957340701301462

URL: http://dx.doi.org/10.1080/08957340701301462

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Do National and State Assessments Converge for Educational Accountability?
A Meta-Analytic Synthesis of Multiple Measures in Maine and Kentucky

Jaekyung Lee

Graduate School of Education
State University of New York at Buffalo

Given the policy imperative of using multiple measures for state education accountability under the No Child Left Behind Act (NCLB), this study examines similarities and discrepancies between the National Assessment of Educational Progress (NAEP) and the states' own math assessment results in Kentucky and Maine, with a focus on 3 major academic performance indicators: proficiency level, achievement gap, and achievement gain. Using meta-analytic techniques, the study synthesizes multiple measures from the two states over the periods of 1992–1996 and 2000–2003. It pinpoints the areas and degrees of the discrepancies and explores contributing factors. It also reports emerging convergence of the NAEP and state assessments under the NCLB.

The No Child Left Behind Act (NCLB), aimed at ensuring high academic standards for all students, requires that schools have all students perform at the proficient level or above by the 2013–2014 school year. The NCLB requires that states include multiple measures of student academic performance to determine yearly state performance. While it is evident that student achievement measured by state assessments is used to evaluate state performance, it is not clear what other measures should be included and how that information should be combined into an overall evaluation of state performance (Henderson-Montero, Julian, & Yen, 2003). Although the NCLB does not prescribe the role of the National
Assessment of Educational Progress (NAEP) for state assessments, the expected use of the NAEP as a multiple measure was to confirm state test results under the purview of the U.S. Department of Education, evaluating the rigor of state standards, growth in student achievement, and the reduction of achievement gaps among concerned subgroups of students (Ad Hoc Committee on Confirming Test Results, 2002; Henderson-Montero et al., 2003).

Demand for stronger systemic accountability for student achievement outcomes requires valid and reliable information on the performance of state education systems as a whole. Although the NAEP and individual state student assessments can be used to inform us of state-level performance, several problems still exist. On one hand, states have difficulty in realigning their student assessment systems and tracking student achievement (Consortium for Policy Research in Education, 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.

NCLB requires that states issue annual report cards that include the following information on statewide student achievement: (a) aggregate achievement data in math and reading/language arts at each proficiency level; (b) the data disaggregated by disability, limited English proficiency (LEP) status, race/ethnicity, socioeconomic level, gender, and migrant status; and (c) most recent 2-year trend data for achievement in required assessment areas by each grade and subject area. This requirement speaks to the importance of the following three pieces of information for performance-based educational accountability: (a) achievement level, (b) achievement gap, and (c) achievement gain.

Although both NAEP and state assessment results may be used as a tool to cross-check and cross-validate state-level performance, previous comparisons of the NAEP and state assessment results showed significant discrepancies in the level of student achievement as well as the size of statewide achievement gains (Klein, Hamilton, McCaffrey, & Stecher, 2000; Lee & McIntire, 2002; Koretz & Barron, 1998; Linn, Baker, & Betebenner, 2002). The percentages of students reaching the proficient level tend to be generally lower on NAEP than on the state assessments. These results were interpreted as implying that for many

---

1The idea of using multiple measures for accountability is consistent with the Standards for Educational and Psychological Testing (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Educational Measurement (NCEM), 1999): “In educational settings, a decision or characterization that will have major impact on a student should not be made on a simple test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision” (pp. 147–148). The same position is found in the Standards for Educational Accountability Systems (Baker, Linn, Herman, & Koretz, 2002).
states, NAEP proficiency levels are more challenging than the states’ own and that state standards are still not high enough (National Education Goals Panel, 1996). Increases in the percentages of Proficient students also tend to be lower on NAEP than on the state assessments. This divergence of trends was interpreted as implying that there are serious questions about the generalizability of gains reported on a state’s own assessment, and hence about the validity of claims regarding student achievement (Linn, 2000). These previous comparisons were often restricted to a single state with a focus on a single performance indicator; there were no systematic analyses combining results on multiple indicators from multiple states. Particularly, those prior studies did not look into possible differences between national and state assessments in their estimation of the achievement gaps, an important indicator of state performance in equity. Moreover, the patterns of NAEP–state assessment discrepancies could have changed recently as a result of NCLB and related changes in state assessments such as greater alignment of state assessment with NAEP.

In light of these concerns, it is necessary to examine whether and how the current NAEP and states’ own student assessments can be used together to inform policymakers of statewide academic performance. It is also critical to examine if the national and state assessments produce consistent results over time, particularly before and after NCLB. Further, it remains to be examined how the patterns vary among states with different types of state testing systems—for example, states that previously had low-stakes state testing (e.g., Maine) and high-stakes state testing (e.g., Kentucky). This requires a systematic comparative analysis of national and state student assessment data in states that can produce valid and reliable information on academic performance, specifically information on the proficiency levels of students, the achievement gaps among different groups of students, and their academic progress. The objective of this study is to investigate discrepancies between national and state assessment results at the state level with focus on these three major academic performance indicators and to explain possible causes of the discrepancies.

DATA AND METHODS

For this study, three major indicators of statewide school system academic performance were examined: (a) the level of proficiency in student achievement, (b) the size of gaps among different groups of students’ achievement, and (c) the amount of progress in student achievement. This study synthesizes measures of the discrepancies between the NAEP and state assessment results in two selected states. It also examines multiple sources of data, including actual assessments and technical reports, to account for the discrepancies.
National and state assessment data from two states, Kentucky and Maine, were examined. These two states were selected because their state assessments shared common characteristics but at the same time provided a sharp contrast of testing environment for this study. Both states (a) put student assessment systems in place early enough to gather baseline data and monitor their progress, (b) made their assessments in line with the goals of their education reform initiatives, and (c) adopted similar performance standards to those in the NAEP. Despite these common characteristics, the two states’ assessments differed significantly in terms of the stakes attached to the assessment results: high-stakes testing in Kentucky versus low-stakes testing in Maine. Primary data sources were from the two states’ NAEP fourth and eighth grade mathematics assessments in 1996 and 2003 (for evaluation of proficiency and achievement gap); 1992 and 2000 NAEP data were also used as supplements (for evaluation of progress). The NAEP state mathematics assessments were taken by representative samples of fourth and eighth graders in the two states. At the same time, the same grades’ math achievement data were also collected from the states’ own student assessments. Kentucky changed its state assessment from Kentucky Instructional Results Information System (KIRIS) to Commonwealth Accountability Testing System (CATS) in 1999; CATS includes multiple tests among which is the Kentucky Core Content Test (KCCT), a criterion-referenced test, that is the successor of KIRIS. Maine has kept the same assessment, the Maine Educational Assessment (MEA), while changing its performance standards in 1999. Because both states changed their state assessments somehow since 1999, separate analyses were conducted for each period, and the results may not be directly comparable to each other. Therefore, the study can give insights into potential problems of discrepancies between national and state assessment results that the states should overcome in meeting the requirements of NCLB.

The NAEP state assessment was administered to a stratified random sample of each state’s fourth and eighth graders, whereas both the MEA and KIRIS/KCCT were given to the virtually entire populations of Maine and Kentucky fourth and eighth graders who were eligible for testing. Table 1 shows how the 1996 and 2003 fourth and eighth grade NAEP samples compare with their corresponding state assessment samples in Maine and Kentucky. The data do not include students who were exempt or absent from testing and whose test scores were missing or not reported for any reasons.

Measures

Several concerns have been raised about what data is required for adequately measuring the performance of a state education system (Laguarda, Breckenridge,
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 Kentucky and Maine meet the above-mentioned criteria. Both Maine and Kentucky assessments modeled their frameworks closely after the NAEP. Test specifications (test blueprint) provide information on the content and format of national and state assessments. Table 2 shows the percentages of questions in 2003 NAEP, MEA, and KCCT Grade 4/5 and Grade 8 math assessments. Compared with the assessment used in the early 1990s, changes were made in two content areas, “number and computation” with less emphasis and “algebra and functions” with greater emphasis, which reportedly reflect the refinement of the NAEP math assessment to conform

<table>
<thead>
<tr>
<th>State/Year/Grade</th>
<th>NAEP</th>
<th>State Assessmenta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maine 1996</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>2,115</td>
<td>16,098</td>
</tr>
<tr>
<td>Grade 8</td>
<td>2,258</td>
<td>17,068</td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>2,879</td>
<td>15,500</td>
</tr>
<tr>
<td>Grade 8</td>
<td>2,861</td>
<td>17,367</td>
</tr>
<tr>
<td>Kentucky 1996</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>2,579</td>
<td>50,527</td>
</tr>
<tr>
<td>Grade 8</td>
<td>2,461</td>
<td>50,664</td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4b</td>
<td>3,451</td>
<td>49,679</td>
</tr>
<tr>
<td>Grade 8</td>
<td>2,833</td>
<td>49,069</td>
</tr>
</tbody>
</table>

Note. NAEP = National Assessment of Educational Progress; MEA = Maine Educational Assessment; KIRIS = Kentucky Instructional Results Information System; KCCT = Kentucky Core Content Test.

bCounts for Grade 5 instead of Grade 4 are reported for 2003 in Kentucky. KCCT math assessment was given to the fifth graders at the elementary level, and there was no exact counterpart of the 2003 NAEP fourth-grade math assessment in Kentucky.
with recommendations from the National Council of Teachers of Mathematics standards (Reese, Miller, Mazzeo, & Dossey, 1997). Reportedly, the curriculum and assessment frameworks for both Kentucky and Maine assessments were based on those employed in creating NAEP tests. Although the percentages were not in exact agreement across all content areas, the overall MEA and KCCT frameworks for mathematics appear to be consistent with the NAEP counterpart. Indeed, they all include four main content areas related to the major mathematical curricular areas of number/computation, geometry/measurement, probability/statistics, and algebra/functions.

One way to verify whether national and state assessments measure the same construct is to compute correlations between test scores from students who took both tests. An early estimation of the correlation between 1992 KIRIS and NAEP reading scores indicated a high level of comparability of KIRIS and NAEP test results (Kentucky Department of Education, 1995). Because the NAEP does not identify individual students, no correlation could be computed between the two tests at the student level. Instead, NAEP restricted data provides a link to the Common Core of Data with common school identification codes, which in turn provides a link to state assessment data. It was possible to match schools common to both NAEP and state assessments and to correlate school mean scores from the two tests. Measurement error and sampling weight need to be taken into account in the estimation of school average achievement on the NAEP. There are five plausible values, resulting from five random draws from the conditional distribution of proficiency scores for each student. Thus, the estimation of NAEP

### TABLE 2
Percentage Distribution of 2003 NAEP, MEA, and KCCT Math Test Items by Content Strand and Grade

<table>
<thead>
<tr>
<th>Content Areaa</th>
<th>Grade 4/5</th>
<th></th>
<th></th>
<th>Grade 8</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NAEP</td>
<td>MEA</td>
<td>KCCT</td>
<td>NAEP</td>
<td>MEA</td>
<td>KCCT</td>
</tr>
<tr>
<td>Number/computation</td>
<td>40</td>
<td>30</td>
<td>45</td>
<td>25</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td>Geometry/ measurement</td>
<td>35</td>
<td>24</td>
<td>25</td>
<td>35</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td>Probability/statistics</td>
<td>10</td>
<td>12</td>
<td>15</td>
<td>15</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>Algebra/functions</td>
<td>15</td>
<td>21</td>
<td>15</td>
<td>25</td>
<td>27</td>
<td>25</td>
</tr>
<tr>
<td>Total percentage</td>
<td>100</td>
<td>95</td>
<td>100</td>
<td>100</td>
<td>95</td>
<td>100</td>
</tr>
</tbody>
</table>

*Note. NAEP = National Assessment of Educational Progress; MEA = Maine Educational Assessment; KCCT = Kentucky Core Content Test.

aSome content areas (i.e., geometry and measurement; probability and statistics) are combined and shown under the same category for the sake of comparison across all three assessments. While Maine’s math standards include a content area called “Discrete Mathematics” which takes 5% of MEA eighth-grade math test items, it is not included here due to the lack of comparable standards in the other assessments and thus the total percentages in Maine do not add up to 100.
school-average math achievement took the average of estimates from separate analyses of the five plausible values, using a within-school sampling weight to account for differential probability of student selection.

The analyses of school-average math test scores from the 1996 NAEP data in conjunction with the 1996 state assessment datasets showed a strong, positive relationship between NAEP and state assessment results in Maine and Kentucky. The school-level correlation between 1996 NAEP and state math scores was .50 (Grade 4) and .52 (Grade 8) in Kentucky and .63 (Grade 4) and .70 (Grade 8) in Maine. These estimates of correlations between NAEP and state assessment scores may have been attenuated due to the restriction of range in school-level aggregate scores; the standard deviation of school-average math scores was less than half of the standard deviation of individual students’ math scores. Thus, the above correlation coefficients have been corrected for restricted ranges: .79 (Grade 4) and .82 (Grade 8) in Kentucky and .89 (Grade 4) and .91 (Grade 8) in Maine.

This study assumes that because the state standards were similar to the NAEP frameworks and there was a reasonably high correlation between the NAEP and state assessment results, the assessment content would be sufficiently similar and the performance could be compared directly. Even with similarities in standards, however, there can still be important differences between the assessments that lead to curricular and instructional variations. Despite considerable agreement at the level of broad content area and basic organization of math, the NAEP and state assessments may reveal differences at the item level in operationalization of content and item format.

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 assessments adopted 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 and continued for the KCCT: Novice, Apprentice, Proficient, and Distinguished. Whereas 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 its performance standards with the MEA proficiency levels until 1999. Despite the lack of standards-assessment linkage prior to 1999, it was reasonable to assume that Maine also set its performance expectation for all
students to the Advanced level on the MEA. For comparison of the assessment results related to achievement levels, this study only examines the level of achievement defined by the states as meeting desired performance standard.²

Meta-Analytic Synthesis of the Measures

For this study, there were two statewide aggregate measures, that is, NAEP and state assessment measures, for each of the three performance indicators (proficiency level, achievement gap, achievement gain) in each of the two states (Kentucky and Maine) at Grade 4/5 and Grade 8 during 1992, 1996, 2000, and 2003. These repeated measures were analyzed separately by each time period, that is, before and after the states changed their assessments (the period of 1992–1996 and the period of 2000–2003).

In combining the NAEP and state assessment measures, this study used meta-analytic techniques with applications of both fixed-effects and random-effects models. Given the limitation of data available from only two selected states at two available grades, a fixed-effects model may be more appropriate because differences between samples that lead to differences in achievement measures are not regarded as random (e.g., if they are regarded as consequences of purposeful design decisions), and the inferences desired apply only to assessments under examination (see Hedges, 1994).

On the other hand, a random-effects model can be appropriate when significant variations among the samples indicate the presence of random effects to be accounted for (Shadish & Haddock, 1994). If the set of data under investigation is conceived as a random sample, a meta-analysis can be viewed as a survey having a two-stage cluster sampling procedure (see Raudenbush, 1994). In this case, the two selected states and two grades may be treated as a set of samples (clusters) that are drawn from the target population of all K–12 grades in all states and then a random sample of students is obtained from the population of all students in each cluster (i.e., selected set of states and grades).³

Using both fixed-effects and random-effects models, this study combined the measures of NAEP–state assessment discrepancies from both Maine and

²Major concerns are that the state’s emphasis on its “Proficiency” standard overlooks the progress made by lower-performing students who meet or exceed the minimum competency level and has a chilling effect on schools and teachers who have moved a significant number of students above that basic level.

³A random-effects model also may be justified on the ground that the fourth- and eighth-grade cohort groups of students tested change every year and thus the achievement data reported each year for students are not static over time. Linn (2001) points out that treating students as a fixed set implies treating the results only as a historical fact rather than as an indication of school quality or a measure against which school progress will be judged by comparison of successive cohort groups.
Kentucky so that the estimates of weighted average effect sizes and standard errors were obtained for each one of the three performance indicators (see Appendix). Further, the study explored potential factors that might explain observed discrepancies between the national and state assessment results. Analyses of the student test data and reviews of technical reports were conducted to provide the information.

This study has limitations because the data were drawn from only two selected states in a single subject area at one or two grades during limited time periods. There are substantial challenges to overcome in conducting a meta-analysis with data from a greater number of states for generalizability. The first problem is the lack of desired state test data (or difficulty of access to the data) on key performance indicators that this study examines, particularly data on the achievement gap by race and poverty during the 1990s; most states did not break down their state test results by subgroups until NCLB. Using only recent data does not allow us to look at possible changes in the results over time. Although the meta-analysis requires calculating the correlations between NAEP and state assessment measures with matched data at the school level in every state, the past school performance data on a state assessment is not readily available for linking to NAEP. The second problem is the difficulty of collecting comparable data and information across states to examine the factors that might affect the NAEP–state assessment discrepancies, because few states make raw test data and technical manuals readily available to outside researchers. Further, a more fundamental difficulty may arise due to the complexity of comparison, as Linn (2001) pointed out that any two state systems differ in terms of multiple factors that would make it difficult to attribute differences in effectiveness to any single factor. With these caveats in mind, the analysis of two purposefully selected cases (Maine and Kentucky) helps illustrate key issues common to different types of states and highlight their differences for comparison.

RESULTS

Indicator 1: How Well Do Students Perform on National Versus State Assessment?

In order to see how students in Kentucky and Maine meet national and state performance standards, I compared NAEP and state math assessment results in 1996 and in 2003. As shown by the 1996 assessment results in Figure 1, the percentage of students performing at or above the NAEP Proficient level is greater than the percentage of students performing at or above the MEA Advanced level. Specifically, the difference is remarkable at Grade 8: 31% of Maine eighth-grade students met the NAEP’s Proficient level in math as of 1996, whereas only 9% of
the students met the MEA’s Advanced level. Whether we base our judgment of Maine students’ performance on the NAEP or MEA achievement levels, we come to the same conclusion—that a majority of the student population in Maine does not perform at Proficiency level across grades. Comparison of 1996 NAEP and KIRIS assessment results also reveal inconsistent performance results in Kentucky. Figure 1 shows the results of the two assessments in 1996, where the percentage of students at or above the NAEP Proficient level is similar to its counterpart at or above the KIRIS Proficient level at Grade 4 but substantially smaller at Grade 8.

The performance standard for the MEA proficiency level appears to have been set at relatively comparable or even higher levels than the standards for NAEP proficiency: the percentage of students at or above the NAEP Proficient level was not much different from its corresponding figures at or above the MEA Advanced level. In contrast, the performance standard for the KIRIS proficiency level appears to have been set at relatively lower levels than the standards for NAEP. The meta-analysis of the NAEP, MEA, and KIRIS assessment results identified inconsistent percentages of Proficient students across the four measures from two states. The weighted average of standardized differences between national and state assessment results on student proficiency rate was .90 for a fixed-effects model and .65 for a random-effects model. This average difference is statistically significant, and the effect size is deemed large (see Table 3).

Figure 2 shows the results of NAEP and state assessments in 2003. Although the performance standard for math proficiency remains higher on the NAEP than on the state assessment in Kentucky, the gap has slightly narrowed between 1996
and 2003. In 2003 the percentages of Proficient students in Grade 8 mathematics were 31 on the KCCT versus 24 on the NAEP. In Maine, where the performance standard for math proficiency was lower on the NAEP than on the state assessment, the gap remains the same after the state changed its performance standards for MEA in 1999. As with Kentucky, Maine also has narrowed the gap a little after NCLB. In 2003, the percentages of Proficient students (meeting or exceeding the standards in Maine’s Learning Results) in Grade 8 mathematics were 18 on the MEA and 29 on the NAEP. The average standardized NAEP–state assessment difference in proficiency rate across the two states for Grades 4 and 8 was about .5. Although the difference dropped to some extent over the 1996–2003 period, the proficiency gap still remains large (see Table 3).

### TABLE 3
Percentages of Maine and Kentucky Fourth and Eighth Graders at or Above Proficient Level on 1996 and 2003 National and State Mathematics Assessments

<table>
<thead>
<tr>
<th>State</th>
<th>Grade</th>
<th>Assessment</th>
<th>Proficiency Rate</th>
<th>NAEP–State Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>Maine</td>
<td>4</td>
<td>NAEP</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MEA</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Maine</td>
<td>8</td>
<td>NAEP</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MEA</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Kentucky</td>
<td>4</td>
<td>NAEP</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KIRIS</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Kentucky</td>
<td>8</td>
<td>NAEP</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KIRIS</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Fixed-effects average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Random-effects average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>Maine</td>
<td>4</td>
<td>NAEP</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MEA</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Maine</td>
<td>8</td>
<td>NAEP</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MEA</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Kentucky</td>
<td>4</td>
<td>NAEP</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KCCT</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Kentucky</td>
<td>8</td>
<td>NAEP</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KCCT</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Fixed-effects average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Random-effects average</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* NAEP = National Assessment of Educational Progress; MEA = Maine Educational Assessment; KIRIS = Kentucky Instructional Results Information System; KCCT = Kentucky Core Content Test.

NAEP–state difference is reported in the unit of logit (see Appendix). Standard error of the difference is shown in parentheses.

*p < .05.*
Indicator 2: How Much Does Achievement Vary Among Student Groups on National Versus State Assessment?

When the performance of a school system is evaluated from a perspective of equity, the size of the student achievement score gap with regard to educationally irrelevant variables (e.g., gender, race, socioeconomic status) becomes an important indicator of a system’s performance. It was examined whether the sizes of achievement gaps between different types of students are consistent between the individual state assessments and the NAEP. Major student background variables were selected that are available in both the national and state assessments and computed standardized gap estimates (see Table 4). A parental education variable classifies students into two categories: high school or more education versus less than high school. A race variable also classifies students into two categories: White versus minority groups. An income variable classifies students into two categories: ones who are ineligible versus eligible for a free/reduced-price school lunch program.

Because the NAEP and state assessments used different scales and they did not allow for direct comparisons of the achievement gap sizes, original measures of the gaps from each of the two assessments were standardized by dividing them by the standard deviations of their corresponding test scores. As the student achievement gaps reported in standard deviation units incorporate any differences in test score distributions as scaling artifacts, any discrepancies between the national and state assessments in the size of achievement gaps among the same student groups require explanation.

As of 1996, the standardized gap estimates in standard deviation units turned out to be slightly smaller on the state assessments than on the NAEP, although
the discrepancies were insignificant in most cases. The only exception to this pattern was a gender gap in Kentucky eighth-grade math, where the gap appeared larger on the KIRIS than on the NAEP. Regardless of the type of assessment in each state, however, it needs to be noted that the score differences between male and female students are relatively small in comparison with racial and social achievement 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 .86 on the NAEP. In Kentucky, the gap between White and minority students was also as large as .60 on the NAEP. Similar estimates of those achievement gaps were obtained from corresponding state assessments. The meta-analysis of the NAEP, MEA, and KIRIS assessment results identified generally consistent results across the gap measures from two states. The weighted average of standardized differences between national and state assessment results on student achievement gaps was .07. Although this average difference was statistically significant at the .05 level, the effect size was too small to be meaningful.

As shown in the lower part of Table 4, the study found continuing patterns of insignificant discrepancies in the math achievement gap estimates between the NAEP and state assessments. None of the NAEP achievement gap estimates reported in 2003 for gender, parental education, race, or income was statistically significantly different from its state assessment counterpart. The weighted average of standardized differences between national and state assessment results on student achievement gaps was .05.

Indicator 3: How Much Has Achievement Improved Toward the Performance Goal on National Versus State Assessment?

In the midst of the standards-based school accountability system, every school system is expected to make continuous academic progress toward a common performance goal set by the state. Under NCLB, states also have to report statewide progress toward the ultimate goal of 100% proficiency by using multiple measures of student performance. The central question is whether the current NAEP and state assessments allow states to consistently keep track of statewide academic performance.

The upper part of Table 5 compares Maine students’ 1992–1996 achievement gains based on the NAEP and MEA math assessment results. Because Maine did not have performance standards and standards-based achievement levels for MEA in 1992, scale scores were used instead to evaluate 1992–1996 achievement gains. Because MEA and NAEP scores employed different scales, a common metric in standard deviation units was established to compute the states’ standardized 1992–1996 gain scores on MEA and NAEP, respectively. Specifically, the gain scores were standardized by dividing them by the standard deviations of students’ scale scores ($SD_{MEA \ Grade \ 4} = 154$, $SD_{MEA \ Grade \ 8} = 132$, $SD_{NAEP \ Maine \ Grade \ 4} = 33$, $SD_{NAEP \ Maine \ Grade \ 8} = 31$).

Table 5 also shows Kentucky students’ 1992–1996 achievement gains based on the NAEP and KIRIS math assessment results. Because both NAEP and KIRIS report gains in the percentage of students meeting their own performance
standards, a common metric in log odds ratio (logit) units was established. Specifically, percentages of students at or above the Proficient level, as obtained from the KIRIS 1992 and 1996 assessment results, were used to compute KIRIS standardized gains, whereas their counterparts from the 1992 and 1996 NAEP state assessment results were used to compute NAEP standardized gains.

As shown previously, there was overall statewide academic improvement in Maine and Kentucky between 1992 and 1996, as measured by the MEA and KIRIS. However, the sizes of state math score gains tend to be considerably greater than are observed in national assessment results (NAEP): approximately

<table>
<thead>
<tr>
<th>State</th>
<th>Grade</th>
<th>Assessment</th>
<th>Raw Gain</th>
<th>Standardized Gain</th>
<th>NAEP–State Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992–1996</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maine</td>
<td>4</td>
<td>NAEP</td>
<td>1</td>
<td>0.03</td>
<td>0.36* (0.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MEA</td>
<td>60</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>Maine</td>
<td>8</td>
<td>NAEP</td>
<td>5</td>
<td>0.16</td>
<td>0.18* (0.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MEA</td>
<td>45</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Kentucky</td>
<td>4</td>
<td>NAEP</td>
<td>3</td>
<td>0.24</td>
<td>0.89* (0.06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KIRIS</td>
<td>9</td>
<td>1.13</td>
<td></td>
</tr>
<tr>
<td>Kentucky</td>
<td>8</td>
<td>NAEP</td>
<td>2</td>
<td>0.16</td>
<td>0.80* (0.07)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KIRIS</td>
<td>15</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>Fixed-effects average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.31* (0.01)</td>
</tr>
<tr>
<td>Random-effects average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.55* (0.17)</td>
</tr>
<tr>
<td>2000–2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maine</td>
<td>4</td>
<td>NAEP</td>
<td>11</td>
<td>0.54</td>
<td>0.28* (0.04)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MEA</td>
<td>5</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>Maine</td>
<td>8</td>
<td>NAEP</td>
<td>–1</td>
<td>–0.05</td>
<td>0.14* (0.04)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MEA</td>
<td>–3</td>
<td>–0.19</td>
<td></td>
</tr>
<tr>
<td>Kentucky</td>
<td>4/5</td>
<td>NAEP</td>
<td>5</td>
<td>0.32</td>
<td>0.01 (0.06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KCCT</td>
<td>7</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Kentucky</td>
<td>8</td>
<td>NAEP</td>
<td>4</td>
<td>0.23</td>
<td>0.06 (0.06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KCCT</td>
<td>6</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Fixed-effects average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.15* (0.02)</td>
</tr>
<tr>
<td>Random-effects average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.13* (0.06)</td>
</tr>
</tbody>
</table>

Note. NAEP = National Assessment of Educational Progress; MEA = Maine Educational Assessment; KIRIS = Kentucky Instructional Results Information System; KCCT = Kentucky Core Content Test.

NAEP–state difference is reported in the unit of logit (see Appendix). All of the gain measures represent changes in the percentage of students at or above Proficient, except for 1992–96 gains for Maine, which represent changes in the average scale scores. Asterisk indicates that the gain is statistically significant at the .05 level. Standard error of the difference is shown in parentheses.
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. A meta-analytic synthesis of the NAEP, MEA, and KIRIS assessment results identified inconsistent results across the achievement gain measures from two states. The weighted average of standardized differences between national and state assessment results on 1992–1996 math achievement gains were .31 and .55 for fixed-effects and random-effects models, respectively. This difference was statistically significant, and the effect size was moderate.

The lower part of Table 5 compares math achievement gains on the NAEP and state assessments between 2000 and 2003. The size of proficiency gain no longer remains larger on the state assessment than on the NAEP. In Kentucky, the NAEP–state assessment gap has narrowed significantly between the mid-1990s and early 2000s. In 2000–2003, Kentucky gains in the percentages of Proficient students in Grade 4/5 mathematics were 5 on the KCCT versus 7 on the NAEP. In Maine, the pattern of achievement gain even reversed on the NAEP and MEA; 2000–2003 gains in the percentages of Proficient students (meeting or exceeding the standards of Maine’s Learning Results) in Grade 8 mathematics were 5 on the MEA and 11 on the NAEP. The weighted average of standardized NAEP–state assessment difference in 2000–2003 math achievement gain across the two states at Grades 4/5 and 8 was small (.15 for the fixed-effects model and .13 for the random-effects model).

Causes of the Discrepancies Between NAEP and State Assessment Results

In the following sections, possible factors that might have affected the discrepancies between NAEP and state assessment results are discussed. Some of the internal factors that are more directly related to psychometric properties of the tests themselves are considered here; they include differences in test content and format, test difficulty and discrimination, performance standards, and standards-setting and test equating methods. Other factors that are external to the tests and can influence the interpretations and uses of test results are also considered. These external factors include differences in testing samples and stakes attached to the test results.

Differences in the definition of performance standards. The NAEP, Kentucky, and Maine assessments all employed four similar performance standards or achievement levels. However, the real issue is one of operational definitions. The definition of standards affects the level of cut scores associated with the standards (Jaeger & Mills, 2001). The differences between NAEP and
state performance results can be explained, in part, by comparing performance-level definitions by subject and grade. The NAEP has both grade-specific and subject-specific definitions of performance levels, whereas 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.

The following definitions of MEA and NAEP math proficiency levels give contrast; the fourth grade specific definition is shown for NAEP, whereas an across-grade definition is shown for the MEA. The NAEP definition of (Grade 4 specific) proficiency is as follows:

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. (Reese et al., 1997)

The MEA definition of (grade-free) proficiency is as follows:

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. (Maine Department of Education, 1996)

It is obvious that the NAEP has more clear and specific definitions of math proficiency than does the MEA. 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 nonroutine (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 nonroutineness of any math problem is a matter of degree and subject to personal judgment. 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.

**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 the Angoff method (National Center for Education Statistics, 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, or 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 judgment method was not reasonable and could yield misleading interpretations (see Pellegrino, Jones, & Mitchell, 1999; Shepard, Glaser, Linn, & Bohrnstedt, 1993; U.S. General Accounting Office, 1993). There were counterarguments to these criticisms, and they pointed out that there is no evidence in the literature to support the contention that the Angoff method requires an impossible cognitive task (see Cizek, 1993; Hambleton et al., 2000; Kane, 1995).

Echoing the earlier criticisms of the NAEP standard-setting process, the MEA Performance Level Guide (1994–1995) from the Maine Department of Education also criticizes the Angoff method as unrealistic and unreliable. It emphasizes the need for a different approach for the MEA in that it employs a totally open-response format (scored on a 0–4 scale). Thus, the MEA standard-setting process utilized a totally different method, the Body of Work method, which involved judges matching actual student work to the predetermined 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. Recently the MEA has used a combination of the Body of Work method (see Kingston, Kahl, Sweeney, & Bay, 2001) and the Contrasting Group method (see Livingston & Zieky, 1982); threshold scores resulting from the two methods were aggregated to obtain the minimum scores for each performance level (Measured Progress, 2003).

The Kentucky standard-setting process for KIRIS shares some common features with Maine’s. 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. As the Kentucky Board of Education adopted new standards for CATS in 2001, a new standards-setting process uses a synthesis of three different methods (Kentucky Department of Education, 2004), including the Contrasting Groups method, the Integrated Judgment method (see Jaeger & Mills, 2001), and the Bookmark method (see Mitzel, Lewis, Patz, & Green, 2001).

Studies show that different standard-setting methods yield inconsistent results (Jaeger, 1989). 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 in Maine and Kentucky. 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.

**Differences in test change and equating.** One factor that can influence the estimation of statewide student achievement gains is the changes made in the two assessments over time. Substantial changes in test content, format, and setting pose threats to making valid comparison of measures collected at different occasions, estimating achievement gains, and drawing inferences about academic progress. While changes in test content tend to be minimal for both national and state math assessments, the relative changes in test format and scoring standards were greater. 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 the multiple-choice items. Likewise, the MEA, which began as a combination of both multiple-choice and constructed-response questions, shifted entirely to constructed-response questions in 1995. This change was later reversed when both states switched to the new assessment forms in 1999 by using a combination of multiple-choice and constructed-response questions.

Less dramatic but notable changes have also been 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, & Jones, 1993). In 1996 NAEP testing, more than 50% of student assessment time was devoted to constructed-response questions. Although both national and state assessments shifted from multiple-choice items to more constructed-response questions, the extent of changes was greater in state assessments than in the NAEP between 1992 and 1996.

Both NAEP and state assessments face simultaneous goals of measuring trends in educational performance and providing information about student achievement on progressive curricular goals. The NAEP uses several procedures to maintain the stability required for measuring trends, while still introducing innovations. Reliable estimation of achievement gains depends on test equating. The NAEP, Kentucky, and Maine assessments all used equivalent scaling and equating methods based on a common set of items referred to as the anchor items. Item Response Theory models were employed across all three assessments for test equating (see Stocking & Lord, 1983). For example, equating for MEA used the anchor-test-nonequivalent-groups design, as described by Petersen, Kolen, & Hoover (1989); the anchor test consisted of a set of common items that were included in two successive test administrations, but the groups of students who took the tests were not equivalent (Measured Progress, 2003).

Nevertheless, there are differences between the NAEP and state assessments in their test-equating frequency and time intervals. NAEP equating was done directly between 1992 and 1996 because there was no equivalent assessment during the time interval. For MEA and KIRIS, which administered 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. The difference in test-equating processes affects the reliability of equating. 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 the 1992 and 1996 state assessment results should become more tenuous as a result of more drastic changes in the format of the test, as well as more frequent test administration and equating.4

4Because equating for the state assessment is done every year to link two adjacent-year assessments (e.g., 1995 and 1996), the comparison of test results from remote years (e.g., 1992 and 1996) is likely to become less reliable because of the accumulation of equating errors. In Kentucky, for example, the standard error of equating for linking 1995 and 1996 KIRIS Grade 4 and 8 math tests with 12 anchor items ranged from .05 to .10 on the IRT theta scale (Kentucky Department of Education, 1997). If we assume that the same amount of error occurs in every year’s equating, the link between 1992 and 1996 state assessment results should become more tenuous as a result of more frequent test equating, as well as more drastic changes in the format of test. However, the discrepancy in gain scores between NAEP and state assessments due to different testing and equating cycles may become of less concern, as NAEP has shifted to biennial test administration scheduling as a result of NCLB, and more frequent testing made it more similar to the state’s annual testing.
Differences in test difficulty and discrimination. Another factor that might influence the estimates of state performance based on national versus state assessments is test difficulty and discrimination. Test items that have a level of difficulty of about .5 maximize the discrimination between high- and low-achieving students. The difficulty of test items should vary among item types. The NAEP included a more varied combination of multiple-choice and constructed-response items, which produces a wider range of item difficulty, than did either the MEA or KIRIS. Specifically, the earlier state assessments, the 1996 MEA and KIRIS, which have relatively limited numbers of only constructed-response items, tend to be very hard, with narrow distributions of item difficulties (see Table 6). The lower item scores of state assessments indicate greater difficulty so that both the MEA and KIRIS may have been more difficult than the NAEP for both low- and high-performing students, most of the state test item scores are below 0.5. Consequently, the MEA and KIRIS were likely to produce smaller achievement gaps than the NAEP.

The distribution of test item difficulty changed significantly in the two states’ assessments when they later shifted to a new assessment form with a combination of multiple-choice and constructed-response questions. The average item difficulty index for the 2003 MEA Grade 8 math was .48 (.51 for multiple-choice and .39 for constructed-response). The average item difficulty of .48 is substantially different from the 1996 MEA Grade 8 math average item difficulty index of .35. The range of item difficulty also became wider. This change in the test difficulty distribution may be attributable to the addition of multiple-choice items, which tend to be relatively easier than constructed-response items.

A similar pattern of change is found in Kentucky. The item difficulty for the 2003 KCCT Grade 8 math ranged from .34 to .81 for multiple-choice items and from .51 to .78 for constructed-response items. This figure also contrasts with the 1996 KIRIS eighth-grade math item difficulty, which ranged from .21 to .60 for constructed-response items. This shows that the state test may have gotten relatively easier and the test also became more diversified in the format with a wider range of item difficulty. This change in the test difficulty does not necessarily imply that the states’ performance standards expected for all students also changed. Nevertheless, the change in the test format and difficulty made the state

\[1\] I want to acknowledge one reviewer’s comment that “presumably, traditional item difficulty and discrimination differences are taken into account by the IRT models when estimating ability. Thus, tests of different difficulties and discriminations would estimate abilities in the same way unless the composition of the tests are such that they measure different abilities.” However, the use of IRT methods in both NAEP and state assessments is by no means a guarantee of their scale equivalence. Even with IRT, the relative performance of the more and less able varies markedly with the absolute difficulty of the test used and the sector of the test-characteristic curve that is operational (Raven, 2005). Cautions should be taken in evaluating the size of achievement gaps from two tests of the same construct if they had different distributions of item difficulties.
<table>
<thead>
<tr>
<th>State/Assessment</th>
<th>Item Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.00–.10</td>
</tr>
<tr>
<td>Maine MEA</td>
<td>0</td>
</tr>
<tr>
<td>Maine NAEP</td>
<td>0</td>
</tr>
<tr>
<td>(0)</td>
<td>(2)</td>
</tr>
<tr>
<td>Kentucky KIRIS</td>
<td>0</td>
</tr>
<tr>
<td>Kentucky NAEP</td>
<td>1</td>
</tr>
<tr>
<td>(1)</td>
<td>(4)</td>
</tr>
</tbody>
</table>

Note. NAEP = National Assessment of Educational Progress; MEA = Maine Educational Assessment; KIRIS = Kentucky Instructional Results Information System. Only common items across test forms are available for the MEA. The total number of MEA test items is 30 and all are polytomously scored constructed-response items. All of the above KIRIS items are polytomously scored constructed-response items. Numbers in parentheses indicate the number of polytomously scored constructed-response items among all NAEP test items; the remainder include multiple-choice items and dichotomously scored constructed-response items. For dichotomously scored items (0, 1 scoring), the item difficulty index is the proportion of students who correctly answered each item (p); the lower index indicates the greater difficulty. For polytomously scored items, the difficulty index is adjusted by dividing its mean by the maximum number of points possible.
assessments more similar to NAEP, and it may have contributed to the convergence of the NAEP and state assessment results.

Minority students and students whose parents have less education or lower income performed significantly lower than their counterparts on both the NAEP and state assessments. 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 either multiple-choice or short constructed-response items at the upper end of the proficiency scale but less information at the lower end of the scale (see Dossey et al., 1993). Therefore, further investigation of the achievement gaps requires understanding how well the level of a given test’s difficulty matches the level of student proficiency for different groups of students.

Differences in testing sample. One factor to consider in understanding the comparability of achievement gaps among student groups is whether the demographic compositions of testing samples are equivalent. Because the 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 the exact same inclusion criteria for their testing and reporting.6

To determine if the student groups compared for the analysis of achievement gap have equal representations in the NAEP versus state testing samples, comparisons were made in the percentages of students as broken down by gender, race, income, and parental education (see Table 7). Both the NAEP and state assessments showed the same distributions of students by all grouping variables; some small differences were within the margin of sampling errors. For instance,

---

6Exclusion rates (including students with disabilities and Limited English Proficiency (LEP) students) were relatively smaller in states’ own assessments than in NAEP. In 1996, the percentage of students excluded from the NAEP assessment in Kentucky was 6 for Grade 4 and 3 for Grade 8, whereas the exclusion rate for KIRIS was about 1 for Grade 4 and Grade 8 (see Reese et al., 1997 for NAEP; Kentucky Department of Education, 1997 for KIRIS). The percentage excluded from the NAEP assessment in Maine was 9 for Grade 4 and 6 for Grade 8, which was higher than corresponding exclusion rates in MEA (see Reese et al., 1997 for NAEP; communication with state assessment director for MEA, B. Maxcy, January 7, 2005).
there was a 2% difference between the assessments for 1996 race variable (11% minority for KIRIS vs. 13% minority for NAEP) in Kentucky, but the numbers were virtually identical considering the NAEP sampling and its standard error for percentage estimate. Any sampling bias that can arise from overrepresentation or underrepresentation of particular groups of students in the population should be considered, particularly when those groups were not separately sampled through stratification.

In tracking and evaluating academic growth statewide over time, the study used a successive cohort group comparison method for both the NAEP and state assessments, that is, comparing the achievement of students at selected grades with that of cohorts of students from previous years at the same grade level. This successive cohort comparison method is used widely in current state accountability systems, and it rests on the implicit assumption that student characteristics are relatively stable from year to year for the students attending a given school system (Linn, 2001). As Table 7 shows, there has been very little change between

### TABLE 7
Percentages of Maine and Kentucky Eighth Graders by Demographic Characteristics in 1996 and 2003 NAEP vs. State Math Assessment Samples

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>Maine MEA</th>
<th>NAEP (%)</th>
<th>Kentucky KIRIS</th>
<th>NAEP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>Female</td>
<td>51</td>
<td>50 (1.14)</td>
<td>49</td>
<td>49 (.96)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>49</td>
<td>50 (1.14)</td>
<td>51</td>
<td>51 (.96)</td>
</tr>
<tr>
<td>Parental education</td>
<td>High</td>
<td>95</td>
<td>95 (.54)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>5</td>
<td>5 (.54)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td>White</td>
<td>NA</td>
<td>89</td>
<td>87 (1.03)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minority</td>
<td>NA</td>
<td>11</td>
<td>13 (1.03)</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>Female</td>
<td>49</td>
<td>50 (1.0 )</td>
<td>48</td>
<td>50 (1.0 )</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>51</td>
<td>50 (1.0 )</td>
<td>51</td>
<td>50 (1.0 )</td>
</tr>
<tr>
<td>Parental education</td>
<td>High</td>
<td>95</td>
<td>96 (1.7 )</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>5</td>
<td>4 (.4 )</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Free/reduced lunch</td>
<td>Eligible</td>
<td>NA</td>
<td>46</td>
<td>43 (1.5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ineligible</td>
<td>NA</td>
<td>54</td>
<td>57 (1.8)</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td>White</td>
<td>NA</td>
<td>86</td>
<td>88 ( .9 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minority</td>
<td>NA</td>
<td>14</td>
<td>12 (.7 )</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* NAEP = National Assessment of Educational Progress; MEA = Maine Educational Assessment; KIRIS = Kentucky Instructional Results Information System; KCCT = Kentucky Core Content Test.

NA: Student background information for this variable is not available in the state assessment data. Standard errors of the NAEP sample percentage estimate appear in parentheses.
1996 and 2003 in terms of the demographic composition profiles of the two eighth grade cohorts in Kentucky and Maine. The between-cohort sample comparability does not appear to be an issue at the state level (at least in the two states), but it still can be a serious threat to an accountability system at the school level. However, the issue may become of less or no concern if the states shift to a quasi-longitudinal or longitudinal comparison method for their school accountability as a result of the NCLB mandate of states testing students at every grade from Grade 3 through Grade 8 since 2006.

Differences in high-stakes testing environment. It is difficult to quantify how high the stakes of testing were and how much influence they 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. 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 incentives to perform well on the state tests.

The states that have high-stakes testing and a strong accountability system would exert greater pressure for schools and students to improve their achievement on the state test. Therefore, the high-stakes testing environment may result in possible inflation of proficiency level. The comparison of the NAEP–state assessment discrepancies in proficiency level between Maine and Kentucky suggests that states with a high-stakes testing (e.g., Kentucky) may have obtained higher proficiency level on their own state assessment relative to NAEP than states with a low-stakes testing (e.g., Maine).

In a similar vein, the greater achievement gains during the 1990s in Kentucky and Maine based on their state assessments might be also related to the impact of state assessments on school curriculum and instructional practices due to the stakes attached to the state test results. Although there can be many other possible reasons for overstated or understated achievement gains (Wise & Hoffman, 2002), high-stakes testing can lead to inflation of academic progress.

Given the high stakes attached to the KIRIS and modest stakes attached to the MEA, it is likely that state assessments, particularly Kentucky’s, show much greater improvement than the national assessment reveals. Indeed, the size of inflated achievement gains relative to NAEP was greater in Kentucky than in Maine. The KIRIS technical manual noted that Kentucky students achieved gains
on both the NAEP and KIRIS, pointing out that “As long as each measure provides an indication of whether changes over time are statistically significant, it is possible to compare trends broadly” (Kentucky Department of Education, 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.

Why did the NAEP–state assessment discrepancy in student achievement gains almost disappear in both states during the early 2000s? Although it is premature to tell whether this change reflects a new trend, one of the possible reasons might be NCLB’s impact on reducing the difference between state assessment and the NAEP. States’ own accountability policies are no longer the only source of high-stakes testing. The role of NAEP is changing as a result of the NCLB. The law requires each state to participate in biennial state assessments of fourth- and eighth-grade reading and mathematics under the NAEP. Similarly, the NCLB requires each Local Education Agency (LEA) to participate, if selected, in the State NAEP. The law does not explicitly require that results from the NAEP be used as evidence to confirm progress on state tests, but its mandate that all 50 states now take part in the NAEP makes such comparisons more likely (see Fuller, Gesicki, Kang, & Wright, 2006; Lee, 2006). Although the NAEP may be used as a tool for the U.S. Department of Education to cross-check and validate state-level or district-level academic progress, increased stakes attached to NAEP results may affect the utility of the NAEP as a confirmatory tool.

**CONCLUSION**

Using student assessments for statewide education accountability purposes requires us to investigate the adequacy and utility of the currently available assessments. Raising questions about the validity of using any single assessment alone for accountability, this study examined the cases of Kentucky and Maine regarding three key indicators of academic performance: students’ proficiency levels, achievement gaps, and achievement gains. Although the study is highly limited in generalizing its findings, with restricted data from only two states in a single subject area, the application of meta-analytic methods to repeated NAEP and state assessment measures sheds light on subsequent studies. Table 8 summarizes the key findings.

First, the large discrepancies between NAEP and state assessments in the proficiency levels may be explained by the fact that the definitions of performance standards and the methods of standard setting were different. There were some similarities between the categories of achievement levels in the NAEP and the corresponding categories in the state assessments. These similarities, however, could mislead one to believe that those two states’ assessment standards are
### TABLE 8
Summary of Discrepancies and Convergence Between NAEP and State Fourth-/Eighth-Grade Math Assessment Results in Maine and Kentucky

<table>
<thead>
<tr>
<th>Academic Performance Indicator</th>
<th>Status and Trend of Discrepancies Between NAEP and State Assessment Results</th>
<th>Plausible Reasons for Observed Differences and Convergence Between NAEP and State Assessment Results</th>
</tr>
</thead>
</table>
| Proficiency levels            | The percentages of students performing at or above Proficient level on state assessments were significantly different from NAEP. These differences were inconsistent between grades and states. Kentucky had a relatively lower performance standard than NAEP, whereas Maine had a higher standard. While the discrepancies have narrowed in both states, they remain large. | • Definition of standards: Despite the same or similar labels of standards, states’ own performance standards were set at different levels than the NAEP performance standard.  
• Standards-setting methods: NAEP and state assessments used different approaches to setting cut points for achievement levels.  
• Testing environment: States with high-stakes testing could exert greater pressure and result in possible inflation of proficiency level. |
| Achievement gaps              | The achievement gaps among racial and social groups were slightly smaller on the state assessments than on the NAEP. The size of this difference between NAEP and state assessments varied among the type of groups compared. The discrepancies in the achievement gaps between the NAEP and state assessments remain small. | • Testing sample: The NAEP sample was equivalent to the state assessment population (in terms of the percentages of different racial and social groups).  
• Test item characteristics: State assessments had narrower distributions of item difficulties and thus less discriminating power than the NAEP. State assessments became more similar to the NAEP, which may have contributed to the convergence of the NAEP and state assessment results. |
| Achievement gains             | The achievement gains were substantially larger on the state assessments than on the NAEP. The size of inflated achievement gains relative to NAEP was greater in Kentucky than in Maine. The differences in achievement gains between the NAEP and state assessments have been reduced in both states. | • Test change and equating: State assessments went through greater changes in test format and more frequent test equating than the NAEP.  
• Testing environment: State assessments had higher test stakes than the NAEP. The high-stakes testing in Kentucky may account for its relatively larger gains on the NAEP than on state assessment. The emerging role of the NAEP as a confirmatory tool for the NCLB accountability could have influenced the NAEP results as well. |

*Note.* NAEP = National Assessment of Educational Progress; NCLB = No Child Left Behind Act of 2001.
consistent with national standards and that the MEA and KIRIS cut points for mathematics proficiency are as high as those of NAEP. Although the NAEP–state assessment gap in proficiency rate dropped to some extent between the mid-1990s and early 2000s, it still remains large enough to be a legitimate concern for cross-validation.

Second, the relatively small discrepancies between the NAEP and state assessments in the achievement gaps may be attributed to the fact that the NAEP testing sample is well representative of the target state testing population in terms of the composition of students. However, there was also some tendency for the size of achievement gaps to be smaller on states’ own assessments than on the NAEP. This may be related to the finding that the NAEP assessment had more discrimination power than the state assessments because the former had a wider distribution of test item difficulties than the latter. This pattern changed in the early 2000s as both states’ assessments became more similar to the NAEP by shifting from the old assessment using only relatively hard constructed-response questions to the new assessment of mixed difficulty with a combination of multiple-choice and constructed-response questions.

Third, the large discrepancies in the sizes of achievement gains may be explained by the fact that the states’ own assessments had greater consequences for schools and thus greater impacts on curriculum and instruction than did the national assessment without stakes attached to the results. Questions are also raised about the reliability of achievement gains on the state assessments, which went through greater changes in testing content and formats and more frequent test equating across the assessments. Although the NAEP–state assessment gaps narrowed substantially during the early 2000s, it remains to be known whether this change is temporary and whether the change is indicative of an increasing convergence of state assessments and NAEP.

This study pinpoints the areas of consistency and inconsistency in the NAEP and state assessment results on state education system performance. Some of the differences between the NAEP and state assessment results may have occurred by design (e.g., the rigor of performance standards, the level of testing stakes), whereas other differences were not planned in advance and cannot be explained systematically (e.g., the degree of changes made in assessment format, the demographic composition of examinees). More in-depth studies are needed in order to identify what has caused the discrepancies between national and state assessment results and what contributed to convergence of NAEP and state assessment results over time. The findings suggest that policy-makers and educators should become more aware of the utilities and limitations of current national and state assessments to produce valid academic performance indicators for accountability.

It remains to be known whether there is a similar converging trend of the NAEP versus state assessment results in other states as well, and whether these changes, if generalizable across the nation, are related to NCLB. The results
might have been different if the data had been collected from all states, which could impact patterns of the discrepancies between NAEP and state assessments. Subsequent studies also need to conduct a closer examination of changes in state assessments over the longer run as a result of NCLB. On one hand, NCLB could make the states augment their existing student assessment systems and align them much closer with NAEP, which has gained greater importance as a tool for the federal government to check and confirm state results. On the other hand, this increasing use of NAEP measures for cross-state comparison or verification purposes also increases the risk of reducing policy goals to the improvement of NAEP test scores and contaminating the results from the NAEP assessment.

ACKNOWLEDGMENTS

This article is based on work supported by the National Science Foundation under Grant 9970853. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. I am very grateful to anonymous reviewers and the editor for their invaluable comments and suggestions on an earlier version of this article.

REFERENCES


As shown in the following statement in the Kentucky Department of Education’s NCLB accountability workbook (Kentucky Department of Education, 2005), it appears that the state education agencies have become increasingly aware of the role of NAEP and the implications of potential NAEP–state assessment gaps:

While it is clearly stated in federal statute that states will not be rewarded or punished based on state NAEP, NAEP data will be a component considered in the validation of the results of state assessments (both at a single point in time and changes over time). Both the educational community and the public at large will use NAEP in this way. It will be most important to understand the relationships between the NAEP curriculum frameworks and Kentucky’s Core Content for Assessment. NAEP will become a more visible assessment component at the national and state levels (p. 61).


APPENDIX
Meta-Analysis of the Discrepancies Between NAEP and State Assessment Results

First, the sample estimator of the standardized state performance indicator and its sampling variance was estimated for each sample by using the formula in Table A below. Logit was used as the metric of an effect size for achievement level (proficiency rate) and achievement gain (gain in proficiency rate). Cohen’s $d$ was used as the metric of effect sizes for achievement gap (average scale score difference between student subgroups).

Second, the differences between national and state assessment results on performance indicators are obtained for each sample by calculating an absolute value of the difference between a pair of standardized performance measures: $d = |z_1 - z_2|$, where $z_1$ is the effect size estimate from NAEP and $z_2$ is its counterpart from the state assessment (MEA or KIRIS/KCCT).

Standard error of the NAEP–state assessment difference is obtained by taking the square root of the sampling variance of the difference for dependent means:

$$s^*_{z_1 - z_2} = \sqrt{s^2_{z_1} + s^2_{z_2} - 2r_{12}s_{z_1}s_{z_2},}$$

where $r_{12}$ is the sample correlation between NAEP and state assessment scores (estimated from the correlation between matched school average scores with correction for the restriction of range).

<table>
<thead>
<tr>
<th>Effect Size Estimate</th>
<th>Definition</th>
<th>Sample Estimator ($z$)</th>
<th>Estimated Variance ($s^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit</td>
<td>Difference between population proportions</td>
<td>$\ln \frac{P_1/(1 - P_1)}{- \ln [P_2/(1 - P_2)]}$</td>
<td>$\frac{1}{P_1(1 - P_1)n_1} + \frac{1}{P_2(1 - P_2)n_2}$</td>
</tr>
<tr>
<td>Cohen’s $d$</td>
<td>Difference between population means</td>
<td>$(M_1 - M_2)/\sigma$</td>
<td>$\frac{n_1 + n_2}{n_1n_2} + \frac{d^2}{2(n_1 + n_2)}$</td>
</tr>
</tbody>
</table>
Third, the average NAEP–state assessment difference is calculated for each performance indicator by averaging the above standardized measures of differences between national and state assessment results with the inverse of sampling variance as weight. The formulas for this weighted average and its variance under fixed-effects and random-effects models are as follows:

1. Fixed-effects model:

Weighted average

\[(D_F) = \frac{\sum V_j^{-1} d_j}{\sum V_j^{-1}},\]

where \(d_j\) is the standardized mean difference between NAEP and state assessment measures of statewide performance from a sample \(j\) and \(V_j\) is the variance of \(d_j\) (within-sample variance).

Standard error of \(D_F = \sqrt{V.}\) where \(V. = 1/\sum V_j^{-1}\)

2. Random-effects model:

Weighted average \((D_R) = \frac{\sum V^*_j^{-1} d_j}{\sum V^*_j^{-1}},\)

where \(d_j\) is a standardized average of the difference between NAEP and state assessment measures of statewide public school student performance from a sample \(j\) and \(V^*_j\) is the variance of \(d_j\). \(V^*_j = V_j\) (within-sample variance; estimation variance) + \(\tau\) (between-sample variance; random-effects variance).

Standard error of \(D_R = \sqrt{V^*.}\), where \(V^*. = 1/\sum V^*_{j}^{-1}\)