DOCUMENT RESUME

ED 468 074 TM 034 389

AUTHOR Lee, Jaekyung; Coladarci, Theodore
TITLE Using Multiple Measures To Evaluate the Performance of Students and Schools: Learning from the Cases of Kentucky and Maine. Research Report. Statewide Systemic Initiatives (SSI) Study.
INSTITUTION Maine Univ., Orono.
SPONS AGENCY National Science Foundation, Arlington, VA.
REPORT NO SSI-RR-2
PUB DATE 2002-07-00
NOTE 42p.; For Research Report Number 1, see TM 034 388.
CONTRACT REC-9970853
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Achievement Gains; Case Studies; Elementary Secondary Education; *Evaluation Methods; Performance Factors; Reliability; *State Programs; Student Characteristics; Testing Programs
IDENTIFIERS International Assessment of Educational Progress; *Kentucky; Kentucky Instructional Results Information System; *Maine; Maine Educational Assessment; Statewide Systemic Initiative

ABSTRACT

This report is the product of the first year of study of the Statewide Systemic Initiatives (SSI) Study on exploring data and methods to assess and understand the performance of states participating in the SSI. The report describes three sets of studies. The first studied valid and reliable measures of student achievement, and how to combine multiple measures of achievement, as well as how results of certification based on a single-measure compare with those based on a multimeasure analysis. The study addressed these questions through analyses of state and local data from two sites in Maine. The second study focused on valid and reliable ways to estimate school performance and identify value-added school effects on student achievement, as well as how the results of school evaluation based on a unilevel analysis compare with those based on a multilevel analysis. The study addressed these questions through multilevel analyses of National Assessment of Educational Progress assessment data in Maine and Kentucky. Results suggest that school-level compositional effects and student-level background characteristics need to be taken into account for fair evaluation of school performance. The third study explored ways to estimate academic progress and how to identify true achievement gains without statistical artifacts. These questions were addressed through the analysis of Maine Educational Assessment and Kentucky Instructional Results Information System assessment data. Results show that to set realistic expectations about schools' academic progress, past trends need to be taken into account, and that combining multiple years of data improves the reliability of school performance measures. (Contains 24 tables and 28 references.) (SLD)

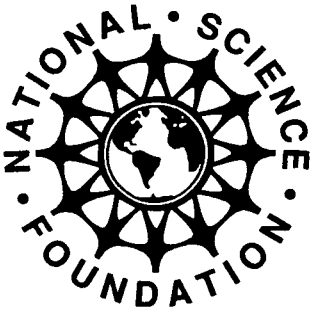
Research Report No. 2
Statewide Systemic Initiatives (SSI) Study

Using Multiple Measures to Evaluate the Performance of Students and Schools: Learning from the Cases of Kentucky and Maine

Jaekyung Lee, Ph.D.
Theodore Coladarci, Ph.D.
University of Maine

July 2002

Sponsored by the National Science Foundation



Research Report No. 2
Statewide Systemic Initiatives (SSI) Study

Using Multiple Measures to Evaluate the Performance of Students and Schools: Learning from the Cases of Kentucky and Maine

by

Jaekyung Lee, Ph.D.
Theodore Coladarci, Ph.D.
University of Maine

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

July 2002

NSF Grant No. REC-9970853

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

July 2002

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

Table of Contents

Preface

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

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

Research Report No. 2 is the product of our first-year SSI research study project, "Exploring Data and Methods to Assess and Understand the Performance of SSI States." During our second project year, we focused on the second research question and conducted a series of studies to address the issues arising from evaluation of student and school performance with multi-dimensional data. First, we examined ways to analyze and combine multiple measures of student achievement including state and local assessment measures. Second, we examined ways to analyze multi-level student achievement data (students nested within schools) and identify the value-added contributions of schools to student learning. Finally, we examined ways to analyze aggregate time-series school performance data and evaluate schools' academic progress over time.

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

Summary

This report consists of three sets of studies. Highlights from each of the studies are summarized below.

Study I

What counts as valid and reliable measures of student achievement? How can we combine multiple measures of student achievement for certification? How do the results of certification based on a single-measure compare with the results from a multi-measure analysis method? Study I addresses these questions through the analyses of state and local assessment measures with data from two sites in Maine. Major research findings from Study I are as follows:

1. Using multiple measures of student achievement for certification requires careful examination and choice of assessments. Assessments should be constructed explicitly to reflect state content and performance standards.
2. It may not be necessary to use multiple measures when measures are highly intercorrelated. However, high intercorrelations among measures are not sufficient for deciding in favor of a single measure. The quality of each measure and its potential consequences on teaching/learning should be considered.
3. Our research suggests that the results of certification vary with the methods of combining multiple measures. A compensatory method with an appropriate weighting scheme may reduce misclassification error.
4. Our research points to the possible hazards of certification based on a single measure (either state or local assessment). Without the assurance of alternative assessments, we must interpret with caution the tendency of the single-measure classifications to underidentify or overidentify students who meet the standard.
5. Using course grades for certification can be problematic as they often include non-achievement information and lack common standards for grading. Nevertheless, standards-based alignment/benchmarking efforts across the state can make classroom assessments more objective and their results comparable.

Study II

What are the valid and reliable ways to estimate school performance? How can we identify value-added school effects on student achievement? How do the results of school evaluation based on a uni-level analysis compare with the results from a multi-level analysis method? Study II addresses these questions through multi-level analyses of NAEP assessment data in Maine and Kentucky. Major research findings from Study II are as follows:

1. Decomposition of variance in the outcome variable shows that the two states have similar distributions of mathematics achievement between the school and student levels. In both states, about 18% of variance exists at the school level and 82% at the student level.
2. Higher performing schools in both states tend to have smaller gaps with regard to one background variable but larger gaps with regard to the other. This indicates that schools are not very effective in addressing both racial and social achievement gaps.
3. Maine schools perform significantly better than Kentucky schools, even after controlling for student and school background characteristics. Despite the average school performance gap, there are no significant differences between the two states' schools in terms of their racial and social gap estimates.

4. Our results suggest that school-level compositional effects (comparing "like with like") as well as student-level background characteristics need to be taken into account for fair evaluation of school performance. However, this can be problematic where there is systematic covariation between school context and school practice variables. Thus, the estimation of school effects requires that we define "school effects" and formulate an explicit model of these effects.
5. Our analysis of school effects involved racial and social achievement gaps as well as average achievement. However, in our cases, much of the observed variability in achievement gaps was sampling variance and, as a result, could not be used as school effect indices.

Study III

What are the valid and reliable ways to estimate schools' academic progress? How can we identify true achievement gains without statistical artifacts? How do the results of school evaluation based on a short-term gains compare with the results from a long-term trend analysis method? Study III addresses these questions through analyses of MEA and KIRIS assessment data in Maine and Kentucky. Major research findings from Study III are as follows:

1. Standardized gain per year is .12 in Maine and .47 in Kentucky in the 1990s. In order to set realistic expectations about schools' academic progress, the past school performance trend needs to be taken into account.
2. Correlations among school average scores for adjacent years are modestly positive (+.40s to +.50s). The correlations become weaker for remote years (proximal autocorrelation). Combining multiple years of data would improve the reliability of school performance measures.
3. Correlations among gain scores with no overlapping years are too small to be significant. Schools that gain more in the current year do not necessarily gain more or less the following years. Despite the lack of stability in school gains based on the successive group comparison method, three-year gain provides a closer estimate of the long-term school growth trend than does one- or two-year gain.
4. Correlations between initial status and gain scores are modestly negative (-.40s to -.50s). This suggests a regression to the mean "status" artifact, by which lower performing schools tend to gain more than higher performing schools. This artifact should be considered in evaluating school progress.
5. Correlations among gain scores for adjacent periods are negative (-.30s to -.50s). Schools that gained more in the previous period tend to gain less in the current period. This suggests a regression to the mean "growth" artifact. This regression artifact also needs to be considered in evaluating school progress.

I. Introduction

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

Methodological challenges are posed by multiple measures as we seek to evaluate student and school performance in more valid and fair ways. There are at least three dimesions of multiple measures to consider: 1) the assessment dimension addresses the type of assessment instrument used to measure student achievement; 2) the level dimension looks at the level at which school system academic achievement is measured and reported; 3) the time dimension concerns when and how many times achievement is measured. Do currently used assessments (i.e., national, state, district, school/classroom) produce consistent achievement measures in these three dimensions? What are the appropriate methods for analyzing such multi-dimensional academic achievement data?

In light of these concerns, we conducted a systematic analysis of student assessment data from Maine and Kentucky, i.e., the National Assessment of Educational Progress (NAEP), Maine Educational Assessment (MEA), Kentucky Instructional Results Information System (KIRIS), Iowa Test of Basic Skills (ITBS), Terra Nova (TN), and classroom assessments, to address the issues of evaluating systemic school reform with multiple measures. We conducted a series of studies to address the issues arising from the evaluation of student and school performance with multi-dimensional data.

In Study I, we examined ways to analyze and combine multiple measures of student achievement including state and local assessment measures. In Study II, we examined ways to analyze multi-level student achievement data (students nested within schools) and to identify the value-added contributions of schools to student learning. In Study III, we examined ways to analyze aggregate time-series school performance data and to evaluate schools' academic progress over time. Although our results arguably do not generalize to all states, they are expected to inform us about desired data and methods for a more systematic and comprehensive evaluation of systemic school reform. Table 1 summarizes the overall research design.

Table 1. Overview of Research Design for Studies I, II, and III

	Study I	Study II	Study III
Research Questions	1. What counts as valid and reliable measures of student achievement? 2. How can we combine multiple measures of student achievement for certification? 3. How do the results of certification based on a single-measure compare with the results from a multi-measure analysis method?	1. What are the valid and reliable ways to estimate the status of school performance? 2. How can we identify value-added school effects on student achievement? 3. How do the results of school evaluation based on a uni-level analysis compare with the results from a multi-level analysis method?	1. What are the valid and reliable ways to estimate schools' academic progress? 2. How can we identify true achievement gains without statistical artifacts? 3. How do the results of school evaluation based on a short-term gains compare with the results from a long-term trend analysis method?
Data Sources	1999 MEA 8th grade math student test score data, 1999/2000 ITBS or TN 8th/9th grade math student test score data, 1999/2000 8th/9th grade math course grades from two sites in Maine	1996 NAEP state assessment 8th grade math student and school data in Maine and Kentucky	1990-1998 MEA 8th grade math school average performance data in Maine and 1993-1998 KIRIS 8th grade math school average performance data in Kentucky
Analytic Methods	Descriptive analysis Correlation/Regression analysis Classification analysis	Descriptive analysis Correlation analysis Multilevel (HLM) analysis	Descriptive analysis Correlation/Regression analysis Trend analysis

II. Study I: Evaluation of Student Achievement with Multi-Assessment Data

A number of state and federal agencies either recommend or require multiple measures to assess student achievement (Ardovino, Hollingsworth, & Ybarra, 2000). However, no criteria about reliability, validity, and weighting in using multiple measures have been set by the governments. Two models are commonly used to combine multiple measures: conjunctive and compensatory (NRC, 1999). Currently available measures of student achievement are often inadequate for evaluation of standards-based school reform, particularly when they rely on norm-referenced standardized tests and use percentile ranks as grade level standards. While local assessments are a potentially valuable source of additional measures, there is often insufficient consistency of the measures across sites. Despite these problems and challenges, districts have devised their own ways to combine multiple measures of achievement, which produces a great deal of variation from district to district (see Jang, 1998; Kalls, 1998; Law, 1998; Novak, Winters, & Flores, 2000).

In the present climate of standards-based education, school leaders in Maine also are being asked to think about assessment in new ways. Student achievement of the state standards, the Learning Results, must be measured by a combination of state and local assessments. Based on these assessments, local educators soon will be expected to "certify" a student's attainment of the Learning Results in order for the student to receive a high school diploma. This site-based, multiple-measures approach to certification is different from other states that rely on a single standardized test for high school graduation.

How should we approach the challenge of combining multiple measures of achievement for arriving at a single judgment of, say, "proficiency" or "meeting the standard"? Specifically, what is an efficient and defensible method for combining multiple measures of achievement? This is the general question that we address in this section.

Data Collection and Analysis

We collected data from two sites in Maine, chosen because of their similarity in community size and proximity to the University of Maine. In both sites, we obtained the following achievement information for each student: (a) the mathematics subscale score on the 8th grade MEA (MEA-M), (b) the mathematics subscale score on the locally administered standardized achievement test (ITBS in Site A and TN in Site B), and (c) the course grade achieved in mathematics. In Site A ($n = 94$), all information was taken in the student's 8th grade year; in Site B ($n = 65$), the standardized achievement test and mathematics grades were obtained in the 9th grade (see Table 2). The MEA-M scores provide the only truly meaningful achievement information for comparing the two sites. From Table 3, one sees that the MEA-M mean for Site B was 17.76 points higher than that for Site A. With a pooled within-group standard deviation of 15.77, this mean difference corresponds to an effect size of $d = 17.76 \div 15.77 = +1.13$.

Creating a Common Scale for Mathematics Course Grade

As can be seen from Table 2, students in each site did not all enroll in the same level of mathematics. Our first task, then, was to create a single variable for "mathematics grade," even though it would comprise grades from different classes. Although we followed the same procedure in both sites, we will illustrate this procedure using data from Site A.

Table 2.When achievement information was collected, by site.

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

Table 3. Distribution of MEA-M mathematics scores in each site.

	MEA-M performance	
course	M	SD
Site A (n = 94)	522.49	14.88
Site B (n = 65)	540.25	16.97
		SDpooled = 15.77

Site A students received a grade, on a 100-point scale, for either general mathematics (n = 59), algebra 1 (n = 29), or geometry (n = 6) (see Table 3). Because we believe that it makes little sense to regard a final grade in general mathematics as being comparable to the same grade in a higher level class, we weighted algebra 1 and geometry grades according to how these two groups of students performed on the MEA-M relative to the general mathematics students (see Table 5). Each of the two mean differences was converted to an effect size:

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

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

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

Analyses and Results

Table 4. Distribution of unweighted mathematics grades for each of three courses (Site A).

course	M	SD
general mathematics (n = 59)	78.24	9.26
algebra 1 (n = 29)	88.17	6.58
geometry (n = 6)	94.33	4.50
		SDpooled = 8.31

Table 5. Distribution of MEA-M mathematics scores for students in each of three mathematics courses (Site A).

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

Correlational Analyses

To examine the relationships among the results of state and local assessments, we obtained student-level within-site correlations among the three measures of student achievement: (a) MEA-M, (b) the mathematics subscale score on the locally administered standardized achievement test (which we refer to as "ITBS/TN"), and (c) the weighted course grade achieved in mathematics ("COURSE").

As Table 6 shows, the three measures of mathematics achievement correlate substantially. Although these correlations are uniformly high, there is some variation in magnitude. Interestingly, COURSE correlates more highly with MEA-M than with ITBS/TN. This is not surprising, insofar as one would expect classroom assessments and the MEA to align with the Learning Results more than would be expected of a commercially available standardized test.

Classification Analyses

Table 6. Correlations among measures of student achievement in mathematics.

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

To explore an efficient and defensible method for combining multiple measures of achievement, we combined the three measures two different ways and compared the results by conducting classification analyses. The two ways of combining measures differed in weighting scheme, but both were based on a compensatory model that allows one's relatively higher performance on one measure to compensate for his or her relatively lower performance on another measure. As with the correlational analyses, these analyses were conducted within site.

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

The cutscores are as follows:

exceeds the standard: 561

meets the standard: 541

partially meets the standard: 521

does not meet the standard: <521

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

Although Maine school leaders soon will be expected to engage in standard setting for their local assessments, the two sites in the present study, like most Maine school districts, have yet to implement standard setting. Consequently, neither COURSE nor ITBS/TN can be directly expressed as a performance level within the context of the Learning Results. However, because MEA-M correlates highly with both ITBS/TN and COURSE (Table 6), we can estimate, using simple regression, the critical cutscore for each of the latter two measures. We began by regressing ITBS/TN on MEA-M and, given the resulting equation, determined the predicted value of ITBS/TN for MEA-M = 541 (i.e., the designated cutscore for "meets the standard"). In Site A, for example, this regression equation is:

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

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

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

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

We then transformed MEA-M, ITBS/TN, and COURSE to z-scores using the standard formula, but with one modification: We replaced the mean with 541 in the transformed MEA-M variable and the estimated cutscore (as described above) in the transformed COURSE and ITBS/TN variables. With this substitu-

tion, the sign of a z-score now indicates the student's performance relative to the MEA-M standard (rather than to the parent variable's mean).

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

The fundamental question is whether the unweighted and weighted composites classified students similarly. That is, when forming an achievement composite, is anything gained by weighting the measures that enter into the composite? As Table 7 shows, there was perfect agreement between the two sets of classifications. This no doubt reflects the relatively uniform correlations among MEA-M, ITBS/TN, and COURSE (Table 6) and, in turn, the relatively uniform component score coefficients that we obtained from the principal components analysis (see Table 8). In short, the results of this analysis indicate that weighting each measure is unnecessary. Thus, if the choice is between weighting or not weighting, the most efficient strategy for combining multiple measures would appear to be the latter. This assumes that correlations among measures are similar (which should be examined empirically) and that the measures are of equal importance. If either assumption does not hold, then weighting would be defensible.

Table 7. Classification similarity: unweighted and weighted composites.

Site A

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

Site B

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

Table 8. Component score coefficients.

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

Table 9. Classification similarity: Unweighted composite and MEA-M.

Site A
(100% agreement)

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

Site B
(92% agreement)

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

Table 10. Classification similarity: Unweighted composite and ITBS/TN.

Site A
(91% agreement)

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

Site B
(91% agreement)

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

Table 11. Classification similarity: Unweighted composite and COURSE.

Site A
(91% agreement)

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

Site B
(91% agreement)

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

A secondary question concerns the level of agreement between the classification based on the unweighted composite and that based on a single measure (see Tables 9-11). Except for the perfect agreement in Site A involving MEA-M, the levels of agreement are fairly consistent, ranging from 89% to 92%. In these later cases, single-measure classification resulted in more students meeting the standard than when classification was based on the composite.

Discussion

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

Having said this, we should acknowledge that high intercorrelations among measures are not sufficient for deciding in favor of an unweighted composite. That is, one also should take into account the announced importance of each measure. For example, if a school district attaches greater importance to a district-wide assessment compared to, say, the standardized test that is annually administered, then the former should receive greater weight—even in the face of a high correlation between the two. Although there are various reasons why local achievement measures may differ in importance, a primary reason is the degree to which a measure aligns—in various respects (e.g., see Webb, 1997)—with the adopted standards. The reliability of assessments also needs to be considered in developing weights.

Our results also point to the possible hazards of classifying student achievement based on a single measure. As Tables 9-11 illustrate, single-measure classification tended to result in additional students identified as meeting the standard. Are these students false positives? Because of two limitations of the present study, we unfortunately do not know. First, unlike MEA-M, which was designed to align with the Learning Results, neither ITBS/TN nor COURSE was constructed explicitly to reflect student attainment of these standards. This clearly is true for ITBS/TN, for no commercially available standardized achievement test is tailored to the standards of any one state (particularly a small state). And although teacher-constructed mathematics assessments (COURSE) in Maine arguably are more responsive to the Learning Results, the task of formally designing classroom assessments to demonstrably align with these standards still looms on the horizon for most Maine school districts. Clearly, in a standards-based climate, the integrity of an achievement composite depends, in part, on the extent to which the component measures are drawing on the same universe of standards. Without this assurance, we must interpret with caution the tendency of the single-measure classifications to putatively overidentify students who meet the standard. Here, too, subsequent research could be illuminating, particularly if the research involves multiple sites that vary with respect to the degree to which each measure is of demonstrable alignment with the announced standards.

A second, and related, limitation of the present study is that neither site had engaged in formal standard setting for either ITBS/TN or COURSE. This is why we obtained regression estimates of ITBS/TN and COURSE cutscores, given each measure's relationship with the MEA-M (for which the cutscore "meets the standard" is known).

III. Study II: Evaluation of School Effects with Multi-Level Data

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

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

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

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

Data and Methods

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

Taking a multi-level organizational perspective and drawing on the relevant literature, we test three models of school effects separately for Maine and Kentucky: Model 1 (no predictors at the student and school levels), Model 2 (predictors at the student level only, with grand-mean centering), and Model 3 (predictors both at the student and school levels, with grand-mean centering). Type A effect is estimated through Model 2 by removing the effect of student background variables. Type B effect is estimated through Model 3 by removing the effects of variables beyond a school's control (e.g., demographic composition). In this study, we consider only race and SES (socioeconomic status) factors. We believe that students' prior achievement (readiness for learning measured at the time of entry into current school) and mobility (length of stay in current school) factors must be considered to estimate authentic school effects but these data are not available in the NAEP.

All analyses were conducted using the HLM 5 program. Table 12 presents descriptive statistics for all variables used in these analyses. MRPCM1 through MRPCM5 are the five plausible values that make up the composite mathematics achievement outcome variable. WHITE is a dummy variable (1 = white, 0 = minority), and SES is a composite factor of parental education level, availability of reading materials at home, and school median income (standardized to have a mean of 0 and a standard deviation of 1 across states).

Model 1

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

Model 2

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

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

where (MRPCM)ij is the composite mathematics achievement of student i in school j; (WHITE)1ij is the indicator of student i's race in school j; (SES)ij is the indicator of student i's socioeconomic status in

Table 12. Descriptive statistics of predictors and outcome variables for HLM analyses of Kentucky and Maine 1996 NAEP 8th grade math data

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

school j; and eij is a Level 1 random effect representing the deviation of student ij's score from the predicted score based on the student-level model. Level 1 predictors are grand-mean centered so that the intercept, β_{0j}, can be interpreted as adjusted mean achievement for school j. This adjustment is chosen to sort out the unique effects of school on achievement after controlling for the influences of student/family characteristics.

The next step in HLM involves fitting an unconditional, or random, regression model at the school level (Level 2). Notice that all Level 1 regression coefficients are regarded as randomly varying across schools, and γ_{00} is the mean value of the school-level achievement outcome beyond the influences of student/ family characteristics. r0j, the school-level residual value from this regression, is used as an indicator of school average performance adjusted for racial and SES mixes of students. Likewise, r1j and r2j are used as indicators of racial and social achievement gaps respectively. The Level 2 (school level) model is

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

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

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

where β_{0j} represents school j's average mathematics achievement adjusted for its composition of students' racial and SES backgrounds; β_{1j} represents school j's racial gap (i.e., the achievement score gap

between white and minority students); and β_{2j} represents school j's social gap (i.e., the extent to which students' SES differentiates their achievement).

Model 3

Model 3 adds two school-level predictors or school aggregate values of student-level predictors. Percent white (PWHITE) and average SES (AVSES) are added to explain between-school variation. r0j, the school-level residual value from this regression, is used as an indicator of school average performance adjusted for racial and social composition effects. Model 3 is

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

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

Results

Model 1 (fully unconditional model)

Decomposition of variance in the outcome variable shows that the two states have similar distributions of mathematics achievement between the school and student levels. In Maine,18% of variance exists at the school level and 82% at the student level; the figures are 17% and 83%, respectively, in Kentucky. A residual school mean from this model is called a Model 1 average. The reliability estimate of these unadjusted school achievement averages is .80 in Maine and .79 in Kentucky, indicating that the sample means tend to be quite reliable as indicators of the true school means.

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

By using race and SES variables as predictors of math achievement at the student level (with grand-mean centering), we obtain adjusted school average achievement that takes into account differences among schools in their students' racial and social mixes. A residual school mean that is obtained after controlling for the effects of student-level predictors, as an indicator of value-added school performance, is called a Model 2 average. The reliability of conditional school means (conditional reliability) becomes lower: .67 in Maine and .62 in Kentucky. As shown in Table 13, Model 2 average is correlated very highly with Model 1 average (r_{me}=.92 and r_{ky}=.87).

The effects of race and SES on achievement are used as indicators of academic inequity, as well as providing the basis for adjusting estimates of school effects. This assumes heterogeneity of regressions among schools, and it models the effects of student's race and SES on achievement as randomly varying at the school level. The within-school racial gap—measured by the estimated average achievement gap between white and minority students within schools—is 12.1 (.41 SD) in Maine and 16.8 (.57 SD) in Kentucky (see Table 14). The within-school social gap, measured by the estimated effect of SES on achievement within schools—is 10.8 (.38 SD) in Maine and 10.6 (.36 SD) in Kentucky (see Table 14). In both states, these gaps are highly significant.

Maine and Kentucky show different patterns of relationships between achievement average and gap estimates (Table 13). In Maine, Model 2 average correlates positively with racial gap (.72) but negatively with social gap (-.63). Conversely, in Kentucky, Model 2 average correlates negatively with racial gap (-.28) but positively with social gap (.57). Higher performing schools in both states tend to have smaller gaps with regard to one background variable but larger gaps with regard to the other. This indicates that schools are not very effective in addressing both racial and social achievement gaps.

We should note that the reliability estimates of racial and social gaps are low: .13 and .21 in Maine, and .30 and .28 in Kentucky. Considering these reliabilities, it appears that both Maine and Kentucky schools vary little in their racial and social gaps. This is attributed to the fact that both states are highly homogeneous in racial composition. However, sufficient variability across schools on racial gap estimates does exist as the homogeneity of variance tests demonstrate significant variation (see the variance component chart in Table 14).

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

School-level predictors of racial and social composition were used to further adjust differences among schools in their average achievement due to composition effects. In Maine, both racial and social composition effects are not significant. This indicates that such school-level adjustment of performance for race and SES factors, in addition to the corresponding student-level adjustment, is not necessary (see Table 14). In Kentucky, only the social composition effect is significant, adding about 7 points to the within-school social gap estimate (see Table 14). Model 3 average (residual school means after controlling for both student and school-level effects of race and SES) correlates .70 with Model 1 average and .94 with Model 2 average (see Table 13).

Pooled HLM Analysis

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

The results of the pooled HLM analyses are summarized in Table 15. First, the comparison of Maine and Kentucky schools without any control for background variables show that Maine schools perform significantly better than Kentucky schools: a gap of 17.18 (Model 1), or roughly 1.2 SD. This gap decreases about 40% when we control for differences in their students' racial and social background characteristics (gap = 9.97, Model 2). When we further control for school composition effects, the Maine-Kentucky school achievement gap becomes slightly smaller but remains statistically significant (gap = 6.18, Model 3). Because Maine schools perform significantly better than Kentucky schools based on both Type A and Type B effect estimates, their greater effectiveness seems to come from sources related to schooling. Students' prior achievement and mobility factors become less important when we compare schools across states (vs. within state). Despite the average school performance gap, it turned out that there are no significant differences between the two states' schools in terms of their racial and social gap estimates.

Table 13. Correlations among school performance indicators

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

Table 14. Summary of HLM Results

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

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

Table 15. Summary of Pooled HLM Results

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

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

Discussion

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

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

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

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

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

Thus, the fundamental issue is not simply a technical choice of estimation methods given the available data. Rather, the estimation of school effects requires that we define "school effects" and formulate an explicit model of these effects. In other words, this approach requires that the model be fully specified: all important variables representing school input, practice, context, and student background would have to be measured and included in the model in order to guarantee that the effects of school practice were unbiased. Nevertheless, school quality variables are generally more difficult to define and measure and the relevant data are expensive to collect (Raudenbush & Willms, 1995).

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

IV. Study III: Evaluation of School Progress with Multi-Year Data

The dominant design for studying progress of schools across the nation is by means of successive-group comparisons (Carlson, 2001). This approach looks at the average achievement gain/loss from one year to the next for successive groups at the same grade level (e.g., comparing the average reading score of 4th graders in 2002 with the average reading score of 4th graders in 2001). Using this approach, one infers a change in the quality of a school or its programs by looking at the performance difference between two groups of students. Inherent weaknesses of this approach are initial group differences and mobility.

Hill (1997) has demonstrated that sampling error makes it difficult to determine which schools are making progress and which are not. Pooling data from multiple grades and/or years may reduce these sampling errors, although more than three or four years of data may be required to draw a valid comparison. Researchers also have found that successive-group comparisons can produce results different from those based on longitudinal comparisons, where performance is followed over time for the same cohort of students (e.g., Carlson, 2001; Dyer et al., 1969). And as Linn and Baker (1999) point out, within-school changes in student body can covary with changes in instruction, complicating the interpretation of data.

Regression artifacts further complicate the evaluation of school progress. Regression to the mean occurs when we examine the difference between two imperfectly correlated measures. Lower performing schools tend to improve their performance status more than higher performing schools. In other words, a school's baseline score (i.e., where the school started) is negatively correlated with the school's gain score (i.e., how much the school improved). We call this phenomenon *regression to the mean status*. Further, when we attempt to look at growth at more than two time points and examine growth in two adjacent periods, we may face another type of regression artifact: *regression to the mean growth*. For instance, if we choose to examine change from year 1 to year 2 and from year 2 to year 3, then the change from year 1 to year 2 will be negatively correlated with the change from year 2 to year 3. It is particularly problematic when one attempts to examine change in the growth between two periods: the "winners" in one period may appear to be "losers" in the other (and vice versa).

The central research question that we explore below is the extent to which current school progress indices reflect real change or a statistical artifact. Specifically, the objective of our study is to (a) investigate regression artifacts in evaluating schools' academic progress and (b) explore methods to overcome their effects in developing school progress measures. To date, little attention has been paid to regression artifacts and their possible implications for school evaluation.

Data

We examined schools in Kentucky and Maine, using the 8th mathematics achievement data collected from the KIRIS for 1993-1998 and the MEA for 1990-1998. Figure 1 shows the distributions of MEA 8th grade math school scale scores from 1990 through 1998. MEA scores range from 100 to 400, with a mean of 250 and standard deviation of 50 in the 1985-1986 base year. Figure 2 shows the distributions of KIRIS 8th grade math school accountability index scores from 1993 through 1998. The KIRIS accountability index score is a weighted composite reflecting the percentage of students at four different achievement levels (0 for Novice, .4 for Apprentice, 1.0 for Proficient, and 1.4 for Distinguished). Thus, KIRIS accountability index scores can range from 0 to 140.

Methods

We computed three different types of school-achievement gain estimates: (a) the one-year gain measure computes the difference between two adjacent-year means (e.g., 1991-1992 gain = 1992 mean minus 1991 mean); (b) the two-year gain measure computes the difference between adjacent two-year moving average scores (e.g., 9091-9293 gain = the average of 1992 and 1993 means minus the average of 1990 and 1991 means); and (c) the three-year gain measure computes the difference between adjacent three-year moving average scores (e.g., 9092-9395 gain = the average of 1993, 1994, and 1995 means minus the average of 1990, 1991, and 1992 means). We also conducted ordinary least-squares (OLS) time-series regression analyses of the school performance data over all years, determining how the long-term trends compare with short-term gains.

Further, we conducted time-reversed analyses to determine whether gain really is a regression artifact (see Campbell & Stanley, 1963; Campbell & Kenny, 1999). In order to remove regression artifacts from gain indices, we conducted regressions of 1-year, 2-year, and 3-year gain scores on their corresponding baseline scores and preceding gain scores, and obtained residualized gains. As Cronbach and Furby (1970) pointed out, residualized gain is a way of removing the effect of pretest status from a posttest score, but it is not a corrected measure of true change because the portion discarded may include some

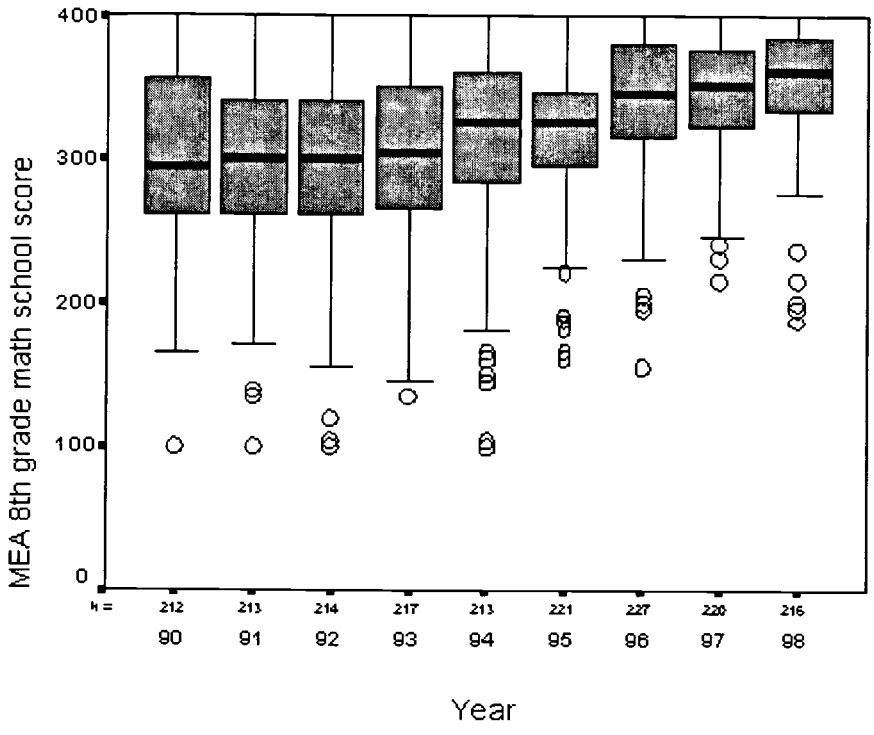


Figure 1. Box Plot of 1990-98 MEA 8th grade math school scores

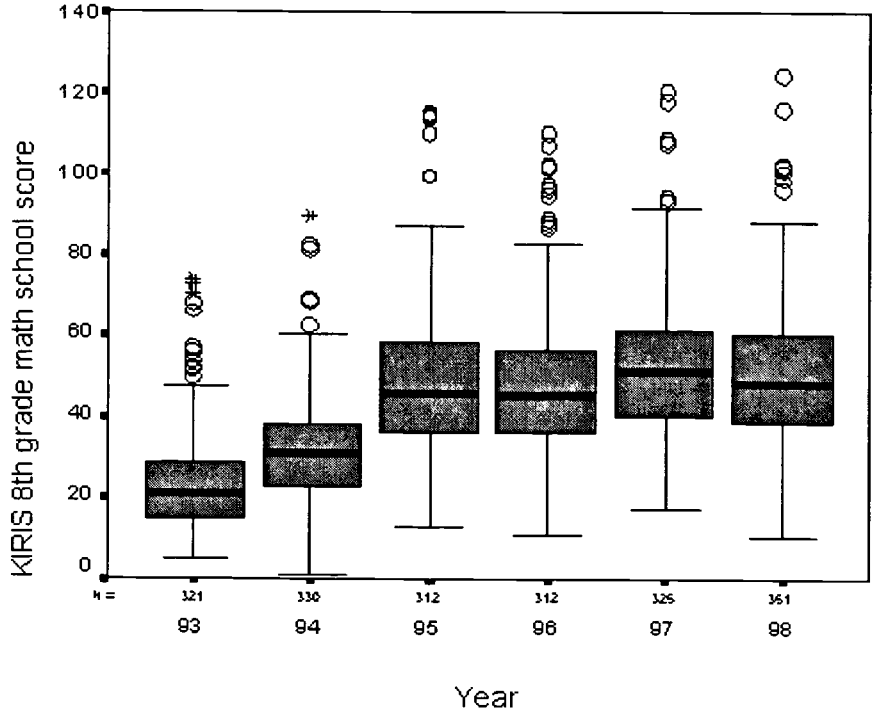


Figure 2. Box Plot of 1993-98 KIRIS 8th grade math school scores

genuine and important changes in the subjects (in this case, schools). Despite the limitations of residualized gain scores (Cronbach & Furby, 1970), we incorporated them in our analyses to explore how they might address regression artifacts and how this adjustment might improve the relationships among raw gain scores.

Analysis of School Means and Their Relationship

Correlations among one-year school means of 8th grade math achievement are modestly positive (Table 16). The correlation coefficients are in the .40s and .50s for adjacent years, and they become weaker for remote years. For example, the correlation between 1990 and 1991 means is .55, while the correlation between 1990 and 1998 means is .34. This indicates proximal autocorrelation in the time-series data (Campbell & Kenny, 1999). These correlations indicate that one-year school mean scores are only moderately stable over the short term, and this stability lessens appreciably over the long term.

Table 16. Correlations among one-year school means of 8th grade mathematics achievement in Maine and Kentucky

	1991 mean	1992 mean	1993 mean	1994 mean	1995 mean	1996 mean	1997 mean	1998 mean
1990 mean	.55	.47	.41	.44	.51	.46	.28	.34
1991 mean		.50	.52	.54	.45	.43	.41	.30
1992 mean			.59	.65	.51	.48	.39	.49
1993 mean				.65 .62	.53 .56	.41 .64	.42 .67	.47 .62
1994 mean					.60 .59	.54 .60	.50 .56	.50 .57
1995 mean						.65 .62	.52 .61	.45 .58
1996 mean							.48 .72	.49 .65
1997 mean								.56 .76

Note. Upper values are for Maine schools in 1990-1998 and lower values are for Kentucky schools in 1993-1998.

Correlations among two-year school means of 8th grade math achievement are in the .60s and .70s (Table 17). The correlations among two-year means are generally higher than the correlations among one-year means, indicating that the former increases stability by combining two years' data. Similarly, correlations among three-year school means of 8th grade math achievement are slightly higher than correlations among two-year school means (Table 18). As one would expect, the correlations among two-year or three-year means are particularly high when the two adjacent periods have common years. However, proximal autocorrelation still prevails: the size of correlation coefficients drop substantially as the two time intervals grow further apart.

Table 17. Correlations among two-year school means of 8th grade mathematics achievement in Maine and Kentucky

	9192 mean	9293 mean	9394 mean	9495 mean	9596 mean	9697 mean	9798 mean
9091 mean	.83	.60	.61	.63	.59	.51	.46
9192 mean		.85	.74	.69	.59	.54	.54
9293 mean			.89	.73	.58	.51	.57
9394 mean				.87 .84	.62 .74	.58 .74	.61 .71
9495 mean					.85 .90	.58 .72	.61 .69
9596 mean						.85 .88	.64 .76
9697 mean							.84 .90

Note. Upper values are for Maine schools in 1990-1998 and lower values are for Kentucky schools in 1993-1998.

Table 18. Correlations among three-year school means of 8th grade mathematics achievement in Maine and Kentucky

	9193 mean	9294 mean	9395 mean	9496 mean	9597 mean	9698 mean
9092 mean	.91	.81	.74	.72.	.60	.59
9193 mean		.93	.85	.72	.60	.60
9294 mean			.93	.81	.64	.64
9395 mean				.91 .95	.77 .89	.69 .79
9496 mean					.91 .85	.82 .86
9597 mean						.91 .94

Note. Upper values are for Maine schools in 1990-1998 and lower values are for Kentucky schools in 1993-1998.

Analysis of School Gains and Their Relationship

Three different types of gain scores were obtained for each school by computing differences among one-year means, among two-year means, and among three-year means. In this section, we examine the relationship among successive gain scores to see whether the schools' gain scores are stable over time. If short-term school achievement gain measures are to serve as a reliable indicator of school effectiveness, then school gains should demonstrate stability as evidenced by positive correlations among successive gains. We understand reliable gains do not guarantee the validity of the gains, but demonstrable reliability is a necessary condition for demonstrating their validity.

Table 19 shows correlations among one-year gains. Correlations among yearly gain scores are very low for close pairs and remote pairs alike. In other words, schools that appear to gain more this year do not maintain their gains the following years. The only exception is the moderately negative correlation between adjacent periods with an overlapping year (see the correlations in central diagonal line in Table 19). For example, the correlation between 9091 gain and 9192 gain is -.45. This is likely to result from the fact that both gains share the 1991 score with different signs, that is, + 91 score in 9091 gain versus - 91 score in 9192 gain. Aside from the artifactual, inverse relationship between adjacent pair of gains, the overall patterns of correlations indicate that there is no stability in the amount of yearly gain scores made by schools.

Table 19. Correlations among one-year gains of 8th mathematics achievement in Maine and Kentucky

	9192 gain	9293 gain	9394 gain	9495 gain	9596 gain	9697 gain	9798 gain
9091 gain	-.45	.08	-.05	-.16	.02	.13	-.12
9192 gain		-.49	.01	-.07	-.05	-.02	.16
9293 gain			-.51	.04	-.10	.07	-.02
9394 gain				-.39 -.34	.10 -.07	-.10 -.12	-.09 .08
9495 gain					-.34 -.57	-.09 .04	-.06 -.06
9596 gain						-.44 -.39	.12 -.07
9697 gain							-.51 -.37

Note. Upper values in each cell are for Maine schools in 1990-1998 and lower values are for Kentucky schools in 1993-1998.

The correlations among two-year gains are somewhat higher than the correlations among one-year gains (see Table 20). However, the higher correlations are observed only when the two periods involve common years. Moreover, the correlations are often negative for the same reason as the one-year gain correlations noted earlier. When we look at two periods with no common years (e.g., 9091-9293 gain vs. 9596-9798 gain), the correlations are barely significant. The correlations among three-year gains may be more stable (see Table 21), but they tend to have the same problems as the correlations among two-year gains.

Table 20. Correlations among two-year gains of 8th mathematics achievement in Maine and Kentucky

	9192-9394 gain	9293-9495 gain	9394-9596 gain	9495-9697 gain	9596-9798 gain
9091-9293 gain	.39	-.48	-.40	-.19	.08
9192-9394 gain		.26	-.35	-.31	-.02
9293-9495 gain			.49	-.12	-.28
9394-9596 gain				.48 .22	-.40 -.45
9495-9697 gain					.32 .42

Note. Upper values in each cell are for Maine schools in 1990-1998 and lower values are for Kentucky schools in 1993-1998.

Table 21. Correlations among three-year gains of 8th mathematics achievement in Maine and Kentucky

	9193-9496 gain	9294-9597 gain	9395-9698 gain
9092-9395 gain	.51	.06	-.30
9193-9496 gain		.68	.30
9294-9597 gain			.69

Note. All values are for Maine schools in 1990-1998; values for Kentucky schools are not shown as their data are available only in 1993-1998.

To examine a long-term trend in schools' academic progress, we ran OLS time-series linear regression to obtain the estimate of annual school gain over the entire period in each state: 9 years in Maine and 6 years in Kentucky. Among the 224 Maine schools with enough data for this regression, 85 schools (38%) had statistically significant gain estimates. The average of all 224 schools' annual gain estimates is 7.72, or .12 standard deviations (based on the 1990 standard deviation). Among the 315 Kentucky schools with enough data for this regression, 140 schools (44%) had statistically significant gain estimates. The average of all 315 schools' annual gain estimates is 5.34, or .47 standard deviations (based on the 1993 standard deviation).

In order to find out how well these different types of gains reflect the long-term trend, we examined the correlations between the former and the latter. Table 22 shows that three-year gain provides a closer estimate of the long-term academic growth trend than does two-year gain, which in turn is better than one-year gain. The correlations between three-year gains and whole-period trends are in the .50s and .60s. This indicates that AYP measures require longer intervals of data to provide a better approximation of the long-term trend. Nevertheless, it is uncertain whether an estimate of even a longer-term trend can provide a satisfactorily reliable and valid assessment of schools' academic progress.

Table 22. Correlations of long-term trend regression coefficient with 1-year, 2-year, and 3-year gain scores

1-year gain		2-year gain		3-year gain	
9091	.12	9091-9293	.44	9092-9395	.62
gain		gain		gain	
9192	.29	9192-9394	.45	9193-9496	.66
gain		gain		gain	
9293	.10	9293-9495	.42	9294-9597	.51
gain		gain		gain	
9394	.17	9394-9596	.34	9395-9698	.47
gain	.34	gain	.44	gain	.58
9495	.22	9495-9697	.30		
gain	.19	gain	.62		
9596	.17	9596-9798	.32		
gain	.22	gain	.56		
9697	.16				
gain	.30				
9798	-.03				
gain	.18				

Note. Upper values in each cell are for Maine schools in 1990-1998 and lower values are for Kentucky schools in 1993-1998.

Regression Artifacts and Residualized Gains

Correlations between initial status and gain scores are modestly negative (ranging mostly from -.40s to -.50s), which suggests that lower performing schools tend to gain more than higher performing schools. This might suggest the well-known regression to the mean artifact. We call this type of regression artifact *regression to the mean status*, which we distinguish from another type of regression artifact explained below. We conducted time-reversed analyses of gain score to determine whether it really is a regression artifact. Schools were classified into three groups (top quartile, middle half, bottom quartile) based on their 8th grade math mean scores in each of two adjacent years, and their performance trajectories were compared.

Figure 3 illustrates regression to the mean status with 1996 and 1997 data from Maine. Forward performance trajectories (solid lines) trace changes in average scores from 1996 to 1997 for three groups of schools that were classified based on their 1996 performance status: high (96 H), middle (96 M), and low (96 L). In contrast, backward performance trajectories (broken lines) trace changes from 1997 to 1996 for

three groups of schools that were classified based on their 1997 performance status: high (97 H), middle (97 M), and low (97 L). For both high and low performing schools, backward trajectories turn out to move in the same direction as their forward counterparts. If backward trajectories are traced from 1996 to 1997, then they appear to move in the opposite directions to forward trajectories. But backward trajectories by definition should be viewed as moving from 1997 to 1996. For instance, high performing schools in 1996 perform less well (going downward) in 1997, but at the same time high performing schools in 1997 also turn out to perform less well in 1996. This time-reversed analysis strongly indicates a regression artifact.

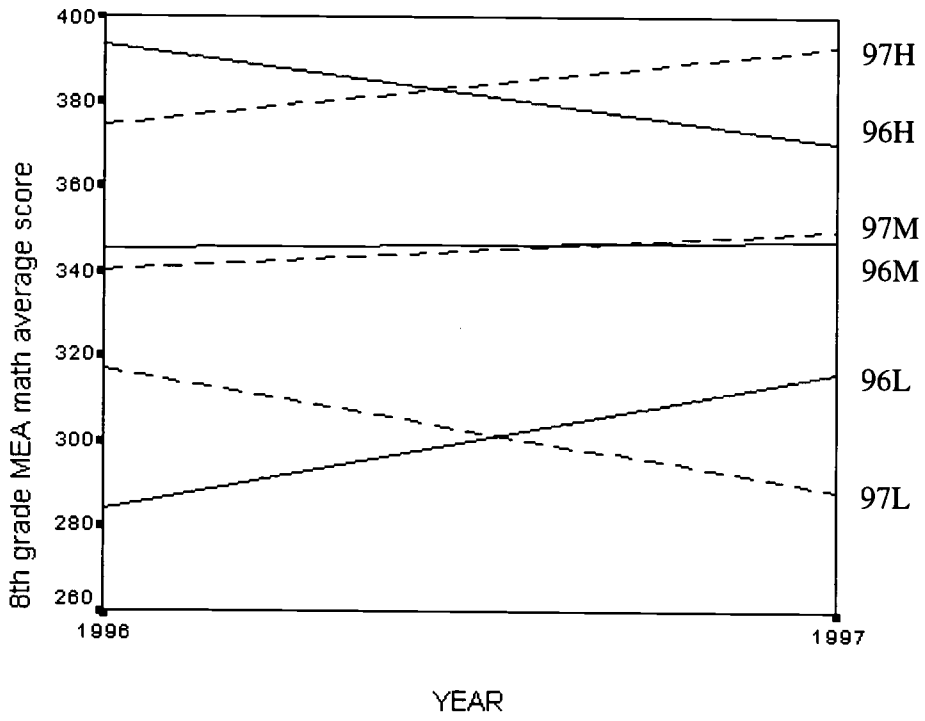


Figure 3. Forward vs. backward trajectories of 1996 and 1997 MEA 8th grade math school scores (YY H=high performing schools in year YY; YY M=middle performing schools in year YY; YY L=low performing schools in year YY)

Correlations among yearly gain scores for adjacent years are also negative (-.30s to -.50s). Schools that gained more in the current year tend to gain less in the next year. This suggests another type of regression artifact. The same is true of two-year and three-year gains, which also show a negative relationship with adjacent year gains (i.e., either two-year or three-year). This type of regression artifact, which we call *regression to the mean growth*, is illustrated in Figure 4. Here, schools were classified into three groups based on their achievement gains in 1995-96 and in 1996-97. Once again, forward versus backward trajectories for high (H) and low (L) improving schools also turn out to move in the same directions. If there had been no regression artifact, forward and backward trajectories would have followed the opposite directions.

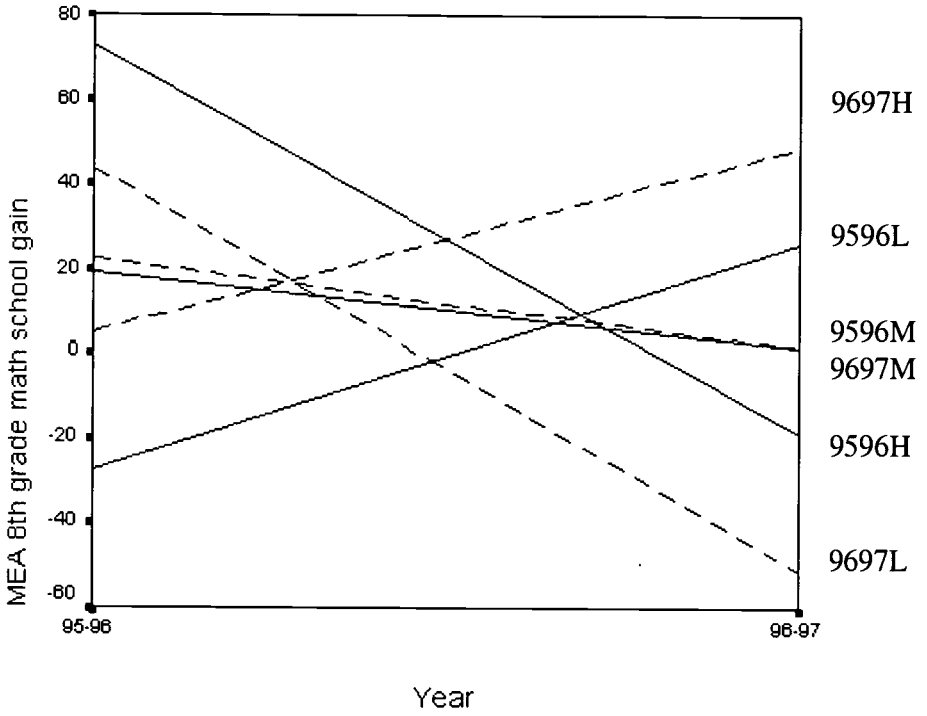


Figure 4. Forward vs. backward trajectories of 1995-96 and 1996-97 MEA 8th grade math school gains (XXYY H=high improving schools between years XX and YY; XXYY M=middle improving schools between year XX and YY; XXYY L=low improving schools between years XX and YY)

We found that the regression-to-the-mean-growth effect operates independently of regression-to-the-mean-status effect. Further, the former is often as strong as the latter. For example, Table 23 shows the results from two separate multiple regression analyses of the 1995-1997 data from Maine schools, one using regular (forward) regression and another using time-reversed (backward) regression. For time-reversed regression, we used current period's gain and baseline scores to predict the previous period's gain. In this case, both forward and backward regression analyses produce similar results: The effects of both predictors are significantly negative. If there had been no regression artifacts, the effects should have weakened substantially or disappeared altogether. We conducted the same analysis for 2-year and 3-year gains, obtaining similar results.

Given these regression artifacts, we attempted to obtain estimates of school gains that are not influenced by such effects. We ran a series of forward regressions as shown in the top part of Table 23 to obtain residuals. Table 24 shows correlations among residualized one-year gains. Values in parentheses show the changes in magnitude compared correlations in Table 19. Although all of the correlations change in a more positive direction, the changes are marginal. Likewise, correlations among two-year and three-year residualized gains show the same tendency of changes in correlation coefficients.

Researchers claimed that regression to the mean is more of a problem for two-wave studies than multiwave studies. Campbell and Kenny (1999) pointed out that the empirical fact of proximal

Table 23. Forward vs. backward regression of 1-year school gain score in Maine

Regression of 9697 gain on 9596 gain and 96 mean					
	Coefficient	Standard Error	Beta	t	p-value
Intercept	150.84	19.200		7.856	.000
96 mean	-.416	.057	-.437	-7.317	.000
9596 gain	-.281	.061	-.273	-4.580	.000
Adjusted R^2 = .348					
Regression of 9695 gain on 9796 gain and 96 mean					
	Coefficient	Standard Error	Beta	t	p-value
Intercept	37.768	23.258		1.624	.106
96 mean	-.177	.067	-.192	-2.641	.009
9796 gain	-.323	.071	-.332	-4.580	.000
Adjusted R^2 = .208					

Note. 9697 gain = 97 mean – 96 mean; 9796 gain = 96 mean – 97 mean; 9596 gain = 96 mean – 95 mean; 9695 gain = 95 mean – 96 mean.

autocorrelation implies that regression to the mean continues after the second wave of measurement. It is true that the largest amount of regression toward the mean usually takes place at lag 1 because that is where the correlation changes most. However, regression to the mean continues because the correlation keeps dropping further after lag 1. Thus, no matter how many time points we look at for an estimation of change or trend, it is necessary to take cautionary steps against regression artifacts.

Table 24. Correlations among one-year residualized gains of 8th mathematics achievement in Maine and Kentucky

	9192 gain	9293 gain	9394 gain	9495 gain	9596 gain	9697 gain	9798 gain
9091 gain	-.11(+.34)	.23 (+.15)	.13 (+.18)	-.01(+.15)	.11 (+.09)	.22 (+.09)	.01 (+.13)
9192 gain		-.13(+.36)	.22 (+.21)	.10 (+.17)	.11 (+.16)	.16 (+.18)	.31 (+.15)
9293 gain			-.26(+.25)	.09 (+.05)	-.03(+.07)	.16 (+.09)	.12 (+.14)
9394 gain				-.17(+.22)	.20 (+.10)	-.10 (.14)	.09(+.18)
				-.34 (-.20)	.12 (+.19)	-.12 (.00)	.13(+.05)
9495 gain					-.16(+.18)	.09 (+.18)	.02 (+.08)
					-.23(+.24)	.15 (+.11)	.07 (+.13)
9596 gain						-.15(+.29)	.24 (+.12)
						.-18(+.21)	.06 (+.13)
9697 gain							-.17(+.34)
							-.19(+.18)

Note. Upper values in each cell are for Maine schools in 1990-1998 and lower values are for Kentucky schools in 1993-1998. Values in parenthesis show changes in correlation coefficients (compared with Table 19) by adjusting for regression artifacts.

Discussion

Both Kentucky and Maine adopted the method of successive-group comparisons (i.e., comparing the performance of different cohorts) to evaluate school progress. Previous studies and the present findings challenge the validity of current progress measures based on this method (see Linn & Haug, 2002). Although the most important source of instability in school performance gain may come from changes in student body and mobility, neither factor is considered in developing and evaluating school progress measures. Indeed, the lower reliability of gain scores due to large sampling error and resulting lower correlations among the gain scores should worsen regression artifacts. But we suspect that these regression artifacts may remain legitimate concerns due to measurement error even if states could switch to a longitudinal-comparison design (i.e., following the performance of the same students).

In summary, our findings show that higher performing schools tend to gain less while lower performing schools tend to gain more. This illustrates the well-known regression to the mean "status" phenomenon. At the same time, our results also reveal regression to the mean "growth" phenomenon. Schools that gained more than others in the past tend to gain relatively less. The former force tends to make higher and lower performing schools appear convergent in their status, whereas the latter force may make more and less improving schools appear convergent in their growth. These two forces as statistical artifacts may confound school progress measures and need to be addressed. At the same time, it is important to realize that any adjustment for regression artifacts cannot cure the underlying problems with current school progress measures that arise from the use of the successive group (cohort) comparison method.

What are the implications of the regression-to-the-mean artifacts for evaluating schools' adequate yearly progress (AYP)? In case of Maine, they set the same quota for all schools, despite the fact that there is substantial variation even among relatively low-performing schools. In case of Kentucky, they set different quota for schools according to the schools' baseline status. Schools that initially performed at a lower level would be required to make a relatively large gains (meeting a higher AYP threshold), while initially higher performing schools would be assigned a relatively lower AYP threshold. Kentucky's approach has the potential to take into account the regression-to-the-mean status artifact, but it is not clear whether this adjustment is reasonable, or how it differs from the residualized gain approach. To examine this, subsequent research should compare individual schools' adjusted gain scores (residuals obtained after controlling for the effect of initial score) with their AYP difference score (difference between school-specific AYP threshold and actual gain made) and with their raw gain.

On the other hand, the effect of regression-to-the-mean growth is not considered by either state's current AYP formula. Once the AYP threshold or quota for each school is set based on the gap between baseline performance and expected performance, it does not change—schools have to meet the same quota every period regardless of how much progress they made previously. In both Kentucky and Maine, the AYP quota for schools do not change as their scores change. A more sensible approach is to reset the AYP quota each year, depending on the amount of progress made toward the goal. In this way schools that improved more than the quota (and thus narrowed the gap more) may be assigned a smaller quota next time. This approach has the potential to address the regression-to-the-mean-growth artifact. Subsequent research should compare individual schools' adjusted gain scores (residuals obtained after controlling for the effect of previous gain) with their AYP difference scores (gap between varying AYP threshold and actual gain vs. gap between fixed AYP threshold and actual gain).

References

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

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

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

Campbell, D. T. & Kenny, D. A. (1999). *A primer on regression artifacts.* New York: Guilford Press.

Campbell, D. T. & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Carlson, D. (2001). The focusing of state educational accountability systems: Which schools do we want to reward or punish—and for doing what? Unpublished draft paper.

Consortium for Policy Research in Education (2000). *State assessment and accountability systems: 50 state profiles.* Retrieved May 20, 2001, from http://www.gse.upenn.edu/cpre/docs/pubs/profiles.html

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—or should we? *Psychological Bulletin,* 74, 68-80.

Dyer, H. S., Linn, R. L., & Patton, M. J. (1969). A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. *American Educational Research Journal,* 6, 591-605.

Hill, R. (1997). Calculating and reducing errors associated with the evaluation of adequate yearly progress. Paper presented at the annual assessment conference of the CCSSO (ED 414307).

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

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

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

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

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

Lee, J. (2000). Using National and State Assessments to Inform the Performance of Education Systems. Paper presented at the annual meeting of AERA, New Orleans (ED 442 871).

Linn, R. L., & Baker, E. L. (1999). Standards-based accountability systems' adequate yearly progress: Absolutes, wishful thinking, and norms. *The CRESST Line*, Spring 99.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, 24(1), 29-36.

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

National Research Council (1999). *High stakes: Testing for tracking, promotion, and graduation*, J. P. Heubert & R. M. Hauser, eds. Committee on Appropriate Test Use. Washington, DC: National Academy Press.

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

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

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

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

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

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

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education.* Research Monograph No. 6. National Institute for Science Education (NISE), University of Wisconsin-Madison. Washington, DC: NISE.

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

TM034389

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

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Using multiple measures to evaluate the performance of students and schools	
Author(s): Jaekyung Lee and Theodore Coladarci	
Corporate Source:	Publication Date: July 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Jaek Lee	Printed Name/Position/Title: Jaekyung Lee, Assistant Professor	
Organization/Address: University of Maine, 5766 Shibles Hall, Orono, ME 04469	Telephone:	Fax:
	E-mail Address: jklee@umit.maine.edu	Date: 7-30-02

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org