

Considerations in Using the Longitudinal School-level State Assessment Score Database

Don McLaughlin

Statistics and Strategies, Palo Alto, CA

December 2005

The National Longitudinal School-Level State Assessment Score Database (NLSLSASD) has achievement scores for over 80,000 public schools in the United States.¹ The nature of the scores varies from state to state, including scores in reading, mathematics, language arts, science, writing, and social science in some or all states, for grades ranging from 3 to 12 and years ranging from 1998 to 2003, plus earlier years for some states. Current activities include adding 2004 and 2005 scores to the database.

The contents of this database have been used in a variety of ways, for studies of school-level correlates of achievement, for search tools for information about schools, for comparison among state assessments using NAEP (the National Assessment of Educational Progress), and for analyses of outcomes associated with federal program interventions.

The NLSLSASD has two attributes that warrant careful consideration in planning for future uses of the data. First, the data are aggregates for sets of students in each school (all students in a grade, and in some cases all students in a grade in a demographic category). Virtually no information is included about the variability of scores within these groups in a school. Second, because this database is founded on the provision of public information by State Departments of Education, its contents are subject to the variety of characteristics of test scores that are the results of individual state assessment program policies, designs, and implementations.

The objective of this paper is to address questions concerning these attributes of the data in the NLSLSASD. By addressing these questions, I hope to provide information that will be helpful to analysts who would use state assessment data for information about correlates of student achievement in American public schools. The results presented here are based on both previous work and new analyses of the NLSLSASD. The paper is organized into six sections, each addressing a methodological question. The first four sections cover characteristics of state

¹ The NLSLSASD is available at <http://www.schooldata.org>.

Copyright © 2005 National Academy of Sciences. All rights reserved. No part of these pages, either text or image may be used for any purpose other than personal use. Reproduction, modification, storage in a retrieval system or retransmission, in any form or by any means, electronic, mechanical or otherwise, for reasons other than personal use, is strictly prohibited without prior written permission.

Opinions and statements included in the paper are solely those of the individual author(s), and are not necessarily adopted or endorsed or verified as accurate by the Planning Committee for the Symposium on the Use of School-Level Data in Evaluating Federal Education Programs or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.

assessment data; the fifth section discusses selected analytical alternatives, and the final section suggests enhancements to the database which would increase the range of potential analyses of it.

1. How different are alternative measures of school achievement?
2. How do results for student-level analyses and school-level analyses differ?
3. What are the values and limitations of pooled within-state analyses?
4. How can state assessment scores be compared between states?
5. How should achievement trends be analyzed?
6. What additions to the database would increase its value?

It is important for analysts to understand that the data in the NLSLSASD cannot, by themselves, provide information about the impact of a program on achievement, they can only provide information about whether participation in a program is a correlate of achievement. Randomized assignment of schools to treatment and control conditions is an essential component of studies which would estimate program impact. By itself, the NLSLSASD can provide statistics which could be explained as program impact (or lack of impact), but such inference is unwarranted without credible rejection of all competing explanations for the same results (e.g., the program was implemented in schools which had a predisposition to succeed or fail).

How different are alternative measures of school achievement?

The NLSLSASD contains scores published by every state, based on the state assessment program in each state. To construct the database, AIR extracted measures from state websites and requested existing electronic files of scores from other states, thereby avoiding the necessity for states to make additional computer runs to make their scores available on the NLSLSASD. AIR performed several tasks on the acquired scores, such as matching the schools to the NCES standard school codes, recomputing percentages meeting standards to render them all cumulative, preparing SAS programs for users to read the data, and naming variables in a systematic fashion.

Nevertheless, the scores from different states are different – different grades, different tests, and different standards. Can achievement score results be combined across states when they are based on different grades, on different subjects, on different tests, or on different statistical summary statistics? While a naive comparison of average achievement scores between states is generally meaningless because each state uses its own scale, the combination of *within-state treatment versus control comparisons* across states, in terms of effect sizes and statistical significance, may be quite meaningful.

In this section, I use the NLSLSASD to estimate the effects of five factors on the outcomes of statistical analyses of test scores: (1) different statistical summary measures, (2) different years, (3) different grades, (4) different tests, and (5) different subjects. A note is included about a technical factor that must be taken into account when comparing school-level results from different states: the distribution of sizes of student samples within schools. States with many very small schools will automatically have less reliable school-level statistics than states with few very small schools.

Different statistical summary measures

School-level achievement scores are reported in different units in different states. The question is to what extent that creates an impediment for using state assessment scores in evaluations. For example, if the design for an evaluation of a school reform calls for, among other things, the comparison of achievement scores in schools implementing the reform versus achievement in other schools in the same state(s), to what extent does it matter that one state's scores are available in one unit and another's is in a different unit? Can the comparisons in the two states be combined into an overall measure of the relation between the reform and achievement? To address this, it is necessary to know how much of an effect differences between computations based on different units can have on the outcomes of analysis.

The units commonly in use, with their abbreviations used here, are:

Mean scale scores	(SS)
Median scale scores	(MS)
Mean raw scores	(RW)
Median percentile ranks	(PR)
Mean normal scale equivalents	(NE)
Percent of students meeting a low level criterion	(C1)
Percent of students meeting a mid level criterion	(C2)
Percent of students meeting a high level criterion	(C3)
Percent of students meeting a very high level criterion	(C4)
Index combining percentages meeting several criteria	(IX)

Obviously, each of these measures has unique properties, but those unique qualities may not have a noticeable impact on the results of an evaluation. The way to evaluate the impact of differences in evaluation results is to examine the correlations among the different score types in states reporting achievement in multiple units. Using a subset of the NLSLSASD, one can compute correlations between different measures of the same construct, between scores for different grades, between scores on different tests, between scores for different years, and between scores for different subjects.

To provide some basic descriptive data on these correlations for this session, I constructed a database of 13,762 elementary school correlation coefficients, across:

the subjects of language arts, mathematics, reading, and science,
grades 3, 4, 5, and aggregate “elementary,”
years 1998, 1999, 2000, 2001, 2002, and 2003,
ten measures (see above),
two tests (in five states), and
49 states and the District of Columbia.

Factorially, these constitute 1,920 possible measures in each state, and potentially more than 20,000 correlations between two scores which vary on one of these factors for each state. The constructed database of correlations has an average of 274 correlation coefficients per state. Each correlation is between two scores on tests that differ on exactly one of the listed characteristics.

Average correlations between pairs of the ten measures are shown in table 1.² For example, the average correlation between the percentages of students meeting a low criterion (C1) and a mid-level criterion (C2), across subjects, tests, grades, and years is .84. The average correlations among mean scale scores (SS), median scale scores (MS), normal curve equivalents (NE), raw scores (RW), percentile ranks (PR), and standards based indexes (IX) range from .97 to .997. The correlations with the percentage of students meeting a mid-level criterion (C2) with these measures are nearly in the same range, varying from .91 to .96. The correlations involving percentages of students meeting low, high, or very high criteria are noticeably lower, ranging from .46 to .88.

These correlations varied little from state to state. The standard deviations of the correlation coefficients across states were very small for the measures other than the extreme standards, .05 or less (see table 1a). Correlations involving the extreme standards varied substantially more from state to state than correlations for other measures, as shown in table 1. Note that a total of 4,038 correlations form the basis for table 1: each correlation is counted for each of its pair of measures in the counts in the bottom row of table 1.

The question that immediately arises is “Which of these correlations are high enough?” There is no single answer to this question because it depends on the use to which the data are to be put.

² The mean correlations shown in this and similar tables were computed by first obtaining the average correlation for each combination (of measure, subject, grade, test, and year) across all states with that combination of measures. The averages of these means were then computed for the table. Using this method virtually eliminated the effects of extraneous factors on the tabulated averages.

As an example of such an effect that is eliminated, suppose science correlations were lower than reading correlations and that science correlations tended to involve normal curve equivalents more than reading correlations did. This would cause correlations with normal curve equivalents to appear smaller than other correlations. To the extent that each combination is found in at least one state, the method used ensures that science and reading correlations contribute equally to the averages for normal curve equivalent correlations.

Certainly, except for the extreme standards (C1, C3, and C4), they appear to be measuring similar characteristics of school-level achievement. To answer this question, it is necessary to consider scenarios in which these scores might be used. Although one can imagine a wide variety of uses of these test scores, the scenario that is most central to the purposes of federal program administrators is the evaluation of educational interventions.

Table 1. Average correlations between alternative measures based on the same test

	C1	C2	C3	C4	IX	MS	NE	PR	RW	SS
C1		.84	.65	.46	.89	.73	.86	.86	.80	.88
C2			.82	.61	.96	.91	.91	.94	.94	.93
C3				.77	.84	.88	.85	.88	.87	.86
C4					.79	.59	–	–	.63	.75
IX						–	–	–	–	.98
MS							–	–	–	.97
NE								1.00	–	.99
PR									.98	.99
RW										1.00
mean	.75	.84	.80	.64	.90	.83	.94	.93	.90	.89
std.dev.	.11	.08	.12	.10	.06	–	.02	.07	.03	.08
n	1,780	1,894	1,777	443	124	68	174	528	110	1,178

NOTE: – indicates combinations that are not available.

Std.dev. is the average between-state standard deviation of correlations.

SOURCE: The National Longitudinal School-level State Assessment Score Database.(NLSLSASD) 2004.

Table 1a. Average correlations between alternative measures based on the same test, omitting correlations with extreme standards

	C2	IX	MS	NE	PR	RW	SS
C2		.96	.91	.91	.94	.94	.93
IX			–	–	–	–	.98
MS				–	–	–	.97
NE					1.00	–	.99
PR						.98	.99
RW							1.00
mean	.93	.97	.93	.97	.97	.97	.96
std.dev.	.05	.03	–	.01	.03	–	.04
n	561	50	30	122	285	58	464

NOTE: – indicates combinations that are not available.

Std.dev. is the average between-state standard deviation of correlations.

SOURCE: The National Longitudinal School-level State Assessment Score Database.(NLSLSASD) 2004.

Although a definitive answer to the question of whether a program has a positive impact requires controlled experimentation, with *randomized* assignment of units (students, teachers, or schools) to alternative treatments, information about the average achievement difference between treatment and control units can be useful for subsequent planning, even if the treatment-control assignment is not randomized. It is only important to remember that any finding of a correlation between treatment and outcome (achievement gains) is open to plausible alternative explanations (such as a bias in the selection of which units get the treatment) and that statistical adjustments cannot completely eliminate those alternative explanations.

Correlational information about the *effect size*, the size of the treatment-control difference relative to typical achievement variation in the population, and *statistical significance*, the likelihood of finding as large a treatment-control difference by chance, are crucial for efficient decision-making about educational strategies in which to invest further study. The NLSLSASD can provide that kind of correlational information. Unfortunately, although the NLSLSASD can provide a very large sample size, the achievement measures in different states are not identical. The practical question is whether they are sufficiently similar that findings of effects in different states can be meaningfully combined into an overall description of the effect size and statistical significance across many states.

First, consider effect sizes. If the effect size for a particular treatment based on one outcome statistic, such as the median percentile rank, is, .5 times the standard deviation of school means for the control group of schools, for example, then based on the correlation of .87 between the two types of measure, one would expect the effect size based on a normal curve equivalent to be between .43 ($=.5 \times .87$) and .58 ($=.5 / .87$).

Next, consider statistical significance. Within each state the basic analysis is to compare achievement in “treatment schools” with achievement in other schools. To simplify the question of combining different measures, I consider the problem of combining results from two states. If two states had the same achievement measure, then differences in the statistical significance between the two states can be attributed either to a difference in sample sizes, which can be computed, or differences in the treatment-outcome context and process in the two states. With different achievement tests, however, finding a significant effect in one state and a non-significant effect in the other might be due to the fact that one of the test statistics had more error (relative to the effects of the treatment) than the other. For example, suppose one state reports the average scale score (SS) in each school and the other state reports the percentage of students meeting a mid-level criterion (C2) in each school.

The effect of that “error” on statistical significance results is related to the correlation between the measures. Specifically, the value of *Student’s t* is reduced by a factor of r when an amount of error is added to a measure that would cause that correlation. Based on table 1, the typical difference between C2 and SS is represented by a correlation of .93, so the use of C2 instead of SS might reduce the value of *Student’s t* by a factor of .93, if SS was the more precise measure, and vice versa. Thus, if the value of *Student’s t* (for the treatment-control difference) in the first state is 3.00,³ one should not be surprised if the value of *Student’s t* in the second state is anywhere between 2.79 ($= 3.00 \times 0.93$) and 3.23 ($=3.00 / 0.93$). However, if the value of *Student’s t* in the second state were, say 1.00, it would be reasonable to infer that the difference between the results in the two states is probably not due merely to differences in the outcome statistics used.

³ For example, if the average achievement in 50 treatment schools is .5 standard deviations higher than in 150 other comparable schools in a state, the value of *Student’s t* is 3.06.

That is, if the correlation between the different types of measures is sufficiently high, then the effects of using different measures on the statistical significance of treatment effects will be confined to a narrow range of effect sizes, especially for large samples. For treatments that are expected to have effect sizes of .5 standard deviations or more, implemented in 50 or more schools in a state, correlations of greater than .7 between types of measures should be sufficient to ignore the differences between types of measures. This computation needs to be carried out for each potential use of the database, but it indicates that generally, except for percentages meeting the extreme standards, the use of different types of measures has little potential for effecting study findings.

Different years

When an educational intervention takes place, it is hoped that it will improve achievement in the schools selected for participation, and that it will change the rank order of the schools' achievement in a state. Simultaneously, both systematic student demographic shifts and school staffing changes can change the rank orderings of schools. How sharp a test of an intervention's effect on achievement is depends on the background level of stability of school-level scores from year to year. Correlates of school-level achievement gains will be easier to find if school-level scores are naturally stable.

The NLSLSASD has 3,014 school-level correlations based on the same test, the same grade, and the same statistical measure, across pairs of years.⁴ The averages of these correlations are shown in table 2. They follow a stable pattern over these years: a correlation of about .77 for a one-year lag and somewhat lower correlations across multiple years.

Table 2. Average correlations between scores on the same test between years

	1998	1999	2000	2001	2002	2003
1998		.78	.74	.71	.65	.65
1999			.76	.73	.68	.66
2000				.76	.73	.68
2001					.77	.72
2002						.77
n	710	996	1,153	1,139	1,063	967
mean correlation by lag	Lag	1 year: .77	2 years: .73	3 years: .69	4 years: .65	5 years: .65
standard deviation		.11	.11	.12	.14	.18

NOTE: Years refer to the spring of the school year.

Standard deviation is the average between-state standard deviation of correlations.

SOURCE: The National Longitudinal School-level State Assessment Score Database.(NLSLSASD) 2004.

These correlations vary substantially between summary statistics and between states. Nebraska results were not included in this analysis because some correlations between different years were negative. In Maine (.41) and Vermont (.46), the cross-year correlations across years in the database are less than .50, while in California (.90), Connecticut (.90), and Pennsylvania (.86),

⁴ For this and subsequent summaries of correlations, percentages meeting extreme standards (C1, C3, and C4) are omitted.

they are greater than .85. Overall, variation in the correlations is well-predicted by taking into account the type of statistical measure, the state, the years involved, the grade, the subject tested, and which of two tests (in a few states): $R^2 = .83$ and the root mean squared error is .05.

The implications of a cross-year correlation of .75 are that difference scores, i.e., year-to-year gain scores, have a standard deviation that is about half ($=2-2r$) as large as the standard deviation of scores for a single year.

Different grades

No Child Left Behind calls for achievement testing starting with third grade and continuing every year through eighth grade, which represents a substantial expansion for most state assessment programs. Before 2003, most states administered tests only in one or two elementary grades. Therefore, the NLSLSASD contains test scores for different grades for different states. The question that arises concerns the extent to which these differences affect the outcome of analyses. Of course, if a particular treatment is targeted at one grade, like grade 5, it makes no sense to measure effects in another grade, like grade 3, but many treatments, such as systemic reforms, target the entire school program, and for these, it is possible that results from one grade in one state can be combined with results for another grade in another state. Our approach to testing this is the same as for the test of the impact of different summary statistics, discussed above. The critical statistic is the correlation between scores at different grades in the same subject in the same school. The average correlations for tests in grades 3, 4, and 5, and an aggregate of elementary grades are shown in table 3.

Table 3. Average correlations for tests of the same subject in different grades

	Grade 3	Grade 4	Grade 5	Aggregate Elementary Grade
Grade 3		.78	.75	.92
Grade 4			.81	.92
Grade 5				.93
mean	.77	.80	.78	.92
standard deviation	.09	.09	.09	.03
N	696	507	693	38

SOURCE: The National Longitudinal School-level State Assessment Score Database.(NLSLSASD) 2004.

Between pairs of the three separate grades, the correlations are between .75 and .91, while the correlations of single grades with elementary aggregates are, as might be expected, much higher. Assuming scores for different grades are just different manifestations of a latent trait (i.e., school-level achievement), these correlations indicate the amount of error in the measures. In that scenario, the average correlation of .78 between grade 3 and 4 scores, for example, indicates that the latent trait (school-level achievement) typically accounts for 78 percent of the variance in the single grade measures and that the correlation of these scores with the latent trait is .88 (= the square root of .78).

There are two important reasons, besides chance variation, that scores for two different grades in a school are different from each other. First, although learning is cumulative, the details and emphasis of the subject matter differs between grades. Second, the scores are for two different cohorts of students. Using the data in the NLSLSASD, it is possible to estimate the relative size of these effects by comparing the correlations between adjacent grades (a) in the same year (but with different cohorts of students) versus (b) in successive years (but with the same cohort of students). In the NLSLSASD, there are 380 sets of correlations involving adjacent grades in the same and adjacent years. In these sets, the average correlation between adjacent grades in the same year is .77, while the average correlation between adjacent years for the same grade is .78. In both cases, the measures being correlated are for different cohorts of students. By contrast, the average correlation between adjacent grades in adjacent years, where the same cohort of students is responsible for both test scores, is .81. The increment in the correlation when the same cohort of students is involved is a stable phenomenon. Because the standard deviation of the difference between these correlations is less than .05, the difference between .81 and either .78 or .77 is statistically significant.

Different tests

An often raised concern is that different states use different tests. In a few states, multiple tests have been used for assessing achievement in a subject, but the development of the NLSLSASD has not focused on the collection of multiple measures. Nevertheless, 61 correlations between pairs of test scores in 5 states (California, Florida, Hawaii, Rhode Island, and Vermont) are included in the database.⁵ The average correlation between the test scores from different tests (of the same subject in the same grade and year, with the same type of summary statistic) is .93, with a between-state standard deviation of .04.

⁵ Scores for Spanish and English versions of the Texas assessment are not included in this comparison because they tend to be based on different sets of students.

This issue often arises in the context of states' changing test between years. To address this, it is necessary to compare the correlation of one test with a different test in the same subject and grade the following year with the correlation (a) between the same two tests in a single year and (b) between two adjacent years for one of the tests. One might expect the effects of both different tests and different years to operate independently to reduce the correlations between measures. The NLSLSASD has data on 46 such matched sets of correlations in the five states listed above, which provide a basis for this comparison. In these 46 cases, the average correlation between two different tests in the same year is .91, the average correlation between two adjacent years for the same test is .76, and the average correlation for different tests between two adjacent years is .74. The similarity of the latter two correlations suggests that different tests are not the most credible explanation for different results in different states or in different years.

Different subjects

Finally, the correlations between pairs of different subjects (e.g., reading, mathematics, language arts, and science) are of interest because for systemic reform, they provide multiple observable measures of a latent trait of school-level achievement. Although specific interventions may focus on specific curriculum subjects, improving the conditions of learning in a school should improve achievement across subjects, and improvements in reading instruction, for example, should have benefits for language arts, mathematics, and science achievement as well.

The NLSLSASD includes scores for a variety of subjects, of which four were selected for these analyses. Average correlations across four subjects, controlling for grade, test, year, and type of summary statistic, are shown in table 4.

Table 4. Average correlations for tests of different subjects in the same grade

	Reading	Mathematics	Language Arts	Science
Reading		.87	.90	.86
Mathematics			.88	.82
Language Arts				.84
mean	.88	.86	.88	.84
standard deviation	.06	.06	.05	.07
n	1,057	1,049	597	439

SOURCE: The National Longitudinal School-level State Assessment Score Database.(NLSLSASD) 2004.

The implication of a cross-subject correlation of .87 is that 87 percent of the variance in these scores is common variance. Thus, it is reasonable to model correlates of school-level achievement using tests in multiple subjects to reduce error variance in the analysis.

School sample size effects on school-level correlations

Finally, it needs to be pointed out that the size of the correlation coefficients depends on the stability of the individual school-level measures, and the stability of school sample means (and percentages of students meeting standards) is related to the sample size within schools. In a state in which the average number of students per grade is 10, one would expect the correlation between grade 3 and grade 4 scores, or between reading and mathematics scores, or between adjacent years, to be lower than in a state in which the average is 25 students per grade.

In analyses comparing correlations of state assessment scores with NAEP scores from the 2003 assessments (McLaughlin, Bandeira de Mello, Blankenship, Chaney, Hikawa, Rojas, William, and Wolman, 2005; McLaughlin, Bandeira de Mello, Blankenship, Chaney, Esra, Hikawa, Rojas, William, and Wolman, 2005), the size of the school sample for NAEP was found to have a significant effect on the correlations of state assessment scores with grade 4 NAEP scores, in both reading and mathematics. Those analyses also found significant negative effects of “extremeness” of the standards and of use of state assessment scores for grade 3 or 5, instead of grade 4, to correlate with grade 4 NAEP scores.

In summary, the fact that assessments differ from one state to the next is not a fatal problem for using state assessment achievement scores for federal program evaluations and policy analyses. The answer to the question is empirical, depending on the correlations that can be expected between different test scores. Based on over 20,000 correlation coefficients for elementary schools in the NLSLSASD, one can conclude that differences among units for reporting, such as average scale scores, percentile ranks, or percentages of students meeting mid-level standards, which tend to be correlated at about the .9 level, are not likely to have significant effects on results. Percentages meeting extremely high or low standards, on the other hand, are not so highly correlated with other test results. Similarly, if two different reading tests (or two different mathematics tests) are administered in a year, it should not affect results substantially.

Scores for different grades in the same school tend to be reasonably highly correlated, at about the .75 to .80 level, although they are not as highly correlated as if the test scores were for the same student cohort, measured in adjacent years. Between adjacent years, test scores for the same grade tend to be correlated at the .77 level, with somewhat lower correlations across greater time intervals. The correlation of adjacent grades in adjacent years, involving the same student cohort, are about .03 greater, on average, than correlations not involving the same student cohort.

From these estimates of correlations one can estimate the level of noise added into national level analyses by the differences between state assessments; because the database includes scores for over 80,000 schools, the level of noise added usually will not be expected to interfere with the identification of relations with the kinds of effect sizes sought for educational programs.

School-level or student-level achievement?

Are school-level scores relevant to education policy analysis? The fundamental objective of education policy is to improve the achievement of students as individuals, and it is reasonable to question whether findings about correlates of achievement from school-level data will tell much about individual achievement. On the other hand, much of educational reform policy is aimed at the school as the critical unit, and most accountability data are reported at the school level. It would therefore be useful to be able to be able to make inferences about achievement trends and achievement gaps from school-level statistics.

The questions of the differences between student-level and school-level statistics cannot be answered purely on theoretical grounds, because the answers depend on empirical phenomena, the most important of which are the relative strengths of between- versus within-school relations between educational processes and achievement. Address these questions requires a database with student-level statistics which can be aggregated to school-level statistics, enabling comparison of the results of alternative analyses.

One such database is a four-state study of 2003 assessments in reading and mathematics. In 2003, the NAEP Validity Studies (NVS) panel put together a database of matched NAEP and state assessment data for over 2,000 students in each of two subjects (reading and mathematics) in two grades (4 and 8) in four states. (The primary purpose of the NVS study was to address validity research questions about the effects of different corrections for nonresponse on results from NAEP.) The basic statistics for these samples are shown in table 5. The state assessment scores for the four states are on different scales, which have been normalized to student means of 250 and student standard deviations of 50 both for ease of comparison and difficulty of identifying the states.

The starting points are that (1) the averages based on student-level and school-level are identical and (2) variances are larger when computed based on students, because school means contain no information about variation within the school. (Note: the NLSLSASD has means for various subgroups within schools, such as students eligible and not eligible for free or reduced price lunch, but it has virtually no information about variation within each subgroup in each school. ⁶) Obviously, means and standard deviations of scores on state assessments cannot meaningfully be compared between states, grades, or subjects; and the relative sizes of between- and within-school standard deviations cannot be directly compared without taking into account the relative sizes of within-school samples in the different rows of the table.⁷

The difference in variances between student- and school-level statistics can be displayed graphically as a comparison of population profiles, which show the distribution of achievement from the lowest percentiles to the highest percentiles of a population. A population profile graphs the score, X , for a unit against the percentile corresponding to the score, $p(X)$; mathematically, the inverse of the cumulative distribution function.

⁶ Actually, for states with multiple standards, one can construct an indicator of within-school variance from the percentages of students in the extreme high and low categories.

⁷ Based on the combination of variance and sample size information, one can infer that there is greater relative between-school variation in achievement, compared to within-school variation, in State 3, and less in State 1, than in the other states in this study. On the other hand, there is little variation in that ratio between grades or between reading and mathematics.

Table 5. Means and standard deviations based on school-level and student-level computations

State	Mean		Standard Deviation		Count	
	School	Student	School	Student	School	Student
Grade 4 Reading						
1	250	250	18.1	50.0	108	3048
2	250	250	20.3	50.0	144	3981
3	250	250	24.8	50.0	165	4192
4	250	250	19.2	50.0	156	4616
Grade 8 Reading						
1	250	250	18.7	50.0	100	2075
2	250	250	20.2	50.0	131	3624
3	250	250	25.4	50.0	130	3466
4	250	250	21.4	50.0	111	3426
Grade 4 Mathematics						
1	250	250	20.7	50.0	108	3093
2	250	250	23.3	50.0	144	3842
3	250	250	22.6	50.0	165	4151
4	250	250	21.1	50.0	156	4500
Grade 8 Mathematics						
1	250	250	19.9	50.0	101	2076
2	250	250	22.1	50.0	131	3535
3	250	250	26.3	50.0	130	3601
4	250	250	21.8	50.0	111	3395

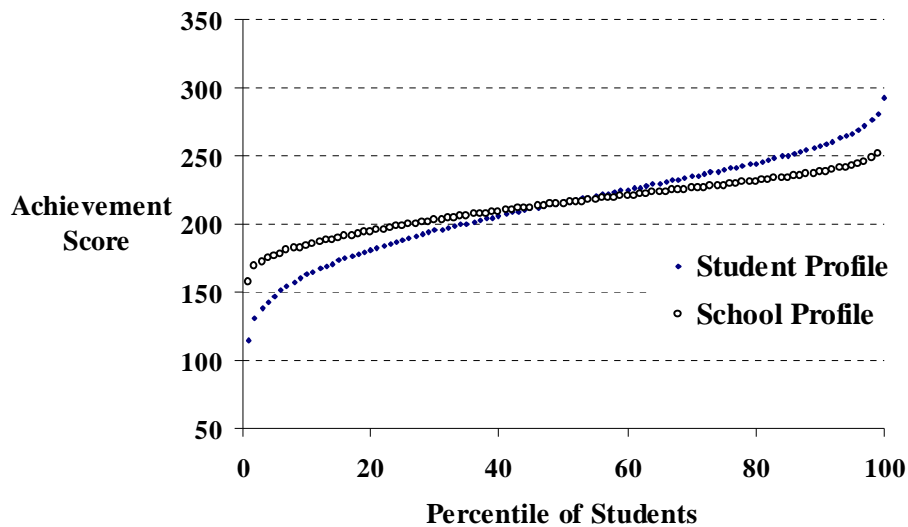
NOTE: Student-level scores are pre-standardized to a mean of 250 and standard deviation of 50 in each state.

SOURCES: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Reading and Mathematics Assessments. Matched state assessment scores.

A population achievement profile for a state can obviously be constructed directly from the distribution of student test scores, but it can also be constructed from school-level averages. By weighting each school's average score by the number of students tested in that school, a school-based population profile can be made to match the student-based profile, but variation within schools is lost.

Comparison of a school-level population profile with the corresponding student-level population profile, based on the same scores, is shown in figure 1. In both cases, the percentiles refer to the student population, because each average school score is weighted by the number of students represented by the school. Thus, compared to the p^{th} percentile in the school-level profile, $(p-1)$ percent of the students are in schools with lower average scores and $(100-p)$ percent of the students are in schools with higher average scores.

Figure 1. Means and standard deviations based on school-level and student-level computations



SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Reading Assessment: Full population estimates.

Although the school-level population profile has the same general shape as the student-level profile, it has much less variation (i.e., it's flatter). Between the 20th and 80th percentiles, the school-level profile varies by only 37 points on the NAEP scale, compared to 63 points for the student-level profile. The median is the same, but in the middle of the population each percentile is only 0.6 points higher than the next lower percentile, compared to 1.0 points for the student-level profile. To emphasize the similarity of the population profile shapes, the same asymmetry between the upper and lower tails of the student achievement distribution can be seen in the both the student profile and the school-level profile: for the student profile, the distance between the 1st percentile and the 21st percentile is 69 points on the NAEP scale (nearly 2 population standard deviations), while the distance between the 80th and 100th percentiles is only 48 points on that scale; and for the school-based profile, the distance between the 1st percentile and the 21st percentile is 38 points on the NAEP scale (about 1 population standard deviation), while the distance between the 80th and 100th percentiles is only 27 points on that scale.

More important than means and variances for analytical purposes are the relations between achievement measures and various factors that might be correlates of achievement. To illustrate the differences in relations observed based on school means versus individual scores, I compare the regression weights estimated for predicting achievement in these four states based on free or reduced price lunch eligibility. The regressions are carried out first based on the school means, as would be possible with the NLSLSASD and then based on student scores; and they are carried out for both state assessment and NAEP to assess whether differences between states are due to differences in the state assessment measure or in the underlying relation between poverty and achievement. Results are shown in table 6.

Table 6. Regression weights for state assessment and NAEP achievement score prediction from free and reduced price lunch data, based on school-level and student-level computations

State	School-level regression		Student-level regression	
	State Assessment	NAEP	State Assessment	NAEP
Grade 4 Reading				
1	-90	-130	-54	-75
2	-142	-160	-86	-94
3	-140	-156	-80	-89
4	-86	-112	-66	-80
Grade 8 Reading				
1	-132	-145	-71	-80
2	-132	-143	-80	-87
3	-147	-172	-86	-99
4	-99	-119	-67	-79
Grade 4 Mathematics				
1	-86	-126	-60	-84
2	-146	-165	-86	-98
3	-112	-169	-69	-94
4	-94	-130	-60	-84
Grade 8 Mathematics				
1	-126	-128	-73	-77
2	-132	-152	-80	-85
3	-149	-182	-80	-106
4	-103	-133	-73	-93

Note: Regression weights are displayed as percentages of a (student-level) standard deviation difference in achievement between free lunch eligible and not eligible students, where reduced price lunch eligible students are scored as .5 eligible.

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Reading and Mathematics Assessments. The National Longitudinal School-level State Assessment Score Database (NLSLSASD) 2004.

A quick glance reveals that the school-level regression weights are larger than their values for student-level regressions.⁸ Generally, the regression weights based on school means are 1.5 to 2 times as great as those based on student scores, for both NAEP and state assessments. The interpretation of this difference is that there is a direct (negative) school-level relation between poverty and achievement, beyond the negative relation for each individual student. The school-level relation may be due to school safety factors, to teachers' pay factors, or to general school resource factors. On the other hand, the negative within-school relation between poverty and

⁸ All of the regression weights in table 6 are statistically significantly different from zero.

achievement may be weakened because it is mitigated by teachers' special attention to problems of children of poverty in their classrooms.⁹

In any case, awareness of the negative between-school relation between poverty and achievement is important for decisions about investment and allocation of educational resources, and that information can be obtained from a school-level database. Of course, to differentiate between indirect relations between community resources and student achievement, mediated by school-level factors, and individual-level relations between poverty and achievement requires a student-level database with accompanying school indicators.

⁹ If there were no direct school-level effect, or within-school mediation, one would expect the two sets of regression weights to be approximately the same. There is no "artifactual" reason for one set of regression weights to be systematically larger than the other.

What are the values and limitations of pooled within-state analyses?

Each state has a different assessment program, which means that scores cannot be compared directly between schools in two different states. The primary design for using the NLSLSASD for federal analytical purposes must therefore be to compare achievement in schools within the same state and then aggregate, or pool, the results of those within-state analyses across states.

The strengths of this approach are (1) that it makes use of achievement scores already collected and (2) that it places a very large representative sample of schools at the disposal of analysts. Given the need to implement expensive true randomized experimental designs in order to find proof of impact of a treatment, it is important to make use of large existing datasets whenever possible to determine which treatments to invest research effort in and to determine parameters of an effective experimental design. The NLSLSASD is designed to facilitate such planning by providing descriptive information about the public schools in all states.

The simplest procedure for conducting pooled within-state analyses is to subtract the average value in each state from all of the observations in that state. That is, all predictors, as well as all outcome variables, are set to a mean of zero in each state. That does not affect correlational analyses within each state (e.g., comparisons of group means, regression analyses, factor analyses, structural equation modeling), but at the same time, it captures no information about covariations of state predictor means with state achievement means.

That is the major limitation of pooled within-state analyses. They are like hierarchical linear modeling without the variation at the higher level. Relations to variables with particularly large variations between states, such as percentages of Hispanic students, are underestimated by pooled within-state analyses.

A solution to this limitation is to replace the zero state mean values on achievement variables with state achievement means on a test that is administered uniformly across states; i.e., the National Assessment of Educational Progress. By creating a synthetic achievement measure whose within-state variation is reflected in state assessment scores and whose between-state variation is reflected in NAEP scores, correlational analyses can be carried out which capture both within- and between-state correlates of achievement.

This is the strategy Gili Drori and I employed in a study of correlates of achievement in public schools participating in the 1994 Schools and Staffing Survey (SASS) (McLaughlin and Drori, 1999). Making use of school-level state assessment scores for 2214 public schools in 20 states previously collected for a NAEP Secondary Analysis Grant, we analyzed the relations between school factors and achievement, controlling for background demographics. In that study a structural equation model was fit both to pooled within-state data and to a synthetic achievement measure. School-level factors included school size (enrollment), class size, normative cohesion, teachers' influence, and school climate. The subject variables were based on combinations of items on the SASS Teacher and Principal Questionnaires. The results of the first (pooled within-state) analysis are shown in table 7, and the results of the second (synthetic achievement) analysis are shown in table 8. The between-state variance estimates were based on the 1994 NAEP reading assessment, which was administered at the state level in grade 4. The NAEP grade 4 between-state variance was extrapolated to grade 8 and grade 12 for the middle and secondary level analyses.

The results of the two analyses are qualitatively the same, with two exceptions. In the equations predicting average student achievement, (1) class size was a stronger and statistically more significant correlate of achievement in analyses which included between-state variation, and (2) school climate was a stronger and statistically significant correlate of middle school achievement when between-state variation was ignored.

Table 7. Pooled within-state SEM associations of organizational and climate factors with student achievement in public elementary, middle, and secondary schools

Independent Effects	Dependent Factor	Elementary (n=1123)	Middle (n=496)	Secondary (n=595)
Student Achievement				
School Size		+0.05	-0.04	+0.28*
Class Size		-0.13	-0.01	-0.17*
Normative Cohesion		-0.04	-0.11*	-0.17
Teachers Influence		+0.00	+0.03	+0.12
School Climate		-0.04	+0.21*	+0.21
R ²		0.63	0.78	0.69
School Climate				
School Size		-0.20*	-0.00	-0.27*
Class Size		+0.17	-0.20	+0.00
Normative Cohesion		+0.24*	+0.36*	+0.22*
Teachers Influence		+0.08	+0.07	+0.25*
R ²		0.63	0.66	0.66
Teachers' Self-Perceptions of Influence				
School Size		-0.10	-0.07	-0.19*
Normative Cohesion		+0.46*	+0.49*	+0.57*
R ²		0.25	0.28	0.44
Normative Cohesion				
School Size		-0.10	-0.11*	-0.25*
R ²		0.13	0.07	0.13
Class Size				
School Size		+0.49*	+0.57*	+0.72*
R ²		0.37	0.34	0.56
Statistical Summary Measures				
GFI (AGFI)		0.97 (0.94)	0.95 (0.91)	0.94 (0.90)
χ^2 (df)		365(100)	270(100)	329(100)

SOURCE: McLaughlin & Drori (1999)

Table 8. SEM associations of organizational and climate factors with student achievement in public elementary, middle, and secondary schools

Independent Effects	Dependent Factor	Elementary (n=1123)	Middle (n=496)	Secondary (n=595)
Student Achievement				
School Size		+0.04	+0.06	+0.32*
Class Size		-0.25	-0.38*	-0.36*
Normative Cohesion		-0.06	-0.01	-0.06
Teachers Influence		+0.03	+0.07	-0.01
School Climate		-0.11	-0.05	-0.08
R ²		0.72	0.87	0.82
School Climate				
School Size		-0.09	-0.14	-0.28*
Class Size		-0.12	-0.18	-0.00
Normative Cohesion		+0.26*	+0.33*	+0.29
Teachers Influence		+0.01	+0.09	+0.13
R ²		0.58	0.62	0.68
Teachers' Self-Perceptions of Influence				
School Size		-0.11	-0.06	-0.20*
Normative Cohesion		+0.39*	+0.35*	+0.31*
R ²		0.22	0.26	0.45
Normative Cohesion				
School Size		-0.02	-0.11	-0.18*
R ²		0.09	0.03	0.09
Class Size				
School Size		+0.47*	+0.57*	+0.63
R ²		0.32	0.42	0.51
Statistical Summary Measures				
GFI (AGFI)		0.96 (0.93)	0.92 (0.87)	0.93 (0.88)
χ^2 (df)		426(100)	395(100)	432(100)

SOURCE: McLaughlin & Drori (1999)

These are results which add a new dimension to the analysis of the Schools and Staffing Survey. This is an example of how a rich school-level database can be linked to achievement scores, providing an extremely cost-effective method for exploring the patterns of correlations between (a) a wide variety of school and teacher factors and (b) reading and mathematics achievement, achievement gains, and achievement gaps. Although these analyses do not, by themselves, prove the effectiveness of any educational interventions, the patterns found will

provide information for intelligent decisions about avenues to pursue to improve achievement in public schools.

Such analyses are not limited to SASS: they can be carried out for any multi-state school-level database, such as the public schools participating in a particular systemic reform initiative. In general, pooled within-state analyses of the relations of state assessment measures of achievement to school-level factors can provide valuable descriptive information about the correlates of achievement in American public schools. By themselves, these analyses may miss important between-state differences, but combining results with NAEP data on between-state differences can fill that gap.

How can state assessment scores be compared between states?

Each state assessment is unique. If third grade students in North Carolina answer 80% of the items on their state reading test correctly and third grade students in South Carolina answer 50% of the items on their state reading test correctly, that tells us absolutely nothing about the comparative reading achievement levels in North Carolina and South Carolina. It may just mean that South Carolina's test has harder items. Even when two states use the same off-the-shelf test, differences in their choices of forms and rules for test administration defeat attempts to compare their results. That was the primary rationale for the expansion of NAEP to provide state-by-state comparisons starting in 1990.

With *No Child Left Behind* came the mandate for each state to define a level of performance (i.e., a cutscore on a test) as a "standard" and to record the percentages of students who meet the standard in each school. Some states set a single standard (e.g., "proficient") in each subject and grade tested and others set two or three standards (e.g., "basic," "proficient," and "advanced"). NAEP was an early leader in this movement, when in 1990 it began to categorize NAEP scores into four categories it named "advanced," "proficient," "basic," and "below basic."

The argument made to justify this was that people could understand what is meant by "50 percent of the students are proficient in reading" better than they could understand "the mean score in reading was 220." Unfortunately, that is a false understanding. The definition of "proficient in reading" is almost always the result of subjective decisions about performance on test items made by small groups of selected experts in paper-filled (not smoke-filled) rooms. In the case of NAEP, the National Academy of Education found the standard-setting procedure to be "fundamentally flawed" (Shepard *et al.*, 1993).

Because each state's standards are set independently, the standards in different states can be quite different, even though they are named identically. Thus, a score in the "proficient" range in one state may not be in the "proficient" range in another state. This leaves state educators to explain to the public that the reason for an apparent difference in student performance across states is merely that each state defines "proficient" differently. States' standards are no more comparable than raw scores. The fact that 80% of third graders in North Carolina meet North Carolina's reading standard and 50% of third graders in South Carolina meet South Carolina's reading standard tells us absolutely nothing about the comparative reading achievement levels in North Carolina and South Carolina. It may just mean that South Carolina set its standard higher.

Using NAEP, it is possible to estimate where the various states' standards fall on the NAEP scale, using an adaptation of equipercentile equating; and one activity in the NAEP State Analysis Project at AIR has been to construct mappings of state standards in reading and mathematics onto the NAEP scale, in order to eliminate the "standards discrepancy" from comparisons between NAEP and state assessment results. In each school selected to participate in NAEP, the point on the NAEP achievement scale which matches the percent reported by the state assessment as meeting the state's standard can be estimated, and these estimates can be aggregated to provide a state-level estimate of the cutscore corresponding to the state's standard, on the NAEP scale.

Each state has set either one or several "standards" for performance in each grade on its reading assessment. Short versions of the states' names for the standards in 2003 are shown in table 9, with the primary standard listed as "standard 3." Comparison of the primary standards on the same scale sheds light on the variations between states in the percentages of students reported to be "proficient," "meeting the standard," or "making satisfactory progress."

Table 9. Short names of state reading achievement performance standards, by state: 2003

State	Standard 1	Standard 2	Standard 3 ¹	Standard 4	Standard 5
Alabama ²					
Alaska		Below Proficient	Proficient	Advanced	
Arizona		Approaching	Meeting	Exceeding	
Arkansas		Basic	Proficient	Advanced	
California	Below Basic	Basic	Proficient	Advanced	
Colorado			Partially Proficient	Proficient	Advanced
Connecticut	Basic	Proficient	Goal	Advanced	
Delaware		Below	Meeting	Exceeding	Distinguished
Dist. of Columbia		Basic	Proficient	Advanced	
Florida		2 Limited Success	3 Partial Success	4 Some Success	5 Success
Georgia			Meeting	Exceeding	
Hawaii		Approaching	Meeting	Exceeding	
Idaho		Basic	Proficient	Advanced	
Illinois		Above Warning	Meeting	Exceeding	
Indiana			Pass	Pass Plus	
Iowa			Proficient		
Kansas	Unsatisfactory	Basic	Proficient	Advanced	Exemplary
Kentucky		Apprentice	Proficient	Distinguished	
Louisiana	Approaching Basic	Basic	Mastery	Advanced	
Maine		Partially Meeting	Meeting	Exceeding	
Maryland			Proficient	Advanced	
Massachusetts		Needs Improv.	Proficient	Advanced	
Michigan		Basic	Meeting	Exceeding	
Minnesota	(2a) Part. Knowl.	(2b) Satisfactory	(3) Proficient	(4) Superior	
Mississippi		Basic	Proficient		
Missouri	Progressing	Nearing Proficient	Proficient	Advanced	
Montana		Nearing Proficient	Proficient	Advanced	
Nebraska			Meeting		
Nevada		Approaching: 2	Meeting: 3	Exceeding: 4	
New Hampshire			Basic	Proficient	Advanced
New Jersey			Proficient	Advanced	
New Mexico		Top 75%	Top half	Top 25%	
New York		Need Help	Meeting	Exceeding	
North Carolina		Inconsist. Mastery	Consistent Mastery	Superior	
North Dakota			Meeting		
Ohio		Basic	Proficient	Advanced	
Oklahoma		Little Knowledge	Satisfactory	Advanced	
Oregon			Meeting	Exceeding	
Pennsylvania		Basic	Proficient	Advanced	
Rhode Island			Proficient		
South Carolina		Basic	Proficient	Advanced	
South Dakota		Basic	Proficient		
Tennessee ²					
Texas			Passing		
Utah ²					
Vermont	Below	Nearly	Achieved	Honors	
Virginia			Proficient	Advanced	
Washington		Below	Met	Above	
West Virginia		Top 75%	Top half	Top 25%	
Wisconsin		Basic	Proficient	Advanced	
Wyoming		Partially proficient	Proficient	Advanced	

¹ Standard 3 represents the primary standard for every state. In most states, it is a criterion for adequate yearly progress.

² Standards-based scores were not available for Alabama, Tennessee, and Utah for 2003.

NOTE: The state standards listed above are those for which assessment data exist in the NLSLSASD.

“Percentile rank,” while not a standard, is needed for comparisons in Alabama, Tennessee, and Utah.

SOURCE: The National Longitudinal School-level State Assessment Score Database.(NLSLSASD) 2004.

By carrying out the equipercntile equating based on the schools participating in NAEP, it is possible to evaluate the accuracy of the equating. Using the NAEP cutscore determined for a state, the percentages meeting the state's standards in each school can be estimated from performance of the NAEP sample of students, and the extent to which these accurately reproduce the percentages reported by the state provides a test of the validity of the equating. One does not expect perfect reproduction of the school-level percentages, due both to sampling variation and measurement error. However, to the extent that NAEP and the state reading assessment are testing different constructs, the error would reduce the validity of the mapping. Therefore, AIR developed a "relative error" measure for the mapping of each state's standard, the ratio of the root mean squared error in reproducing the school-level percentages to the error expected due to sampling and measurement error. A value of 1.0 on the relative error scale corresponds to equating NAEP with itself, and larger values correspond to estimates of reliable differences between the constructs measured by NAEP and a state assessment.

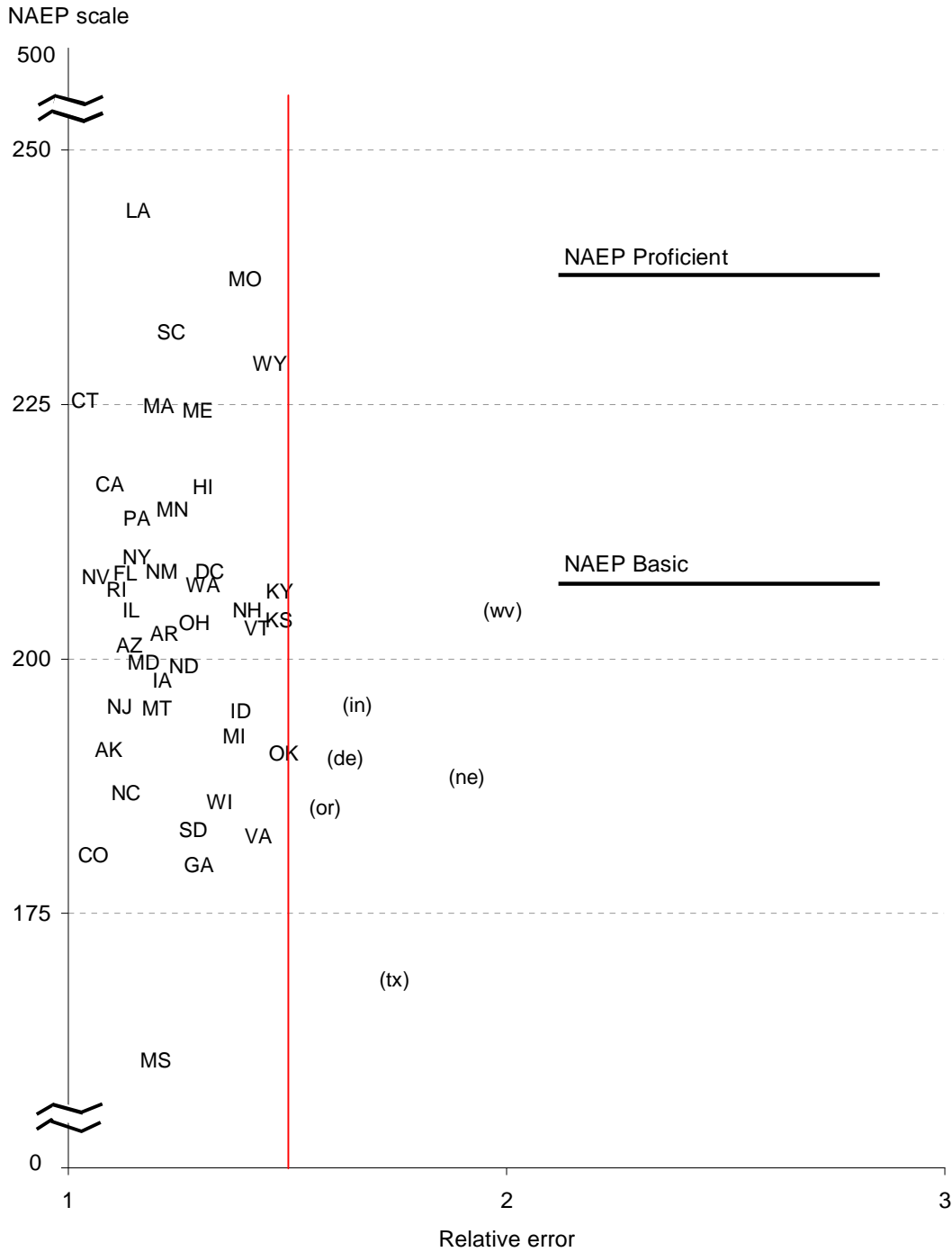
The results for grade 4 reading are shown in figure 2. The variation in reading proficiency standards between the states is so large that they render the term "standard of proficiency in reading" meaningless. Louisiana has set its primary reading standard where only 22 percent of the fourth graders in the nation can meet it, while across the river, Mississippi has set its standard where 87 percent of the nation's fourth graders could meet it.

Figure 2 includes all but three states, which are omitted because the NLSLSASD has percentile rank scores but no standards-based reading scores for schools in Alabama, Tennessee, and Utah in 2003. For all but six of the 47 included states (the six indicated in parentheses in figure 2), the estimated relative error is less than 1.5, indicating that more than 2/3 of the reliable variance in NAEP and state assessment school means is in common; that is, that the mappings of the standards are meaningful. For the six states to the right of the arbitrary criterion line of 1.5, less confidence in the validity of the mappings onto the NAEP scale is warranted, because the amount of error in the scores is larger than would be expected based on sampling and measurement error.

It should be emphasized that the placement of states in figure 2 only shows where their primary reading standards are set. It conveys no information about the relative reading achievement of fourth graders in these states. In fact, there is no correlation between states that set high standards and states that have high average performance on NAEP ($R^2=.0024$) (McLaughlin et al., 2005a).

The bottom line is that percentages of schools meeting standards cannot be compared between states – even though states may be testing very similar constructs of reading achievement, the subjective nature of the standard-setting procedures have created differences between states in the meaning of "meeting the standard" which are very large. To compare scores between states, one must rely on NAEP.

Figure 2 NAEP scale equivalents of primary state reading achievement standards, for grade 4 or adjacent grade, by relative error criterion: 2003



NOTE: Relative error is a ratio measure of reproducibility of school-level percentages meeting standards. The vertical line indicates a criterion for maximum relative error. The six states displayed in lowercase letters to the right of the vertical line have relative errors greater than 1.5; the variation in results for individual schools in these states is large enough to call into question the use of these equivalents.

SOURCE: McLaughlin *et al.* (2005a)

Two issues concerning the analysis of achievement trends

Comparison of achievement in treatment versus control schools is a major potential use of the NLSLSASD. The credibility and validity of those comparisons depend on both the quality of the data and the appropriateness of the analytical methods used. Some differences between analytical methods ordinarily have only minor effects on outcomes but others frequently have substantial effects. In the course of working group meetings sponsored by the U.S. Department of Education in 2001 and 2002, two important issues were debated with respect to these analyses: (1) how trends should be computed, and (2) how demographic information about student populations should be taken into account.

Definitions of Trends

Two issues arose with respect to the definition of trends. The first issue applies to two-year gains. The question is whether the first year score should simply be subtracted from the second year score to obtain a gain measure or whether the first year's measure should be used as a covariate in analyses of the second year's measure. The two primary arguments for the latter approach are that "gain scores are unreliable" and that analysis of covariance empirically estimates the optimal weight to be used in subtracting the first year's score from the second year's score. Both of these arguments must be rejected, however. First, because school-level scores are correlated across years by about .70, gain scores are not unreliable. Second, computations which use the first year's score as a covariate are based on the assumption that it is measured without error, but the first year's score generally has about the same amount of error as the second year's score. The bias caused by this departure from assumptions usually leads to underestimation of gains and overestimation of the sizes of coefficients related to other variables that are correlated with achievement. The bottom line is that the dependent variable for analyses of achievement gains should be the gain, computed as a subtraction of the first year's score from the second year's score.

When the scoring of a test is changed between two years, a simple subtraction would, of course, be likely to yield an incorrect answer. In that case, each year's scores should be transformed to normal scores (i.e., by transforming the rank order (percentile) of each school into the corresponding normal deviate before subtracting one year's score from another year's).¹⁰

The second issue concerns trends over more than two years. Standard methods for trend analysis parameterize trends as linear functions of time. These analyses yield estimates of the rate of change per year. In reality, although demographic changes in school populations may be gradual, changes in school administrations and introduction of reforms and other interventions are generally discontinuous events, occurring in particular years. It is likely to be much more informative to report the amount of improvement in particular years, relative, say, to the funding for a program, than to report linear rates of gain. For example, with scores for years 0, 1, 2, and 3 after funding, it is useful to measure gains from year 0 to year 1, year 1 to year 2, and year 2 to year 3. It is also useful to measure gains between year 0 and year 3, but without including the intermediate years' data in the analysis and without inferring that the gains represent an average rate of gain per year for three years.

¹⁰ Analyzing the percentile ranks directly without the normalization will, in almost every case, yield the same conclusions as analysis of the normal scores but will not give as much weight to differences in the extreme percentiles as the analysis of normal scores would.

Demographic controls

As mentioned in the introduction, without randomized assignment of schools to treatment and control conditions, the only information that can be provided by the NLSLSASD is correlational, not causal. However, one way to rule out particular alternative causal explanations is to select control schools which are like treatment schools on one or more observable covariates, such as the percentage of students who are eligible for free or reduced price lunch (i.e., a poverty indicator). The control can be in terms of selection of a particular “control” school to match each treatment school or in terms of statistical adjustments. Each of these methods has a disadvantage: for matching, it is generally impossible to find exact matches; and for statistical adjustments, deviations from the model assumptions can bias results. Combining the two methods, by matching categories of similar schools and then applying statistical adjustments within the categories, is preferable. In any case, analysts must take care not to overinterpret their results, because both statistical matching and statistical adjustments are based on observable measures, which should not be expected to completely eliminate the effects of a latent trait such as community poverty on student achievement.

One approach to solving this problem is to create an outcome variable that is largely insensitive to demographic effects. For example, poverty is a very strong correlate of average achievement scores across schools, but it is only weakly correlated with gains from one year to the next.

If a demographic adjustment is needed, it is useful to create a statistic which is readily interpretable by non-statisticians. In some conditions this can be achieved by using the set of control schools to estimate what the achievement measure would be in a school with any particular profile of covariates. The profiles of covariates can then be used to estimate an expectation of what achievement might have been in treatment schools if they had not had the treatment. Subtracting the expected achievement from the observed achievement in each treatment school yields a residual achievement effect that may be a correlate of the treatment-control difference (i.e., an “effect” of treatment). Of course, limitations of this procedure exist due to (a) the incompleteness of the adjustment due to omitted variables, (b) errors in measurement of the observed covariates, and (c) differences between treatments and controls in the distribution of the covariates and in the functional relation between covariates and achievement. As an example of the last of these problems, if the natural relation between poverty and achievement is stronger (i.e., the curve steeper) at low levels of poverty than at high levels of poverty, and if all treatment schools have poverty levels higher than almost all control schools, then the procedure will overestimate the positive “effect” of the treatment when controlling for poverty in this way.

In general, analyses of the data should start from questions that need to be answered, provide summary statistics that are as sharply relevant to the questions as possible, and point out how what is not known which could change the results and modify the interpretations of the data.

What additions to the database would increase its value?

In the preceding sections, I have tried to provide some information about the kinds of things for which school-level state assessment scores are useful and about issues of how these data should be analyzed. In this final section, I describe four types of information which, if available, would significantly increase the range of valid analyses which could be carried out using these data.

Fiscal information

Addition of information about expenditures at the school level would facilitate analyses to determine the extent to which either infusion of funds or draining of funds is correlated with student achievement. Presently, the F-33 survey provides information about expenditures at the school district level, but not at the school level. Thus, at present it is necessary to aggregate school-level achievement data in the NLSLSASD to the district level in order to estimate the correlation of funding and achievement at the district level. However, this is likely to be insufficient for policy purposes because of the substantial variation of resources and other conditions for learning between schools within districts, especially within large metropolitan districts.

Several cautions need to accompany the addition of school-level financial data to the database. First, analyses of the cost of educational resources need to be carried out, to adjust for the fact that the same dollars purchase differing amounts of educational resources in different locales. Second, interpretations of analyses of the relations between funding and achievement must be tempered with the understanding that these relations have alternative causal explanations. Among these are (1) the fact that parents in affluent neighborhoods often both provide more educational resources in their homes than parents in other neighborhoods, while also providing more financial support for their children's schools; and (2) the fact that schools with high-achieving students are more attractive to teachers and may therefore be able to hire and keep the premium teachers without offering premium salaries. Third, resources for schools come from variety of sources, and attempts to interpret relations between funding and achievement as if those alternatives were not available can lead to conclusions which are not helpful to schools.

Teachers' qualifications

No Child Left Behind aims to collect information on the qualifications of teachers in the nation's public schools, as a way to estimate how important variation in that factor between schools is in explaining variation in achievement among schools. Student-teacher ratios have long been available for public schools, but there is no systematic database of teacher qualifications at the national level. Information about teacher qualifications would enable a variety of new analyses of the NLSLSASD.

Accommodation validity

Since 1995, there has been a dramatic increase in the availability of testing accommodations for students with disabilities and English language learners, such as one-on-one testing and extended time. Unfortunately, there has not been an accompanying body of research on how to score accommodated tests so that their reading and mathematics achievement scores are comparable to reading and mathematics achievement scores of students who are not offered the accommodations. In many cases, states combine accommodated with non-accommodated scores in reporting school averages, and the ranks of schools in a state may be affected by the percentages of their students who are offered testing accommodations.

As long as some students receive testing accommodations and others do not, there is a need both for rigorous reporting of testing accommodations provided for each student in state

assessments and for rigorous research on the appropriate scoring rules for accommodated test performance. Accommodations are meant to eliminate barriers to the demonstration of achievement in subjects like reading and mathematics, but it is difficult to develop testing accommodations for students with learning disabilities which remove the testing barriers, or lack of enabling skills, while leaving the substantive, or target, domain of the test unchanged. Ultimately, this research requires clarification of the definitions of achievement domains, so that results of accommodated testing can be scored in a manner that corrects for alterations in substantive domain requirements while not affecting the enabling function of the accommodations.

Small sample suppression

For confidentiality reasons, as well as in recognition of the unreliable nature of averages based on one or two or three students, state assessment programs suppress scores (a) in very small schools and (b) for very small subgroups within schools. This is especially important for the measurement of achievement gaps between two demographic groups, one of which is much smaller than the other. Although in most states, more than 90 percent of Black, Hispanic, and economically disadvantaged students are included in published state assessment results, in a few states the coverage is less than 70 percent of the minority students, due to small sample suppression.

There are four major alternatives to address the suppression problem: (1) to contract with states to provide the data in a confidential manner, without suppressing any small sample averages, (2) to call for the combination of small sample data into larger groupings, such as across grades or across years, (3) to apply a randomization rule which would guarantee that while most small sample figures would be accurate, at least a few would be randomly altered, assuring that no student's score would be definitely identifiable, or (4) to design analyses to adjust for suppression, imputing likely scores from other information and carrying out sensitivity analyses to determine the extent to which conclusions might be affected by suppression.

Reference

- McLaughlin, D., Bandeira de Mello, V., Blankenship, C., Chaney, K., Hikawa, H., Rojas, D., William, P., & Wolman, M. (2005a) *Comparison between NAEP and state reading assessment results: 2003*. Report under review by the National Center for Education Statistics.
- McLaughlin, D., Bandeira de Mello, V., Blankenship, C., Chaney, K., Esra, P., Hikawa, H., Rojas, D., William, P., & Wolman, M. (2005b) *Comparison between NAEP and state mathematics assessment results: 2003*. Report under review by the National Center for Education Statistics.
- McLaughlin, D. & Drori, G. (1999) *School-level correlates of academic achievement: Student assessment scores in SAS public schools*. Washington, D.C: U.S. Department of Education, (NCES 2000–303)
- Shepard, L., Glaser, R., Linn, R. & Bohrnstedt, G. (1993) *Setting performance standards for student achievement*. Report of the National Academy of Education Panel on the evaluation of the NAEP Trial State Assessment: An evaluation of the 1992 achievement levels. National Academy of Education, Stanford, CA.