

Comparison of State Elementary School Mathematics Achievement Standards, Using NAEP 2000

Don McLaughlin
Victor Bandeira de Mello

American Institutes for Research

Presented at the American Educational Research Association
New Orleans, April 2002

Slide 1

Victor Bandeira de Mello and I and our colleagues have been developing a national school-level state assessment score database. This database now has scores for more than 80,000 public schools, in 49 states, plus the District of Columbia and Puerto Rico.

Slide 2

The database includes scores for various years in different states, but for most states we have scores for the 1998-1999 and 1999-2000 school years, and we are in the middle of adding scores for the 2000-2001 school year. We have added Federal program identifiers for the schools, and with the approval of NAEP, we have matched the schools that participated in the 2000 grade 4 State NAEP mathematics assessment with the corresponding schools. Generally, we have matched about 100 schools in each of the 40 states that participated in the NAEP assessment. These represent a random sample of schools in each state.

We have explored the use of these matched schools to create a scale for comparing mathematics achievement standards in different states.

Slide 3

Turning right to the results, before getting mired down in the methodological details and sources of error, we can see how several states' standards compare to each other, as well as to the NAEP achievement levels. This slide shows the range of NAEP performance from 220 to 290. 220 represents average grade 4 NAEP mathematics performance, and the standard deviation of NAEP performance is 30 to 40 points.

Maine's "exceeds the standard" is the most difficult of the standards to reach, according to this comparison, followed "advanced" in Louisiana, Wyoming, Massachusetts, Kansas, Missouri, and South Carolina. These same states' "proficient" standards cluster around the NAEP "proficient" standard, along with Georgia's and New York's "exceeds the standard" and Rhode Island's "meets the standard."

You may notice the vertical red line. The x-axis in this chart shows the "standard error" of the estimate of each state's standard. (The "standard error" is not an error by the state in setting its standard but the amount of error we introduced in placing these standards on the same scale.) We computed the "standard error" as the standard deviation of our estimates of the standard based on each matched school. Thus, for

example, Maine's standard is estimated based on some schools to be 9 points higher than 277 and on others to be 9 points lower than 277. As an average for the state, the standard error of the mean would be on the order of 1 NAEP point in most cases.

Based on some calculations I will get to, I came to the conclusion that there is a threshold standard error near 12 and that estimates with greater standard errors (to the right of the red line) are not very reliable. Lack of reliability is mostly a function of limitations of our database, although it can also be a function of the reliability and communality of the tests being compared.

The names applied here to these standards are based on four steps of transformation – (1) state assessment webpage translations of the levels; (2) our staff's reading of the state assessment webpages; (3) encoding the standard names into SAS variable labels; and (4) my shortening for the purpose of including many on one page. There are a wide variety of names, in addition to basic, proficient, and advanced, including California's PR75, the percent of students achieving the 75th percentile on the national norms for the test. Moreover, these are 2000 norms, and some states are in the process of defining or redefining standards. I apologize for any mislabeling that has resulted and welcome suggestions for improving the accuracy of these short names.

Most of the standard errors in this chart are to the left of the red line.

Slide 4.

At the lower levels of achievement, the standards are much less well-defined. This may be due to lower reliability of NAEP scores in the lowest range of the scale, but it is just as likely to be due to lower reliability of the state assessments at these levels. In fact, one standard included in this chart, Massachusetts's "tested," is just an indication of the percentage of students for whom inclusion in their assessment was appropriate.

Large standard errors do not necessarily indicate that NAEP is testing something different from the state assessment. For example, the placement of the Texas "passing" standard is due to the fact that that standard is relatively easy, yielding a highly skewed distribution of percentages, with many above 90 percent. Using a more stringent performance standard on the same Texas (TAAS) scale, which is not in our database, would undoubtedly yield closer matches to NAEP.

Slide 5

Putting these two slides together, we have a display of all of the standards in our database for grade 4 mathematics in 2000. These are not all of the states with standards, of course. Some have tests in grades 3 and 5, and others have standards defined in ways different from the percent of students achieving a score higher than a specified cutpoint.

Slide 6

What are achievement standards?

Achievement standards are a way to make test results "meaningful" – *if* we know what kids should know and be able to do. This becomes a partitioning of the test score distribution into regions of scores that we say "meet the standard" or "not". This is

usually done in terms of a student's score, with the school-level statistic being the percentage of students who "meet the standard."

The school itself then can have a standard for what is a satisfactory percentage of students meeting the student standard. In some states, school standards are stated in terms of a score, such as a national percentile, that 50% of the students should achieve. For example, "a median score at or above the 60th percentile" is called "exemplary" in TN, "excellent" in UT.

Some states are focusing their standards more on how much students learn each year rather than on their accumulated knowledge and skills. This "value-added" approach sets standards in terms of whether test score gains from one grade to the next are sufficient to ensure that students do not fall further behind each year and that students who are behind move upward, toward the middle of the score distribution. It is impossible to relate these two types of standard, "rate of learning" and "level of achievement," without a comprehensive theory of cognitive skill growth.

Why do different states have different achievement standards?

(1) Because each state is responding to a variety of pressures to set standards, and the accepted methodology is to set standards in terms of test scores, and states use different tests. Because "teaching to the test" with standardized "multiple choice" tests is considered bad form, states are left with no alternative except to develop their own standards.

(2) Because there is a long history of distrust of national standards for educational achievement, states would not accept a common standard if it were available.

In any case, because standard setting procedures are sensitive to a wide variety of factors, including the values and experiences of the individuals participating in setting the standards, it is not surprising that the results might be quite different in different states.

Why would we want to compare the achievement standards in different states?

As long as states are not competing with each other, there is no compelling reason to compare achievement standards of different states. NAEP, based on random samples, provides periodic estimates that can be used to compare achievement, and these comparisons have, for the past twelve years, generally been considered fairly low stakes. With the new Federal focus on improving achievement in schools across the country, this is changing. State education agencies are developing long-term plans for school reform based on identifying schools "that don't meet standards," helping them, and if the help is ineffective, taking severe actions. The difficulties that states are likely to find themselves in implementing these plans may well be related to how their standards compare to the standards in other states. The farther current achievement is from meeting the standards, the greater will be the pressure and the frequency of failure and intervention.

How can we compare the achievement standards in different states?

The trick is to find a common test, related to the tests in the different states, that is administered to a representative random samples of students in each state. NAEP comes to mind. In this presentation, we focus on school standards that are defined in terms of the percentage of tested students with scores above some threshold.

Slide 7

The method of generating the scale position for a standard based on a single school is to start with the percent achieving the state's standard in the school (say 60 percent) and find the place on the NAEP scale where the same (60) percent of the performance distribution is above. Using standard NAEP files, we do this separately for each of five plausible values and compute the mean of those values.

Note that different state schools, with different means and variances, are treated as equivalent if they have the same (say, 60) percent of students achieving the standard.

Slide 8

If NAEP and a state assessment are testing similar mathematics skills, then the schools that score higher on the state assessment should be the same as the schools that score higher on NAEP. In fact, we should get the same estimate of the standard, except for random error, from different schools. If one school has 60 percent meeting the standard and another has 40 percent meeting the same standard, then we would expect that there would be a particular NAEP scale value that was exceeded by 60 percent of the NAEP sample in the first school and by 40 percent of the NAEP sample in the second school.

Slide 9

On the other hand, for various reasons, we may have three schools, all with 60 percent achieving the state's standard but with differing NAEP distributions. That will yield three different estimates of the standard. This may be due to measurement error or to differences between the NAEP sample in a school and the set of students tested on the state's own assessment; or it may be due to fundamental differences between the sets of skills required to do well on the state's test and on NAEP. In any case, when this error is small, it lends credence to comparisons based on the mapping onto the NAEP scale, and when it is large, it detracts from the credibility of those comparisons.

Slide 10

I would now like to turn to several issues in the comparison of standards between states.

What are the sources of error in comparisons?

The sources of error in the comparisons are of two types: measurement error and population representation (or sampling) error.

Measurement Error: There are two kinds of measurement error, systematic and random. Are the tests testing the same thing, and are the tests reliable at the school mean level? If they are not testing the same thing, then there will be large amounts of error in the association of one or the other of them with NAEP. However, even if they are testing the same thing, they, and NAEP are not perfectly reliable. Certainly they are more reliable as measures of a school than they are as measures of individual students, but some random error remains.

Population Representation Error: Is the NAEP sample representative of the school? One concern about the relations between NAEP and state assessments is the difference in exclusion rules. To avoid that issue, I included the "excluded" NAEP students for these analyses, using a full-population representation algorithm to impute the

performance of excluded students with disabilities or limited English proficiency. Another question is whether the NAEP sample is large enough to reduce sampling error to a reasonable level. For some states, we only have data in grades other than 4 or years other than 2000. Does the year make a difference? Does the grade make a difference?

How accurate are the comparisons?

If we ignore the lower level standards, the average standard deviation is on the order of 10 points on the NAEP scale. That means, of course, that estimation for individual schools is too unreliable to report, but if we divide by the square root of the number of schools in a state's NAEP sample, the average standard error is on the order of 1 NAEP point.

To give some context, I tried several simulations.

Slide 11

As a test of the amount of sampling and measurement error involved in using NAEP samples, we split the NAEP sample in half in each school, used one half to compute the percentage achieving a standard (we picked 213, the basic cutpoint), and used the other half to estimate the scale value that would be achieved by that percentage. The standard errors, when adjusted for the half-sample inflation, were in the range from 6 to 12, similar to the range observed for many of the higher level standards.

Slide 12

To estimate the impact of NAEP measurement error, we used the NAEP sample to estimate itself, without splitting the sample in two. The only variability in this case is due to variability in the distributions of the five plausible values. As you can see, there is very little variability in school-level statistics due to the NAEP measurement factor alone.

Slide 13

The average standard error for standards that map onto levels above 200 on the NAEP scale is 10.9. If we just considered error due to NAEP measurement error, as represented by variation among plausible values for an individual, the error is tiny (1.3 points). However, the simulation that combined NAEP measurement and sampling error yielded a standard deviation of 7.6 (actually twice that, but an adjustment for the half-sample size was employed). The difference between 7.6 and 10.9 is the effect of the fact that the state assessment and NAEP are not perfectly parallel.

10.9 is noticeably smaller than the value that would have been obtained if the tests were completely uncorrelated (19.8); and it is smaller than the overall standard deviation of NAEP school means.

Slide 14

How closely correlated are the state assessment and NAEP school means?

Many of the correlations are greater than 0.70, but many are substantially lower. It is clear that some correlations are lower because they are at extreme points of distributions, such as the lowest level in Massachusetts and the highest level in Missouri.

Others are low because of limitations of the data in our database. In particular, the Nebraska results are questionable, because we were not able to include a uniform set of scores for all schools in Nebraska in the current version of the assessment database. The reasons that others are low are yet to be studied.

In order to determine the extent to which low correlations are a function of the NAEP samples in the states, we again used the half-sample simulations.

Slide 15

We can compare the correlations of NAEP with state assessments with correlations of NAEP with itself. (The correlations with state assessments are the maxima in each row in the preceding table.) We see that the ratios of the state-NAEP correlations to the best they could be expected to be, indicated by the NAEP-NAEP correlations, are greater than .85 for most states. These correspond to standard errors of 12 or less (the redline in the initial charts). Based on these results, I would not attach any substantial meaning to the results I showed for Nebraska, Texas, and Michigan; and the results for Missouri, South Carolina, and Vermont are on the borderline.

How closely aligned are the assessments?

There is a difference between alignment and correlation. Two tests that are reasonably highly correlated with a third may not be very highly correlated with each other. However, if there is reason to believe, due to similarity of content, that the two high correlations represent the same communality, then the correlation is more likely to be “transitive”. We would not use NAEP **math** to set up a scale for comparing state **reading** standards, even with high correlations. On the other hand, two tests considered closely aligned, but not highly correlated, cannot be linked to a common scale. Both correlation and alignment are necessary for a strong linkage.

Why are some standards higher than others?

Clearly, there is variation in the performance standards set by different states. That might be due to setting standards more stringently in states where performance is already high, in order to create a challenge for students and teachers everywhere. Or it might be due to the various random fluctuations and distortions introduced by faulty standard-setting methods. We found that when we correlated the NAEP state means with the points at which we estimate that the states set their proficient levels, there was little correlation. That is, states with higher average achievement were not setting standards higher.

Slide 16

On the other hand, the percentages of students who achieved the proficient (or related) standards in different states were clearly related to the place where the standards were set: the harder the standards are, the fewer are the numbers of student meeting those standards. All of these 16 standards might be considered to represent “proficiency,” but they clearly represent different definitions of proficiency. There may be more variation in standards than there is in student performance over states.

What about state assessment standards for other grades?

Some states had grade 3 or grade 5 math assessments in 2000 but no grade 4 assessment. In others, all we have are aggregate elementary school percentages. We can carry out the same computations with those data, and they should yield the same results, *if . . .* there is no systematic difference across grades and years. That is, if schools in which 60 percent of students achieve the standard in grade 4 are the same as schools in which 60 percent of students achieve the standard in grade 5, then either grade could be used for the link to grade 4 NAEP. For states in which there is only a grade 5 score, we cannot test this, but in states with both a grade 4 and a grade 5 score, we can.

Slide 17

Here we see the values set for the middle California standard, based on 9 different sets of state assessment scores (and one set of NAEP scores). Unfortunately, from a statistical perspective, but fortunately from a policy perspective, these figures show systematic variation in the estimate of the standard.

The estimates are lower in later years, meaning that the state test became easier for the students in later years (i.e., performance was improving). The estimates were lowest in grade 3, indicating that performance in grade 3 was relatively higher than in grades 4 and 5. Conceivably this could be due to the lower class sizes in grades 3 and below (but not 4 and 5) in California.

A note about the standard errors: they are smallest for the population that included the same students in the same grade, but not by very much.

To verify the pattern of scores in California, we can compare the percentages of students in the 9 populations achieving the standard.

Slide 18

Comparing the columns, there was about a 9 percent increase in meeting the standard between 1999 and 2000, and another 2 percent increase from 2000 to 2001. Comparing the rows, 6 percent more third graders were achieving the national median than fourth or fifth graders.

Nevertheless, here are estimates for states that had grade 5 assessments or for which we only have aggregate elementary school scores.

Slide 19

The pattern for grade 5 is generally similar to that for grade 4, with the exception that the standard errors are higher, and we have a slightly different mix of states.

Slide 20

For a few states, all we have are aggregate elementary school percentages meeting standards.

Slide 21

Finally, for a few other states, we have standards defined in terms of the percentile (or other score) achieved by half the students in the school. Montana's "advanced" level (the 76th percentile) is far above its "proficient" level (the 45th

percentile); but Tennessee has four levels very close to each other (ranging from the 45th to the 60th percentile).

Is there a value in setting standards higher?

The basic assumption in standards-based reform is that setting challenging performance standards will contribute to the improvement of education. That assumption can be empirically tested, and a next step in the comparison of standards will be to find out whether gains are greater in states that set higher standards, either absolutely or relative to their students' current levels. It is quite possible that providing reports of test performance has its greatest motivating effect if the standards are set at a point where percentages achieving standards are most sensitive to instructional enhancements, whether those standards are stringent or moderate.

Slide 22

In summary, these are the major points I want to leave you with.

States have very different mathematics standards.

Comparison of standards is possible, using NAEP.

The higher the standard, the lower the percentage of students achieving it.

Errors exist in comparisons, due to measurement error, differences in skills covered, and differences in populations assessed.

The amount of that error can be estimated by observing the variability of estimated standards across schools.

The standard error for higher level standards is roughly 10 points on the NAEP scale (0.3 NAEP standard deviations); and it is greater for lower level standards.

Comparison of State Elementary School Mathematics Achievement Standards, Using NAEP 2000

Don McLaughlin
Victor Bandeira de Mello
American Institutes for Research

American Educational Research Association
New Orleans

April 2000

Database for Comparing Mathematics Achievement Standards

State Assessments

**National School-Level State Assessment Score
Database**

**80,000 Public Schools, in 49 states, DC, and
Puerto Rico**

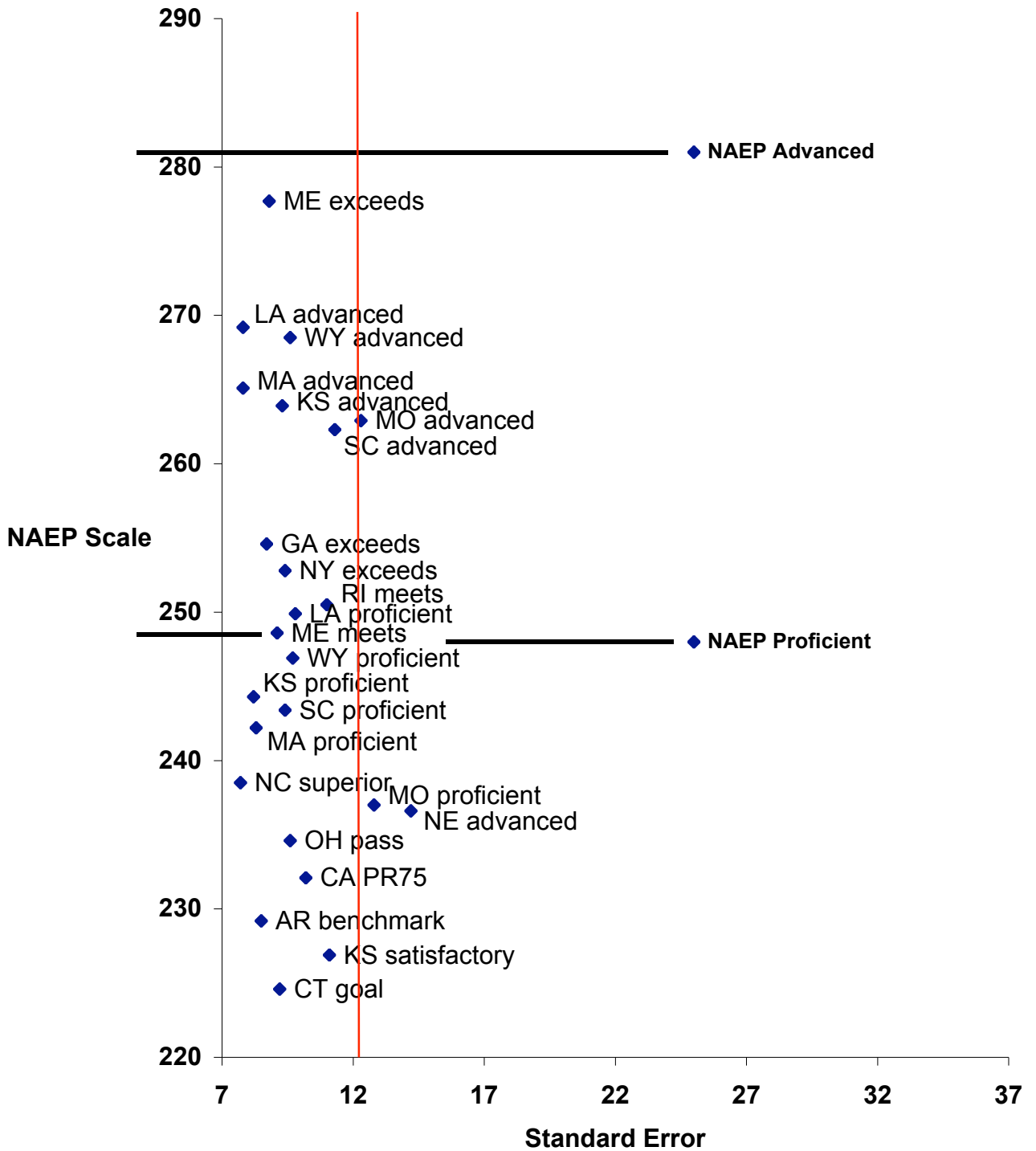
**1998-1999, 1999-2000, 2000-2001, and some
earlier years**

NAEP

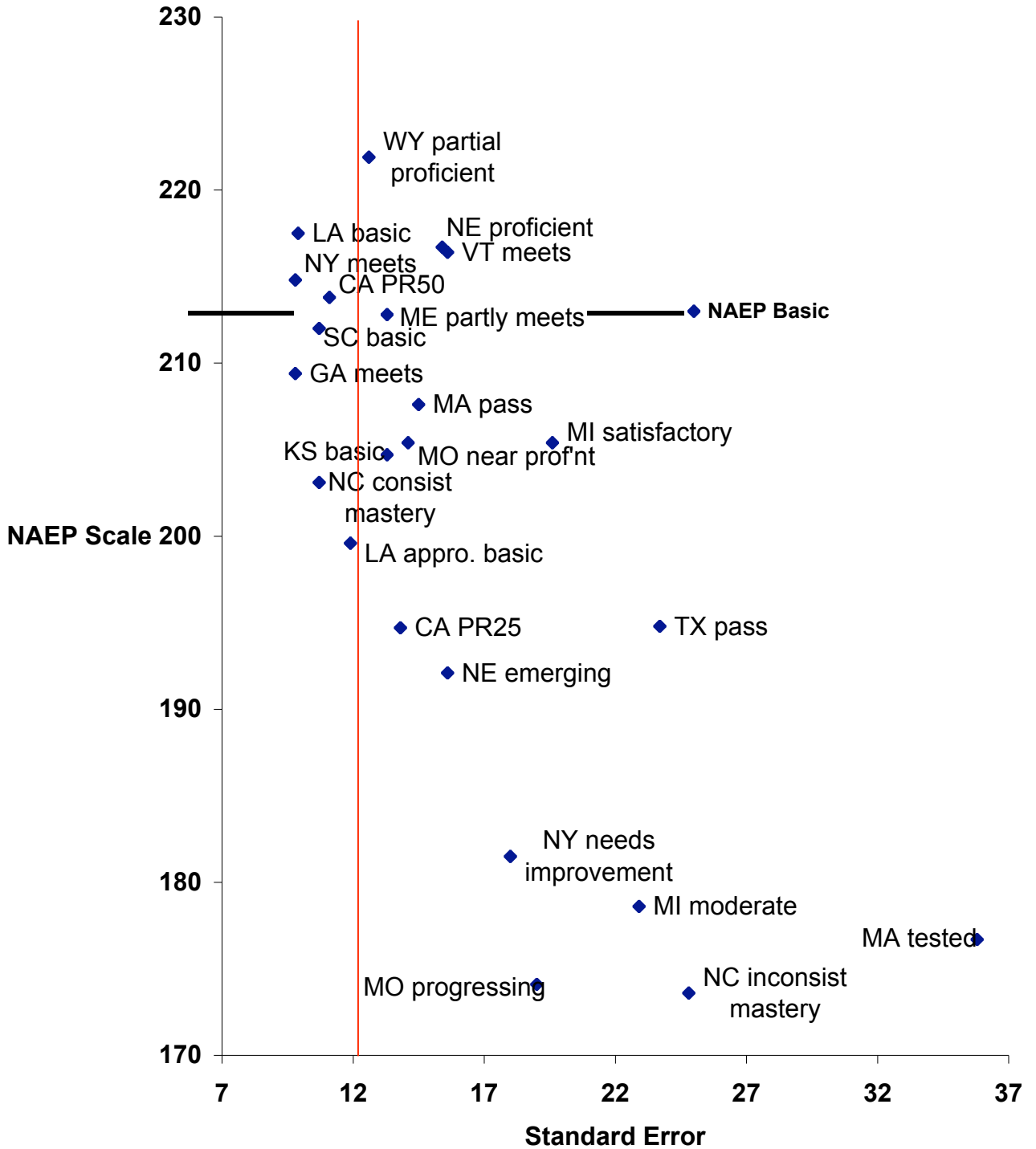
**Grade 4 Mathematics Assessment in 2000, in
40 States and DC**

**Approximately 100 Schools Matched to the
State Assessment Database in Each State**

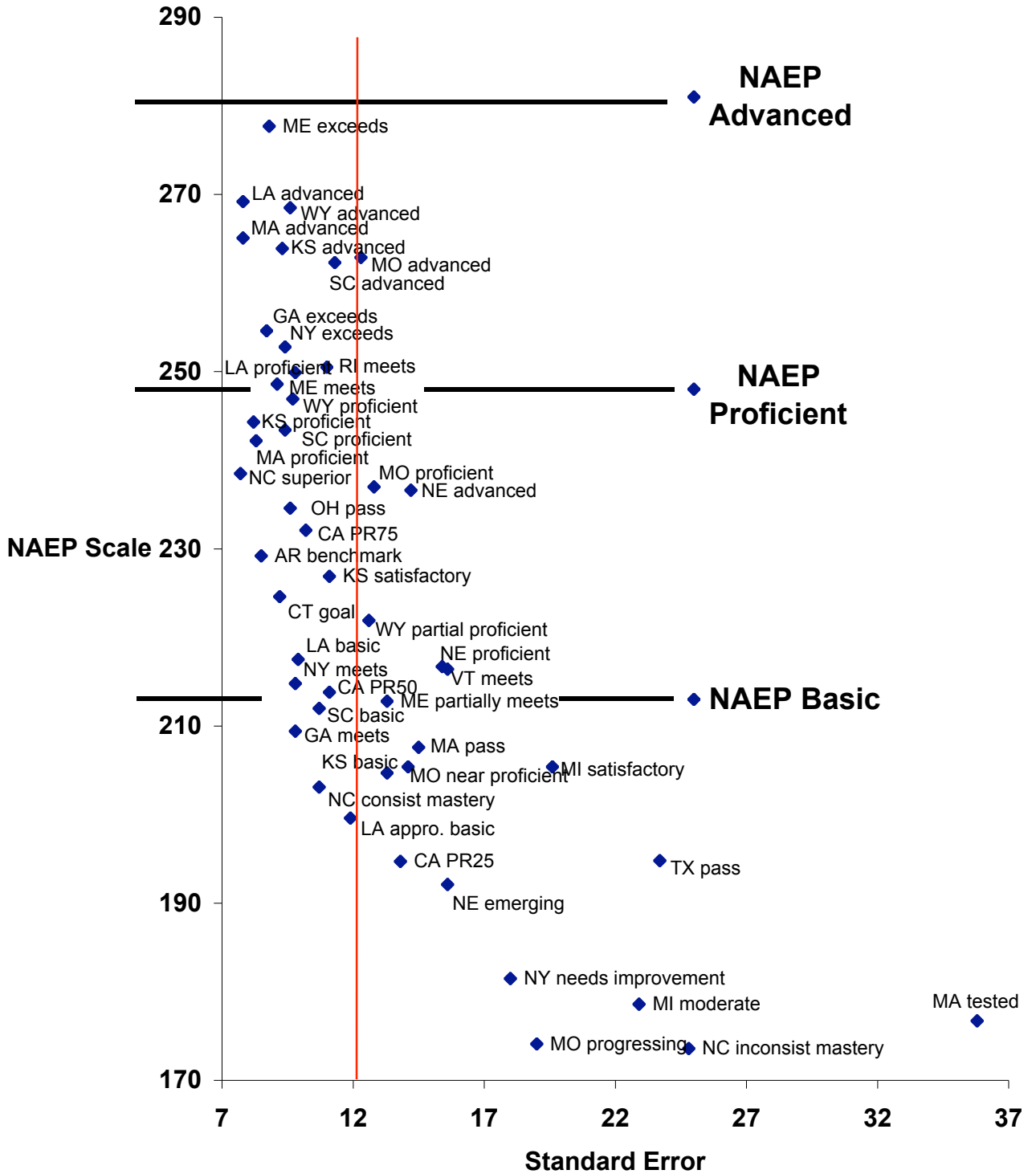
NAEP Equivalents of Math Standards: (Higher Levels) Grade 4, 2000



NAEP Equivalents of Math Standards: (Lower Levels) Grade 4, 2000



NAEP Equivalents of Math Standards: Grade 4, 2000



Basic Questions about School Achievement Standards

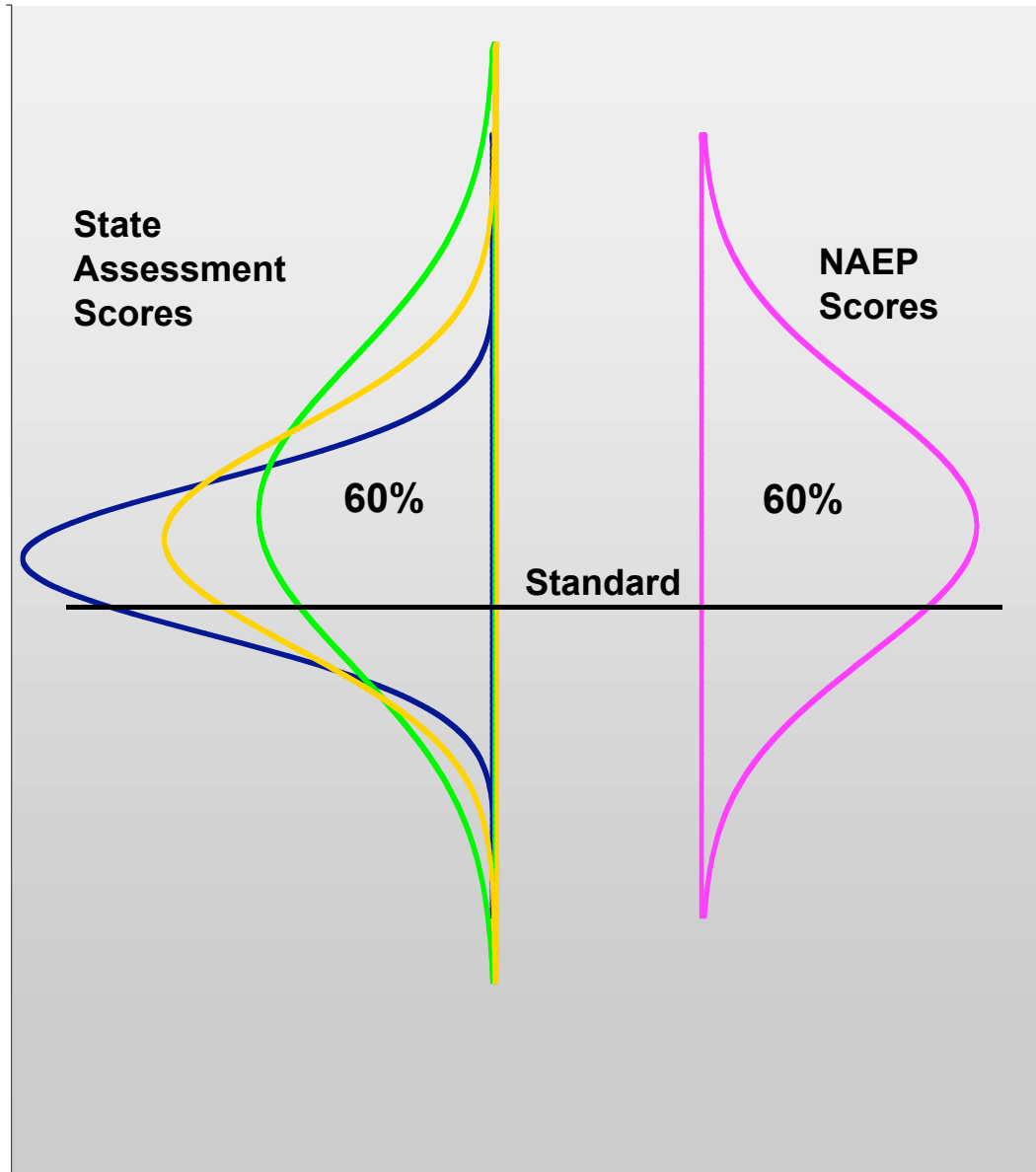
What are achievement standards?

Why do different states have different achievement standards?

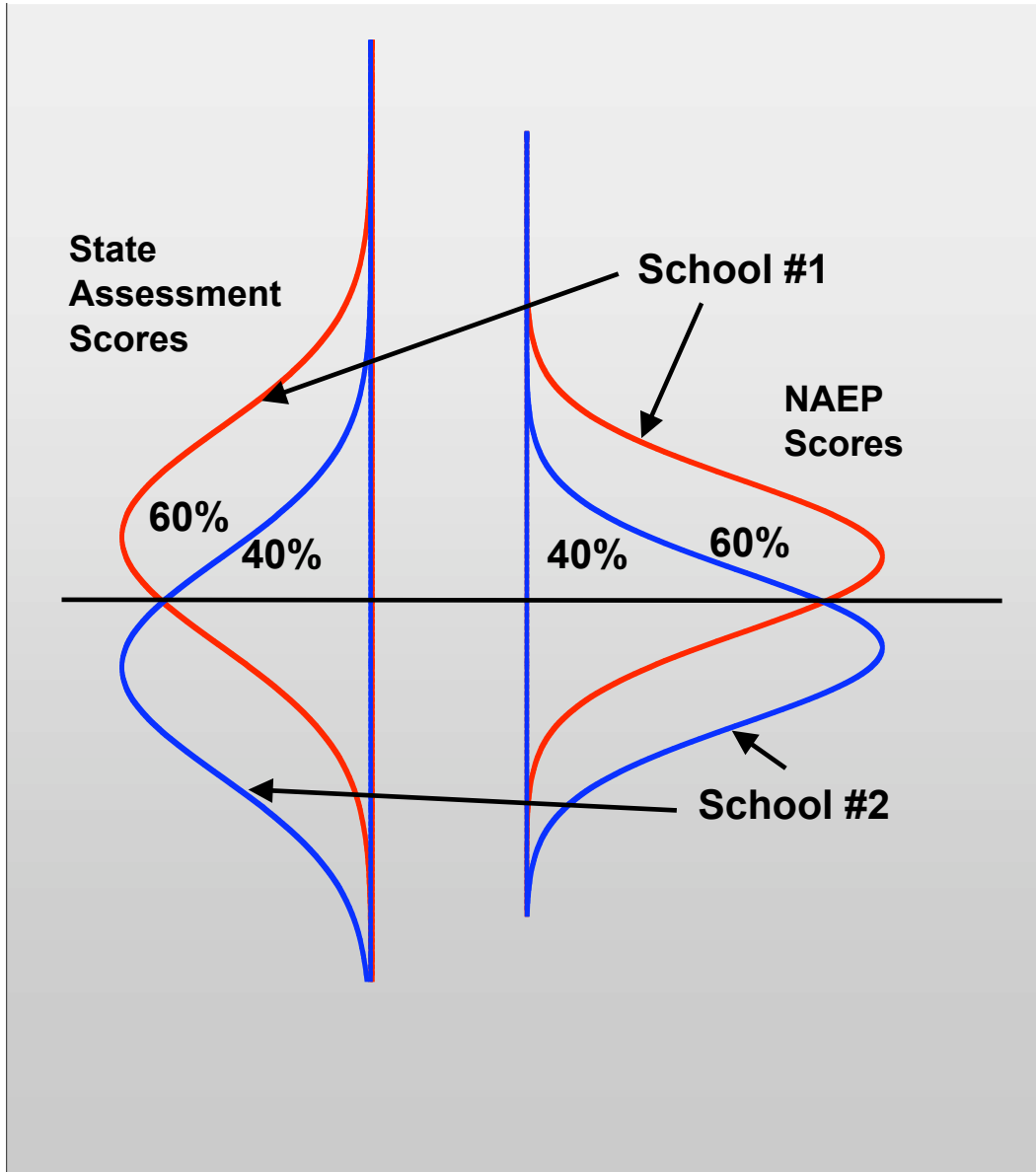
Why would we want to compare the achievement standards in different states?

How can we compare the achievement standards in different states?

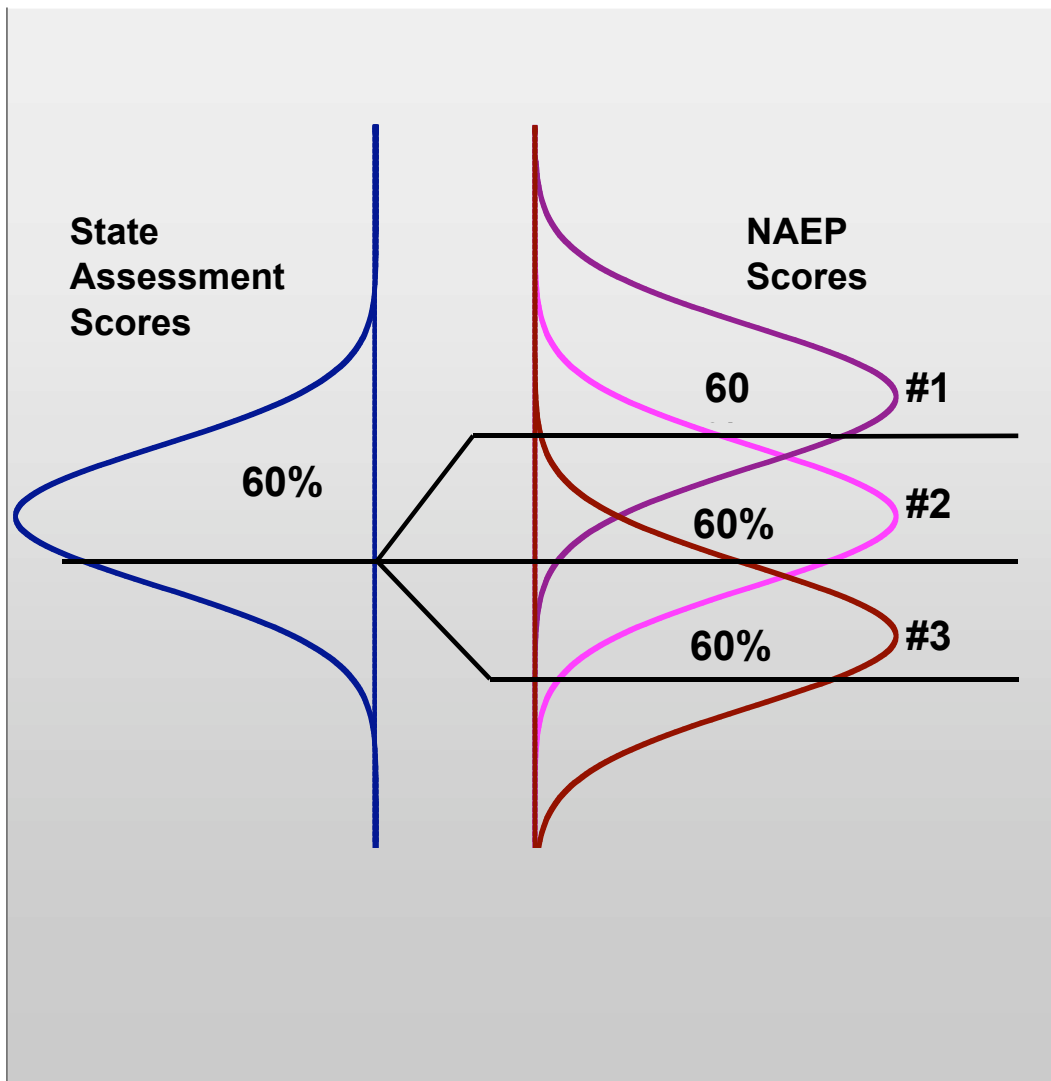
Setting NAEP Scale Score for State Performance Standard



Different Schools with Different Percentages Map a Standard onto the Same NAEP Score



**Three Schools with Same State Scores but
Different NAEP Scores Map a Standard onto
Different NAEP Scores**



Issues in Comparing Standards

What are the sources of error in comparisons?

How accurate are the comparisons?

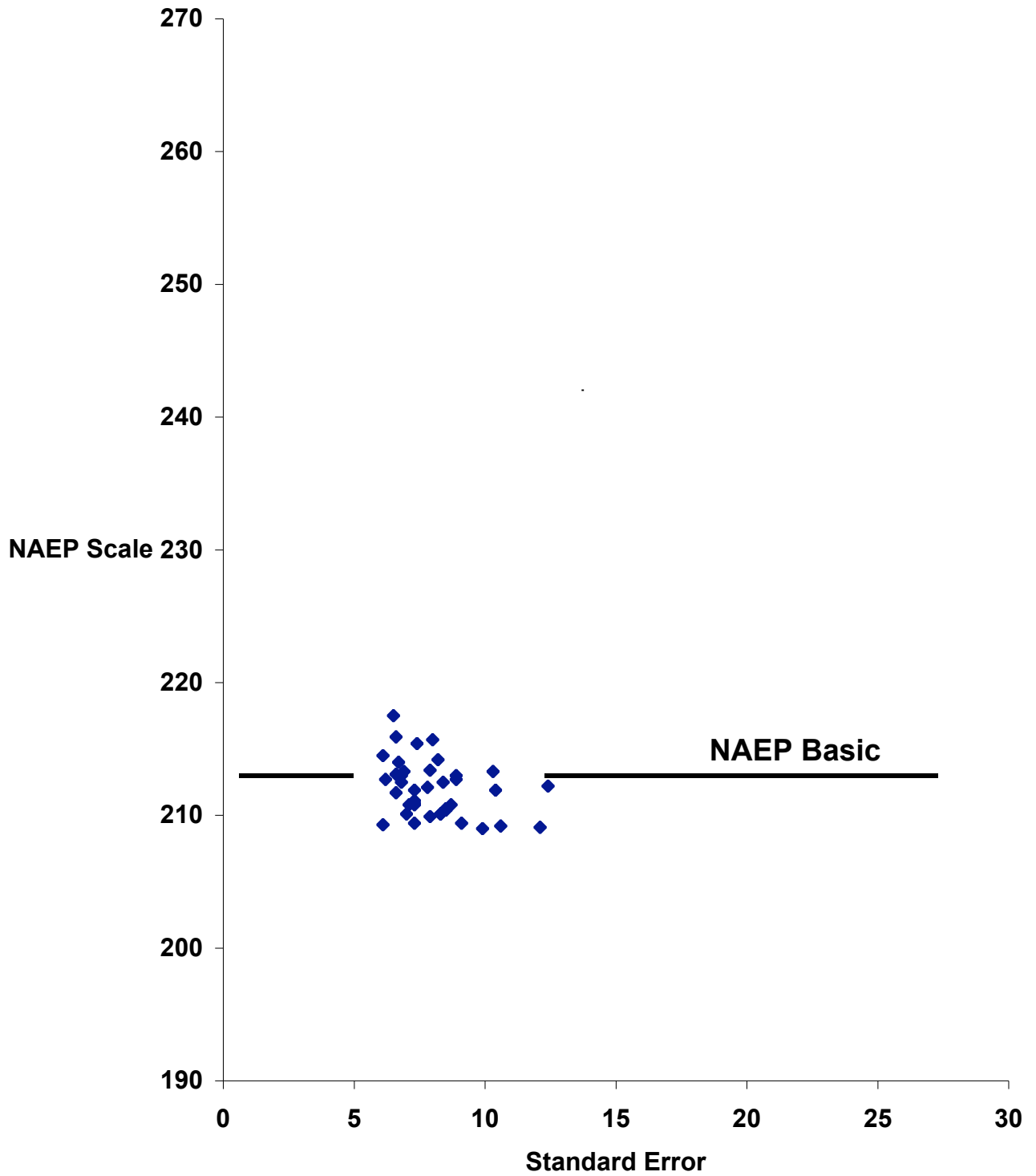
How closely correlated are the assessments?

How closely aligned are the assessments?

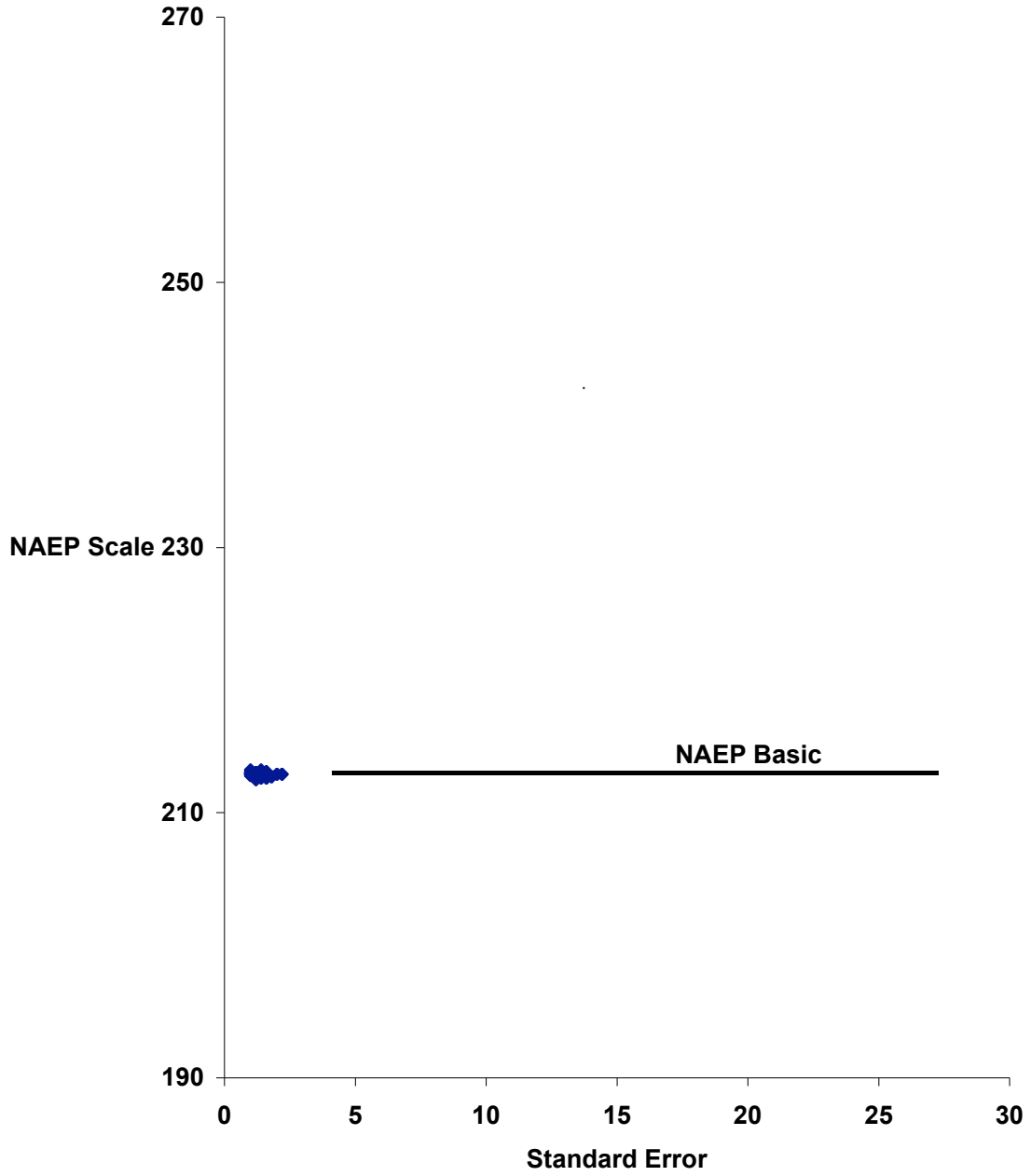
Why are some standards higher than others?

Is there a value in setting standards higher?

**Simulated Estimation of NAEP Basic Standard:
(Sampling and Measurement Error) Grade 4, 2000**



**Simulated Estimation of NAEP Basic Standard:
(Measurement Error) Grade 4, 2000**



Comparative Standard Errors

Simulations of Estimating the NAEP Basic Level		Error in Estimating State Standards		
Measurement Error	Measurement and Sampling Error	Observed	Hypothetical if Measures Uncorrelated	Std. Dev. of NAEP School Means
1.3	7.6	10.9	19.8	14.9

Notes:

All figures are standard deviations of school-level estimates, for Grade 4 Math 2000. State standards estimated to be less than 200 are excluded.

The simulations used NAEP data to estimate the percentage of the plausible value distribution in a school greater than 213.

The sampling error simulation used one-half of the records in a school to estimate the percentage of the NAEP distribution greater than 213 and the other half as the basis for translating the percentage (back) to the corresponding scale value. The value shown is reduced to adjust for the half-sample size.

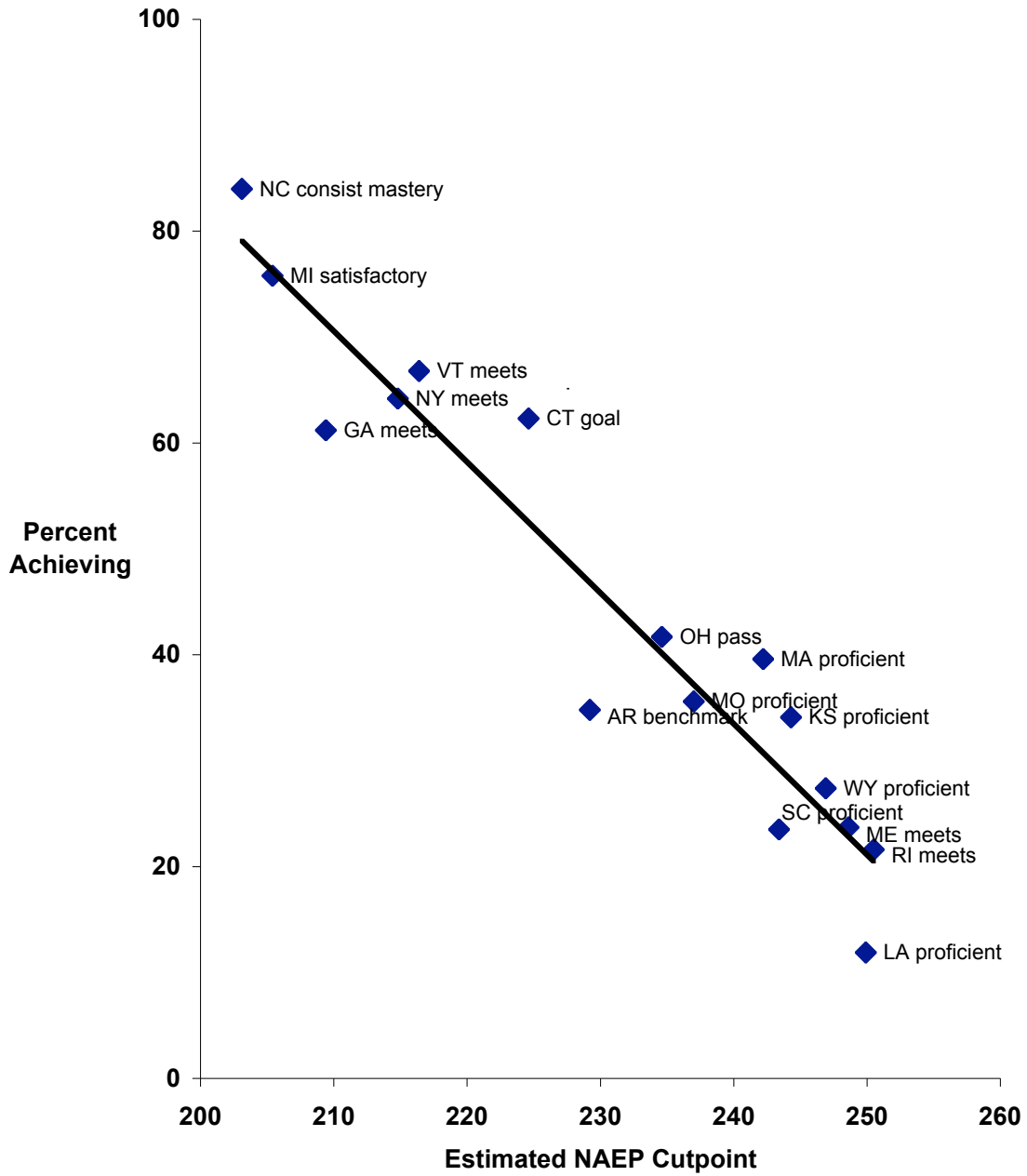
Correlations between Percentages Achieving Standards and NAEP Means

	Lowest Level	Level 2	Level 3	Highest Level
Arkansas	0.78			
California	0.78	0.82	0.79	
Connecticut	0.87			
Georgia	0.84	0.79		
Kansas	0.75	0.73	0.76	0.49
Louisiana	0.80	0.79	0.66	0.38
Massachusetts	0.10	0.78	0.84	0.72
Maine	0.58	0.57	0.39	
Michigan	0.55	0.55		
Missouri	0.68	0.67	0.56	0.11
North Carolina	0.41	0.63	0.77	
Nebraska	0.43	0.32	0.37	
New York	0.74	0.86	0.77	
Ohio	0.77			
Rhode Island	0.66			
South Carolina	0.77	0.75	0.62	
Texas	0.54			
Vermont	0.55			
Wyoming	0.60	0.56	0.35	

Comparative Correlations between Percentages Achieving Standards and NAEP Means

	NAEP to NAEP Correlation	Highest State to NAEP Correlation	Ratio
Arkansas	0.90	0.78	0.86
California	0.92	0.82	0.89
Connecticut	0.90	0.87	0.96
Georgia	0.86	0.84	0.98
Kansas	0.88	0.76	0.87
Louisiana	0.87	0.80	0.92
Massachusetts	0.86	0.84	0.98
Maine	0.59	0.58	0.99
Michigan	0.87	0.55	0.64
Missouri	0.87	0.68	0.78
North Carolina	0.75	0.77	1.02
Nebraska	0.85	0.43	0.50
New York	0.89	0.86	0.97
Ohio	0.86	0.77	0.89
Rhode Island	0.88	0.66	0.75
South Carolina	0.88	0.77	0.87
Texas	0.86	0.54	0.63
Vermont	0.76	0.55	0.72
Wyoming	0.67	0.60	0.89

Percent Achievement of State Math Standards: Grade 4, 2000



**Variation in Estimated Standards,
by Grade and Year
Example: California PR50**

	1999	2000	2001
Grade 5	219 (12)	213 (12)	211 (12)
Grade 4	219 (14)	214 (11)	211 (12)
Grade 3	215 (13)	208 (14)	206 (14)

Notes:

Standard errors are shown in parentheses.

NAEP 2000 Grade 4 was used to anchor all cases.

A higher estimated standard indicates that fewer students in the cell exceeded the state standard.

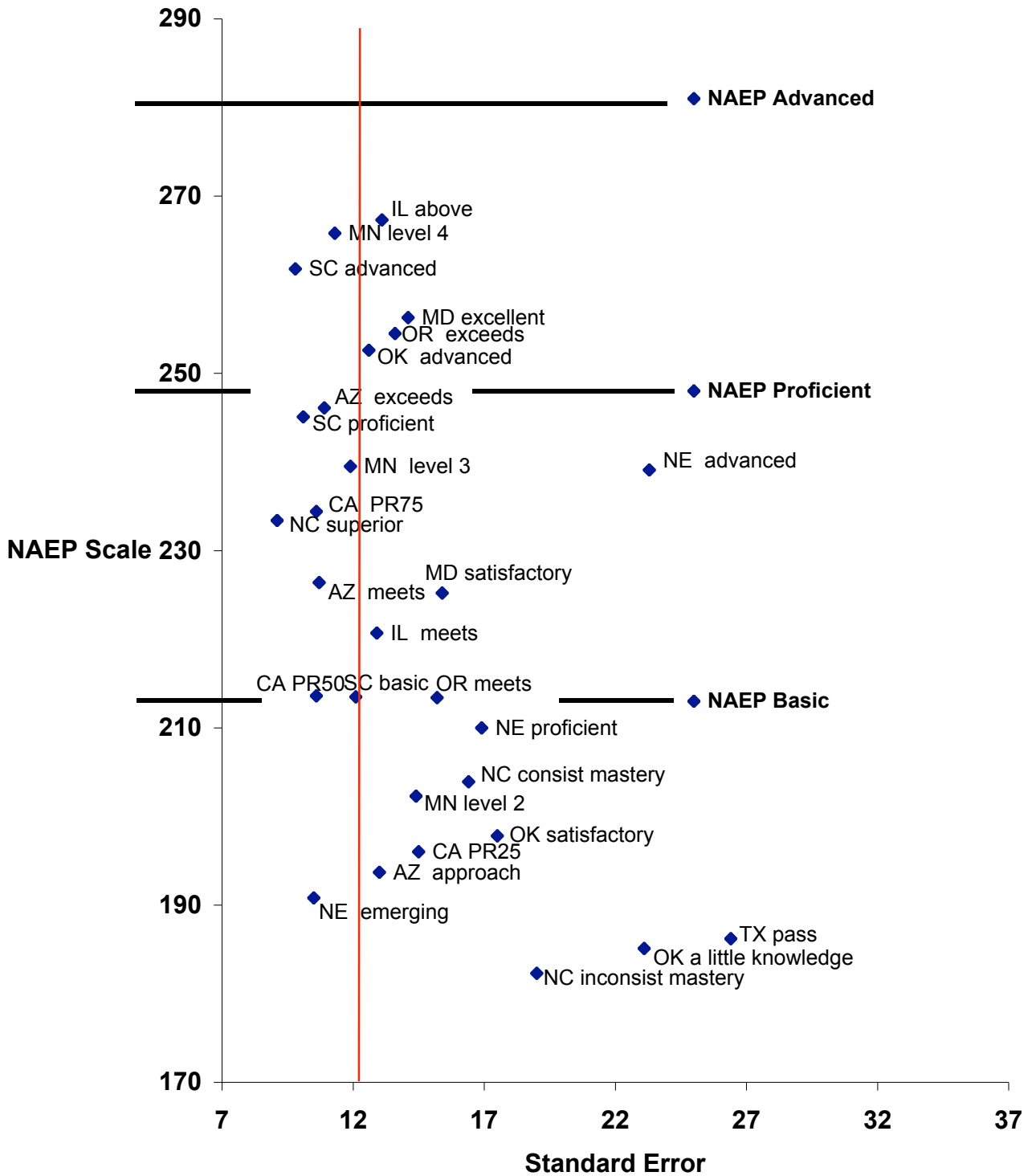
**Variation in Percent of Students with
Scores at or above Standard,
by Grade and Year
Example: California PR50**

	1999	2000	2001
Grade 5	43.9	52.1	54.6
Grade 4	44.0	51.9	54.4
Grade 3	49.2	59.1	61.0

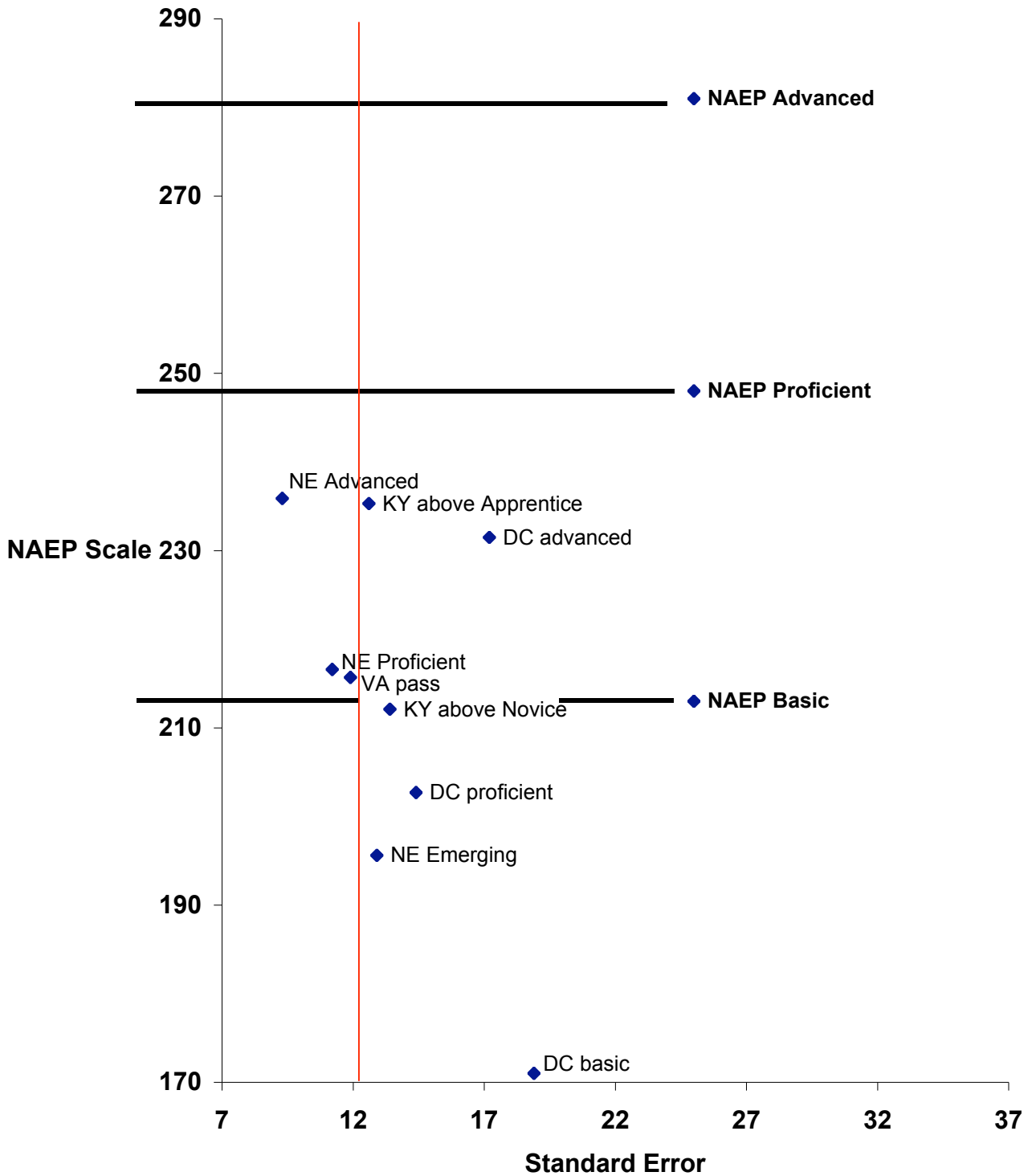
Notes:

Values are averages based on the sample of schools participating in State NAEP 2000.

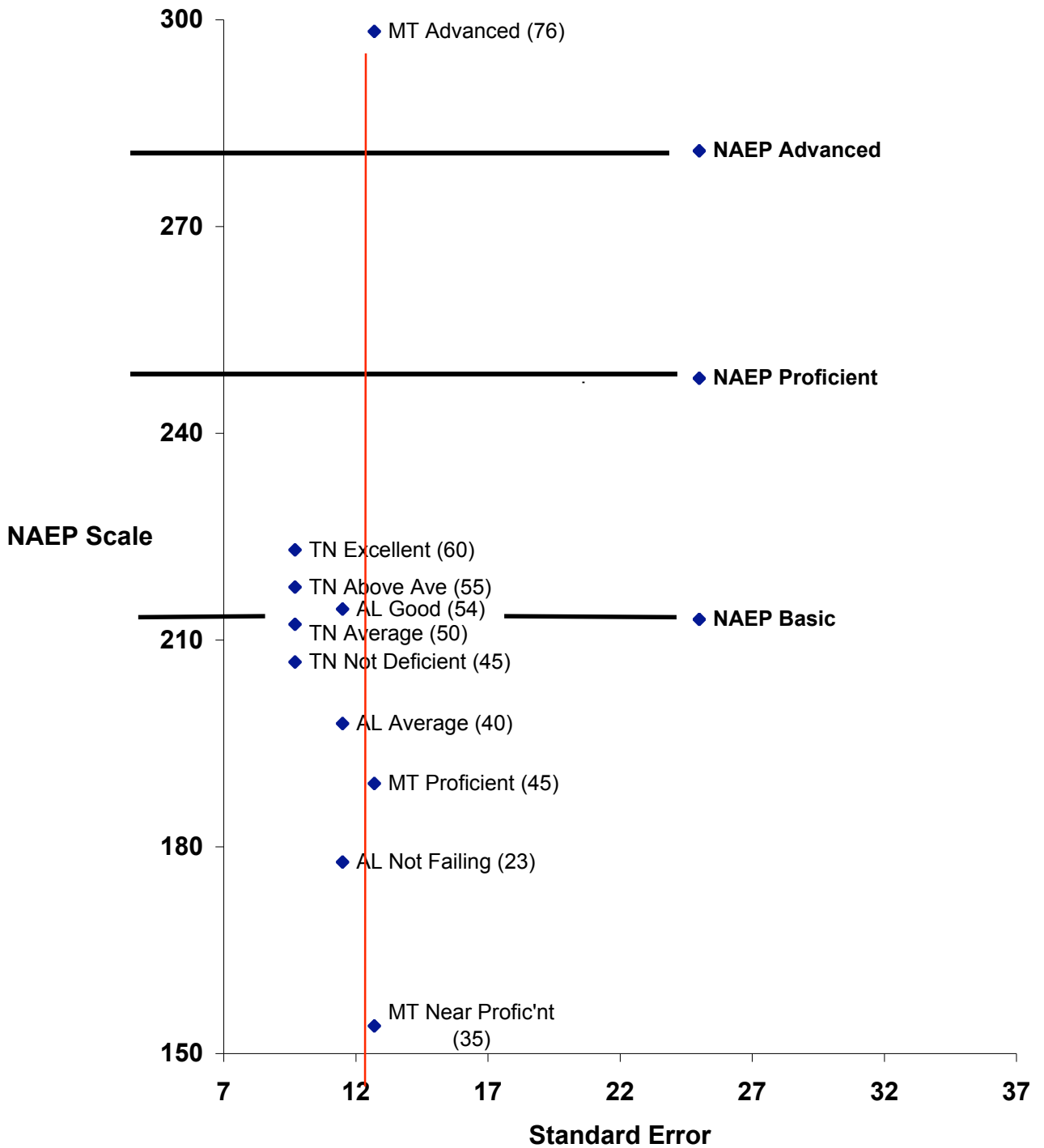
NAEP Equivalents of Math Standards: Grade 5, 2000



NAEP Equivalents of Math Standards: Elementary Grade Aggregate, 2000



NAEP Equivalents of Math Standards Expressed as Median Percentiles: Grade 4, 2000



Summary

1. States have set very different mathematics achievement standards
2. Comparison of standards is possible, using NAEP.
3. The higher the standard, the lower the percentage of students achieving it.
4. Errors exist in comparisons, due to measurement error, differences in skills covered, and differences in populations assessed.
5. The amount of that error can be estimated by observing the variability of estimated standards across schools.
6. The standard error for higher level standards is roughly 10 points on the NAEP scale (0.3 NAEP standard deviations); and it is greater for lower level standards.