Contents for a New NAEP Report: The Five Largest States

by
Paul E. Barton

Prepared for the National Assessment Governing Board

February 2009

Paul E. Barton is an education writer and consultant. He is the former director of the Policy Information Center at Educational Testing Service, and has worked with NAEP data for many years.

CONTENTS

Introduction

This paper addresses the possible content and organization of a National Assessment Governing Board (NAGB) “Mega Report” based on the most populous five states and the largest metropolitan areas in each of them.

It is not a “design” but is much more than an outline, since it discusses the content the report could encompass and why a particular measure should be included, names the sources of information used, and illustrates with data so the reader can “see” what a chart or a table might look like and what either might convey.

Averages and Achievement Levels

It makes sense, I think, to begin the report in a traditional way with averages and achievement levels. This is such standard practice that I can add nothing to the "how" of doing that. The five states would be compared on the basis of averages and with the US as a whole. Reporting would be comprehensive here, perhaps more so than with some other types of comparisons suggested later, depending on the size of the report NAGB wants—something of readable length, I assume, that does not give the appearance of a reference tome but is comparable to other very readable and attractive NAEP reports.

Data would be by gender, race/ethnicity, and school lunch eligibility for each subject and grade. The presentation would allow each state to be compared by each subgroup. Later, for achievement, I will suggest that states be put in alphabetical order and not ranked; I explain this more in another section. In any event, with only five states involved, it would be easy enough to see how they rank.

The major metropolitan areas can be included with the same subgroup breakouts. And because these big metropolitan areas may be expected to have very large differences from the rest of the state, it would be illuminating to subtract the metro area data out and compare it with data from the rest of the state. As the number of metropolitan areas expand, this may eventually be possible for most of the states.

Trend data should be included, too, with thought given to how far back to go—I assume to 2000, at least, or to the assessment closest to it.

The averages would be followed by the achievement levels, much on the same lines. As for the length of the document, achievement levels are bulkier in that there are three levels for each subgroup. This will influence the size of the document. Perhaps the data for the metropolitan area compared with the rest of the state could be limited to just the averages, although if a choice has to be made, NAGB may want to use the achievement levels for this instead of the averages.

At some point, perhaps at section end, a transition to the next section can explain why averages and achievement levels are important: the average uses *all* student scores and is the most inclusive, but in an era of "standards," achievement levels are set to give guidance about how much students should know and be able to do. But given the wide distribution of scores in any one grade in the US, other ways of looking at achievement and comparing subgroups are necessary to see the full picture—and subsequent sections will do that.

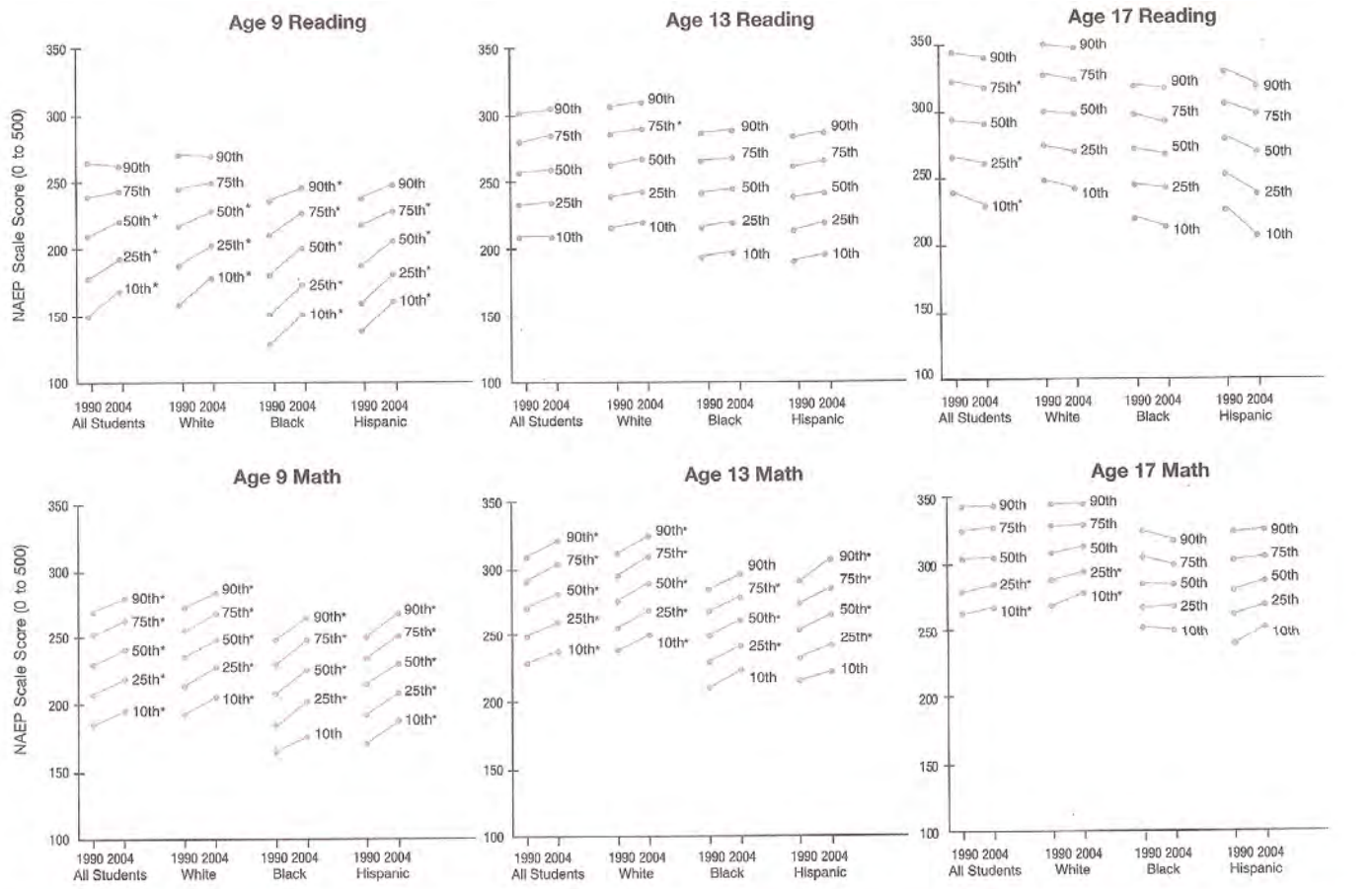
Reporting Both Level and Trend by Percentiles

Although it is necessary and desirable to use averages, their usefulness to convey meaning becomes more limited as the distribution of student scores widen. For example, if the average age of fourth graders is known, one can correctly picture a class where all students are about the same age. If the average height is known, more variation might be expected, but still, one could picture a reasonably useful image of the height of a fourth grader. If a class of student scores is averaged, there may not be a huge variation, especially if the school draws from a homogeneous neighborhood; for example, working class families. However, if scores in a large school district, state, or the nation are averaged, the result says little about the student population, because the variation is huge in what students know and can do.

NAEP has been making information available that shows the variation in scores at the tenth, twenty-fifty, fiftieth, seventy-fifty, and ninetieth percentile. But there is little emphasis on this, and it is not published in a way that allows observation of *change over time* in student scores, nor are the distributions for the different ages/grades shown side by side to drive home the width of the distributions. Below is a chart reproduced from a report I co-authored with Richard J. Coley, and released in 2008 by the ETS Policy Information Center. The chart uses long-term trend assessment and therefore is in terms of age, not grade; if done by grade, it would look basically the same. It shows scores at different percentiles and trends over the period from 1990 to 2004 in reading and mathematics for the three ages.

Figure 1

Percentile Distribution of NAEP Reading and Mathematics Scores, by Age and Racial/Ethnic Group, 1990 and 2004



* Indicates statistically significant difference from 1990 to 2004

Source: NAEP special tabulations prepared by ETS.

What first jumps out is the wide distribution of achievement in any one grade. Also apparent is *where* in the distribution the change in scores is taking place. This is particularly important as the discussion intensifies concerning whether cut-points used in the sanctions system are resulting in schools concentrating on the "bubble": students just below the cut-point. In mathematics for ages 9 and 13, it is striking that lines ascended all up and down the distribution, although not all the changes were statistically significant. And this happened among all three racial/ethnic subgroups of students. At age 17, the picture changed when scores for all subgroups are almost all flat.

The reading story is less uniform, with some increases at age 9, a flatness at age 13, and some downward sloping lines at age 17, although only a few of these are statistically significant.

The *most* striking thing is the large overlap in scores over the three ages. Laying a straight edge across the chart will show, for example, that about a fourth of 17-year-olds read no better that about a tenth of the 9-year-olds. This is also true in math.

This chart, I believe, adds a depth of understanding of the level of achievement and the trends, all on one page. I suggest one for each subject for each state with trend data for a five- or six-year period.

Measuring Inequality

In the prior section, a variation is seen in student achievement scores so great that there is overlap among 9-, 13-, and 17-years-olds. This is illustrated with the data drawn from the NAEP Long Term Trend. Visually, the wide spread of scores and the overlap is seen. But how can comparisons be made in quantitative terms? Is the degree of score inequality changing? The inequality in averages by race/ethnicity and income is known, but how does the variation differ among subgroups? And how is that changing over time? Are students becoming more or less unequal over time?

The most used measure for income inequality in the Gini Coefficient, the measure used for the statement that the US has the greatest income inequality among developed nations. I am not sure what the best approach would be for measuring this for education and comparing the US to other countries. This is obviously possible on some basis, given the international assessments and the linkages made to NAEP scores.

For purposes of illustration, I have chosen a simple approach for which data is readily available. I have looked at the score differences between the tenth and ninetieth percentile of students on Grade 8 NAEP for 2002 and 2007 (2003 in the case of Illinois). This may provide enough of a basis for judging whether to push this concept further, and it gives a new perspective in reporting educational achievement and progress.

The table below compares each of the five states and the US in total and by race and ethnicity. School lunch eligibility could be used to compare by poor/non poor. Is there more or less variation in achievement among poor students?

Table 1

Difference Between Scores at the 10th and 90th Percentile
NAEP Grade 8 Reading

	Total	White	Black	Hispanic	Asian American
National					
2002	85	77	83	85	90
2007	88	78	83	88	89
California					
2002	91	79	87	88	89
2007	95	83	93	90	96
Florida					
2002	87	79	89	86	-
2007	86	82	88	84	69
Illinois					
2003	84	76	78	88	84
2007	83	78	78	78	79
New York					
2003	81	72	81	72	96
2007	87	73	84	93	83
Texas					
2003	86	73	85	82	80
2007	83	74	77	79	80

Source: National Assessment of Educational Progress, downloaded 1/14/2009

Overall, the spread in scores stays in the same ballpark (ignore any small differences of several points as I do not have a standard error calculation). Also, the spread among White students is generally lower than among the other subgroups. Among the highest achieving subgroup, Asian American students have score spreads as wide, and sometimes wider, than other subgroups. This may be a surprise. The Policy Information Center in 1997 published *Diversity Among Asian American High School Students*, pointing out that Asian American achievement showed a lot of variation and it was a disservice to the Asian community to ignore this because it overlooked many needy Asian children.

The change is small in the five-year period, although some change is present. For the total student population, California has the most variation—higher than the US total. Clearly, this way of looking at student achievement shows a dimension not revealed in averages or in the percent above or below an established cut-point.

Can we get some fix on the significance of this spread? What does it mean to have a difference of 75 to 90 scores points between students at the tenth and ninetieth percentiles *in the same grade*? For some guidance, see Table 2 below.

Table 2

Average NAEP Scores in 2007 Reading in Grades 4 and 8

	Grade 4	Grade 8	Difference
White	231	272	41
Black	203	245	42
Hispanic	205	247	42
Asian American	232	271	39

Source: National Assessment of Educational Progress, downloaded 1/22/2009

In this table, one is struck by the similarity of the score differences over these four years of school. In the prior section on growth, something similar was seen in growth of a cohort from Grade 4 to Grade 8 where the differences were already present in Grade 4. As for the light the above table sheds on the question of the significance of the spread of scores in Grade 8, as seen in Table 1 on page 10, the spread is about *double* the four-year difference between scores for Grades 4 and 8. Roughly, the spread of scores in the same grade is worth about eight grades. The question becomes: What does it mean in the United States for students to be "on grade level," an oft-used term with over 20,000 listings on Google.com.

Measuring Growth in School

NAEP, playing an ever-increasing role in education, is respected as *the* authoritative source of knowledge about what students know and can do, and how this changes over time. NAEP is thought of generally as a measure of how well the nation's formal education system is doing and how that is changing over time. NAEP is put to this use, although I can recall nowhere in which NAEP has set out distinctions about where progress is expected to take place and who is responsible for making it.

With few exceptions, but with more now than in the past, the history of standardized testing has been to measure what students know and can do at a point in time; for example, at the end of the eighth grade. Until 1963, when Robert Glaser introduced Criterion Referenced Testing, almost all school-based testing, with such names as the Iowa Test of Basic Skills, was "norm-referenced" testing. Much of it still is. Tests were used to see how students compared with one another, and how schools compared with other schools or states with other states. The intent was not to measure in some absolute sense what students knew and could to in relation to some standard of what they should know and be able to do. Glaser, on the other hand, introduced tests that measured where students were in relationship to some defined goal or standard.

End-of-year testing measures what cognitive abilities and knowledge students have acquired since birth. If the testing was at the end of the eighth grade, it measured what the student had learned in life in general, and in grades 1-7 as well as in Grade 8, even if the student had in the past attended other schools. This type of testing, which was already in existence, became the testing program used in the new test-based accountability era for judging whether schools, in any one year, had been effective with its students that year. NAEP has been relied upon, to various extents, to report about state progress and to confirm the results that states reported in raising scores or in reaching a "proficient" level of achievement. (In a prior paper commissioned by NAGB while the No Child Left Behind (NCLB) Act was being shaped in Congress, I opposed any legislation that would put NAEP in a formal role of involvement with the test-based sanctions system.)

So, like other testing, in an assessment near the end of the eighth grade, NAEP captures what students learned from any source, in school or out. It does not assess what students learned *while they were in the schoolroom.* In a growing number of states, a "value added" approach is under development to do that. I have suggested, over the last 15 years that in addition to its regular report, NAEP report on a cohort basis student *gains* from Grade 4 through 8. This gets much closer to measuring the learning that takes place as a result of what schools do with students and what students do in school. It is not an absolute measure, however, since it covers what happens when students are at home and in summer experiences when some students gain knowledge and some students lose it, but it is a closer measure of what schools do than is measuring total knowledge at a point

in time. I have been involved in using NAEP data in this way in reports published by the ETS Policy Information Center.[1]

Using NAEP to report student gain in knowledge has a respectable pedigree. Two documents were critical in reshaping NAEP for the first time when it was transferred to ETS; this was first reflected in the 1984 assessment. One document was *Measuring the Quality of Education*, the evaluation funded by foundations and carried out by Willard Wirtz and Archie LaPointe in 1983. It recommended that instead of a report just at the test exercise level, a scale be established on which assessment results for all three ages/grades could be reported. It also suggested that assessments be given at least every four years so there could be a basis for reporting student growth over four years from Grades 4 through 8, or from Grade 8 through 12; I comment on this below.

This use of NAEP was included in the ETS redesign of NAEP. The design document, *A New Assessment For a New Era*, was prepared by the two best-known psychometricans and education measurement experts in the country, and perhaps in the world—Frederick Lord and Samuel Messick—and another leading researcher, Albert Beaton, who became the first research director of NAEP at ETS. The only times this "growth" approach has been used was in the two reports referenced above, which set forth the qualifications involved in reporting NAEP on this basis.

The first report found that state rankings based on regular NAEP reporting of eighth grade math knowledge was quite different from state rankings of how much students *gained* in math knowledge from Grades 4 through 8. In fact, the rankings were almost inverted. I use this information in speeches, stating that among participating states, Maine had the highest math scores in Grades 4 and 8 and Arkansas had the lowest scores. However, students in Maine gained 52 scale points from Grades 4 through 8, *and so did students in Arkansas*. Then I ask: Which state does a better job? Puzzled looks always cross people's faces. Of course, students in Maine enter Grade 4 with more knowledge of math than do students in Arkansas, and students from Maine enter Grade 1 better prepared, also.

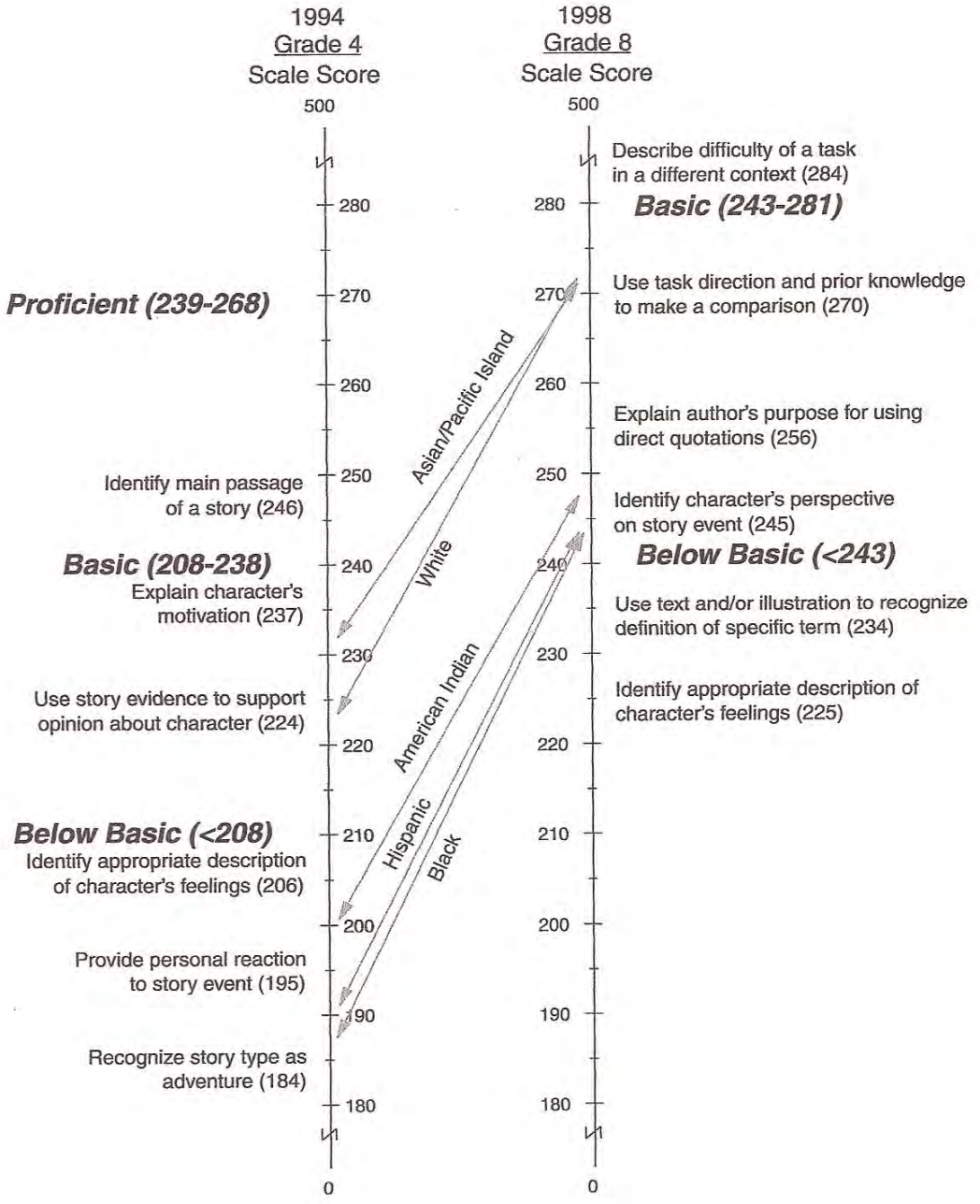
In that first report, there were no statistically significant differences between scale point gain in knowledge by race and ethnicity on a national basis. In the second report, Black and Hispanic students gained more than did White and Asian American students; this was confirmed later in a slight narrowing of the gaps in regular NAEP reporting based on race and ethnicity.

I recommend perusal of these two reports and the way the data are presented, and will be glad to make copies available to the staff and the Committee. In Coley's 2003 report, he combined much information in one chart, utilizing the Item Mapping of NAEP as well as the achievement levels. Below the chart from that report is reproduced.

[1] Barton and Coley, *Growth in School: Achievement Gains from the Fourth to Eighth Grade*, ETS policy Information Center, 1998, and Coley, *Growth in School Revisited*, ETS Policy Information Center, 2003.

Figure 2

NAEP Reading Item Map Showing Growth
From Grade 4 in 1994 to Grade 8 in 1998, by Racial/Ethnic Group



Note: The position of an item on the scale represents the scale score attained by students who had a 65 percent probability of successfully answering the item. (The probability was 74 percent for 4-option questions and 72 percent for 5-option questions.)

Source: Patricia L. Donahue et al., *NAEP 1998 Reading Report Card for the Nation and the States*, Washington, DC: National Center for Education Statistics, 1999.

A line for each subgroup average goes from the scale for Grade 4 Reading to the scale for Grade 8 Reading four years later. The average White student can likely use "story evidence to support opinion about character" in the fourth grade, and by the eighth grade can "use task direction and prior knowledge to make a comparison." An average Black fourth grader is likely to be able to "recognize a story as adventure," and by the eighth grade can "identify a character's perspective on a story event."

I have also looked at gain by state in terms of the differences for White and Black students. It is not published, but I would be glad to share the bar chart that I developed. In general, the gains in reading for Black students are typically a little more than the gains for White students. In mathematics, the gains for White students are a little more than the gains for Black students (No tests of statistical significance were done, although this was done in the two reports referenced above).

Similar comparisons could be made for gains from Grade 8 to 12. In doing this, however, two new considerations enter. One is that about one-fourth of eighth grade students drop out by the end of the twelfth grade, and that has to be considered when interpreting the results. The second consideration is that the rate of leaving school may have changed, and that would have to be checked. Although the graduation rate does change, it is reasonably steady at the national level over stretches of four years or so—to the extent that it can be measured accurately.

It is a wholly separate matter, but I later bring up the possibility of putting Grade 12 scores in the context of the percent who graduate from high school. If "Educational Progress" is being reported, whether or not students complete school is a very significant element of progress. Also, the differences in achievement as measured by NAEP in Grade 12 varies over time and among the states in how many make it to Grade 12; low scorers in Grade 8 may have removed themselves from school by the time students are assessed in Grade 12. What does that mean when comparing Grade 12 scores among the states? The original NAEP assessed students who had left school, so this was part of the initial concept of NAEP. A literacy assessment that goes back to 1986 is available for young adults, but not at the state level—although *estimates* are available on literacy by state, as recently extended from the NAALS assessment.

Returning to measuring gain in school as distinct from total knowledge, this has been moving forward at the state level. NAEP led the nation in assessing educational progress, and could consider taking the lead in this aspect of progress. A real jump to the front would be for NAEP to give an assessment to a sub-sample of students at the beginning of the year and at the end, as a trial in measuring gain in school, thus creating a model for the states to follow. No base of knowledge exists for how much various categories of students learn in Grade 8, and there is no basis for concluding what is typical or what a standard might be.

Benchmarking the NAEP Scale

I have long had an interest in developing ways to present the results of NAEP assessments that help the reader comprehend the data. Numbers on a scale are very abstract, and the only easy thing to see is trend—whether some numbers go up or down. Progress occurred around the mid-1980s when the effort was made to "anchor" the NAEP scales. This meant that committees looked at the exercises students got right at different score points, and specifically at 50-point intervals on the scale, and generalized as to what students at those intervals could do. This was described in a sentence and a short paragraph at 50-point score intervals. The 1992 National Adult Literacy Survey also did this, with a description of what respondents could do within a score interval rather than at a particular point on the scale.

This anchoring approach was dropped when the three achievement levels were adopted, both for NAEP and the 2003 literacy survey. Achievement levels convey something different: What student *should* know and be able to do. NAEP is an assessment of what students *can* do, which the anchoring was designed to do. The report should make it as easy as possible for the reader to understand this.

Although I think it was a step backward for NAEP to abandon the "anchor point" approach, it was a step forward when NAEP started using Item Maps showing actual exercises that students were likely to be able to do at different score levels. I think it would help to expand the number of items shown in the Mega Report. However, I also think there is merit in characterizing what students can do based on looking at the set of exercises, either around a scale interval or for the scores between two scale intervals, as did the 1992 literacy assessment. This is consistent with the designation of cut-points considered to be standards the nation is urged to reach.

In an NCES publication in 1993-1994, I was co-author of *Interpreting NAEP Scales*, by Gary Phillips, et al. That publication contains a section on a number of different approaches available and how they may be done, including an approach on Achievement Levels.

In the section on "Benchmarking the NAEP Scales," I put on the left side of a vertical achievement scale one-sentence descriptions of what students could do at 50-point intervals, as NAEP reports now show the item mapping. On the right side of the scale, I put examples of what different populations of students score, on average, on the NAEP scale. An example that would apply at Grade 12 would be "Students Who Took the AP test," and "Students From the Top Ten Schools," a reference to the top ten in the NAEP sample. I have reproduced the two illustrations below from this report. They are hypothetical; I do not have the numbers to put on the charts.

Figure 3

Hypothetical Examples of High Achievement Benchmarks on the National Assessment of Educational Progress

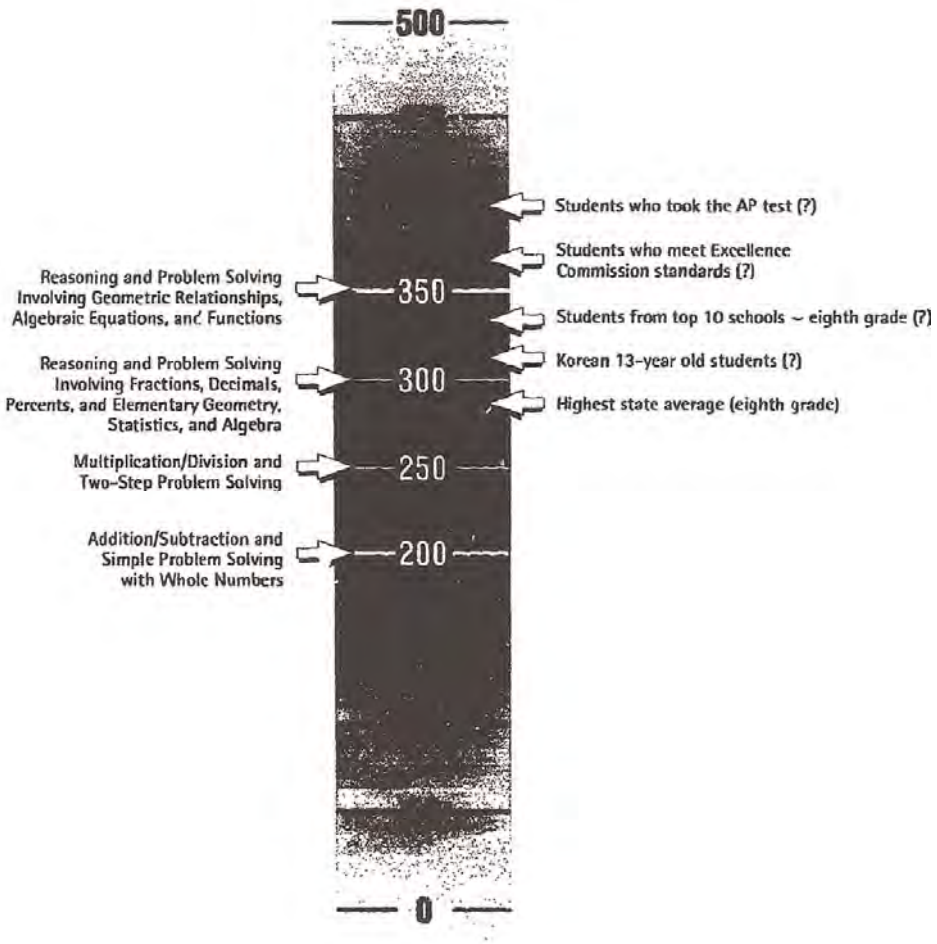
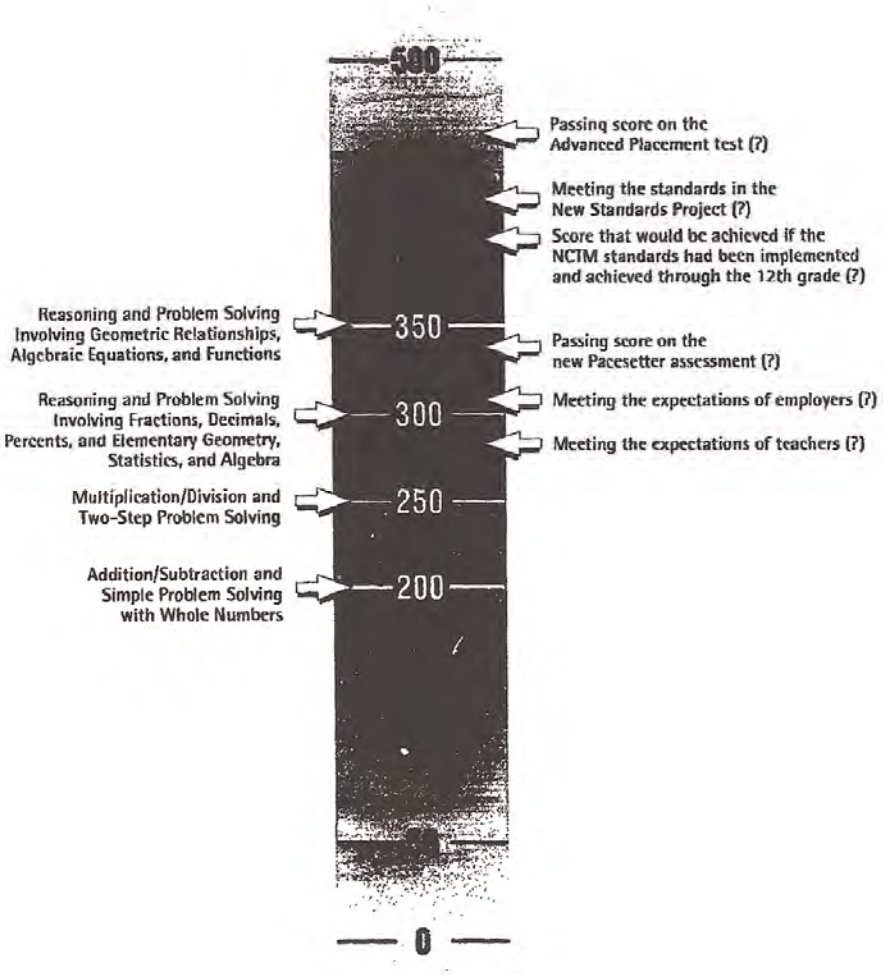


Figure 4

Hypothetical Examples of Accomplishment Benchmarks on the National Assessment of Educational Progress



When I did a paper for the Commission on Twelfth Grade NAEP, I drew on this work in suggesting how to do something similar (*Grading the 12th Graders).* Among others, there were two sets of benchmarks. One set was for the average NAEP scores of students who took and passed placement tests at a set of colleges chosen to be examples at different categories of colleges, such as "Selective Liberal Arts Colleges" or "Research Universities." Another set was for different occupations, which would take too much space to explain here. A decision on this seems to be heading toward doing "equating tests" to establish "predictive validity" of NAEP scores for students able to pass such placement examinations. I see many problems with this approach, which I discuss in another short paper prepared for the National Assessment Governing Board. (See *12th Graders and All Their Futures* at www.nagb.org in 20th Anniversary Conference papers.)

I recommend a benchmarking approach to help readers navigate the meaning of NAEP assessments. It would take a few people some effort to arrive at a useful and feasible set of benchmarks. A few examples follow, some of which may be on the charts above.

- The average scores of students who took AP courses in calculus (we know who from the background questions)
- The average scores of a high-scoring and a low-scoring country (we have these from the linking studies with TIMMS)
- The average scores of a top tier of schools in the sample, and the average scores of students in a bottom tier (the top and bottom five percent, for example)
- The average scores in the highest- and lowest- scoring states.

In terms of the scale score chart, I would put the item maps on one side (or better, use some version of the old anchor points) and the benchmarks on another. I assume that there also would be a desire somehow to identify the achievement levels.

Progress Toward State-Set Standards

It has been possible to look at state achievement and trend in two ways. One is the state's test and the standards it has set for proficiency; the other is NAEP and the standards it has set for proficiency.

Mapping 2005 State Proficiency Standards Onto the NAEP Scale, the NCES project that resulted in doing exactly what the title says, created a new set of possibilities. So far, NCES and others have reported the state standards (the cut-point for "proficient" in NAEP terms) in the 34 states where available data made this possible, comparing them to NAEP proficiency standards and typically making two points. One is that there is a lot of variation among the states and the other is that state standards fall below, or well below, the NAEP cut-points for proficiency.

In other work, I have gotten calculations of where these state standards are in terms of *percentiles* on the NAEP scale. From this, I know the percent of students in a state who reach or exceed the state standard on the NAEP scale. This answers a different question, one of how high the state has aimed in setting a standard that would raise achievement *according to the state's own score distributions*. In these terms, a high NAEP cut-point in a high-achieving state may not be more ambitious for that state than for a state with a lower cut-point, but with lower scores on NAEP. States vary widely in their average NAEP scores, depending on variation in their financial and human capital as well as variation in the quality of their instruction and the rigor of their curriculum.

To illustrate, the average NAEP scores in North Carolina and South Carolina, states next to each other, are almost identical. However, for eighth grade reading, South Carolina set its proficiency cut-point at 276 on the NAEP scale and North Carolina set its cut-point at 217—a great difference. South Carolina's cut-point is at the seventieth percentile of students in the state and North Carolina's cut-point is at the twelfth percentile. So in South Carolina, just 30 percent of students are at or above the proficient level, as compared with 88 percent in North Carolina.

The five states can be compared on the basis of their proficiency cut-point scores and the percentage of their students who at or above the cut-point for proficiency, as show below for eighth grade reading in 2005.

Table 3

Grade 8 Reading 2005

State	Proficiency Cut-Point	Percent At or Above
California	262	39
Florida	265	43
Illinois	245	73
New York	268	49
Texas	225	83

The Mega Report also can show the trend in each state in the percent of students meeting the state cut-point as measured on the NAEP scale—from 2002 to 2009, for example—showing the total and for each race/ethnic subpopulation. The latest state cut-point available can be used; if it was different in 2002, the current cut-point is operative. My available information for the percentiles was performed only on eighth grade reading for 2005; these would have to be identified for each year. And there is also the question of what subjects and grades to show, as well as whether to show intermediate years. Such decisions will rest on how large the report is to be, balancing among the different sets of comparisons. Also, there may be a design for a composite chart that will minimize the space it all takes.

The same comparisons could be made for major metropolitan areas in the state.

An additional question is whether to present the data for each state using its *own* test scores, with it own cut-points. This will bring out whether the trend is different in what the state measures and uses, and what NAEP shows. Such comparisons are an attention-getter. Using NAEP to confirm state results would push the use of NAEP further than it goes now. This may be at the borderline of what I have argued against in the past—keeping NAEP as far away from the accountability system as possible to protect it from the effects of such use. There is, of course, the question of whether the state test is more in sync with the curriculum in use and what is taught in the state than is NAEP. It is not necessarily true, as many studies have pointed out, that there is alignment between the state test, the state content standards, and the delivered curriculum. Making a comparison of the degree of alignment between the state test and the state content/curriculum, and between the NAEP assessment and the state content/curriculum, would be a considerable undertaking.

State and International Comparisons

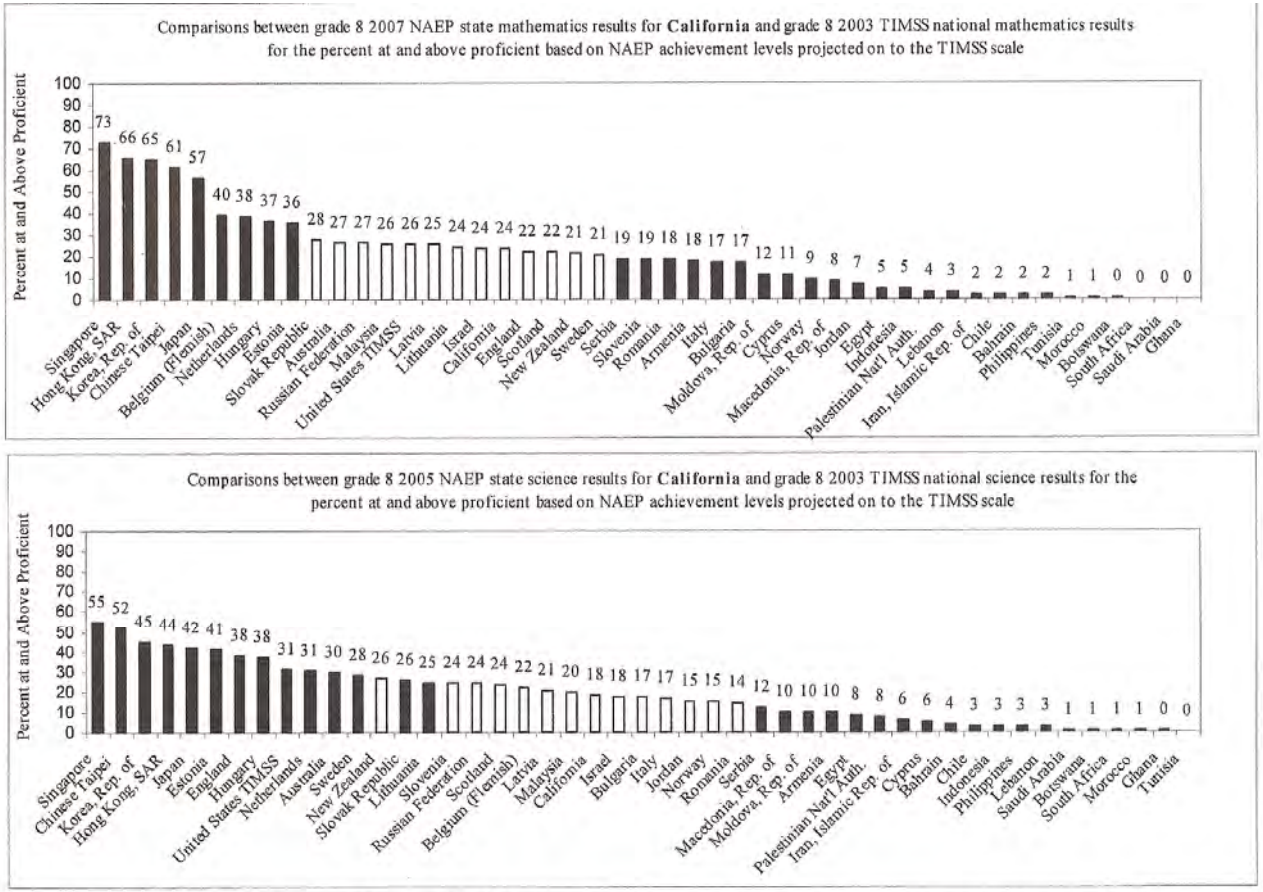
The opportunity is now available to show the US and states in terms of where they rank internationally. The analysis, *Chance Favors the Prepared Mind: Mathematics and Science Indicators for Comparing States and Nations,* November 2007, was sponsored by NCES and performed by Gary W. Phillips of the American Institutes of Research. The report uses mathematics and science data collected in 2005 and 2007, and TIMMS data for Grade 8 in 2003. Comparisons are made in terms of the percent at or above the proficient level. I assume the Mega-State Report could use data from the 2009 assessments for this purpose.

As an example, I have reproduced the comparisons graphed in the report for California. A graph can be done for all five states, perhaps with one chart using only a portion of the nations in the study. New York can be seen next to Estonia and the Slovak Republic, and Florida between Malaysia and the Russian Federation. (Massachusetts would be up among the economically developed countries, and Mississippi likely would be among some of the under developed countries.) Such a picture could illuminate the discussion in the United States of setting a common or national achievement standard. The US is a collection of geographic areas that spans the range of countries from the well-developed to the under-developed, depending on the definitions used.

Including this would enlarge the dimensions of NAEP reporting on the states.

Figure 5

Comparisons between NAEP results for California and TIMSS results in Mathematics and Science



Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007

Indicators of Conditions for Educational Progress: Using NAEP Background Questions

From one perspective or another, I have watched or been involved with the development of background questions, and the uses—or non-uses—of them. A number of perspectives exist on their need and uses. These include playing a role in the scores themselves, as in conditioning; being material for research projects, such as the excellent work that David Grissmer has performed; and being included in NAEP reports alongside achievement scores. Recently, of course, they have been little used in regular NAEP reports.

My own views on background questions were expressed in a report I did for NAGB a number of years ago, emphasizing the need to be clear about the criteria for what should be excluded and getting the number of questions down to a more manageable number. As for what was to be included in NAEP reports, I argued that they should not be used to determine—in regular NAEP reporting as opposed to outside research uses—the relationship between the answers and NAEP scores. In simple comparisons, one cannot establish such relationships. This, I assume, is the reason they were finally dropped from regular NAEP reporting.

What I recommended then—and recommend now—is that NAEP think of these questions as ways to construct indicators, drawing on the whole body of research in education, cognition, etc., that tries to convey what life experiences, conditions, and school factors are correlated with actual educational achievement. If NAEP collects data through its background questions to permit constructing such indicators—also drawing on outside research findings—it would be constructive to show the levels and trends of such correlates in NAEP reports. It could help clarify both what current achievement is and what trends are favorable or unfavorable to educational progress, in terms of student scores. Thus, the scores themselves, and the indicators, could together constitute educational progress and its prospects. The outside body of research would establish relationships to achievement, and not by making comparisons in NAEP reports between background factors and achievement.

To this, of course, one must know what student experiences and conditions are correlated with cognitive development and school achievement. For this, I draw on the work I did for *Parsing the Achievement Gap: Baselines for Tracking Progress,* a 2003 report that addressed the question of whether these gaps mirror the gaps in actual student achievement. Working from syntheses that others had performed and from individual research undertakings, the report involved identifying matters the research community has reached "a reasonable degree of agreement" about as being factors correlated with educational achievement. The qualifications and limitations of my work are set forth in the report, which called for some highly regarded research organization to do a more thorough job of identifying and working with these correlates to carry the work forward. That never happened, and an update titled *Parsing the Achievement Gap II* is now in press. These reports marry the correlates and educational achievement with statistics

showing what the gaps are in these school and non-school life experiences and conditions. The second report shows trends in the size of these gaps over the prior five or six years by race/ethnicity and income.

In my recommendation for the use of information from background questions, I will draw on this past work, which identified 14 correlates in the first report and 16 in the update. The support for them is in the two reports, which I will not repeat here due to the length. If NAGB decides to use this approach in the Mega Report, it could have some outside reviewers comment on the list I developed, realizing that: 1) seldom is there complete agreement; 2) research is a continuous process of thesis and antithesis; and 3) close calls sometimes need to be made, as my list does a couple of times.

The 14 correlates follow, and I have gone through all of the 2009 questionnaires to see what matches might be possible with NAEP data (NAEP data were used for some of the correlates in the *Parsing* reports).

SCHOOL FACTORS
Curriculum Rigor
High School Curriculum
AP participation
Teacher Preparation
Certification
Preparation in Discipline
Teacher Experience
Teacher Absence and Turnover
Teacher Absence
Teacher Turnover
Class Size
Internet Access
Computer Ratio
Fear and Safety at School
Street Gangs at School
Physical fights
SCHOOL AND NONSCHOOL
Parent Participation
NON SCHOOL
Frequent School Changing
Low Birth Weight
Environmental Damage
Lead Exposure
Mercury Poisoning
Hunger and Nutrition
Talking and Reading to Young Children
Television Watching
The Parent-Pupil Ratio
Summer Achievement Gain/Loss

Below, I show what my gleaning of the 2009 background questions produced as Indicators of Conditions for Educational Progress. I have not looked at the questionnaires for years past; those involved in NAEP will likely know what is available from the past. I suggest that the Mega Report show trend for the years for which there is a question, and not insist upon only those that have trend for the same assessment years, because the questions have varied over the years. If a year includes a question for the first time, then use it, expecting that the question will be continued in future years. I suggest reporting as follows for each indicator.

1. Rank the states by the indicator, leading with the indicator most favorable to progress in achievement; for example, the least television watching, or the most teachers with the most experience.

2. Compare states to the US average.

3. Show both the state as a whole and the major metropolitan area in the state.

4. For each state, show the student population as a whole and broken down by gender, race/ethnicity, and school lunch eligibility.

The questionnaire items that I have matched to the correlates follow.

Curriculum Rigor

AP and IB participation (for the two states in the pilot Twelfth Grade NAEP Assessment) or the percentages of students taking the highest levels of math in the eighth grade (from student questionnaires)

Teacher Preparation

1. Certification: from the teacher questionnaires

2. Preparation in discipline taught: from teacher questionnaires, whether the teacher has a minor or major in the discipline being taught.

3. Experience: From teacher questionnaire

Teacher Absence and Turnover

1. Absence: from the school questionnaire

2. Teacher Turnover: from the school questionnaire about the teachers who started the year who are still there at the end of it.

Class Size

From the teacher questionnaire. Although there has been a huge amount of research over decades showing that class size matters, there is also debate, including the results of the Tennessee STAR project. However, there would be little debate, from the standpoint of equality, that classes should not be larger for minority students than for majority students. The "reducing class size movement" has spread among the states and much of the policy debate is about the relative returns to using different approaches, since reducing class size is expensive.

Classroom Technology

There is much information in the student and teacher questionnaires. I suggest a try at making an index from a set of individual answers. If that is not possible, I suggest picking two examples of technology that represent the most advanced use. In *Parsing*, I used Internet access and computer-to-student ratio. NAEP has a lot of information on technology. The research is not strong on the correlation of technology with achievement, but it favors the more advanced uses in delivering instruction.

Television Watching

NAEP has dropped the question (I'm not sure when.) but information is available elsewhere.

Fear and Safety at School

I did not identify a correlate in the NAEP questionnaires. However, *Parsing* has statistical information from other sources. In 1992, NAEP included in the student questionnaire a question about whether disruptions in the class interfered with learning.

Frequent School Changing

Information is in the school questionnaire on how many years the student had been in that school.

The Parent-Pupil Ratio

This is the term I used in two reports on *America's Smallest School* to denote whether a student is from a one-parent or a two-parent family. Quite a while ago, NAEP dropped this question because of sensitivity concerns. In the meantime, the proportion of students in one-parent families has continued to grow among all population subgroups; among Black students, only about one in three children live with two parents. The research is plentiful and clear that having two parents matters, on the average, in student achievement. The recognition of the importance of having fathers in the home has grown. Candidate Obama gave a speech on it. I have argued for reinstating the question and tracking the trends. In the second report on the family, I developed a proxy measure from

the two NAEP questions on the amount of education the mother and the father have, with the difference in the "don't knows" representing the difference in whether both parents were in the home. I correlated the proxy measure with a census measure of one-parent families and, finding a very high correlation, used it in the report. Including this factor in a multiple regression analysis added considerably to the correlation between a set of economic/social factors and student achievement at the state level, as well as in the high school graduation rate.

The above list has stayed with correlates that can be turned into indicators using NAEP background question data, except for Fear and Safety at School, and Television Watching, for which I found nothing in the questionnaires. I listed them for use of non-NAEP data, for this would permit including all the direct school factors among the list of correlates. I also have added a couple of non-school correlates: student mobility and two-parent families, as student conditions during the period when students are in school.

Structure, Teaching, and Pedagogy

NAEP in its extensive set of background questionnaires to students, teachers and schools has created a unique body of data on what goes on in schools, and it gets this data every couple of years. With the questions emerging from committees of experts about teaching and learning within each subject matter assessed, care has gone into deciding what is important to know about the education system. This body of information has not been used in NAEP reporting, although it is available for those who know about it and want to get it from the NAEP website. However, these data are used primarily by members of the research community and possibly by professional associations, if at all. The information is not likely reaching very many people.

As discussed above in the indicators section, there has been a lack of agreement on how to use the results of these questions. Some people on the committees that pose the questions could be expected to have beliefs about the effectiveness of some of the educational practices about which they are interested. They know, of course, that NAEP will measure student achievement, and they likely see a possibility of finding out about effectiveness and relative effectiveness of different types of practices or approaches. This may be possible if a research project were undertaken to learn this; it would involve developing controls for all the other factors that affect student scores to isolate the effect differences in particular practices might have. An example is using NAEP data to compare charter schools and public schools though contracts NCES has made with research organizations.

However, this cannot be done in the periodic reporting of NAEP results. Scores of students taught one way cannot be compared with scores of students taught another way. Having the data available, however, does provide a base for doing serious research. From these data, much can be learned about how the education system is structured. Common practices can be learned, and uncommon practices, how practices are changing over time, and how one state or district compares with another.

Below are some examples of concerns that appear to generate considerable interest in the education community, and in some cases, in the general public interested in education.

Adequate Resources. From the teacher's point of view, the question might be, "Which of the following statements best describes how well your school system provides you with the materials and other resources you need for mathematics instruction?" In the early 1990s, this question was asked, and I found that, after controlling for parental education, there was a high correlation with the answer of "I have all the resources I need" and the variation in state test scores. With some work on this, it might be a candidate to put into the indicators section of the Mega Report. In any event, there is good reason to have interest in what teachers say. Do teachers have all, most, some, or none of the resources needed, and how does this compare and change over time?

Different Approaches to Different Students. This is always of interest, particularly when it is looked at from the standpoint of the degree of different treatments given in subpopulation groups, income, race/ethnicity, and different levels of student achievement. I do not know whether there are theories about when such different treatments should be used, and under what circumstances. Subject matter committees likely are aware of such things. The following are the choices: different "standards" for some students, different supplements, different classroom activities, different teaching methods, and different pacing of teaching.

Types of Resources Available. Although there is a general question about whether teachers have the resources they need to teach, some are specific to the subject; for example, eighth grade science. These include textbooks, lab equipment, computer labs, audiovisual materials, measurement instruments, and science kits. Such information is likely of considerable interest to state education departments and members of committees in legislatures.

Math Eighth Grader Study. This could be based on the student questionnaire. As high school math gets beefed up, and with Algebra 1 being pushed back to eighth grade, this is a way of tracking what is happening and the extent to which it is changing. There are nine choices, including basic or general math, introduction to algebra, first-year algebra, second-year algebra, and geometry. This would create a lot of interest, I think, in comparing states, districts, different population subgroups, and tracking change over time.

Math Twelfth Grader Study. This could be based on the student questionnaire, with 14 choices of courses taken from Grade 8 to 12, ranging from general math to Algebra II to Trigonometry to Calculus.

The Mega-States Report would be a good place to introduce the idea of presenting such material from the background questionnaires. States and districts could be surveyed to see what would be of most interest to them. Of course, such information will be available for only two of the five states.

For an excellent example of how these data can be used, see *The Fourth Grade Classroom*, Richard Coley and Ashaki Coleman, ETS Policy Information Center, 2004.

The Context for Making Educational Progress

Every school, district, and state school system exists within a context of conditions that support and affect teacher ability to teach and student ability to learn. When scores of students in one school are compared with another, or one state with another, or one nation with another, the information is important, but it is limited in explaining why achievement is at a particular level or and why it is lower or higher in one place than in another.

These differences have to do with resources with which to build adequate schools, equip them, and attract well-qualified teachers to them. The availability of resources has to do with the wealth of the community, the district, and the state as a whole. And that has to do with natural resources, commercial/industrial structure, per capita income, and inequality in family income. All these things relate, to some extent, to the desires and motivations of a particular population to tax itself to support adequate schools. A large factor in all of this is the level of education of the citizens and their health and well being.

With that in mind, I suggest using a limited number of key indicators that permit comparing the five states with each other, and with the nation, in terms of their contexts for making educational progress. I do not suggest, in the preceding sections, ranking the states in terms of their student achievement scores; these scores provide incomplete information about the relative will to improve and the effort expended to do so. In any event, with only five states, it is easy to see how they compare in student achievement. However, it would be fine to rank them on the resource and capability measures, which could be consulted in relation to student achievement.

Fortunately, this does not have to be an exercise of exhaustive search, for there is a recent publication with the highest of credentials. *The Measure of America: American Human Development Report 2008-2009* was written, compiled, and edited by Sarah Burd-Sharps, Kristen Lewis, and Eduardo Borges Martins, and is a joint publication of the Social Science Research Council and Columbia University Press.

Another recent report on a single but important indicator is *Indirect County and State Estimates of the Percentage of Adults at the Lowest Literacy Level for 1992 and 2003* by Layla Mohadjer, et al, published by the National Center for Education Statistics in January 2009. The project took the data from the 1992 and 2003 National Adult Literacy surveys and used the relationships between adult literacy and the population characteristics of those assessed in the survey to estimate the level of adult literacy at many geographic levels—one being the state and another being the metropolitan area. The report, in addition to providing a specific percentage of those lacking Basic Prose Literacy Skills, gives a range called the “credible interval,” a term similar to a confidence interval, so as not to give importance to small differences.

Below are these literacy estimates for 2003 for the five states, followed by the larger set of indicators.

State	Percent
California	23
Florida	20
Illinois	13
New York	22
Texas	19

These data are available for 1992, and so a comparison could be made. The credibility intervals could be displayed; there would be little significant difference among these states, except perhaps for Illinois. But that means something. The estimates are also made by county, so something approximating large municipal areas should be possible.

Regarding the indicators in *The Measure of America*, I am sure the design team would want to go through all of them and make choices, but some suggestions follow. A principal one is a composite index made up of six indicators called the American Human Development Index, pages 32-33 of the report: Life Expectancy at Birth, At Least Senior High School Diploma, At Least a Bachelor's Degree, Graduate or Professional Degree, School Enrollment, and Median Earnings.

Here are the five state HD indexes.

Table 4

Human Development Indexes

State	Index
Texas	4.57
Florida	4.96
Illinois	5.42
California	5.62
New York	5.81

In addition to this index, the report also has three others: Education, Health, and Income, with scores for each of the states (p.163). I suggest including those.

The inclusion of these indexes would provide a comprehensive picture of the wellbeing of the people in each state, and its economic resources in terms of the income of its citizens and human capital.

The information is also included by Congressional Districts, beginning on page 164. Perhaps this could be used to report data by the five metropolitan areas included in the five states.

Other Considerations

Several factors and considerations are referenced in the work statement, or arise from perusing regular NAEP reports.

Accommodation and Exclusion Rates. This has always been a difficult matter to address. I have looked at the last NAEP reports of how this is handled and can think of no reason to do differently for the Mega Report. No doubt, the guidance given in the NAEP reports about how to view differences is less than satisfactory to the reader: Differences among the states should "be considered" when comparing scores; "the effect is not precisely known" but performance "could be affected if exclusion rates are comparatively high." I am sure all this has been given much thought and assume that no better information is now available for this new report.

Rank Order. I have commented on this a couple of places in the text. I would not rank by performance on achievement measurements; with five states, this is easy enough to see, and I believe it has been NAEP practice to list alphabetically. I point out that there are a lot of factors beside effort and teaching quality that shape student achievement. However, in the context section, in matters such as indices of Human Development, these may be ranked in order. It might be useful to put the state achievement scores alongside these context rankings.

A Composite of Academic Achievement. It sound attractive to develop a composite, but on balance, I do not recommend it. Inevitably, apples are compared to oranges, or at least oranges to tangelos. While the appearances of having the same scale scores for each subject and of setting cut-points with the same level such as "proficient" would seem to make a combination reasonable, it is unclear how much a particular level of math knowledge reported by NAEP *equals* a level of science or reading knowledge. Comparing differing mixtures that make up a single number will likely be misleading, although there would be no way of knowing how much. From psychometricians, I have heard some possibilities discussed for doing this; for example, having the students in each subject also answer reading questions, and then use, in some way, their reading achievement to create some comparability among several subjects. The terminology was "anchoring" the achievement scales. NAGB might want to have some discussion among the experts about the possibilities of this. Albert Beaton and Ina Mullis would likely remember such discussions in which Sam Messick was involved.