Charter advocates use misleading NAEP statistics to misrepresent successful public schools as failures

Charter proponents point to what they say is the low percentage of proficient students in public schools, based on the NAEP test. However NAEP proficiency levels have been repeatedly identified as ridiculously high benchmarks. For example, studies have shown that even the top scoring countries in the world would have more than half of their students failing the proficiency bar on NAEP.

EDUCATION WEEK
Published: November 29, 2006

‘Proficiency for All’ Is an Oxymoron

Accountability should begin with realistic goals that recognize human variability.

By Richard Rothstein, Rebecca Jacobsen, & Tamara Wilder

The No Child Left Behind Act requires all students to be proficient by 2014. This is widely understood to be unattainable because 2014 is too soon. But there is no date by which all (or nearly all) students, even middle-class students, can achieve proficiency. “Proficiency for all” is an oxymoron.

—Peter Lui

The federal education legislation does not define proficiency, but refers to the National Assessment of Educational Progress. Although the Bush administration winks and nods when states require only low-level skills, the law says proficiency must be “challenging,” a term taken from NAEP’s definition. Democrats and Republicans stress that the No Child Left Behind law’s tough standards are a world apart from the minimum competency required by 1970s-style accountability programs.

But no goal can be both challenging to and achievable by all students across the achievement distribution. Standards can either be minimal and present little challenge to typical students, or challenging and unattainable by below-average students. No standard can simultaneously do both—hence the oxymoron—but that is what the No Child Left Behind law requires.

As the Harvard University professor Daniel Koretz, an expert on educational assessment and testing, has noted, typical variation in performance between those with lower and higher achievement is not primarily racial or ethnic; it is a gap within groups, including whites. Performance ranges in Japan and Korea, whose average math and science scores surpass ours, are similar to the U.S. range. If black-white gaps were eliminated in the United States, the standard deviation of test scores here would shrink by less than 10 percent. It would still be impossible to craft standards that simultaneously challenged students at the top, middle, and bottom.

The No Child Left Behind Act’s admirable goal of closing achievement gaps can only sensibly mean that achievement distributions for disadvantaged and middle-class children should be more alike. If gaps disappeared, similar proportions of whites and blacks would be “proficient”—but similar

proportions would also fall below that level. Proficiency for all, implying the elimination of variation within socioeconomic groups, is inconceivable. Closing achievement gaps, implying the elimination of variation between socioeconomic groups, is daunting but worth striving for.

Not only is it logically impossible to have "proficiency for all" at a challenging level. The law and NAEP stumble further. Their expectations of proficiency are absurd, beyond challenging, even for students in the middle of the distribution. The highest-performing countries can't come close to meeting the No Child Left Behind Act's standard of proficiency for all. "First in the world," a widely ridiculed U.S. goal from the 1990s that was supplanted by this federal legislation, is modest compared with the demand that all students be proficient.

States, no matter how well-intentioned, cannot perform psychometric miracles that are beyond the reach of federal experts.

We can compare performance in top-scoring countries with NAEP's proficiency standard. Comparisons are inexact—all tests don't cover identical curricula, define grades exactly the same, or have easily equated scales. But rough comparisons can serve policy purposes.

On a 1991 international math exam, Taiwan scored highest. But if Taiwanese students had taken the NAEP math exam, 60 percent would have scored below proficient, and 22 percent below basic. On the 2003 Trends in International Mathematics and Science Study, 25 percent of students in top-scoring Singapore were below NAEP proficiency in math, and 49 percent were below proficiency in science.

On a 2001 international reading test, Sweden was tops, but two-thirds of Swedish students were not proficient in reading, as NAEP defines it.

How did we get standards so divorced from reality, even for students in the middle of the distribution? Few Americans realize how unscientific the process for defining proficiency was—and must be. NAEP officials assembled some teachers, businesspeople, parents, and others, presented these judges with NAEP questions, and asked their opinions about whether students should get them right. No comparison with actual performance, even in the best schools, was required. Judges' opinions were averaged to calculate how many NAEP questions proficient students should answer.

From the start, experts lambasted this process. When officials first contemplated defining proficiency, the NAEP board commissioned a 1982 study by Willard Wirtz, a former U.S. secretary of labor, and his colleague Archie Lapointe, a former executive director of NAEP. They reported that "setting levels of failure, mediocrity, or excellence in terms of NAEP percentages would be a serious mistake." Indeed, they said, it would be "fatal" to NAEP's credibility. Harold Howe II, a former U.S. commissioner of education responsible for NAEP's early implementation, warned the assessment's administrators that expecting all students to achieve proficiency "defies reality."

In 1988, Congress ordered NAEP to determine the proficient score. Later, U.S. Sen. Edward M. Kennedy's education aide, who wrote the bill's language, testified that Congress' demand was "deliberately ambiguous" because neither congressional staff members nor education experts could formulate it precisely. "There was not an enormous amount of introspection," the aide acknowledged.

Others urged NAEP to wait. In 1991, Gregory Anrig, then the president of the Educational Testing Service, which administered NAEP, suggested delaying proficiency definitions until they could be properly established. Chester E. Finn Jr., an influential member of the NAEP governing board,

responded that by delaying reports on how few students were proficient, “we may be sacrificing something else—the sense of urgency for national improvement.”

Once achievement levels were set, the government commissioned a series of evaluations. Each study denounced the process for defining proficiency, leading to calls for yet another evaluation that might generate a better answer.

The first such evaluation, conducted by three respected statisticians in 1991, concluded that “the technical difficulties are extremely serious.” To continue the process, they said, would be “ridiculous.” Their preliminary report said that NAEP’s willingness to proceed in this way reflected technical incompetence. NAEP fired the statisticians.

Congress then asked the U.S. General Accounting Office for its opinion. The GAO found NAEP’s approach “inherently flawed, both conceptually and procedurally.” “These weaknesses,” it said, “could have serious consequences.” The GAO recommended that NAEP results not be published using percentages of students who were allegedly basic, proficient, or advanced.

In response, the U.S. Department of Education commissioned yet another study, this one by the National Academy of Education. The panel concluded that procedures for defining proficiency were “subject to large biases,” and that levels by which American students had been judged deficient were “unreasonably high.” Continued use of NAEP proficiency definitions could set back the cause of education reform because it would harm the credibility of NAEP itself, the panel warned.

Finally, the Education Department asked the National Academy of Sciences to weigh in. It concluded, in 1999, that the “process for setting NAEP achievement levels is fundamentally flawed” and “achievement-level results do not appear to be reasonable.”

All this advice has been ignored—although now, every NAEP report includes a congressionally mandated disclaimer, buried in the text: “Achievement levels are to be used on a trial basis and should be interpreted with caution.” The disclaimer adds that conclusions about changes in proficiency over time may have merit, but not about how many students are actually proficient. Yet the same reports highlight percentages of students deemed below proficient or basic, and these, not the disclaimer, are promoted in NAEP’s press releases.

A curiosity of the No Child Left Behind legislation is that while it imposes sanctions on schools where all students are not proficient, it also acknowledges that NAEP proficiency definitions should be used only on a “developmental basis,” until re-evaluated. No re-evaluation has been performed.

Although the legislation implies that proficiency is as NAEP defines it, the law permits states to set their own proficiency levels. States use their own judges to imagine how students should perform. Widely differing conclusions of judges in different states is proof enough of how fanciful the process must be. States, no matter how well-intentioned, cannot perform psychometric miracles that are beyond the reach of federal experts.

State definitions now result in many states’ reporting far higher percentages of proficient students than NAEP does. Some states define proficiency in NAEP’s below-basic range. More will do so if the No Child Left Behind law’s requirement of proficiency for all continues.

Even then, the demand for proficiency for all cannot be met because of the inevitable distribution of ability in any human population. The federal law exempts only 1 percent of all students. From what we

know of normal cognitive distributions, this means that students with IQs as low as 65 must be proficient; these cognitively challenged young people must do better in math than 60 percent of students in top-scoring Taiwan. Were proficiency standards lowered to NAEP's basic level, children with IQs as low as 65 would be expected to perform better than the 22 percent of Taiwanese students whose achievement is below NAEP's basic score.

Discussions of reauthorizing the now almost 5-year-old law typically propose to "fix" it: by crediting gains as well as levels, extending deadlines past 2014, fiddling with minimum subgroup sizes, giving English-learners more time. None of these can save the law unless we jettison the incoherent demand that all students be proficient.

We could design accountability with realistic goals that recognize human variability. Although research and experimentation is needed to determine practical and ambitious goals, we can imagine the outlines.

We might, for example, expect students who today are at the 65th percentile of the test-score distribution to improve so that, at some future date, they perform similarly to students who are now at the 75th; students who today are at the 40th percentile to perform similarly to those who are now at the 50th; and students who are at the 15th percentile to perform similarly to those who are now at the 25th. Such goals create challenges for all students and express our intent that no child be left behind.

Such goals would perhaps have to vary for subpopulations, ages, regions, and schools. The system would be too complex to be reduced to simple sound bites and administered by the highly politicized federal Department of Education.

The No Child Left Behind Act cannot be "fixed." It gives us a "sense of urgency for national improvement" at the price of our intellectual integrity, and an unjustified sense of failure and humiliation for educators and students. It's time to return to the drawing board.

Richard Rothstein is a research associate of the Economic Policy Institute, in Washington. Rebecca Jacobsen and Tamara Wilder are Ph.D. candidates in politics and education at Teachers College, Columbia University. Research for the study on which this essay is based was supported by the Teachers College Campaign for Educational Equity.

Vol. 26, Issue 13, Pages 32,44

The Washington Post

The Answer Sheet
A School Survival Guide for Parents (And Everyone Else)

Posted at 04:00 AM ET, 11/04/2011

NAEP: A flawed benchmark producing the same old story

By Valerie Strauss

This *was written by James Harvey, executive director of the National Superintendents Roundtable. Harvey, who helped write the seminal 1983 report "A Nation at Risk," is the author or co-author of four books and dozens of articles on education and has been examining the history of NAEP as part of his doctoral studies at Seattle University.*

By James Harvey

The latest results from the National Assessment of Educational Progress were released this week and can be summarized quickly: New NAEP numbers tell the same old story. Fourth- and eighth-grade students have inched ahead in mathematics but only about one third score at the proficient or higher level in reading.

Proficiency remains a tough nut to crack for most students, in all subjects, at all grade levels. NAEP routinely reports that only one third of American students are proficient or better, no matter the subject, the age of the students, or their grade level. But no one should be surprised.

NAEP's benchmarks, including the proficiency standard, evolved out of a process only marginally better than throwing darts at the wall.

That's a troubling conclusion to reach in light of the expenditure of more than a billion dollars on NAEP over 40-odd years by the U.S. Department of Education and its predecessors. For all that money, one would expect that NAEP could defend its benchmarks of Basic, Proficient, and Advanced by pointing to rock-solid studies of the validity of its benchmarks and the science underlying them. But it can't.

Instead, NAEP and the National Assessment Governing Board that promulgated the benchmarks have spent the better part of 20 years fending off a consensus in the scientific community that the benchmarks lack validity and don't make sense. Indeed, the science behind these benchmarks is so weak that Congress insists

that every NAEP report include the following disclaimer: “NCES [National Center for Education Statistics] has determined that NAEP achievement levels *should continue to be used on a trial basis and should be interpreted with caution*” (emphasis added).

Proficient Doesn’t Mean Proficient

Oddly, NAEP’s definition of proficiency has little or nothing to do with proficiency as most people understand the term. NAEP experts think of NAEP’s standard as “aspirational.” In 2001, two experts associated with NAEP’s National Assessment Governing Board (Mary Lynne Bourque, staff to the governing board, and Susan Loomis, a member of the governing board) made it clear that:

“[T]he proficient achievement level does not refer to “at grade” performance. Nor is performance at the Proficient level synonymous with ‘proficiency’ in the subject. That is, students who may be considered proficient in a subject, given the common usage of the term, might not satisfy the requirements for performance at the NAEP achievement level.”

Far from supporting the NAEP “proficient” level as an appropriate benchmark for student accomplishment, many analysts endorse the NAEP “basic” level as the appropriate standard.

Criticisms of the NAEP Achievement Levels

What is striking in reviewing the history of NAEP is how easily and frequently its governing board has shrugged off criticisms about the board’s standards-setting processes.

In 1993, the National Academy of Education argued that NAEP’s achievement-setting processes were “fundamentally flawed” and “indefensible.” The Government Accounting Office in 1993 concluded that “the standard-setting approach was procedurally flawed, and that the interpretations of the resulting NAEP scores were of doubtful validity.”

The governing board was so incensed by a report it received from Western Michigan University in 1991 that it looked into refusing to pay the university’s prominent assessment experts before hiring others to take issue with the report’s conclusions.

The governing board absorbed savage criticism from the National Academy of Sciences in 1999. Six years after the National Academy of Education report, the National Academy of Sciences concluded that:

“NAEP’s current achievement level setting procedures remain fundamentally flawed. The judgment tasks are difficult and confusing; raters’ judgments of different item types are internally inconsistent; appropriate validity evidence for the cut scores is lacking; and the process has produced unreasonable results.”

In fact, reported the National Academy of Science panel, “the results are not believable” largely because the NAEP results flew in the face of other evidence.

Too few students were judged to be advanced, thought the panel, when measured against other indicators of advanced work, such as completion of Calculus or participation in Advanced Placement.

Fully 50% of 17-year-olds judged to be only basic by NAEP ultimately obtained four-year degrees. Just one third of American fourth graders were said to be proficient in reading by NAEP in the mid-1990s at the very time that international assessments of fourth-grade reading judged American students too rank Number Two in the world.

For the most part, such pointed and critical comments from eminent authorities in the assessment field have rolled off the governing board and NAEP like so much water off a duck's back.

As recently as late 2009, the U.S. Department of Education received a report on NAEP that it had commissioned from the Buros Institute at the University of Nebraska. The institute is named after Oscar Krisen Buros, the founding editor of *Mental Measurements Yearbook*. The report noted, "Validity is the most fundamental consideration in developing and evaluating tests."

The Institute then took NAEP to task for, among other things, lacking a "validity framework," ignoring any program of organized validation research, unprofessionally releasing technical reports years after NAEP results had been announced to the public, and the fact that "notably absent [are] clearly defined intended uses and interpretations of NAEP." The Institute went on to recommend:

"… [a] transparent, organized validity framework, beginning with a clear definition of the intended and unintended uses of the NAEP assessment scores. We recommend that NAGB continue to explore achievement level methodologies…. [W]e further recommend that NAGB consider additional sources of external validity [such as] ACT or SAT scores…and transcript studies…to strengthen the validity argument."

In short, for the last 20 years it has been hard to find any expert not on the U.S. Department of Education's payroll who will accept the NAEP benchmarks uncritically.

NAEP and International Assessments

The NAEP benchmarks might be more convincing if most students elsewhere could handily meet them. But that's a hard case to make, judging by a 2007 analysis from Gary Phillips, former acting commissioner of NCES. Phillips set out to map NAEP benchmarks onto international assessments in science and mathematics.

Only Taipei and Singapore have a significantly higher percentage of "proficient" students in eighth grade science (by the NAEP benchmark) than the United States. In math, the average performance of eighth-grade students could be classified as "proficient" in six jurisdictions: Singapore, Korea, Taipei, Hong Kong,

Japan, and Flemish Belgium. It seems that when average results by jurisdiction place typical students at the NAEP proficient level, the jurisdictions involved are typically wealthy — many with “tiger mothers” or histories of not enrolling low-income students or those with disabilities.

Complexity and Judgment

None of this is to say that the NAEP achievement levels are entirely indefensible. Like other large-scale assessments (Trends in International Math and Science Survey, the Progress on International Reading Literacy Survey, and the Program on International Student Assessment), NAEP is an extremely complex endeavor, depending on procedures in which experts make judgments about what students should know and be able to do and construct assessment items to distinguish between student responses. Panels then make judgments about specific items and trained scorers, in turn, bring judgment to bear on constructed-response items, which typically make up about 40 percent of NAEP items.

In summary, three important facts about NAEP have been downplayed, ignored, or swept under the rug and need to be acknowledged and addressed.

First, NAEP’s achievement levels, far from being engraved on stone tablets, are administered, as Congress insists, on a “trial basis.” Second, the NAEP achievement levels are inherently based on judgment and not science. While it is not entirely fair to say that this is little better than throwing darts at the wall, it is fair to say that this is little better than educated guesswork. Third, the proficiency benchmark seems reachable by most students in only a handful of wealthy or Asian jurisdictions.

Enough questions exist about these achievement levels that Congress should commission an independent exploration to make sense in straightforward language of the many diverse definitions of proficiency found in state, national, and international assessments. A national assessment that puts proficiency beyond the reach of students throughout the Western world and most of Asia promises not to clarify our educational challenges but to confuse them.

Follow The Answer Sheet every day by bookmarking http://www.washingtonpost.com/blogs/answer-sheet. And for admissions advice, college news and links to campus papers, please check out our Higher Education page. Bookmark it!

By Valerie Strauss | 04:00 AM ET, 11/04/2011