'Proficiency for All' – An Oxymoron

By

Richard Rothstein, Rebecca Jacobsen, and Tamara Wilder

Paper prepared for the Symposium, "Examining America's Commitment to Closing Achievement Gaps: NCLB and Its Alternatives," sponsored by the Campaign for Educational Equity, Teachers College, Columbia University, November 13-14, 2006

Richard Rothstein (riroth@epi.org) is a Research Associate of the Economic Policy Institute. Rebecca Jacobsen (rjj7@columbia.edu) and Tamara Wilder (tew2101@columbia.edu) are Ph.D. candidates in Politics and Education at Teachers College, Columbia University.

Research for this paper was supported by the Campaign for Educational Equity, Teachers College, Columbia University. Views expressed in this paper, however, are those of the authors alone, and do not necessarily represent positions of the Campaign for Educational Equity or of Teachers College. We are grateful for the advice and assistance we have received from scholars and policy experts (James Guthrie, Walt Haney, Daniel Koretz, Robert Linn, Lawrence Mishel, Senta Raizen, Michael Rebell, Bella Rosenberg, Jesse Rothstein, Christopher Weiss) and government technical experts (Eugene Owen, Susan Loomis, Larry Feinberg, Gary Phillips, Kelley Rhoney). None of these are responsible for our failure to follow their advice or heed their cautions in all cases, and so the errors of fact or interpretation that remain are the sole responsibility of the authors.

Introduction and Summary

No Child Left Behind (NCLB) requires all students in grades 3 through 8, in each racial, ethnic, and socio-economic group, and whether they have special needs or are native English speakers, to be proficient in math and reading by 2014. This is widely understood to be unattainable, but educators and policy makers are insufficiently aware of the causes of our looming failure. Many of the law's supporters believe that the goal of 'proficiency for all' can't be reached primarily because there is too little time between now and 2014 for schools to improve sufficiently, and that the problem can be fixed by making the deadline more distant to allow more time to improve. For this symposium, we have been asked to consider whether such a goal can be reached; if so, how long it might take if, in fact, 2014 is too soon; and if the goal is unattainable no matter how distant, how we might establish more reasonable school goals for narrowing the achievement gap and raising the achievement of all children.

We conclude that the problem is more fundamental than a mis-estimate of how long it might take for all students to achieve proficiency. There is *no date* by which all (or even nearly all) students in any subgroup, even middle-class white students, can achieve proficiency. Proficiency for all is an oxymoron, as the term 'proficiency' is commonly understood and properly used.

In the following pages, we show why this is impossible, in several steps. First, we attempt to discern the meaning of 'proficiency' in NCLB, and conclude from the language and structure of the legislation that it intends all students to be proficient as defined by the National Assessment of Educational Progress (NAEP). Although the U.S. Department

of Education has looked the other way as many states have claimed compliance with NCLB by requiring only low skill levels to pass standardized tests, the law explicitly requires standards of proficiency to be "challenging," a term taken directly from NAEP's achievement level descriptions.

We show that by ignoring the inevitable and natural variation amongst individuals, the conceptual basis of NCLB is deeply flawed; no goal can simultaneously be challenging to and achievable by all students across the entire achievement distribution. A standard can either be a minimal standard which presents no challenge to typical and advanced students, or it can be a challenging standard which is unachievable by most below-average students. No standard can serve both purposes – this is why we call 'proficiency for all' an oxymoron - but this is what NCLB requires.

NCLB's admirable, though difficult goal of closing the achievement gap can only sensibly mean that the distributions of achievement for disadvantaged and middle class children should be more similar. If there were no achievement gap, for example, similar proportions of white and black students would be 'proficient' and similar proportions of white and black students would achieve below that level as well. 'Proficiency for all,' which implies the elimination of variation *within* socioeconomic groups, is inconceivable. Closing the achievement gap, which implies elimination of variation *between* socioeconomic groups, is extraordinarily difficult, but worth striving for.

We demonstrate that the inevitable distribution of student outcomes is such that if all, not only some, students were to reach NAEP's challenging academic standard of proficiency, impossible gains would be required. By comparing NAEP results to scores on international exams, we show that even the top-performing countries in the world are

far from being able to meet a standard of 'proficiency for all,' as NAEP defines it. Indeed, 'first in the world,' a widely ridiculed U.S. education goal from the 1990s that was supplanted by NCLB, is actually much more modest than NCLB's goal of 'proficiency for all'.

It is only in the last 15 years that NAEP results have been reported in terms of proficiency and other achievement levels. We describe the shift from NAEP's original scale and norm-referenced results to this more recent, criterion-referenced reporting. Discussing the methods used by the federal government to develop current NAEP achievement levels, we show that definitions of proficiency are fraught with subjectivity. Even if well-intentioned, making judgments of what students *ought to be* capable of, rather than basing judgments on observations of what actual students can achieve, yields results that the federal government itself acknowledges should be “interpreted with caution.” The movement away from scale and norm-referenced score reports has resulted in the politicization of standardized testing.

The problems we describe cannot be fixed by lowering NCLB’s expectation, for example, lowering it to one that all students must achieve NAEP’s basic level, not proficiency. Such a reduction would effectively return NCLB to the ‘minimum competency’ accountability standard of the 1970s that NCLB was explicitly designed to reject because it created no incentives to develop the critical thinking skills that today’s graduates should possess. Even so, this basic standard still cannot be applied to 99 percent of all students, as NCLB demands. As the performance of 'first in the world' countries demonstrate, many students would still fail a requirement that all students have basic levels of achievement.

The irresponsibility of NCLB's expectation of 'proficiency for all' should not lead to the abandonment of goals for the improvement of student achievement, nor does it suggest that public education systems should not be accountable for realizing challenging degrees of improvement. We describe a simple statistical procedure, inspired by 'benchmarking' practices employed in the business world, which can be used to establish strenuous but realistic goals for improved achievement by students at all points in the distribution. Benchmarking permits a sophisticated return to norm-referenced measures of academic achievement, something not new to education but which has been abandoned in the NCLB legislation.

We conclude by describing reforms in education and youth development that might be necessary to raise achievement and to narrow achievement gaps, substantially. Because unacceptably low average achievement for disadvantaged children is established in our current education and social system by age three, and because skill developed at later ages depends on investments in skill at earlier ages, we describe a 19-year program that might bring a birth cohort of children to maturity with high levels of performance. Remedial and compensatory programs may contribute to higher achievement for cohorts already moving through the system, but probably cannot succeed in the realization of goals that inspired the framers of *No Child Left Behind*.

NCLB and the NAEP Standards

NCLB states that all children shall "reach, at a minimum, proficiency on challenging State academic achievement standards and state academic assessments," and that these standards must "contain coherent and rigorous content" and "encourage the

teaching of advanced skills."[1] The law does not further define "challenging" standards, but it is reasonable to infer that such a standard challenges typical children to achieve at a higher level than their past performance. This inference is supported by the law's requirement that the National Assessment of Educational Progress (NAEP) be administered biennially in math and reading to a sample of fourth and eighth grade students in each state, providing a standard by which state judgments about proficiency can be compared. Furthermore, NCLB uses language to describe proficiency that parallels that of NAEP, whose definition of proficiency is "demonstrated competency over challenging subject matter."[2] As Christopher T. Cross, appointed by the Department of Education in 2002 to coordinate rulemaking for NCLB, recently noted, NAEP "is supposed to be the benchmark for states, and that is why its use was expanded" in the act.[3]

The NCLB requirement that proficiency be "challenging" can also be traced to an influential series of articles on "systemic school reform" in the late 1980s and early 1990s that had an important influence on the development of federal accountability. In these, Marshall Smith and Jennifer O'Day proposed a program to create schools with "coherent and challenging instructional programs, that genuinely engage all, or at least most of their students."[*] They called for new standardized tests for accountability purposes that would "stand as a serious intellectual challenge for the student."[4] The reform goal of "challenging content for all children," Smith and O'Day wrote, should take on "an aura of official policy;" and although NAEP is not explicitly aligned with any state's curriculum,

[*] Marshall Smith was education advisor to Governor Bill Clinton when the latter co-chaired the National Governors Association education task force at the 1989 Charlottesville Education Summit where federal education goals were adopted; Dr. Smith then chaired the task force on education standards established by federal law in 1991 to develop a national accountability system, and went on to serve as President Clinton's deputy secretary and undersecretary of education.

"we expect that it will be moderately sensitive to effects of curricula that emphasize challenging content."[5]

NCLB specifies that NAEP achievement level definitions shall only be used on "a trial basis" until the Commissioner of Education Statistics evaluates them and determines that they are "reasonable, valid, and informative to the public."[6] Yet nearly five years later, there has been no significant reconsideration of historic NAEP definitions of achievement levels, so it is again reasonable to infer that NCLB's implicit definition of proficiency is consistent with NAEP criteria.[*] In the NAEP administrations immediately prior to the adoption of NCLB, only 22 percent of fourth graders in public schools nationwide were deemed proficient in math and 27 percent in reading. For eighth graders, only 25 percent were deemed proficient in math and 29 percent in reading.[7†]

This gives us a rough way to estimate how much improvement would be required for all students in all subgroups to be proficient. At present (the most recent data are from 2005), 71 percent of all eighth graders in public schools are below proficiency in reading on the NAEP. For the typical student, becoming proficient would require a gain of 0.6 standard deviations.[8‡] In other words, by 2014 the median student would perform similarly to a student who is at about the 72nd percentile of performance today.[§] For a

[*] As we discuss below, this requirement for a re-evaluation of NAEP achievement levels has been part of the Elementary and Secondary Education Act for 12 years, and ignored throughout that period.

[†] Data for fourth graders in reading, and for fourth and eighth graders in mathematics, are from NAEP administrations in 2000. Data for eighth graders in reading are from NAEP 1998. NAEP was not given for eighth grade reading in 2000. Data are for all public school students, including those who took the test with accommodations. These data include the percent of all students whose scores were above the proficient cut score, including those whose scores were above the advanced cut score.

[‡] These and similar estimates in this paper are approximations because the distributions of test scores are not perfectly normal and therefore the median (or typical) student may not be identical to the mean (or average) student. Our estimates, however, are calculated from the mean, assuming perfect normality. In 2005, the proficiency cut score was 281 in reading, the mean score was 260, and the standard deviation was 35.

[§] Throughout this paper, we adopt a convention of describing percentile ranks as ascending with improved performance. In other words, the best-performing 1 percent of students are described as being at or above

student whose performance is below the median, but still similar to that of most same-age students (i.e., those who are below the median but still performing better than the lowest-performing 16 percent of all students), becoming proficient would require a gain of up to 1.6 standard deviations.* In other words, a student who is now at the 16th percentile in today's achievement distribution would also perform similarly to a student who is now at the 72nd percentile. Approximately one-sixth of all students would require a gain even greater than 1.6 standard deviations.

World-Class Standards

Let's examine another approach to estimating proficiency. In the 1994 legislation, Goals 2000, a Congressionally mandated objective was that U.S. students should be "first in the world in math and science" by the year 2000. Many education reformers, even those who boasted of having the highest expectations, later acknowledged that this goal was absurd. As the federal government's National Education Goals Panel, established to monitor progress towards these goals, acknowledged, the first-in-the-world aim "led to a certain amount of derision and sarcasm."[9] We don't need to be first in the world, reformers seemed to reason in 2001; all we require is to be minimally proficient. NCLB's expectation that all students should be proficient seemed to be a more modest and achievable goal than first-in-the-world standing.

---

the 99th percentile, and the poorest-performing 1 percent of students are described as being at or below the 1st percentile.

* Students who perform "similarly" to most same-age students are defined here, consistent with conventional terminology, as those who are between one standard deviation below and one standard deviation above the mean, or students who perform better than approximately the poorest-performing 16 percent of students, but not as well as approximately the best-performing 16 percent of students.

Yet this expectation has matters backwards. Reaching proficiency for all is an even higher and more unreachable aspiration than being first in the world, because even first-in-the-world educational systems have a wide range of performance. No matter how much more time were permitted to achieve NCLB's goal, all American students would not be proficient, even if the United States became demonstrably the world's highest performing nation.

We can compare these slogans: 'proficiency-for-all' versus 'first-in-the-world.' In 1993, the National Center for Education Statistics (NCES) computed an approximate equation of performance between American students on the eighth grade NAEP test, given in 1992, and an international exam, the Second International Assessment of Educational Progress (IAEP), given the previous year.[*] This comparison requires assuming that NAEP and IAEP tests are similar in content and in scaling, and so is not usable for any precise purposes. We describe it here only to provide a very rough idea of how foolish is the goal of proficiency for all.

According to these experimental data, Taiwan was first in the world in math in 1991. If Taiwanese 13 year-olds had taken the U.S.' NAEP exam the following year, their estimated average NAEP score would have been 285, compared to American eighth graders' average score of 262.[10] But NAEP defines eighth graders as proficient if they achieve a score of 299, not only far higher than the U.S. average score, but considerably higher than the average Taiwanese score as well.[11] Although Taiwanese students were first in the world in math, *approximately 60 percent of them scored below what NAEP*

[*] The International Assessment of Educational Progress was funded by the National Science Foundation and administered by the Educational Testing Service for the U.S. Department of Education, National Center for Education Statistics. NCES referred to its equating of the two tests as "experimental;" we use the term "approximate" instead, to avoid suggesting that NCES conducted an actual experiment using the two tests.

defines as proficient.[12*] Thus, *even if the United States were first in the world in math, we would still be far from meeting the NCLB goal of all students being proficient.*

According to more recent (2003) data from the Third International Mathematics and Science Survey (TIMSS[†]), American eighth graders had an average scale score of 504 in math and 527 in science, compared to scores in the highest scoring country (Singapore) of 605 and 578, respectively.[‡13] Yet still, approximately 25 percent of students in Singapore are below what NAEP defines as proficient in math, and 49 percent are less than proficient in science. We display these comparisons in Figures 1 and 2, below. In Korea, the second highest scoring country in math and third highest scoring country in science, one-third are less than proficient in math and 60 percent are less than proficient in science. In Chinese Taipei, the second highest scorer in science, 53 percent of eighth grade students are less than proficient. And in Hong Kong, the third highest scorer in mathematics and the fourth highest scorer in science, one-third are less than proficient in math and 62 percent are less than proficient in science.[§]

---

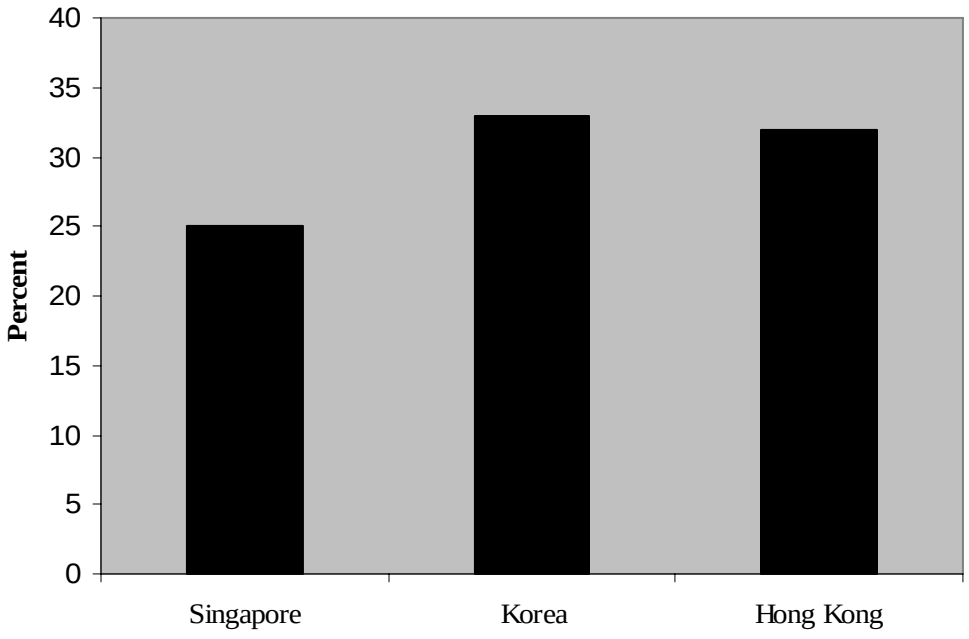
[*] The estimate reported here, that about 60 percent of Taiwanese eighth graders were less than proficient in math, comes from an Educational Testing Service study using the initial NAEP proficiency cut score, set in 1992, of 294. With the proficiency cut score subsequently redefined as 299, a larger than 60 percent share of Taiwanese eighth graders would have been deemed below proficiency. We emphasize again that these estimates are approximate, and can be considered accurate in order of magnitude, but not precise. The particular extrapolations reported here are based on the data reported by NCES that the 50th percentile Taiwanese score on the IAEP is equivalent to 286 on the NAEP scale; the U.S. proficiency cut score on NAEP was defined as 294; and the 75th percentile score for Taiwan on the IAEP is equivalent to 310 on the NAEP scale. The largest share of students to reach the equivalent of NAEP advanced status was 8 percent of Chinese students, but this was a small sample of only the most elite Chinese students; next largest were Korean students, 6 percent of whom reached the equivalent of the NAEP advanced level (Pashley and Phillips 1993, Table 5, p. 26; Table 4, p. 25).

[†] TIMSS was administered by the International Association for the Evaluation of Educational Achievement (IEA).

[‡] Singapore is not really comparable to other countries; it is a city-state, much of whose working class commutes on a daily basis from Malaysia, the country where its children attend school. If the achievement of other countries was also based on testing only (or predominantly) their middle classes, scores more appropriately comparable to Singapore's might be obtained.

[§] These approximate comparisons of TIMSS 2003 in mathematics and science with NAEP 2003 in mathematics and NAEP 2005 in science were calculated using a method demonstrated by Robert L. Linn (2000) when he compared TIMSS 1994-95 to NAEP 1996. Professor Linn estimated where NAEP cut

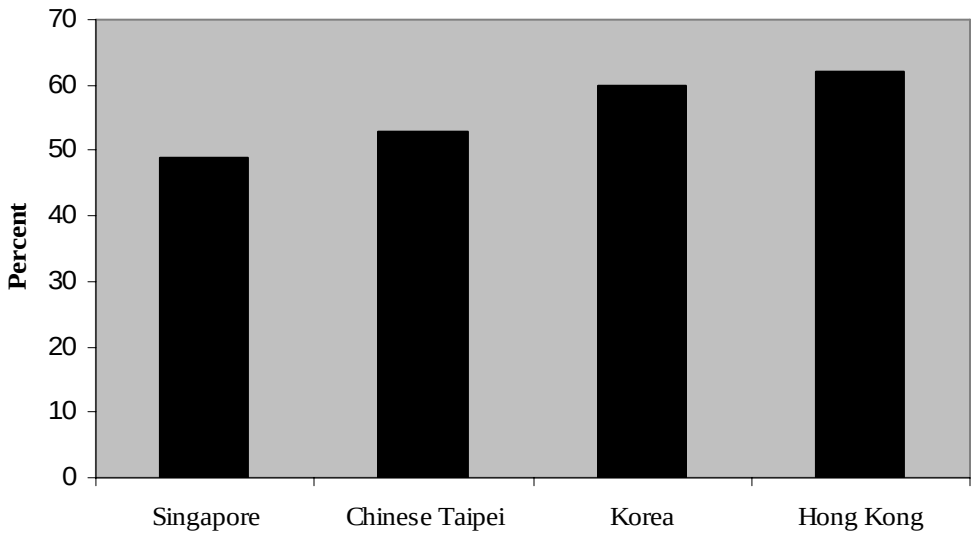
Figure 1. Percent of Students Predicted to Score Below NAEP 8th Grade Math Proficiency Cut Score



Source: TIMSS 2003: Gonzales et al., 2004; NAEP Cut Scores: Reese, et al., 1997
Note: Percentages estimated using Linn's (2000) linking method

scores would fall on the TIMSS scale, assuming that the percent proficient or above would be the same for U.S. students on the eighth grade TIMSS mathematics assessment as it was on the eighth grade NAEP mathematics assessment. In 2003, 27 percent of U.S. eighth graders were at or above the NAEP proficiency cut score. Using Professor Linn's linking method, the approximate equivalent of the NAEP proficiency cut score on the TIMSS 2003 is the score that only 27 percent of U.S. students reached, or the score that corresponds to the 73rd percentile in the U.S. distribution. We estimated the percent below this proficiency standard for each country from the predicted percentile score of a student in that country scoring one point below the estimated NAEP cut score on the TIMSS scale.

Figure 2. Percent of Students Predicted to Score Below NAEP 8th Grade Science Proficiency Cut Score



Source: TIMSS 2003: Gonzales et al., 2004; NAEP Cut Scores: Grigg, et al., 2006.
Note: Percentages estimated using Linn's (2000) linking method

On the Progress in International Reading Literacy Study (PIRLS), a 2001 reading test administered by the International Association for the Evaluation of Educational Achievement (IEA), America's 10 year-olds scored ninth highest in the world – the highest scoring countries were Sweden, the Netherlands, England, Bulgaria, Latvia, Canada, Lithuania, and Hungary, all of which, including the U.S., were closely bunched together – the average U.S. performance was only 0.2 standard deviations below that of Sweden.[*14] But on NAEP's achievement level report, only 30 percent of U.S. 10-year olds were deemed proficient in reading the next year.

[*] On the IEA scale, the U.S. mean was 542 and the Swedish mean was 561. The scale was constructed so that the standard deviation of test scores was 100.

We repeat here the caution that applications of NAEP proficiency levels to international tests are only rough approximations to suggest orders of magnitude, and are not technically defensible for precise uses. Having said that, *by comparing the NAEP scale to scores on this international reading test, we estimate that about two-thirds of all Swedish students, the highest-scoring students in the world, were not proficient in reading as NAEP defines it.*[*]

In short, being first in the world is a very modest aspiration compared to NCLB's expectation that all students will be proficient. Proficiency-for-all is a standard that no country in the world comes close to meeting, nor is it one that any country can reasonably expect to meet.

NAEP's Proficiency Definition is Inconsistent with Other Achievement Data

Other data we have on student achievement provide further evidence that NAEP cut scores for achievement levels are unreasonably high.[†] For example, the NAEP definitions tell us that in 2000, the number of twelfth grade students who performed at the advanced level in mathematics was equal to only 1.5 percent of all U.S. 17 year-olds.[‡] Yet as Figure 3 shows, in the same year, nearly double that number (2.7 percent) of all 17 year-olds were awarded college credit in calculus because they passed highly demanding

[*] We estimated this as follows: 29 percent of U.S. students were proficient on the NAEP in 2001. The published 75th percentile score on the PIRLS was 601 (Mullis et al. 2003, Ex. B.1). From this, we estimate that the 71st percentile score was about 588. Swedish students with a scale score of 588 would have been at the 66th percentile ranking of all Swedish students. The estimates rely, as noted, on unsupportable assumptions, for example that that the distributions and difficulty of NAEP and PIRLS are equivalent.

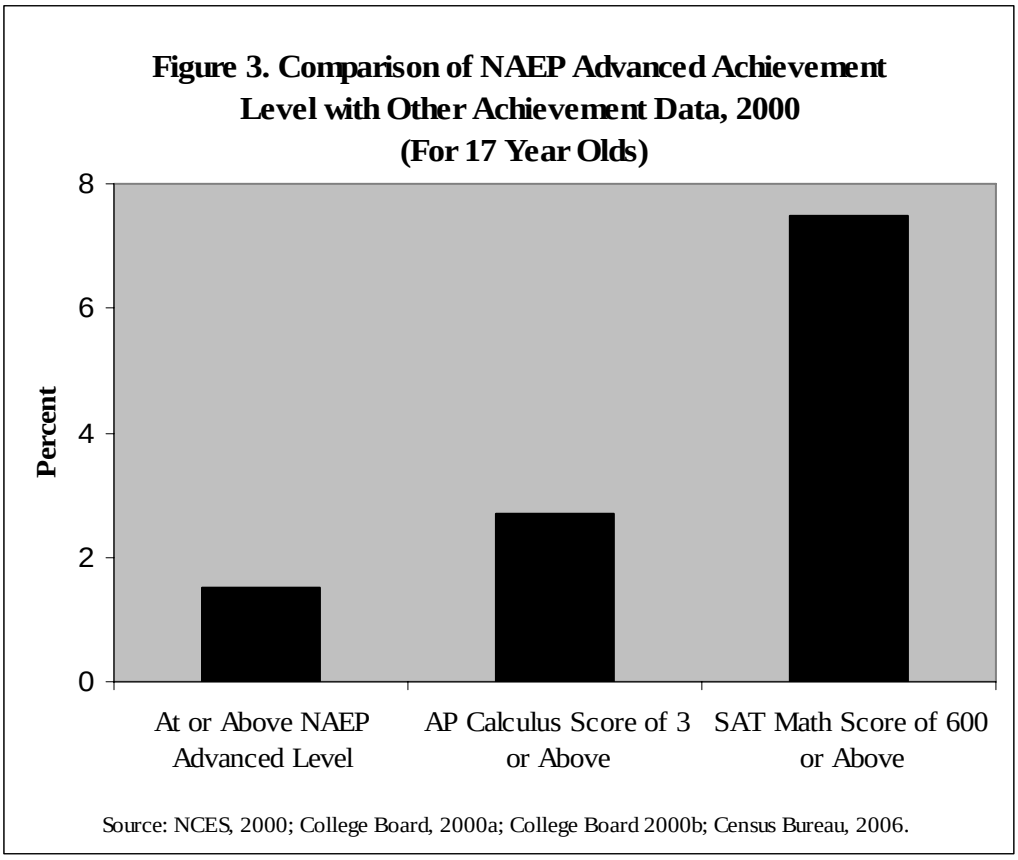
[†] One analysis concluded that only the cut scores for basic and proficient performance are too high, not those for advanced performance (GAO 1993). However, as the following discussion indicates, it is also plausible that the advanced level is too high. One of us has previously published criticisms of the NAEP proficiency standards that are substantially similar to those expressed here. The following discussion has appeared, in substantially similar form, in Rothstein (1998, pp. 71-74), and in Rothstein (2004 pp. 86-90).

[‡] This estimate takes the number of twelfth graders who performed at the advanced level, divided by the Census Bureau (2006) report of the size of the 17 year-old cohort in the year 2000. Other estimates in this paragraph are calculated with a similar methodology.

advanced placement exams - designed to measure not merely mastery of the high school curriculum, but mastery of beginning college mathematics itself. Advanced placement exams are given in only some U.S. high schools and, if available to all students, more than 2.7 percent of all 17 year-olds would likely have achieved passing scores.

Similarly, in 2000, 8 percent of all 17 year-olds, five times the number deemed to be at the advanced level by NAEP, scored over 600 on the SAT math test,[*] a score which most college admissions officials consider reflects advanced math achievement; the actual number of students who achieve at this level could be half again as high as 8 percent, because this number does not account for the fact that the SAT is not taken by many high-scoring college-bound students in states where the ACT is more common.[†]

[*] Estimated from: 304,870 of 2000 high school graduates scored over 600 on math SAT; the total population of 17 year-olds in July, 2000, was 4.04 million (College Board 2000; Census Bureau 2006).
[†] The "half again" (50 percent) estimate is very rough, reached because the number of college bound high school seniors taking the ACT is about 85 percent of the number taking the SAT. (In 2000, approximately 1.1 million seniors took the ACT and 1.3 million took the SAT.) However, some students take both the ACT and SAT. We assume that the level of math achievement is approximately similar in states where the SAT is dominant and in states where the ACT is dominant.



Grade-Level Standards

The administration uses the term 'grade level' interchangeably with proficiency, although 'grade-level' is not a term that appears in NCLB. For example, a recent Department of Education letter to state governments threatens sanctions for states that have not complied with NCLB requirements, warning that the NCLB "goal of having all students reach *grade-level standards* by 2013-14 is an urgent one" and so delays would not be tolerated.[15] And Secretary of Education Margaret Spellings now describes NCLB in this way: "We've set a historic goal to ensure every child—regardless of race, income, or zip code—can read and do math at grade level. And we've given ourselves a deadline to do it by 2014 because parents have waited long enough."[16] Nowhere has the

Department defined what 'grade level' means, or how it might differ from 'proficiency' as required by the Act.

As conventionally used, however, the term 'grade level' describes performance that is considerably below the standard of proficiency as defined by NAEP or NCLB. Grade-level performance usually means the average performance of students currently in a given grade. It is usually established by administering a standardized test to a national random sample of students in that grade. The average score is, by definition, grade-level performance in the base year in which the national sample was tested. And also by definition, approximately half of the students in the nation perform below grade level to some degree.[*] Increasing numbers of students, of course, can get above a previously-established grade level standard if achievement rises subsequent to the base year. But no matter how high average achievement becomes, approximately half of all students will demonstrate below grade-level performance for the year in which it is measured. When the Department of Education posits a goal of having all students at grade level, it presumably intends a contemporary grade-level standard, not an historic and obsolete one. If this is its intent, then all students at grade level is a logical impossibility.

[*] Approximately half, not precisely half, because the distribution of scores in the national random sample will not likely be perfectly normal. Educators often use the term 'grade level' more loosely, to describe not only students who are precisely at the average performance level for their grade, but to describe all students who perform better than the lowest-scoring 25 percent of students in their grade, but not as well as the highest-performing 25 percent of students in their grade (Rosenberg 2004). Using this description, half of all students, by definition, perform at grade level. Note that this description of grade-level performance is similar, but not identical to our description in this paper (see page 17, below) of student performance of students which is "similar to that of most same-age students," or within one standard deviation of the mean; i.e., the approximately 68 percent of students who perform better than the lowest-scoring 16 percent of students in their grade, but not as well as the highest-performing 16 percent of students in their grade. Yet whether the term 'grade-level performance' is taken to describe performance precisely at the median, or to describe the performance of the middle 50 percent or 68 percent of students, it cannot be used as the Department of Education now uses it, to describe a proficiency standard calculated without reference to the actual performance of today's students.

Even if we interpret the NCLB goal as all students achieving grade level standards, using standards of 2001 when NCLB was drafted, the task will be out of reach. Looking again at international comparisons, approximately one-fourth of students in Taiwan, the highest scoring nation in eighth grade math in 1991, were below the *average* performance of U.S. students. In other words, one-fourth of Taiwanese students performed below grade level for the U.S. eighth grade in 1991. And in 2003, approximately 10 percent of eighth grade students in Singapore, the highest scoring country in the TIMSS assessment, were similarly below grade level for the U.S. in that year.[*17]

Inevitable Individual Variability – Why 'Proficiency for All' is an Oxymoron

As these illustrations show, achieving proficiency for all is not simply a matter of adjusting the 2014 goal date by a few years. We claimed earlier that the slogan of 'proficiency for all' is an oxymoron, confusing a minimum standard, one that all (or almost all) students should meet, with a challenging standard, one that requires typical students to reach beyond their present level of performance. Even if schools were to improve so that typical students achieved a challenging level of performance, below-average students, even if challenged, would not reach the same level. If a challenging standard were achievable by below-average students, it would no longer be a standard that was challenging for typical students.

Think of it this way. Imagine you are teaching a class with a typical range of students. There are many 'C' students, but there are also quite a few 'B's, a few 'A's, a few

[*] Calculated by estimating the standard deviation from the mean, using the Singapore score distribution, of a Singapore student's score that was equivalent to the U.S. average score, and then by estimating the probability (10.38) of a Singapore student achieving below that score.

'D's and maybe even an 'A+' student in the mix. Imagine that your principal tells you to ignore the 'A's and 'A+'s – you won't be held accountable for whether they learn much new this year. But for the rest of the students, you are told you must teach effectively enough so that you can conclude the year with a test which has two characteristics: First, all students will pass. Second, the test will be hard enough so that both your "B" students and your "D" students will find it a challenge to get a passing grade. We defy you to design such a test, no matter how effective your teaching may be. It is a logically impossible task.

There is no aspect of human performance or behavior that is not achieved in different degrees by individuals in a large population. There is an average level of math performance for eighth graders, but some perform above or below that level. There is an average level of teaching ability for eighth grade math teachers, but some perform above or below that level. There is an average susceptibility to influenza, an average pace to run a mile, an average height and weight for adults, an average inclination to attend church each week, an average skill in the operation of motor vehicles. In each of these areas, some individuals are considerably above average, many are slightly above average, many are slightly below average, and some are considerably below average. In most of these areas, the distributions are close to what statisticians call normal (when plotted, the resulting graph looks bell-shaped), but perfect normality is not the rule. In general, however, for distributions that are close to normal, we say that roughly two-thirds of all humans perform reasonably similarly on any characteristic – statistically speaking, we say that the approximately one-third who perform slightly below average are within one standard deviation of the mean, and the approximately one-third who perform slightly

above average are also within one standard deviation of the mean. But this still leaves about one-sixth who are considerably below average, as well as about one-sixth who are considerably above.

In its administration of NCLB, the U.S. Department of Education barely acknowledges this human variability. It permits the lowest performing 1 percent of all students to be held to a vague "alternate" standard of proficiency, and the next lowest performing 2 percent to be held to a "modified" standard of proficiency, which still must lead to "grade level" achievement and to a regular high school diploma.[18] Let's be clear about what this means: Under NCLB, children with I.Q.s as low as 65 must achieve a standard of proficiency in math which is higher than that achieved by 60 percent of students in Taiwan, the highest scoring country in the world (in math), and a standard of proficiency in reading which is higher than that achieved by 65 percent of students in Sweden, the highest scoring country in the world (in reading).[*]

[*] Initially, Department regulations permitted schools to exempt only students with the most severe cognitive disabilities, not to exceed 1 percent of all students, from regular accountability testing. Such students were still required to show adequate yearly progress towards a "modified" grade level proficiency standard, established by a similar process to that used for defining proficiency for all students (Lee 2003). In other words, the Department assumed that students who were approximately 2.3 standard deviations and more below average should have been expected to perform reasonably similarly to average and above-average students. This group includes students who are classified as mildly mentally retarded, expected under NCLB to meet a somewhat modified "grade level" proficiency standard. (Students are classified as mildly mentally retarded if they have I.Q. scores between 50 and 70. A little less than half of such students, those with I.Q.s below 65, are in the bottom 1 percent of all students in cognitive ability. A little more than half of mildly mentally retarded students have I.Q. scores between 65 and 70. These students typically can be expected to finish high school with academic achievement up to a sixth grade level [Gurian 2002].) The expectation that grade level proficiency could be modified only for the bottom 1 percent was so egregious that in 2005 the Department responded to complaints by proposing that an additional 2 percent of students (including, most probably, students with I.Q.s between 65 and 72) could be assessed based on “modified achievement standards” (Saulny 2005; DOE 2005a, 2005b), and it now characterized the standard required of the bottom 1 percent as an "alternate," not modified, achievement standard. However, the proposed rule specifies that the modified standards should still be "aligned with grade-level content standards, but are modified in such a way that they reflect reduced breadth or depth of grade-level content," yet would not preclude such "a student from earning a regular high school diploma" (DOE 2005b). Such language leaves it entirely unclear how the achievement standards can actually be modified, and suggests that ineffective and unaccountable school practices are the cause of even the most able of these students typically achieving only sixth grade academic levels and failing to earn regular high school diplomas. The Department's

Nonetheless, Secretary of Education Margaret Spellings rejects calls for substantial modifications to NCLB, claiming the law is "99.9 percent pure." And although this may have been a flippant comment, she later clarified her remark, saying that she "meant to convey only that no changes were needed to its 'core principles.'"[19]

Determining Realistic Goals for Improvement

More reasonable supporters of contemporary school accountability policies acknowledge that not all children are alike and there is an inevitable distribution of achievement, but say that, at least, the distribution can be narrowed by targeted policies that raise the performance of students at the bottom at a faster rate than performance improves for students overall. Then the gap between a proficient and a minimum standard need not be so great as it is today.

In addition to raising all parts of the distribution (i.e., shifting the distribution 'to the right'), it is also possible to narrow the distribution somewhat, but probably not by

---

proposed rule states that requiring such students to achieve adequate yearly progress toward grade-level standards "would provide a safeguard against leaving children behind due to lack of proper instruction" (DOE 2005b). The Department's rule also asserts that students between the bottom 1 percent and bottom 3 percent can achieve proficiency, but these students may need more time to do so, which is the only reason for using temporary modified standards: "[W]e acknowledge that, while all students can learn challenging content, certain students, because of their disability, may not be able to achieve grade-level proficiency within the same time-frame as other students..." (DOE 2005b). As Daniel Koretz (2006b) has observed, "The proposed regulations are impractical, I think. They call for standards and assessments that are 'modified in such a manner that they reflect reduced breadth or depth of grade-level content.' But if, each year, you cover grade level material in less breadth or depth, over time, students will fall behind grade level." Three states (Kansas, Louisiana, and North Carolina) have now begun to use modified assessments for these students (Samuels 2006), with panels of teachers and other "educational stakeholders" establishing proficiency cut scores for students with I.Q.s as low as 50 (NCBOE 2006, p. 19). The proposed U.S. Department of Education rule, permitting a total of 3 percent of students to be held accountable for alternate, or modified, but standardized, proficiency standards, was published in the *Federal Register* on December 15, 2005, but has not yet been formally adopted, although the comment period closed on February 28, 2006. Even this new 3 percent exemption will doubtlessly be modified in any NCLB re-authorization that passes Congress. But the fact that we are even debating whether children with below-normal mental capacity should achieve a standardized definition of proficiency is breathtaking. Even if the intent of the proposed regulation were clear, it would still hold schools accountable for getting students with IQ scores as low as 72 to proficiency according to the regular, unmodified, grade level standards.

very much. As Daniel Koretz points out in an important new paper, the typical variation in children's achievement, the gap between children at lower and higher levels of academic competence, is not primarily a racial or ethnic gap; it is a gap within race and ethnic groups, including whites. The range of student performance in Japan and Korea, more homogeneous societies than ours whose average math and science scores surpass those of the United States, is similar to the range here. Professor Koretz estimates that if the black-white gap were entirely eliminated, the standard deviation of U.S. eighth grade math and reading scores would shrink by less than 10 percent. Perhaps some additional shrinkage would result if we were able to reduce the race-neutral achievement gap by family income, but even so, most of the existing variability in student performance would remain.[20] It would still be just as impossible to craft a standard which was a simultaneous challenge to students at the top, the middle, and the bottom of the distribution.

One way to establish boundaries on what might be reasonable expectations for improvement would be to examine historical precedent. It is generally agreed that since about 1963, U.S. student achievement has gone through three distinct phases. At first, and until the late 1970s, achievement declined. Then, until the late 1980s, achievement rose. And from about 1990 to the present, math scores have continued to climb while reading scores have been mostly stagnant or have declined slightly.[21]

The test score decline of the 1960s and '70s was considered very significant, a national crisis. The necessity of arresting this decline, a "rising tide of mediocrity," was an important motivation behind the *Nation At Risk* report of 1983.[*] How large was the decline? As the report stated, on the College Board's SAT, average verbal scores fell by

[*] In actuality, when *A Nation At Risk* was issued, the score decline had already ended and, indeed, scores had been rising again for several years. But the National Commission on Excellence in Education, author of the report, was apparently not aware of this development at the time.

about 50 points and math scores by about 40.[22] Overall, average test scores dropped by a similar amount, about 0.4 standard deviations.[23] Social, cultural, and economic factors were responsible for some of this decline – for example, children from larger families typically achieve at lower average levels than children from smaller families, perhaps because children from large families get less adult attention, and the score decline corresponds to the period when baby boomers moved through schools.[24] So perhaps (this is just a guess) the decline in average achievement attributable to a deterioration in school quality was about 0.2 standard deviations. In other words, at the end of this period, typical students (i.e., those at the 50th percentile in a ranking of all students by their performance) scored similarly to students who were at about the 34th percentile at the beginning of the period, and if we controlled for non-school factors, we might say that deteriorating school quality caused a typical student to fall to about the 42nd percentile.

So if we thought a school improvement program could cause a student achievement gain equal in size to the decline caused by the deterioration of school quality some four decades ago, we might aim for a situation where, 15 years hence, typical students achieve at about the level that students at the 58th percentile achieve today.

The second phase, from about the mid 1970s to the late 1980s, saw student achievement (on NAEP math tests) rise by 0.2 standard deviations in the twelfth grade, and by more than 0.3 standard deviations in the fourth and eighth grades.[25] David Grissmer has estimated that the black-white gap was cut in half during this period, and that about half of the cut was attributable to family factors (smaller black families, and higher levels of black parental education). This would leave about half of the decline in the black-white gap otherwise unexplained, possibly attributable to school

improvements.[26] Using this period of improvement as a yardstick, we might try to improve schools again at a similar rate, to get typical students, 15 years from now, up to achievement levels of today's students who are in the mid-50s in percentile ranks. Keep in mind, as always, that this aspiration describes the movement of typical students, those at about the middle of a national distribution. Some will improve to higher levels, and some will regress to lower ones.

In behavioral science, an intervention designed to improve human performance is generally considered effective but small if it improves average performance by 0.2 standard deviations; medium if it improves average performance by 0.5 standard deviations; and large if it improves average performance by 0.8 standard deviations.[27] In other words, at the conclusion of a moderately successful intervention, average individuals will perform at the level that individuals who were previously at the 69th percentile were performing. Other individuals, those above and below average, would also perform at a correspondingly higher level. But such interventions are rare, especially if measured over reasonably proximate time spans.[*]

However, in education, where good experimental controls are absent, large effect sizes are less probable than in fields like experimental psychology or medicine.[28] In the field of education, it seems reasonable to classify an effect size of 0.5 as quite large; a

[*] Certainly, the effectiveness of medical doctors is more than one standard deviation higher than it was one hundred years ago, survival rates for most diseases are more than one standard deviation above what they were one hundred years ago, and life expectancy is more than one standard deviation longer than it was one hundred years ago. Occasionally, a technological breakthrough has a short-term result of such magnitude. Survival rates of HIV patients jumped by more than a standard deviation with the development of antiretroviral drugs, as did survival rates from heart disease with the development of surgeries such as pacemakers and by-passes. "If the statistics of 1940 had persisted, fifteen thousand mothers would have died [in childbirth] last year (instead of fewer than five hundred)—and a hundred and twenty thousand newborns (instead of one-sixth that number)" (Gawande 2006). But breakthroughs are rare, and cannot be the model for ongoing educational reform efforts which, absent unforeseeable breakthroughs, must be incremental.

more practical standard for successful educational reform would be one that shifted average performance by something like a quarter to a third of a standard deviation within the foreseeable future, or one that enabled typical students to perform at the level that students who were previously at the 60th to 63rd percentile were performing.

If the United States was to revamp its schools and child welfare practices to be first in the world in math, this would require that typical American eighth graders (those who were at the 50th percentile rank in a national distribution of student achievement) would perform at the level that eighth graders who were at the 90th percentile of math achievement performed before the intervention began.[*29] This would be an upward shift in average performance of about 1.3 standard deviations, a magnitude of accomplishment that would be extraordinary in any field. And, as noted above, even if all students' achievement improved under such a regime, about half would still perform below the new, higher average level.

We can again translate this to classroom reality. Any teacher understands how it works. She can set high expectations for student performance, and may successfully elicit high average performance for her class. A teacher, for example, might expect 'C' students to raise their sights so that they produce 'B' work. But this teacher would never make 'B' level work the minimum passing grade. She would understand that some students would earn only a 'C' and would still be eligible to advance to the next grade, even under the strictest of no-social-promotion policies.[†]

---

[*] Estimated by calculating, for the 2003 TIMSS math assessment, the percentile rank of a U.S. student whose score was the same as the mean score for the highest scoring country (Singapore).

[†] Professors teaching graduate students do make 'B' a passing grade, but all students do not achieve this: many are not admitted to graduate study because they do not represent the top of the distribution of academic skill in this particular field, and some who are admitted cannot earn a minimum 'B' grade and never receive a Ph.D.

NCLB's Goal Confusion

What NCLB has done is the equivalent of demanding not only that 'C' students become 'A' students nationwide, but that 'D' and 'F' students also become 'A' students. As noted above, this confuses two distinct goals – that of raising the performance of typical students, and that of raising the minimum level of performance we expect of all, or almost all students. Both are reasonable instructional goals. But given the nature of human variability, no single standard can possibly describe both of these accomplishments. If we define proficiency-for-all as the minimum standard, it cannot possibly be challenging for most students. If we define proficiency-for-all as a challenging standard (as does NCLB), the inevitable patterns of individual variability dictate that significant numbers of students will still fail, even if they all improve. This will be true no matter what date is substituted for NCLB's 2014.

Categories of performance in NAEP reflect these realities. Cut scores are established not only for proficiency, but for advanced and basic performance as well.[*] Continuing to use eighth grade mathematics as a representative illustration, reasonable goals for school and student improvement might be to reduce the 32 percent share of public school students who now perform below the basic level, and increase the 28 percent share who now perform at or above the minimally proficient level.[30] A single goal, however, cannot serve both purposes, as NCLB requires.

[*] For advanced performance in eighth grade math, students must score at least 0.935 standard deviations above the mean, and for basic performance students must be below the mean, but no more than 1.02 standard deviations below (cut scores from Loomis and Bourque 2001a; standard deviations from IES 2006).

In response to this argument, some note that commonplace tests do exist in which all test takers are expected to be proficient. A state written driver's license exam is, it is said, an example, and if we can expect all drivers to be proficient at understanding the rules of the road, we can also expect all fourth graders to be proficient in math.[*] This analogy fails for two reasons. First, not *all* test takers pass the written exam to get a drivers' license, although almost all do, eventually, having taken the test multiple times, something not permitted in contemporary school accountability systems.[†] Second, anyone who has taken a drivers' license test knows that the level of difficulty is extraordinarily low. Passing can be assured by devoting only a few minutes to review of the state manual published for this purpose. Nobody would call a state driver's test "challenging." If it were indeed challenging, the goal of having everyone pass would be no more in reach than the goal of having all fourth graders proficient in math. If we define proficiency low enough (perhaps something less than the basic level on NAEP), it is certainly possible to achieve the NCLB goal of having almost all students proficient by some future date. But such a standard would not be challenging to most students, and would do little to spur typical students to perform at higher levels than they do today.

[*] Typical is a Washington State Department of Education website with "frequently asked questions" about the state's testing and accountability program: "Think of the [state test] like the test you take to earn a driver's license. It doesn't matter what the average score on the test is or whether some drivers scored above or below you. What matters is whether you can show you have the driving skills and knowledge of traffic laws to 'meet the standard' and get a license" (WASL 2006). A parent involvement program promoted by the U.S. Department of Education describes NCLB's testing requirements like this: "If teachers cover the subject matter required by the standards and teach it well, students should do well on the test. It's like taking a driver's test. The instructor covers all the important content the state wants you to know and much more" (Project Appleseed 2006).

[†] In part, they eventually pass because their skills improve with more study. In part, test takers become familiar with the specific questions asked by the test, even if they don't become better at understanding traffic rules overall. And in part, they eventually pass because there is an element of chance involved in selecting the answer to any multiple choice question, and both the first failing score and the later passing score are statistically indistinguishable. In these respects, a driver's exam is similar to school achievement tests.

The Unintended Consequences of a Shift from Norm-Referenced to Criterion-Referenced Reporting

The origins of this confusion between a minimum and an aspirational standard can be traced to a shift in how we describe student achievement. In the past, certainly until 30 years ago, it was usual to describe achievement in norm-referenced terms. After administering a standardized test to a nationally representative sample of students in a given grade, psychometricians calculated a median score and then reported each student's score, and each subgroup of students' average score, in reference to this national mean.

Such norm-referenced reports took a variety of forms. In cases where the public was familiar and comfortable with a test, scale scores were reported with no further explanation. An example of a norm-referenced test with which we are fully comfortable is the college admission test, the SAT, where instead of reporting that a given student is at the 84th percentile, or approximately one standard deviation above the mean in the base year when the test was first normed, we say that the student has a score of 600. By definition, 600 is the score of an 84th percentile test taker in the original national sample (and 500, the mean score of the sample). The scale score of 600 has no absolute significance; it is only a norm-referenced score, a convention with which the public is familiar. When we are told that a student scored 600 on the SAT math test, most of us, comfortable with this norm-referenced scale, have an intuitive understanding of what that student can do.[*]

[*] This is actually an oversimplification, and no longer strictly true. Last year, the SAT was modified, a third (writing) test was added, and the verbal test changed and renamed the "critical reasoning test." The new writing test was established, consistent with previous practice, with a mean of 500, but the standard deviation was established as 110, not 100. The critical reasoning and math tests were not renormed at this time, with the result that 500 on these tests represents the mean score in 1991, while 500 on the writing test

Initially, NAEP was also reported only in scale scores, with the possibility of norm-referenced interpretations, but because the public was unfamiliar with the NAEP scale, and because NAEP tried to get too fancy by employing different scales for different grade levels, NAEP scale score reports have no intuitive meaning for the public.* Unlike the SAT, which uses the easily remembered number 500 for average performance, and the easily remembered 400 or 600 for performance that was one standard deviation below or above average in the base year (1991) when the SAT was last normed, NAEP scale scores for public schools are now, for example: 237 for average fourth grade math performance (and 266 for one standard deviation above); 278 for average eighth grade math performance (and 314 for one standard deviation above); 217 for average fourth grade reading performance (and 253 for one standard deviation above); 260 for average eighth grade reading performance (and 295 for one standard deviation above); and so on.[31] Because neither the public nor even relatively sophisticated policy makers could ever become familiar with these needlessly complex conventions, Congress sought an alternative.

But while the desirability of defining performance levels as an alternative to scale scores was advocated by many policy makers, little consideration was given to the complexities involved. In the 1988 Congressional reauthorization of NAEP, the Senate bill instructed the National Assessment Governing Board (NAGB), established to administer the test, to "identify feasible achievement goals for each age and grade in each

---

represents the mean in 2005. With these new more complicated properties, scores on the SAT may become less intuitively meaningful.

* NAEP scales for higher grade levels use a higher, but overlapping series of numbers. This unfortunate complication leads to a misunderstanding, even by relatively sophisticated educators, that the scales can be combined into a continuous series, and that a fourth grader, for example, who achieves a score equal to that of the typical eighth grader, is capable of doing "eighth grade work." Such a conclusion, however, is unwarranted. Fourth graders who achieve such a score are not answering the same questions as typical eighth graders.

subject area under the National Assessment." The House bill made no mention of achievement goals but the final bill that emerged from the conference committee somehow substituted the word "appropriate" for "feasible," so NAGB was now instructed to identify *appropriate*, not feasible, achievement goals.[32]

We've Been Warned about Irresponsible Achievement Levels, but Proceeded Anyway

There is a considerable difference between feasible goals, which must be grounded in reality, and appropriate goals, which can mean anything the goal-setters choose. When NAGB attempted to carry out Congress' intent, it asked Terry Hartle who, as Senator Edward Kennedy's chief education staff member, was a drafter of the bill, to explain what it meant. Hartle testified at a NAGB hearing that Congress' choice of language was "deliberately ambiguous" because neither Congressional staff nor education experts were able to formulate it more precisely. "There was not an enormous amount of introspection" on the language, Hartle reported.[33]

A few experts protested at the time. One was Harold Howe II, former U.S. Commissioner of Education, who had played an important role in developing the NAEP some 20 years before. Howe wrote to the Commissioner of Education Statistics,

> …[M]ost educators are aware that any group of children of a particular age or grade will vary widely in their learning for a whole host of reasons. To suggest that there are particular learnings or skill levels that should be developed to certain defined points by a particular age or grade is like saying all 9th graders should score at or above the 9th grade level on a standardized test. It defies reality.[34]

Nor was Mr. Howe the first to sound such a warning. Six years before, when momentum was first building for NAEP reports that went beyond scale scores, the federal

governing body for NAEP, then called the Assessment Policy Committee, asked three foundations (Carnegie, Ford, and Spencer) to finance a year-long study of NAEP and how it should be improved. The foundations commissioned former U.S. Labor Secretary Willard Wirtz and his colleague, Archie Lapointe, to conduct the study and they, in turn, convened an advisory council including the president of the Educational Testing Service, prominent scholars, a corporate (IBM) official, and the deputy director of Great Britain's comparable educational assessment program.[35]

The Wirtz-Lapointe report, presented to the Assessment Policy Committee in 1982, recommended that NAEP develop descriptions of what students know and can do if they achieve scale scores at various levels. But the report warned that NAEP should not go further and define passing points or cut scores which, the report said, "would be easy, attractive, and fatal… Setting levels of failure, mediocrity, or excellence in terms of NAEP percentages would be a serious mistake… [t]he ultimate conclusions as to the levels of student achievement that are to be considered good or bad must be left to the users of the Assessment information [at the local or possibly the state level]," the report concluded. "[T]he making of judgments about what is 'good' or 'bad' is reasonable and responsible only in terms of particular educational environments."[36]

Most policy makers, however, endorsed the idea of a defined achievement level that would indicate whether U.S. students passed or failed the NAEP exam or, in the common phrasing of the time, whether students actually did as well as they ought to do. So in 1990 NAGB, while retaining the NAEP scale scores, adopted criterion-referenced NAEP reporting as well, acknowledging, with no apparent embarrassment, that

"appropriateness is a matter of taste."[*37] Although initially NAGB intended to define only one achievement level, proficiency, it eventually decided to establish three points on each NAEP scale to describe achievement levels – basic, proficient, and advanced – and reported group scores (there are no individual scores on NAEP) as either below basic, at least basic but not yet proficient, at least proficient but not yet advanced, and at least advanced.

For twelfth graders, proficiency was defined as the level of performance that *all* students should achieve; as NAGB put it: "At grade 12 the proficient level will encompass a body of subject-matter knowledge and analytical skills, of cultural literacy and insight, that *all* high school graduates should have…;"[38] or, as NAGB policy further explained it, "the knowledge and skills all students need to participate in our competitive economy… and the levels of proficiency needed to handle college-level work."[†39] At the time NAGB made this pronouncement, approximately 29 percent of all 17 year-olds eventually went on to graduate from college. NAGB's unexamined assumption that the NAEP proficiency standard could be defined at a level that would more than triple this rate to something like 100 percent is another illustration of the fanciful thinking underlying the achievement level process.[‡]

---

[*] "As well," because there is a commonplace misunderstanding that tests are either norm-referenced or criterion-referenced. As this NAEP illustration shows, scores on a standardized test that is designed to assess a wide range of student performance can be reported either in norm-referenced or criterion-referenced terms. NAEP is not a norm-referenced or criterion-referenced test. Its results can be reported in either convention. The same is true of state tests used for accountability purposes under NCLB, although many of these tests are now only reported to the public in criterion-referenced terms, although originally the same, or very similar tests were reported only in norm-referenced terms.

[†] In 1994, a new generic definition of proficiency was adopted by NAGB, which retained from 1990 the phrase "competency over challenging subject matter," but which no longer included a reference to "all high school graduates" (Brown 2000, p. 15). However the new definition did not result in lowered cut scores; on the contrary, at least some cut scores increased (see footnote on p. 9, above).

[‡] In 2001, the 1991 cohort of 17 year-olds would have been 27 year of age. In 2001, 29 percent of the 25 to 29 year-old age group had earned a bachelor's degree (NCES 2004, Table 8). Some qualifications are needed: not all 17 year-olds who have the level of proficiency needed to handle college-level work may

Subjectivity in Determining Cut Scores

Today, many state standardized tests, used for measurement purposes under NCLB, use similar terminology about the competitive economy and college readiness to describe their cut scores.

There is nothing scientific about establishing these cut scores. There are several methods for doing so.[40] But all available methods require subjective decisions of panels of judges who decide what constitutes proficiency for a particular subject and grade – although the federal and state governments are rarely so candid as NAGB was initially in stating that appropriate achievement levels are simply a matter of taste.

One common method is to ask each judge to imagine what a barely proficient student can generally do, and then estimate, for each question on a test, the probability that such a student will answer the question correctly.[*] When each judge's estimates for such a probability for each question on the test are averaged together, and all the judges' average estimates are averaged together, the result is the minimum test score (in percent correct) that a student must achieve to be deemed proficient. This is the method used to set the proficiency cut score for NAEP. Similar exercises were used to define basic and advanced performance.[†41]

The National Assessment Governing Board, consisting of 26 governors and other state education officials, classroom teachers, teachers' union officers, school

---

enroll in college, and not all those who enroll may graduate. Although not all those who enroll have the necessary proficiency to handle college-level work, we assume that those who eventually graduate do so.

[*] Another way to pose the same question is to ask judges to imagine a group of 100 barely proficient test-takers, and to estimate how many of them will answer particular questions correctly.

[†] In the psychometric literature, this is referred to as the "Angoff method," or, when used to establish multiple cut scores (basic, proficient, advanced), the "modified Angoff method."

administrators and academic experts, hired contractors who, in turn, appointed panels of teachers, professors, business leaders and other citizens to decide which NAEP questions a student should be expected to answer correctly if that student were deemed to be at the basic, proficient, or advanced levels.* The panelists were given no standard by which to make these judgments except their own opinions. NAGB's effort, in 1991, to use such panels to establish cut scores in mathematics were discarded because the panelists' judgments were so inconsistent. When NAGB formed panels of judges in 1992 to set NAEP cut scores in all subjects, it also established a new panel to re-define the cut scores in math.[42]

In the case of mathematics, for example, NAGB established three panels of 20 judges, one for each of the grade levels tested – fourth, eighth, and twelfth. Of the 60 judges, 33 were schoolteachers, 9 were other educators, and 18 were members of the general public.[43] After each panelist decided what percentage of NAEP questions a basic, proficient, or advanced student should answer correctly on each test, the percentages established by all panelists were then averaged together.[44] There was wide variation in the panelists' opinions, confirming that an average might incorporate great subjectivity.

It might be hoped that those making these judgments have in mind students they have known who get adequate grades, but if so, the judgments will likely be flawed. Few teachers or other educators have had deep experience with a fully representative group of students. Most teachers spend their careers in similar communities where students'

* NAGB required that 55 percent of these judges be classroom teachers, and another 15 percent be other educators (curriculum specialists, principals). The remainder could be non-educators such as parents of school children or business executives who employ recent high school graduates. The contractor, American College Testing (ACT) selected all the teacher judges from nominees of school district superintendents, teacher union officers, and private school executives. The non-teacher educator judges were selected from nominees of faculty of schools of education. The non-educator judges were selected from nominees of local Chambers of Commerce, mayors, and chairs of school boards (ACT 1992).

demographic characteristics and average ability levels are different from those of students who live in other types of communities.

Yet even if it were reasonable to expect that the judgment of teachers and other educators who have had experience with a representative group of pupils in the targeted grades would be valid regarding whether students would be likely to answer particular questions correctly, there is no reason to defer to their judgment (or taste) regarding whether answering those questions correctly should be deemed proficiency. Defining proficiency is a subjective process which does not mostly rely on experience.[*]

It is apparent that where the cut point is determined to lie is a function of the identity and qualifications of the judges upon whose subjective opinions the decision rests. A manual by the Educational Testing Service (ETS) states that the judgments should be "made by persons who are qualified to make them," but provides no specific guidance regarding what the qualifications might be.[45] Had NAGB considered the question of qualifications too carefully, it might have abandoned the process because, regardless of qualifications, "appropriateness is a matter of taste." Ultimately NAGB was more interested in getting judges who were representative of various constituencies

[*] We do successfully define proficiency for professional certifications. Physicians, accountants, hairdressers and others obtain licenses to practice only after satisfying boards of examiners that they possess proficiency in their fields, and passing examinations is part of these processes. But the professional boards that establish cut scores on such examinations usually attempt to maintain existing professional standards in the licensing of new practitioners. These boards do not use cut scores as a way of radically raising the existing level of professional practice. As a result, the boards can rely on experience to determine what competent physicians, accountants, or hairdressers can actually do; they do not use cut scores to require newly licensed practitioners to perform at radically higher levels than most existing practitioners. In practice, cut scores on such licensing examinations are often changed based on supply and demand factors: if there is a shortage of job seekers in a profession, licensing boards lower the cut scores for passing the test, confirming that even for licensing exams, the concept of proficiency has no objective meaning (Glass 1978). Further, there is considerable selectivity in the pool of candidates. NCLB achievement levels must apply to all students. But not all young people are qualified to enter training to become candidates in medicine, accounting or hairdressing, so the variability in performance among candidates is much narrower than among all students. And then, professional schools are expected to weed out students who were admitted, but are not likely to pass the licensing exam at the end of their training.

(teachers, business leaders, etc.) than it was in worrying too much about how judges could be qualified to make valid judgments about what students ought to be able to do, as distinct from what most students are actually able to do.

These subjective judgments, while well intentioned, would lead to overestimates of proficient performance even if judges had personal experience with a fully representative group of students. This overestimation occurs because when teachers and educators, as well as members of the general public, think about proficiency, they don't only have in mind students they have known who get adequate grades. Rather they tend to think of a performance level that is higher than students actually achieve, but one that they hope students will achieve or think students should achieve. It is a rare teacher who considers that her students' average performance should not have been higher than it was, if only the students had tried a little harder, parents could have been persuaded to be a little more supportive, the teacher had organized the curriculum a little differently, or some distracting event had not occurred during the school year. So it is not surprising that the NAGB judges established definitions of NAEP achievement levels that were unreasonably high, despite the judges having gone through several days of training designed to avoid that very result.

In 1978, long before NAGB began to define cut scores for NAEP, a measurement expert warned that judges will almost invariably set unrealistically high criteria. He described the mental process that judges typically apply as one of "counting backwards from 100%":

> An objective is stated and a test item is written to correspond to it. Since the objective is felt to be important – or else it wouldn't have been stated – its author readily endorses the proposition that everyone should be able to answer the test question based on it;… But reason and experience prevail

> and it is quickly recognized that perfection is impossible and concessions must be made for mental infirmity, clerical errors, misinformation, inattention, and the like. Just how great a concession should be made becomes distressingly arbitrary….[46]

If this is an accurate description of how NAGB judges approached their task, it is not surprising that NAEP proficiency is defined unreasonably stringently. The mental burden of proof, as it were, falls on deviations from perfection.

Following the work of the first panels in 1991, NAGB determined that the math cut scores that had been established were in fact too high, so it simply reduced the judges' decisions.[*47] NAGB's action was arbitrary, as were the judges' decisions themselves. We noted above that, according to the proficiency definitions ultimately adopted by NAGB, 25 percent of students in Singapore, the highest-scoring country in the world in TIMSS eighth grade math, were less than proficient. Had the panelists' own judgments been maintained by NAGB, then approximately 32 percent of Singapore's students would be deemed less than proficient.[†]

After the first achievement levels for math were established, NAGB conducted a forum to hear public comments about the new standards. Mary Harley Kruter, one of the judges who participated in the process to establish cut scores, was the mathematics

---

[*] NAGB "decided that the mathematics standards were too stringent…. The cutscores for all grades and levels were set one standard error below the original overall composite cutscore computed from panelists' ratings."

[†] Some information is available regarding judges' actual decisionmaking. A report to NAGB by American College Testing (ACT), the organization contracted in 1992 to conduct the exercises to establish achievement levels, sheds some light on the process. The report describes how ACT judges established achievement levels for the NAEP writing assessment. In order to try to get the judges to be more consistent, in their own ratings and with each other's, the judges were shown how their ratings compared with others and were then asked to go through subsequent rounds of reading and rating each writing sample. The ACT report includes a chart displaying the second-round decisions of each judge on eighth grade writing samples. This chart shows that: 1) there was wide variation from judge to judge in determining whether particular samples were basic, proficient or advanced; and 2) most judges found large differences between basic writing and proficient writing, but little difference between proficient and advanced writing (ACT 1992, p. 12, Figure 2, "Example of Round 2 Interjudge Consistency Feedback"). The cut scores eventually adopted by NAGB for eighth grade writing (Basic, 114; Proficient, 173; Advanced, 224), however, are more nearly equidistant (Loomis and Bourque 2001b).

education project director for the National Academy of Sciences. At the NAGB forum, Ms. Kruter testified that her panel had too little time to make reasonable judgments about the cut scores: "We were uncomfortable that we did not do the best job we could do," she said. "It was a rushed process." Greg Anrig, president of the Educational Testing Service which was administering the NAEP, urged NAGB to delay employing achievement levels until it could be certain that they had been established properly.

In response, Chester E. Finn, Jr., a former NAGB chairman and still a board member, explained why the board was unwilling to delay the use of cut scores to report on the percentage of students who are proficient: If we delay, Mr. Finn stated, "we may be sacrificing something else - the sense of urgency for national improvement."[48]

NAGB itself has issued contradictory statements regarding how seriously its achievement levels should be taken. In 1990, it stated that the proficiency level represented merely "acceptable" achievement.[49] More recent NAGP publications acknowledge that the definitions are indefensible, although NAGB continues to use them. As a NAGB report put it in 2001:

> 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. Further, Basic achievement is more than minimal competency. Basic achievement is less than mastery but more than the lowest level of performance on NAEP. Finally, even the best students you know may not meet the requirements for Advanced performance on NAEP. [50]

In the early 1990s, NAGB, Congress, and the Department of Education all commissioned studies to evaluate the achievement level setting process and the validity of the results. Each study concluded that the achievement levels were flawed and urged NAGB to discontinue their use, or to use them only with the most explicit warnings about

their unscientific nature. The government's response to each of these studies was to commission yet another study, hoping that a different group of scholars would emerge with a more favorable conclusion.

The first of these studies, by three well-known and highly respected statisticians, was conducted in 1991, following the initial efforts of judges to establish math cut scores.[*] According to the statisticians' preliminary report, "the technical difficulties [with NAGB's achievement level definitions in math] are extremely serious" and to repeat the process in new standards setting exercises for other subjects would be "ridiculous." The statisticians concluded that NAGB was technically incompetent and that Congress should reconstitute it with members who had more psychometric sophistication. NAGB's response was to cancel the statisticians contract before the final report could be submitted.[51]

But the statisticians views had been publicized, so the House Education and Labor Committee asked the General Accounting Office (GAO) to decide whether the statisticians were right in their indictment of NAGB's standards-setting process. In 1993, the GAO released its report entitled *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations*, and concluded that defining cut scores is not a task that lay people can reasonably perform. For example, the GAO found that NAGB panel members, whose judgments were averaged regarding the probability of students at different achievement levels answering each test item correctly, could not properly distinguish between easier and more difficult test items. For example, judges had a tendency to classify open-ended items as difficult, when they were not necessarily so, and multiple choice items as easy, even when they were not. As a result students whose

[*] The authors were Daniel L. Stufflebeam, Michael Scriven, and Richard M. Jaeger.

NAEP scores are only at the basic level actually answer correctly more easy questions than the standards-setting panelists predicted. The same is true of students whose scores are at the proficient level. Students at the advanced level, in contrast, answer fewer difficult questions correctly than they are expected to do, but pump up their average scores by answering a higher percentage of easier questions than they are expected to do. Therefore, the GAO concluded, the cut scores for basic and proficient students could have been set considerably lower than they were, based on the NAGB panel's own standards. This would result in much larger numbers of students being deemed proficient.[52] The GAO reached these conclusions:

> We conclude that NAGB's… approach was inherently flawed, both conceptually and procedurally, and that …the approach not be used further until a thorough review could be completed…
>
> These weaknesses are not trivial; reliance on NAGB's results could have serious consequences. For example, policymakers might conclude that since nearly 40 percent of 8th grade students did not reach the basic level…, resources should be allocated so as to emphasize fundamental skills for most classes. Since many students who scored below 255 [the cut score for basic performance] were in fact able to answer basic-level items (according to our analysis), this strategy could retard their progress toward mastering more challenging material…
>
> In light of the many problems we found with NAGB's approach, we recommend that NAGB withdraw its direction to NCES that the …NAEP results be published primarily in terms of levels. The conventional approach to score interpretation [i.e., reports of scale scores] should be retained until an alternative has been shown to be sound.[53]

Indeed, the GAO's warning of the "serious consequences" to follow from use of NAEP achievement levels predicted almost precisely how a decade later, NCLB, based on use of such levels, has caused a distortion in the curriculum for lower scoring students, leading to an undue emphasis on basic skills that "retard[s] their progress toward mastering more challenging material."

In response to the GAO's criticism, the Department of Education acknowledged that "one reason the judges may have set such high standards is that they did not have the disciplining experience of comparing their personal estimates of what students at a given level will do with what students like those at that level actually did" [emphasis in original].[54] Then the Department of Education commissioned its own study of the NAGB achievement levels, to be performed by a National Academy of Education (NAE) panel.* Confirming the GAO's findings, the NAE panel concluded that the procedure by which the achievement levels had been established were "fundamentally flawed," were "subject to large biases," and that the achievement levels by which American students had been judged deficient were set "unreasonably high."[55] The NAE recommended that the method used for establishing NAEP achievement levels should be abandoned and that the achievement levels themselves should not be used. In fact, the NAE panel stated, continued use of these standards could set back the cause of education reform because it would harm the credibility of NAEP itself.[56]

Still not satisfied, the Department of Education next contracted with the National Academy of Sciences to conduct another evaluation of NAEP. The Academy's panel held a conference in 1996 on the achievement level-setting process, and published its conclusions three years later. The "process for setting NAEP achievement levels is fundamentally flawed," the Academy report stated. "[P]rocesses are too cognitively complex for the raters, and there are notable inconsistencies in the judgment data... Furthermore, NAEP achievement-level results do not appear to be reasonable compared with other external information about students' achievement."[57]

* The panel was chaired by Robert Glaser and Robert Linn, and its investigation conducted by Lorrie Shepard.

None of this advice has been followed. In the 1994 re-authorization of the Elementary and Secondary Education Act (ESEA), of which NCLB is the subsequent re-authorization, Congress acknowledged these judgments of the scientific community by instructing that the achievement levels should be used only on a "developmental basis" until the Commissioner of Education Statistics re-evaluates them and determines that the levels are "reasonable, valid, and informative to the public."[58] As noted above, similar language remains in NCLB. However, the only re-evaluation that has been performed was that of the National Academy of Sciences, noted above, which reiterated the prior studies' condemnations of NAEP achievement levels. A result of that re-evaluation has been that NAEP reports now include disclaimers about the validity of the proficiency levels being used. Yet the same NAEP reports continue to use them, while government officials continue to issue pronouncements about the percentages of students who are not proficient, without mentioning the disclaimers. For example, recent NAEP reports include a caution, buried in the text, defending the use of achievement levels only for observing trends, i.e., changes in the percent of students who achieve proficiency over time, but not for validating the percentages at any given point in time. The caution concludes by offering no defense of achievement level definitions other than the fact that government officials continue to use them:

> As provided by law, NCES, upon review of congressionally mandated evaluations of NAEP, has determined that achievement levels are to be used on a trial basis and should be interpreted with caution. However, NCES and NAGB have affirmed the usefulness of these performance standards for understanding *trends* in achievement. NAEP achievement levels have been widely used by national and state officials [emphasis added].[59]

NCLB imposes sanctions on schools and school districts for failing to meet levels of proficiency on state tests that, although lower in many cases than NAEP levels, were established using similar processes. Irrespective of the actual level of state cut scores, NCES asserts that achievement levels established in this way should only be used "on a trial basis and …interpreted with caution," and then only for purposes of understanding trends over time, not for purposes of judging how many students are truly proficient at any given time.

Because the establishment of criteria is necessarily subjective, no matter how well informed the opinions of judges may be, an almost inevitable consequence of a decision by both the federal and state governments to shift to reporting of performance in criterion rather than norm-referenced terms, has been the politicization of standardized testing. When proficiency criteria were established for NAEP in the early 1990s, the criteria were made unreasonably high because policy makers wanted to spur school reform by demonstrating (or exaggerating) how poorly American students perform. In none-too-subtle language, the General Accounting Office concluded that NAGB established these standards, despite their lack of scientific credibility, because

> the benefits of sending an important message about U.S. students' school achievement appeared considerable, and NAGB saw little risk in publishing scores and interpretations that had yet to be fully examined... NAGB viewed the selection of achievement goals as a question of social judgment that NAGB, by virtue of its broad membership base, was well suited to decide.[60]

Political, not scientific, considerations continue to explain NAGB's stubborn refusal to abandon achievement level cut scores which have no scientific or scholarly credibility. In 2000, NAGB commissioned a review of the controversy by James Popham, a nationally respected psychometrician. Acknowledging that the cut scores are widely

regarded as being too high, Professor Popham noted that resistance to lowering them was based on a belief that doing so "would present a clear admission to the world that the Nation's much touted pursuit of *demanding* levels of student performance was little more than public-relations rhetoric. [Lowering the cut scores] would forever damage NAEP's credibility because it would be seen as little more than a self-serving education profession's adjust-as-needed yardstick." Nonetheless, Professor Popham concluded, "if not modified, [the achievement levels policy] may make NAEP an educational anachronism within a decade or two."[61]

The Subjectivity of State Proficiency Standards

When proficiency criteria have been established by state officials for purposes of accountability under NCLB, political considerations have also prevailed. Several states have established proficiency cut scores which are in NAEP's below-basic range.[*] These relatively low criteria ensure that more schools can escape NCLB sanctions. Colorado has two cut scores for proficiency: one is termed 'proficient' for purposes of compliance with NCLB, but the same cut score is termed only 'partially proficient' for purposes of compliance with state educational accountability policies. [62] Other states, such as South Carolina, have set criteria that are relatively high, presumably for reasons similar to those of federal officials who set NAEP standards.[63] Eighth grade mathematics students in Montana, who achieve far above their state's proficiency standard, can walk across the border to Wyoming and, with the same math ability, fall far below proficiency in that

[*] Of eight states whose fourth grade reading standards were examined by the Northwest Evaluation Association, four (California, Colorado, Iowa, and Montana) have proficiency cut scores which are below NAEP's basic cut score (see Table 4 in Kingsbury et al. 2003 vs. Table 1 in Perie, Grigg and Donahue 2005; the data are not from the same year, however.)

state.[64] In Missouri, 10 percent fewer students are deemed proficient on the state eighth grade math test than on NAEP, while in Tennessee, 66 percent more students are deemed proficient on the state test than on NAEP.[65]

South Carolina's case is particularly interesting. In the mid-1990's, the state's students were four times as likely to be deemed proficient by South Carolina's own accountability system as were proficient on the NAEP.[66] Today, however, the state's students are no more likely to be proficient by state than by NAEP criteria.[67] This does not reflect a change in student learning, only a change in arbitrary definition.

Capricious state standards in other states produce anomalies of their own. In the 1990s, for example, Massachusetts established cut scores on its state test that resulted in only 28 percent of its eighth graders deemed proficient in science. But at approximately the same time, Massachusetts' eighth graders scored higher, on average, than students in every country in the world (except Singapore) on the TIMSS.[68]

One state, Louisiana, found that its proficiency definition was so high that adequate yearly progress under NCLB could not be fulfilled, so it simply decreed that, for NCLB purposes, its basic cut score would be considered a challenging standard of proficiency.[69] As Robert Linn has observed, "the variability in the stringency of state standards defining proficient performance is so great that the concept of proficient achievement lacks meaning."[70]

To summarize: the goal of all students, in all subgroups, achieving proficiency by 2014, or by any subsequent date, is not achievable because:

a) inevitable variation in student performance makes it logically necessary that all students cannot be at or above a level that typical students find "challenging;" and

b) the concept of proficiency in multiple academic subjects and grade levels is impossibly subjective, so subjective that basing an accountability system upon it, involving sanctions and rewards, will almost inevitably impose these sanctions and rewards either on too many or too few schools, depending on the political objectives of the standards-setting process.

Aiming Lower – for Basic, not Proficient Performance

Can the problems we have described be fixed by reducing NCLB's expectations, for example, by abandoning the demand that all students achieve 'challenging' standards of performance, and instead accepting that all students should only be required to achieve at something more like the 'basic' level on NAEP? In other words, should NCLB concede to states that have sabotaged NCLB's intent by setting very low standards of proficiency?

Unfortunately, such an attempt to rescue NCLB's intent is inadvisable from a policy perspective, and also unworkable. Requiring all students to achieve above a basic, as opposed to proficient, cut score has been tried before. In the 1970s, in response to federal requirements that states assess student performance to show that federal aid was being properly utilized, states adopted standardized testing regimes in which mostly basic skills were tested.[71]

Educators and policy makers soon developed contempt for such exams, and they were abandoned after about a decade of use. The National Commission on Excellence in Education, in its influential 1983 indictment of American schools, *A Nation at Risk*, had

this to say: "'Minimum competency' examinations (now required in 37 states) fall short of what is needed, as the 'minimum' tends to become the 'maximum,' thus lowering education standards for all."[72] Accountability for only basic skills, the Commission found, had created incentives to deliver curriculum which did not challenge typical students. The Commission recommended that minimum competency exams be abandoned, to permit creation of curriculum emphasizing more advanced skills.

A few years later, Marshall Smith's and Jennifer O'Day's proposal for accountability and testing aligned with challenging curriculum (systemic school reform) was an explicit rejection of the minimum competency movement in American education. Minimum competency tests, they wrote, "emphasized recognition of facts, word analysis, mathematical computation skills, routine algorithmic problem solving, and little else." Such tests, Smith and O'Day noted, were aligned with the curriculum for disadvantaged children because, with the incentives provided by such tests, teachers only taught such basic skills. Test scores of the most disadvantaged children did increase in such a system, but the academic needs of students higher in the achievement distribution were ignored: "because the scores of more well-to-do and majority students did not change during this time, the achievement gap narrowed."[*] Smith's and O'Day's call for systemic school reform was for "the coherence and clarity of the back-to-basics movement [to be] replaced with a similar coherence and clarity in support of the new, challenging content."[73]

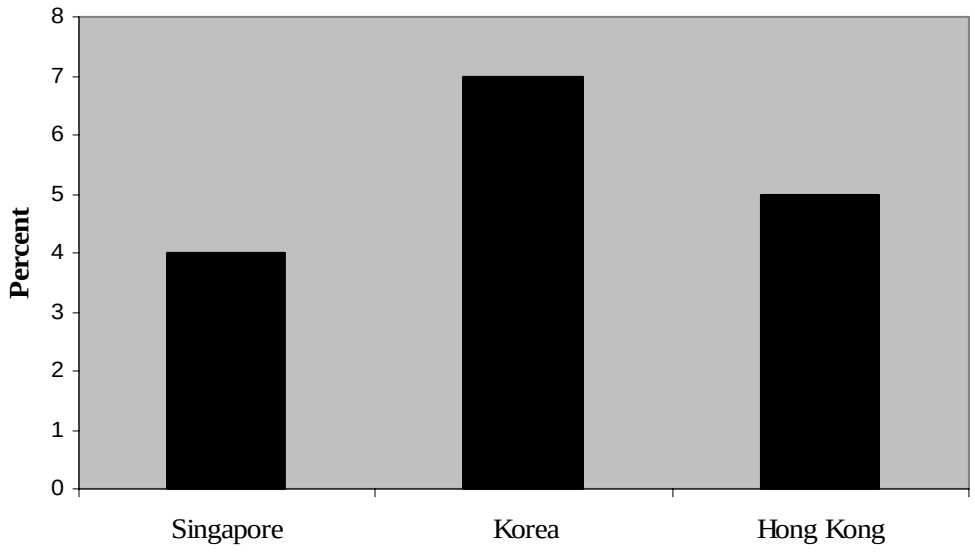
[*] At about the time that Smith and O'Day were challenging the value of curriculum that emphasized only basic skills, others were beginning to notice that a significant portion of the increase in test scores during the 'minimum competency' period was not real, and reflected teaching to tests, excessively narrowed curricula, and some cheating. See, for example, Cannell (1988 and 1989) and Koretz (1988). In the present paper, we do not discuss score inflation in high stakes testing systems, which is as inevitable in NCLB-mandated tests as it was in the standardized tests of the 1970s (Koretz 2006a). It is, however, yet another flaw in accountability systems, such as NCLB, that rely exclusively on high-stakes tests.

If we were to forget these lessons from a generation ago, and modify NCLB to require only a basic level of achievement, not a challenging standard of proficiency, the logical flaws would still apply that are inherent in attempts to apply a single standard to students across the full range of the ability distribution. Standards that are a challenge for students at the bottom of the distribution cannot be a challenge for students higher up.

NAEP's basic level of performance would also be unworkable as NCLB's minimum standard because even NAEP's basic cut scores are too high for too many students. We noted that many students in the highest-scoring countries in the world achieve at less than proficiency as NAEP defines it. Fewer, though still a significant proportion of students in the highest-scoring countries, achieve at less than basic as NAEP defines it.

The smallest share of students below basic in the highest-scoring countries was found on the 2003 TIMSS administration. As Figure 4 shows, in mathematics, only 4

Figure 4. Percent of Students Predicted to Score Below NAEP 8th Grade Math Basic Cut-Score

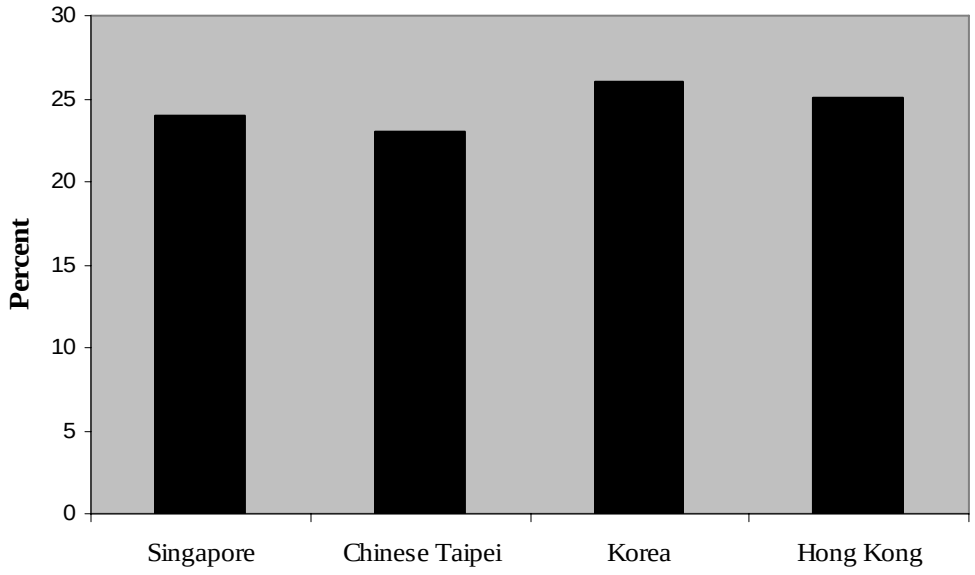


Source: TIMSS 2003: Gonzales et al., 2004; NAEP Cut Scores: Reese, et al., 1997.
Note: Percentages estimated using Linn's (2000) linking method.

percent of eighth graders in Singapore, 5 percent of those in Hong Kong, and 7 percent of those in Korea were below basic. Note, however, that this is at least four times as many below-basic students as would be permitted by NCLB if its requirement were lowered to this level, for NCLB permits only 1 percent of students to perform below its minimum standard.

However, the mathematics TIMSS results were an exception. Figure 5 shows that

Figure 5. Percent of Students Predicted to Score Below NAEP 8th Grade Science Basic Cut-Score



Source: TIMSS 2003: Gonzales et al., 2004; NAEP Cut Scores: Grigg, et al., 2006.
Note: Percentages estimated using Linn's (2000) linking method

the problem is more severe on other international tests we examined. In science, on the same 2003 TIMSS administration, 24 percent of eighth graders in Singapore, 23 percent of those in Chinese Taipei, 25 percent of those in Hong Kong, and 26 percent of those in Korea performed below NAEP's basic cut point.

In the comparison of mathematics scores on the 1991 International Educational Assessment and the 1992 NAEP, 22 percent of students in highest-scoring Taiwan performed below NAEP's basic level, as did 19 percent of those in runner-up Korea.*

And recall that on the 2001 international comparison of 10 year-old reading, about two-thirds of top-scoring Swedish students performed below what NAEP calls

* As noted above (see note on page 9), NAGB raised the cut scores for NAEP subsequent to the time these 1991 IEA and 1992 NAEP comparisons were computed. Using NAEP's current definitions of basic performance, more than 22 percent of Taiwanese students and more than 19 percent of Korean students would score below basic on NAEP if they were to take the NAEP exam.

proficient. If NCLB were only to require NAEP's basic standard, 26 percent of Swedish 10 year-olds would still fail to meet that expectation.

Thus, with the exception of mathematics achievement on the TIMSS, the best scoring countries in the world still seem to have from one-fifth to one-fourth of their students performing below the basic standard. If, in the best of circumstances, U.S. educational outcomes improved to be the best in the world, a similar proportion of American children would be expected to fail NCLB's standard, even if this standard were lowered to NAEP's definition of basic achievement, solely because of the normal distribution of human ability and performance.

Note that these expected normal distributions apply to very large populations, much larger than those in typical schools. Thus, for example, while we can reasonably say that, nationwide, approximately 20 percent of the nation's children should be expected to score at about 87 or below on an I.Q. test, larger shares of students in many typical schools should be expected to score at that level or below (and smaller shares in many other typical schools). Thus, even if NCLB's standard were reduced to basic, and even if U.S. child development and educational institutions were improved to be the best in the world, large numbers of students would continue to fall short of the accountability requirement.

Of course, if NCLB's standard were lowered to NAEP's basic standard, the share of students who passed would increase. And as this happened, the share of students who were unchallenged by the new standard would also increase. We might then see new commissions and studies denouncing how schools were removing challenging material from their curricula because they were only being held accountable for basic skills. There

is no way around the simple logic we discuss in this paper: a standard that more students can pass will challenge fewer students; a standard that challenges more students will be one that fewer students can pass.

A Return to Norm-Referenced Reporting

If we abandon the goal of proficiency-for-all by any date, we are not left without means to track the performance of schools, or means to judge whether growth is satisfactory on subjects measurable by standardized tests.[*] The alternatives, however, in our judgment, require a return to norm-referenced reporting.

Norm-referenced reporting should not be offensive to education policy-makers who frequently urge public education to adopt practices used in the business world. There, norm-referenced standards are considered the ideal way to set goals and measure progress. When sophisticated business managers want to reduce the number of defective parts produced each week, or increase the number of phone calls handled by an operator each hour, or ship a greater number of orders each day, they don't convene panels of experts to fantasize about the ideally productive assembly line, call center, or warehouse. Rather, they assemble data on how many defective parts, calls, or shipments their successful competitors produce. In management theory, the term used to describe this practice is "benchmarking." The benchmark is the standard of performance set by a more efficient competitor, and which the benchmarking manager aspires to achieve. As one management text summarizes the practice: "Benchmarking results in process practices and measurable goals based on what the best in the industry is doing and is expected to

[*] But not all curricular areas are measurable by standardized tests, nor are all that are measurable actually measured. For a discussion of the curriculum narrowing that results from accountability for standardized test scores, see Rothstein 2006; Rothstein and Jacobsen 2006.

do. The approach contrasts sharply with the rather imprecise, intuitive estimates of what needs to be done [that] characterize current searches for productivity."[74]

Thus, in the business world, criterion-referenced goals are considered "imprecise and intuitive estimates." In contrast, firms using norm-referenced goals, or benchmarks, "can and should learn from others and constantly measure themselves against the best in the industry."[75] When many firms in an industry adopt benchmarking as a strategy, "continuous improvement" is said to result. This is because each firm that achieves, or slightly exceeds, its benchmark sets a new standard for the firm or firms which had previously served as a model. And as these firms too achieve their benchmarks, they set a new standard for the first firm, which must again revise its goal. But there is a realistic limit to this process. No benchmark can be set beyond what has actually been achieved by firms in the same industry, in the same markets, and with the same external constraints. Quoting again from the management text, a "target incorporates in it what realistically can be accomplished within a given time frame… Considerations of available resources, business priorities, and other operational considerations convert benchmark findings to a target, yet steadily show progress toward benchmark practices and metrics."[76]

A European management text warns against a tendency of American businesses to seek unachievable goals or to be the best in the world, instead of choosing a benchmarking target that is realistic: "[I]t is counterproductive to look to examples which are way above [the] benchmarking organization [in performance]… We would therefore suggest that good examples should be found with the aim of creating sufficient improvement for the organization[,] not to overreach itself, but to initiate change in terms

of real improvement… Let us therefore modify the extremely high ambitions of some of our American colleagues and recommend 'sufficient' rather than maximum improvement. This does not imply any reduction of ambition. In our opinion it leads instead more rapidly and efficiently to the goal of continuous improvement."[77]

For our educational goals, can we adopt this business philosophy of benchmarking for realistic, but continuous improvement?

An Alternative Goal System

We discussed earlier the confusion between an expectation for minimally acceptable performance that all, or almost all students should meet, and proficient performance, a level of achievement that is challenging for typical students. We concluded that the goal of proficiency for all ignores this distinction.

An alternative way to establish a goal for performance, therefore, would be to expect a higher percentage of students to achieve proficiency than presently do so. This may seem straightforward and be intuitively appealing, but there are fatal difficulties with this approach. Using figures that are merely illustrative, here is how such an alternative might work.

It should rely on the benchmarking principle, as should any goal-based accountability system. One way to do this might be to say that we want the percentage of students in each subgroup who achieve proficiency at some future time to rise to the percentage of students who now achieve proficiency in the highest-scoring three-fourths of that subgroup. At the present time, for example, in NAEP eighth grade math, 37 percent of white students are deemed proficient and 9 percent of black students are

deemed proficient.[78] For the share who are proficient to rise to the share of the highest-scoring three-fourths of these groups today, 49 percent of white and 12 percent of black students would have to achieve proficiency.[*] We would next have to determine how rapidly schools might improve, with appropriate reform interventions, so that the average percent proficient for all students would be equal to the average percent proficient for the top three-fourths of students today. Imagine we determine that this can be accomplished by 2014. With these data, policy might demand that the share of white students who are proficient should increase by 12 percentage points and the share of black students who are proficient should increase by 3 percentage points by 2014.

Using the same method, if we were to set a more ambitious target, the percentage of students who are proficient in the top two-thirds of each subgroup rather than the top three fourths, we would expect the share of white students who are proficient to rise by 19 points and the share of black students who are proficient to rise by 5 points.

The fatal difficulties with this approach should now be mostly obvious. Certainly, making the top two-thirds or top three-fourths the target is arbitrary. Further, because (as described above) the definition of proficiency is arbitrary and perhaps even capricious, retaining this standard as the basis of an accountability system leaves the system open to continued political manipulation. If, at a future time, the definition of proficiency is reconsidered, it will be difficult to maintain any longitudinal account of how progress toward the goal is being made. The same problem would arise when new tests are established, with their own cut scores to establish proficiency, even if the standard-setting process attempts to make the new cut score equal in difficulty to the old.

[*] 37 / 75 = .49; .09/.75 = .12

Also, relying upon definitions of proficiency for accountability decisions distorts instruction. When progress is measured solely by changes in the percentage of students who score above a pre-defined proficiency point, schools are given perverse incentives to concentrate instruction only on those students who are just below the proficiency point, ignoring the needs of students who are already above (or far below). As a result, test scores may improve for only some students who are at one point in the distribution – the only one that matters for accountability purposes - while other students' scores may decline or remain stagnant. Then, average scores can go *down* while the number of students who pass the proficiency point goes up.[79] Indeed, the achievement gap itself can seem to be narrowing (measured by percent proficient) while the true gap, in average scale scores, widens.[80]

Such an accountability system is likely to be so complex that few policy makers or even experts will be able to understand it. This is because the difference between the average percent proficient in the top scoring three-fourths of schools, and the average percent proficient in all schools, is likely to be quite different from subject to subject, from subgroup to subgroup, and from grade to grade. The goal of a gain of 12 points in percent proficient for whites and 3 points in percent proficient for blacks will not necessarily apply to other grades and subjects, if the benchmarking principle is to be preserved.

Another difficulty is political. It is hard to imagine how we could establish different standards for different sub-groups when standards are expressed in this way. It would almost certainly be politically unacceptable for national policy to state that, by

2014, we want 49 percent of white students, but only 12 percent of black students, to be proficient.

Establishing Reasonable Accountability Targets for Student Achievement

These problems can be ameliorated if, rather than establish a target for percent proficient, we use relative performance measures. This requires abandoning achievement-level reporting, and returning to scale scores reported in norm-referenced terms. Such a system would expect students in each demographic group to perform at a higher level than they presently do, by establishing benchmarks based on what demographically similar students, in best practice conditions, actually do achieve.

To sketch out what such targets might look like, we propose a thought experiment. Let's assume that we wanted to establish a goal for performance in eighth grade mathematics and we determined that, for each racial, ethnic, socioeconomic, language, and disability subgroup, an effect size of 0.3 would be a challenging but reasonable goal, and could be achieved with intensive school improvement interventions in 10 years. As noted earlier, an effect size of 0.3 means that the average student in each subgroup would perform at a level that is today achieved by students at about the 62nd percentile in that subgroup. This is a very substantial and ambitious goal, but perhaps not beyond what we can reasonably expect to achieve after sustained and effective reform efforts. It is based on rates of educational improvement actually achieved in the recent past with particular interventions (though not systemwide), and on a rate that behavioral science generally has found to be realistic in many cases. However, it would be irresponsible to base an accountability system on such precedent alone. Considerably

more research and experimentation would be required before we could say with any certainty that such a goal could be reached for particular grades, subject areas, or demographic groups, or that 10 years is the appropriate time frame to expect it to be reached.*

An advantage of investing in the further development of such a norm-referenced standard is its transparency and easy availability for public discussion and debate. Public discussion of proficiency standards are impossible to conduct, because there is no way for members of the public to determine what the cut score on a test should be to denote proficiency. With the norm-referenced approach we suggest, however, democratic discussion is a simple matter. Those who believe that an effect size of 0.25 or 0.4 is realistic (or an improvement for the average student to about the 60th or 66th percentile from the base year, respectively), based on their evaluation of other reforms in education or in other areas of social policy, can propose an alternative goal.

Robert Linn has suggested that "past experience is the first place to look" to know if a goal is realistic and if, for example, the best tenth of low-income schools have registered increases in percent proficient of 4 percent a year, then NCLB could establish such a required rate of increase for all low-income schools.[81] While this approach is more reality-based than expecting all students to achieve fanciful definitions of proficiency by an arbitrary date, it makes sense only for the very short term; there is no reason to think that rates of improvement currently achieved by some schools can be continued into the future, either by these schools or by others. Because a baseball player has improved his batting average from .270 last year to .280 this year by being more selective about which

* This is not the place to sketch out such experimentation in detail, but it would be desirable to have an existence proof that some randomly selected school districts, after implementation of a carefully designed improvement program, actually could achieve systemwide gains with effect sizes of 0.3 in 10 years.

pitches to chase and which to take, it does not follow that this or any player should be able to improve his average by 10 points each year indefinitely. Using past rates of improvement seems reality-based, but does not truly apply the benchmarking principle.

Past rates of improvement do not provide a good basis for ongoing targets. Lower-achieving students may sometimes be able to improve faster than average if they have more room to grow (in other words, because they do not risk bumping-up against a ceiling of maximum possible performance); higher-achieving students may sometimes be able to improve faster because, still far from a ceiling, they have more human capital with which to accelerate learning. Demanding a 0.3 effect size for mean scale scores, however, expects no more of the typical student during the target period than is presently being achieved by about 38 percent of all students. If this were adopted as an ongoing principle, it would be self-limiting in that, as the achievement of the top 38 percent of students reached practical limits, the expectations for typical students would also be limited by reality.

The system we propose could be refined further, although we should be cautious about making it so complex that it is not understandable by relatively sophisticated policy makers and members of the public. One important area of refinement is the need to ensure that improvement is experienced by the full range of students. If this is not done, then incentives could be created to concentrate efforts on only some students, to improve the average. For example, the average scale score on NAEP could go up because disproportionate gains were made by the most able students while fewer or no gains were made by the lowest-scoring students. Such gains would not advance the cause of educational equity.

Such a result can be avoided by disaggregating the goal. We have suggested that a goal might be an effect size of 0.3, or moving typical students, who are presently at the 50th percentile, up to about the 62nd percentile in the contemporary distribution. If we are concerned that this improvement in the average is being driven by instructional attention only to students deemed easiest to improve, we could establish goals not only for 50th percentile students, but also for students presently at the 25th and 75th percentiles. Employing the standard of a 0.3 effect size, we might determine that the goal for the next decade will have been met only if, for example, students now at the 50th percentile rank perform at the level of students who are presently at the 62nd percentile, if students who are now at the 25th percentile perform at the level of students who are presently at the 35th percentile, and if students who are now at the 75th percentile perform at the level of students who are presently at the 84th percentile. Such an accountability standard would ensure that instructional improvement occurs across the spectrum of student ability.*

This approach to educational improvement goals is not new. It was an aspect of Goals 2000, the national education goals that were enacted by Congress in the 1994 re-authorization of ESEA, only to be abandoned by NCLB in 2001. We noted above that one aspect of Goals 2000, that the United States would become first in the world in math and science, was foolhardy. But a more reasonable aspect was the requirement that "[t]he academic performance of all students at the elementary and secondary level will increase significantly *in every quartile*, and the distribution of minority students *in each quartile* will more closely reflect the student population as a whole" (emphasis added).[82] Any

* Another essential refinement in such an accountability system should be to ensure the use of additional accountability measures that relied on the assessement of skills not amenable to standardized testing. Such an accountability system is not the subject of this paper, but we have discussed it elsewhere (Rothstein 2006; Rothstein and Jacobsen 2006).

sensible set of improvement goals should abandon NCLB's insistence on a single proficiency standard for all students, and return to this notion of progress in each quartile of the student achievement distribution.*

Implications for Equity

We next consider what the implications of such a program might be for equity. Presently (using eighth grade NAEP mathematics results for purposes of illustration), the median black student performs at the 27th percentile of the all-student distribution, the black student who performs at the 25th percentile of the black student distribution is at the 11th percentile of the all-student distribution, and the black student who performs at the 75th percentile of the black distribution is at the 49th percentile of the all-student distribution.[83]

* One reviewer of an early draft of this paper properly observed that a return to norm-referenced reporting is not essential for what we propose, and that the program recommended here could as well be defined in criterion-referenced terms. To do so, three criteria might be established, Criterion A, Criterion B, and Criterion C, perhaps corresponding to the present achievement of students at the 25th, 50th, and 75th percentiles of the national distribution. Then, if the ten-year goal were an effect size of 0.3, it could be expressed as an expectation that the percentage of students passing Criterion A would increase from 75 percent to 84 percent; the percent passing Criterion B would increase from 50 percent to 62 percent; and the percent passing Criterion C would increase from 25 percent to 35 percent. As a further refinement, A, B, and C could be labeled "basic," "proficient," and "advanced." Our reviewer suggested that such criterion-referenced goals would have the advantage of not challenging the widespread attachment of contemporary educators and policymakers to the notion that we should measure whether students meet standards, not how they perform relative to one another. As the example just given illustrates, the attachment to criterion-referenced reporting, as opposed to norm-referenced reporting, can be only a matter of terminology, not substance. We agree that our reviewer's suggestion might make our proposal more politically appealing. If such criteria were used, however, the essential commitment must be preserved that no single criterion can be established as a goal for all students. And if three criteria were employed, it is essential that A, B, and C be somewhere in the middle of the distribution, for example at the 25th, 50th, and 75th percentiles in present performance, as we suggest. If the use of criteria tempted policymakers to substitute fantasy for reality (as is done in present policy) when establishing cut scores, for example, if A, B, and C were set at the current 32nd, 72nd, and 98th percentiles, this would do no good at all.
The value of norm-referenced reporting is that the use of percentile rankings instead of criteria reminds policymakers and the public of the inevitable distribution of outcomes from students, even following the most successful school improvement program.

Improvements for black students with an effect size of 0.3 would bring the median black student to the 36th percentile in the all-student distribution; would bring the black student who performs at the 25th percentile of the black student distribution to the 16th percentile of the all-student distribution, and the black student who performs at the 75th percentile of the black distribution to the 60th percentile of the all-student distribution.

The magnitude of the achievement gap would not be affected substantially. After improvement with an effect size of 0.3 for both black and white students, the gap in achievement of typical black and typical white students would remain steady at 36 percentile points. For black and white students who were at the 25th percentiles of their respective subgroup distributions, the gap would grow from 28 to 32 percentile points. For those who were at the 75th percentiles of their respective subgroup distributions, the gap would narrow from 32 to 28 percentile points.

That the gap would not be substantially affected should not be surprising. If school improvement generates equivalent gains for both black and white students, gaps may well not change. After all, black and white students presently enter the school years with a pre-established gap from early childhood that is, by many measures, not significantly different from the gap at the end of schooling, because both black and white students gain in ability at similar rates during the school years.[84] Narrowing the gap, as opposed to raising the absolute achievement of black and white students, will require addressing the pre-established gap that exists at the beginning of school.

A More Ambitious Program for Equity

If we truly wanted to narrow the achievement gap substantially, while also improving the achievement of middle class white children, we must do more than hold schools accountable for improvement. Accountability is a necessary, but not sufficient component of reform. Also necessary is a program of improvements likely, not only to generate effect sizes of 0.3 or more in the foreseeable future, but also to enable disadvantaged children to enter school more ready to learn.

Such a program would undoubtedly be costly, but must consist of more than an indiscriminate increase in school spending. It should be based on the best research presently available about what is required to dramatically raise student outcomes.

What would we do, how long would it take, and how much would it cost? What follows is another thought experiment, in this case one which ignores political realities.

The most compelling evidence for effective policies to raise outcomes for disadvantaged children comes from studies of early childhood interventions. We clearly distinguish programs for early childhood (pregnancy to four years of age), pre-kindergarten (for four year-olds), and kindergarten (for five year-olds). The importance of a healthy birth for later development is well known. Children born with low weight, for example, have poorer academic achievement and more special education placement and behavioral problems.[85] High quality early childhood programs have great power to alter lifelong outcomes.[86] Recently, advocacy of the priority of investment in early childhood has been given a boost by the arguments of Nobel laureate (in economics) James J. Heckman.

Professor Heckman, collaborating with, among others, Jack P. Shonkoff, co-editor of an important National Academy of Sciences study on the neurobiology of early childhood development,[87] argues what should be an obvious point: both academic and non-cognitive achievement follows "hierarchical rules."

> Later attainments build on foundations that are laid down earlier… [C]ognitive, linguistic, social, and emotional competencies are interdependent; all are shaped powerfully by the experiences of the developing child… Although adaptation continues throughout life, human abilities are formed in a predictable sequence of sensitive periods,

with prenatal development and early childhood the most influential.

> By the third grade, gaps in test scores across socioeconomic groups are stable by age, suggesting that later schooling and variations in schooling quality have little effect in reducing or widening the gaps that appear before students enter school….
>
> At current levels of resources, society overinvests in remedial skill investments at later ages and underinvests in the early years.
>
> Although investments in older disadvantaged individuals realize relatively less return overall, such investments are clearly beneficial. Indeed the advantages gained from effective early interventions are sustained best when they are followed by continued high-quality learning experiences."[88]

Heckman's analysis suggests a way to think about the extent to which it is possible to achieve both equity and excellence, and how long it would take to do so. Consider the following 19-year program for the cohort of disadvantaged children to be born next year, which we call "year 1."

In year 1, programs should ensure that all disadvantaged pregnant women receive adequate prenatal care. This would not fully fulfill the model's requirements, because less healthy births are not only predicted by adequate medical care but also by freedom from stress, adequate nutrition and similar characteristics less common for disadvantaged

women. Nonetheless, social policy could make an important impact in areas that are easier to influence, like adequate medical care.

Year 2 should bring the provision of high-quality early childhood care, including routine and preventive pediatric care, for children up to one year of age. In year 3, such care should be provided for these children in their second year, in year 4 for their third year, and in year 5 for their fourth year. In year 6, we should provide an adequate pre-kindergarten program.

Reform of educational and other institutions of youth development should continue in the same fashion. Pediatric care should continue, with year 7 seeing the addition of kindergarten classes with sufficiently high teacher-child ratios and highly qualified teachers. For each subsequent year, highly qualified teachers should be assured, it being possible to start increasing class sizes for the cohort as it enters fourth grade in year 11. High quality after-school and summer programs should be provided for the cohort, beginning with prekindergarten.

In forthcoming work, one of us has costed-out such a program of adequacy, which he presently estimates will eventually (by year 19 of the program just described) boost what the nation currently spends on the education of children from birth to age 18 by about 40 percent (before medical insurance reimbursements).[*]

This is not a big increase. Nineteen years provides a very gradual time frame in which to realize such a 40 percent increase, and the pace of increase would be no greater than that experienced by elementary and secondary education during the last half century.[89]

[*] Richard Rothstein has done the modeling work for such a program in collaboration with Whitney C. Allgood.

In one respect, however, this estimate understates the needed increase in finances, because it assumes that we do nothing to attempt to compensate for the causes of low achievement in cohorts born prior to year 1. Yet in another respect, the 40 percent growth estimate is probably much too high, because it does not account for the savings to be realized in present expenditures were such a program to be implemented.

Special education costs would certainly be reduced. Compensatory education expenditures for older youth in the program and subsequent cohorts might be less necessary. Perhaps class sizes could be increased in the later grades once a cohort was better prepared earlier in life. Perhaps it would be easier and thus less expensive to attract qualified teachers in the later grades, once a cohort was more adequately prepared for grade level work. Other governmental expenditures would also be offset. For example, the costs of controlling crime (including prisons) and of welfare would fall. More productive workers would generate higher tax receipts. Such savings have been estimated elsewhere, including by James Heckman, and we will not describe them in detail here.[90]

The proposal design for a stepped program, beginning in year 1, with adequacy added for each subsequent year, is necessary because, as James Heckman and his colleagues note, "skills beget skills, success breeds success, and the provision of positive experiences early in life is considerably less expensive and more effective than the cost and effectiveness of corrective intervention at a later age."[91]

It is not possible to model an adequate system of education and youth development by proposing interventions for older children who have not benefited from earlier adequacy. This is not to say that we should not attempt to compensate, for older children, for the lack of adequate programs in the early years when it most mattered. But

such compensation, while limiting the damage, cannot itself make all American children proficient or, as a more modest ambition, make the United States 'first in the world.'

Bibliography

ACT (American College Testing). 1992. "Description of Writing Achievement-Levels Setting Process and Proposed Achievement Level Definitions. 1992 National Assessment of Educational Progress." Presented to the National Assessment Governing Board, October 21. ERIC # ED 351 697.

ACT. 2000. *2000 National Score Report*. http://www.act.org/news/data/00/00data.html

Alonso, Juan Diego, and Richard Rothstein. Forthcoming. *Where the Money Goes. Changes in the Level and Composition of Education Spending, 1967-2005.* Washington, D.C.: Economic Policy Institute.

Barnett, Steven, 1995. "Long Term Effects of Early Childhood Programs on Cognitive and School Outcomes." *The Future of Children* 5 (3), Winter, 25-50.

Bracey, Gerald W. 2005. "Tips for Readers of Research: Handle Pass Rates with Care." *Phi Delta Kappan* 87 (4), December: 333-334.

Brown, William. 2000. "Reporting NAEP by Achievement Levels: An Analysis of Policy and External Reviews." In Mary Lyn Bourque and Sheila Byrd, eds. *Student Performance Standards on the National Assessment of Educational Progress: Affirmation and Improvements*. Washington, D.C.: National Assessment Governing Board. http://www.nagb.org/pubs/studentperfstandard.pdf

Camp, Robert C. 1989. *Benchmarking. The Search for Industry Best Practices that Lead to Superior Performance.* Milwaukee, WI: Quality Press. American Society for Quality Control

Campbell, Jay R., Catherine M. Hombo, and John Mazzeo. 2000. *NAEP 1999 Trends in Academic Progress. Three Decades of Student Performance*. NCES 2000-469, August, Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement,

Cannell, John Jacob, 1988. "Nationally Normed Elementary Achievement Testing in America's Public Schools: How All 50 States are Above the National Average," *Educational Measurement: Issues and Practice*, vol. 7, no. 2, Summer.

Cannell, John Jacob, 1989. *How Public Educators Cheat on Standardized Achievement Tests: The 'Lake Wobegon' Report.* Albuquerque: Friends for Education.

Census Bureau. 2006. Population Estimates Program, Population Division, U.S. Census Bureau (on-line), http://www.census.gov/popest/estimates.php, accessed September 16, 2006.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

College Board (College Board and Advanced Placement Program). 2000a. 2000 National Summary Reports. College Entrance Examination Board.

College Board, 2000b. Total Group Report. National Report. 2000 College Bound Seniors. A Profile of SAT Program Test Takers. College Entrance Examination Board

Commission (The National Commission on Excellence in Education). 1983. *A Nation At Risk: The Imperative for Educational Reform. A Report to the Nation and the Secretary of Education*. Washington, D.C.: U.S. Government Printing Office. April.

Dillon, Sam, 2006. "As 2 Bushes Try to Fix Schools, Tools Differ." *The New York Times*, September 28.

DOE (U.S. Department of Education). 2005a. *Raising Achievement of Students with Disabilities.* December. http://www.ed.gov/admins/lead/speced/achievement/factsheet.pdf

DOE (U.S. Department of Education). 2005b. "34 CFR Parts 200 and 300. Title I – Improving the Academic Achievement of the Disadvantaged; Individuals With Disabilities Education Act (IDEA) - Assistance to States for the Education of Children with Disabilities. Proposed Rule." *Federal Register* 70 (240), Thursday, December 15: 74624-74638.

Education Week. 2006. "Chat Wrap Up: The Changing Federal Role in Education." *Education Week* 26 (10), November 1: 33.

Elliott, Emerson J. 1992. "Letter from Emerson J. Elliott, Acting Assistant Secretary, U.S. Department of Education, to Eleanor Chelimsky, Assistant Comptroller General, General Accounting Office. March 25." Reprinted in GAO 1993, p. 66-73

GAO (U.S. General Accounting Office). 1993. *Educational Achievement Standards. NAGB's Approach Yields Misleading Interpretations.* GAO/PEMD 93-12. Washington, D.C.: General Accounting Office.

Gawande, Atul. 2006. "The Score. How Childbirth Went Industrial." *The New Yorker*, October 9.

Glass, Gene, 1978. "Standards and Criteria." *Journal of Educational Measurement* 15 (4), Winter: 237-261.

Gonzales, Patrick, et al. (Juan Carlos Guzmán, Lisette Partelow, Erin Pahlke, Leslie Jocelyn, David Kastberg, and Trevor Williams). 2004. *Highlights From the Trends in International Mathematics and Science Study (TIMSS) 2003*. U.S. Department of Education, Institute of Education Sciences. December. NCES 2005–005.

Grigg, W., M. Lauko, and D. Brockway. 2006. *The Nation's Report Card: Science 2005*. Washington, D.C.: US Department of Education, National Center for Education Statistics, May. NCES 2006-466

Grissmer, David W., et al. (Shelia Nataraj Kirby, Mark Berends, and Stephanie Williamson). 1994. *Student Achievement and the Changing American Family*. Santa Monica, CA: Rand.

Gurian, Anita. 2002. "About Mental Retardation." NYU Child Study Center (on-line), March 19. http://www.aboutourkids.org/aboutour/articles/about_mr.html#introduction. Accessed on September 25, 2006.

Hack, Maureen, Nancy K. Klein, and H. Gerry Taylor, 1995. "Long-Term Developmental Outcomes of Low Birth Weight Infants." *The Future of Children* 5 (1), Spring: 176-196.

Heckman, James J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science* 312, June 30: 1900-1902.

Heckman, James J., and Dimitriy V. Masterov, 2004. *The Productivity Argument for Investing in Young Children*. Working Paper 5, Invest in Kids Working Group. Committee for Economic Development, October 4. http://www.ced.org/docs/report/report_ivk_heckman_2004.pdf

Hoff, David J., 2002. "States Revise the Meaning Of 'Proficient." *Education Week* 22 (6), October 9: 1, 24-25

Horn, Catherine, Miguel Ramos, Irwin Blumer, and George Madaus. 2000. *Cut Scores: Results May Vary.* Boston College: National Board on Educational Testing and Public Policy, Monograph Volume 1, Number 1, April.

IES (Institute of Education Sciences, National Center for Education Statistics). 2006. *NAEP Data Explorer* (on-line). http://www.nces.ed.gov/nationsreportcard/nde/criteria.asp. Accessed September 16, 2006.

Karlof, Bengt and Svante Ostblom. 1993. *Benchmarking: A Signpost to Excellence in Quality and Productivity*. New York: John Wiley & Sons.

Kingsbury, G. Gage, et al. (Allan Olson, John Cronin, Carl Hauser, Ron Houser). 2003. The State of State Standards. *Research Investigating Proficiency Levels in Fourteen States*. Northwest Evaluation Association, November 21. http://www.nwea.org/research/national.asp and http://www.nwea.org/assets/research/national/State%20of%20State%20standards%20-%20complete%20report.pdf

Klein, Alyson. 2006. "Spellings: Education Law Needs Only a Soft Scrub." *Education Week* 26 (2), September 6: 35.

Knudsen, Eric I., et al. (James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff). 2006. "Economic, Neurobiological, and Behavioral Perspectives on Building America's Future Workforce." *PNAS* (Proceedings of the National Academy of Sciences of the United States of America) 103 (27), July 5: 10155-10162

Koretz, Daniel 1986. *Trends in Educational Achievement.* Washington, D.C.: Congress of the United States, Congressional Budget Office.

Koretz, Daniel 1987. *Educational Achievement: Explanations and Implications of Recent Trends*. Washington, D.C.: Congress of the United States, Congressional Budget Office.

Koretz, Daniel, 1988. "Arriving in Lake Wobegon: Are Standardized Tests Exaggerating Achievement and Distorting Instruction?" *American Educator*, Summer, 8-15, 46-52.

Koretz, Daniel. 2005. Harvard Graduate School of Education, Course Lecture, "Understanding Today's Educational Testing," October 17.

Koretz, Daniel. 2006a. "The Pending Reauthorization of NCLB: An Opportunity to Rethink the Basic Strategy." Invited Paper for Civil Rights Project/Earl Warren Institute Roundtable Discussion on the Reauthorization of NCLB. Washington, D.C., November 16, 2006.

Koretz, Daniel. 2006b. Personal correspondence (with Richard Rothstein), September 26.

Lapointe, Archie E., and Stephen L. Koffler.1982. "Your Standards or Mine? The Case for the National Assessment of Educational Progress." *Educational Researcher* 11 (10), December: 4-11

Lee, Jaekyung. 2006. *Tracking Achievement Gaps and Assessing the Impact of NCLB on the Gaps: An In-Depth Look Into National and State Reading and Math Outcome Trends.* Cambridge, MA: The Civil Rights Project at Harvard University. http://www.civilrightsproject.harvard.edu/research/esea/nclb_naep_lee.pdf

Lee, Stephanie. 2003. *No Child Left Behind and Students with Disabilities.* Presentation to the Office of Special Education Programs Staff, March 20.

Linn, Robert L. 2000. “Assessments and Accountability.” *Educational Researcher,* 29(2), 4-16.

Linn, Robert L. 2003. "Accountability: Responsibility and Reasonable Expectations. 2003 Presidential Address." *Educational Researcher* 32 (7), October, 3-13.

Linn, Robert L. 2004. *Rethinking the No Child Left Behind Accountability System.* Paper prepared for a forum on No Child Left Behind sponsored by the Center on Education Policy, July 28.

Linn, Robert L. 2006. *Educational Accountability Systems.* CSE Technical Report 687. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. June.

Linn, Robert L., Eva L. Baker, and Damian W. Betebenner, 2002. "Accountability Systems: Implications of Requirements of the No Child Left Behind Act of 2001." *Educational Researcher* 31 (6), 3-16.

Livingston, Samuel A., and Michael J. Zieky. 1982. *Passing Scores. A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Educational Testing Service

Loomis, Susan Cooper, and Mary Lyn Bourque, eds. 2001a. *National Assessment of Educational Progress Achievement Levels, 1992 – 1998 for Mathematics.* July. Washington, D.C.: National Assessment Governing Board. http://www.nagb.org/pubs/mathbook.pdf

Loomis, Susan Cooper, and Mary Lyn Bourque, eds. 2001b. *National Assessment of Educational Progress Achievement Levels, 1992 – 1998 for Writing.* July. Washington, D.C.: National Assessment Governing Board. http://www.nagb.org/pubs/writingbook.pdf

Magnuson, Katherine A., and Jane Waldfogel. 2005. "Early Childhood Care and Education: Effects on Ethnic and Racial Gaps in School Readiness." *The Future of Children* 15 (1), Spring: 169-196.

Mullis, Ina V.S., Michael O. Martin, Eugenio J.Gonzalez, and Ann M. Kennedy. 2003. *PIRLS 2001 International Report: IEA’s Study of Reading Literacy Achievement in Primary Schools.* International Association for the Evaluation of Educational Achievement. Chestnut Hill, MA: Boston College. http://timss.bc.edu/pirls2001i/pdf/p1_IR_book.pdf

NAE (National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessments). 1993. *Setting Performance Standards for Student Achievement. A Report of the National Academy of Education Panel on the Evaluation of the*

NAEP Trial State Assessments: An Evaluation of the 1992 Achievement Levels. Stanford, CA: National Academy of Education

NCBOE (North Carolina Board of Education) 2006. "SBE Meeting 08/2006. Recommended Interim Academic Achievement Standards (Cut Scores) and Descriptors for the NCEXTEND2 EOG Writing Assessments Grades 4 and 7." http://www.ncpublicschools.org/sbe_meetings/0608/0608_hsp/hsp0608.pdf

NCEE (The National Commission on Excellence in Education).1983. *A Nation At Risk: The Imperative for Educational Reform. A Report to the Nation and the Secretary of Education.* Washington, D.C.: U.S. Government Printing Office. April

NCES (National Center for Education Statistics). 2000. "National Assessment of Educational Progress (NAEP). 2000 Mathematics Assessment." http://www.nces.ed.gov/nationsreportcard/nde/viewresults.asp.

NCES (National Center for Education Statistics). 2004. *Digest of Education Statistics-2003*. NCES 2005-025. Washington, D.C.: U.S. Department of Education, Institute of Education Sciences.

NCLB (No Child Left Behind Act of 2001). 2001. http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_bills&docid=f:h1enr.txt.pdf

NEGP. "The National Education Goals Panel. Building a Nation of Learners. Goal 3: Student Achievement and Citizenship." http://govinfo.library.unt.edu/negp/page3-7.htm

Nores, Milagros, et al. (Clive R. Belfield, W. Steven Barnett, and Lawrence Schweinhart). 2005. "Updating the Economic Impacts of the High/Scope Perry Preschool Program." Educational Evaluation and Policy Analysis 27 (3), Fall

O'Day, Jennifer A. and Marshall S. Smith, 1993. "Systemic Reform and Educational Opportunity." In Susan H. Fuhrman, ed., *Designing Coherent Education Policy. Improving the System*. San Francisco: Jossey-Bass.

Olson, Lynn. 2006. "Department Raps States on Testing." *Education Week* 25 (42), July 12: 1, 36-37

Pashley, Peter J. and Gary W. Phillips. 1993. *Toward World Class Standards: A Research Study Linking International and National Assessments.* Educational Testing Service, ETS-24-CAEP-01.

Pelligrino, James W., Lee R. Jones, and Karen J. Mitchell, eds. 1999. *Grading the Nation's Report Card.* Washington, D.C.: National Academies Press

Perie, Marianne, W. Grigg, and G. Dion. 2005. *The Nation's Report Card. Mathematics 2005*. U.S. Department of Education, National Center for Education Statistics, October. NCES 2006-453. http://nces.ed.gov/nationsreportcard/pdf/main2005/2006453.pdf

Perie, Marianne, W. Grigg, and P. Donahue. 2005. *The Nation's Report Card. Reading 2005*. U.S. Department of Education, National Center for Education Statistics, October. NCES 2006-451. http://nces.ed.gov/nationsreportcard/pdf/main2005/2006451.pdf

Phillips, Meredith, 2000. "Understanding Ethnic Differences in Academic Achievement: Empirical Lessons from National Data." In David W. Grissmer and J. Michael Ross, eds. *Analytic Issues in the Assessment of Student Achievement.* Washington DC: National Center for Education Statistics.

Popham, James. 2000. "Looking at Achievement Levels." In Mary Lyn Bourque and Sheila Byrd, eds. *Student Performance Standards on the National Assessment of Educational Progress: Affirmation and Improvements*. Washington, D.C.: National Assessment Governing Board. http://www.nagb.org/pubs/studentperfstandard.pdf

Project Appleseed (Project Appleseed, the National Campaign for Public School Improvement). 2006. "Frequently Asked Questions and Answers for Families and Communities." http://www.projectappleseed.org/nclbtesting.html. Accessed September 22, 2006.

Reese, Clyde M., et al. (Karen E. Miller, John Masseo, John A. Dossey). 1997. *NAEP 1996 Mathematics Report Card for the Nation and the States. Findings from the National Assessment of Educational Progress.* Washington, D.C.: US Department of Education, Office of Educational Research and Improvement, February. NCES 97-488.

Reichman, Nancy E. 2005. "Low Birth Weight and School Readiness." *The Future of Children* 15 (1), Spring: 91-116

Rosenberg, Bella. 2004. *What's Proficient? The No Child Left Behind Act and the Many Meanings of Proficiency*. Washington, D.C.: The American Federation of Teachers, May. http://www.aft.org/pubs-reports/downloads/teachers/WhatsProficient.pdf

Rotherham, Andrew. 2006. *Making the Cut: How States Set Passing Scores on Standardized Tests*. EducationSector, July. http://www.educationsector.org/analysis/analysis_show.htm?doc_id=385844#PDF

Rothman, Robert. 1991. "NAEP Board Urged To Delay Standards-Setting Plan." *Education Week*, January 16.

Rothstein, Richard. 1998. *The Way We Were? The Myths and Realities of America's Student Achievement*. New York: Century Foundation Press.

Rothstein, Richard. 2004. *Class and Schools. Using Social, Economic, and Educational Reform to Close the Black-White Achievement Gap.* New York: Teachers College Press.

Rothstein, Richard. 2006. *Equity in What? Defining the Goals of American Education for which We Seek Equity*. Tisch Lecture, Teachers College, Columbia University, January 30. http://www.tc.columbia.edu/news/article.htm?id=5467

Rothstein, Richard, and Rebecca Jacobsen. 2006 (forthcoming). "The Goals of Education (working title)." *Phi Delta Kappan* 88 (4), December.

Rothstein, Richard, and Karen Hawley Miles. 1995. *Where's the Money Gone? Changes in the Level and Composition of Education Spending*. Washington, D.C.: Economic Policy Institute.

Salganik, Laura Hersh, et al. (Richard P. Phelps, Leonard Bianchi, David Nohara and Thomas M. Smith). 1993. *Education in States and Nations: Indicators Comparing U.S. States with the OECD Countries in 1988.* Washington, D.C.: US Department of Education, Office of Educational Research and Improvement, October. NCES 93-237.

Samuels, Christina A. 2006. “Regulations on ‘2 Percent’ Testing Awaited.” *Education Week* 26 (3), September 13: 31-32.

Saulny, Susan. 2005. "U.S. Provides Rules to States For Testing Special Pupils." *The New York Times*, May 11.

Shonkoff, Jack P., and Deborah A. Phillips, eds., 2000. *From Neurons to Neighborhoods. The Science of Early Childhood Development.* Washington, D.C.: National Academy Press.

Smith, Marshall S. and Jennifer A. O'Day, 1991. "Systemic School Reform." In Susan H. Fuhrman and Betty Malen, eds., *The Politics of Curriculum and Testing*. Bristol, PA: Falmer Press, 233-267.

Spellings, Margaret. 2006. "Secretary Spellings Delivered Remarks at Education Trust Dispelling the Myth Award Ceremony." November 6. Press Release. http://www.ed.gov/news/pressreleases/2006/11/11062006.html

Valentine, Jeff C. and Harry Cooper. 2003. *Effect Size Substantive Interpretation*

Guidelines: Issues in the Interpretation of Effect Sizes. Washington, DC: What Works Clearinghouse.

Vinovskis, Maris. 1998. *Overseeing the Nation's Report Card. The Creation and Evolution of the National Assessment Governing Board (NAGB)*. National Assessment Governing Board, U.S. Department of Education. http://www.nagb.org/pubs/95222.pdf

WASL (Washington Assessment of Student Learning). 2006. "Frequently Asked Questions about the WASL." http://www.wasl2006.com/faq/. Accessed September 22, 2006.

Wirtz, Willard, and Archie Lapointe. 1982. *Measuring the Quality of Education. A Report on Assessing Educational Progress*. Washington, D.C.: Wirtz and Lapointe.

Wurtz, Emily. 1999. *National Education Goals: Lessons Learned, Challenges Ahead*. Washington, D.C.: National Education Goals Panel. December.

Yoshikawa, Hirokazu. 1995. "Long-Term Effects of Early Childhood Programs on Social Outcomes and Delinquency." *The Future of Children* 5 (3), Winter, 51-75.

Zajonc, R.B. 1986. "The Decline and Rise of Scholastic Aptitude Scores: A Prediction Derived from the Confluence Model," *American Psychologist*, vol. 41, no. 8 (August), pp. 862-863.

Endnotes

[1] NCLB, Title 1, Sec. 1001; Title 1, Part A, Subpart 1, Sec. 1111 (b)(1)(D)
[2] e.g., Perie, Grigg and Dion, p. 2.
[3] Education Week 2006
[4] Smith and O'Day 1991, p 236, 244.
[5] O'Day and Smith 1993, p. 262-3, 301.
[6] NCLB, Title VI, Part C, Sec 411 (e)2(c).
[7] IES 2006.
[8] IES 2006.
[9] Wurtz 1999, p. 19.
[10] Salganik et al. 1993, Table 9a, page 56.
[11] Reese et al. 1997. Figure 3.2, page 44.
[12] Pashley and Phillips 1993, Table 4, p. 25.
[13] Gonzales et al. 2004. Tables 3 and 9.
[14] Mullis et al. 2003, p. 24 and Exhibit 1.1, p 26.
[15] Olson. 2006 (emphasis added).
[16] Spellings 2006.
[17] Gonzales et al. 2004. Tables C2 and C12
[18] DOE 2005b.
[19] Klein 2006; Dillon 2006.
[20] Koretz 2006a.
[21] Campbell, Hombo, and Mazzeo 2000; Koretz 1986; Grissmer et al. 1994.
[22] Commission 1983, pp 8-9.
[23] Koretz 1986, 2006a.
[24] Zajonc 1986; Koretz 1987; Rothstein 1998.
[25] Linn 2000; Koretz 2005.
[26] Grissmer et al.1994
[27] Cohen 1988, p. 13.
[28] Cohen 1988, p. 13; Valentine and Cooper 2003, p. 5.
[29] Gonzales et al. 2004, tables 3 and C12.
[30] IES 2006.
[31] IES 2006
[32] Vinovskis 1998, p. 42.
[33] Vinovskis 1998, p. 42.
[34] Vinovskis 1998, p. 43.
[35] Lapointe and Koffler 1982, p. 4. Wirtz and Lapointe 1982, p. x.
[36] Wirtz and Lapointe 1982, p. 32, 33, 38, 33.
[37] Vinovskis 1998, p. 44.
[38] Vinovskis 1998, p. 45, emphasis added
[39] Vinovskis 1998, p. 44.
[40] Rotherham 1996
[41] Livingston and Zieky 1982.
[42] GAO 1993, p. 15.
[43] Loomis and Bourque 2001a.
[44] GAO 1993, p. 12.
[45] Livingston and Zieky 1982
[46] Glass 1978.
[47] Loomis and Bourque 2001a, p. 3.
[48] Rothman 1991.
[49] Popham 2000, p. 164.
[50] Loomis and Bourque 2001a, p. 2.
[51] Vinovskis 1998, p. 46-47

[52] GAO 1993, p. 31-32.
[53] GAO 1993, p. 38.
[54] Elliott 1992, p. 71.
[55] NAE 1993, p. xxii, 148.
[56] NAE 1993, p. xxiv.
[57] Pelligrino, Jones, and Mitchell 1999, p. 7.
[58] Vinovskis 1998, p. 56.
[59] Griss, Lauko and Brockway 2006, p. 5.
[60] GAO 1993, p. 57.
[61] Popham 2000, p. 176
[62] Kingsbury et al. 2003.
[63] Kingsbury et al. 2003.
[64] Kingsbury et al. 2003.
[65] Linn 2006.
[66] Linn 2000, Fig. 6, p. 10.
[67] Lee 2006, Table B-1, p. 68.
[68] Horn et al. 2000, pp 24-25
[69] Hoff 2002
[70] Linn 2006.
[71] Linn 2000.
[72] NCEE 1983, p. 20.
[73] O'Day and Smith 1993, p.258, 262.
[74] Camp 1989, p. 21.
[75] Camp 1989, p. 21.
[76] Camp 1989, p. 15.
[77] Karlof and Ostblom, p. 21.
[78] Perie, Grigg and Dion 2005.
[79] Linn, Baker, and Betebenner 2002; Koretz 2006a; Rothstein 2006.
[80] Bracey 2005.
[81] Linn 2003, p. 4, 12; Linn 2004, p. 5.
[82] NEGP.
[83] Perie, Grigg, and Dion 2005.
[84] Rothstein 2004, p. 57; Phillips 2000.
[85] Hack, Klein and Taylor 1995; Reichman 2005.
[86] Barnett 1995; Yoshikawa 1995; Magnuson and Waldfogel 2005.
[87] Shonkoff and Phillips 2000.
[88] Heckman 2006. Also, Knudsen et al. 2006.
[89] Rothstein and Miles 1995; Alonso and Rothstein (forthcoming).
[90] Nores et al. 2005; Heckman and Masterov 2004; Heckman 2006.
[91] Knudsen et al. 2006, p. 10161.