

'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.[§]

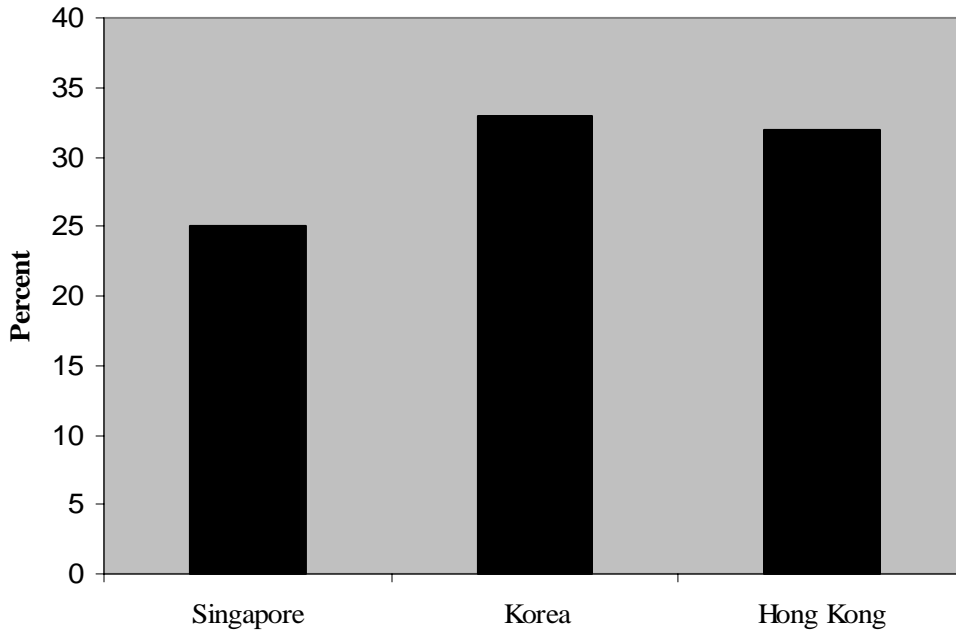
* The estimate reported here, that about 60 percent of Taiwanese eighth graders were less than proficient in math, comes from an Educational Testing Service study using the initial NAEP proficiency cut score, set in 1992, of 294. With the proficiency cut score subsequently redefined as 299, a larger 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

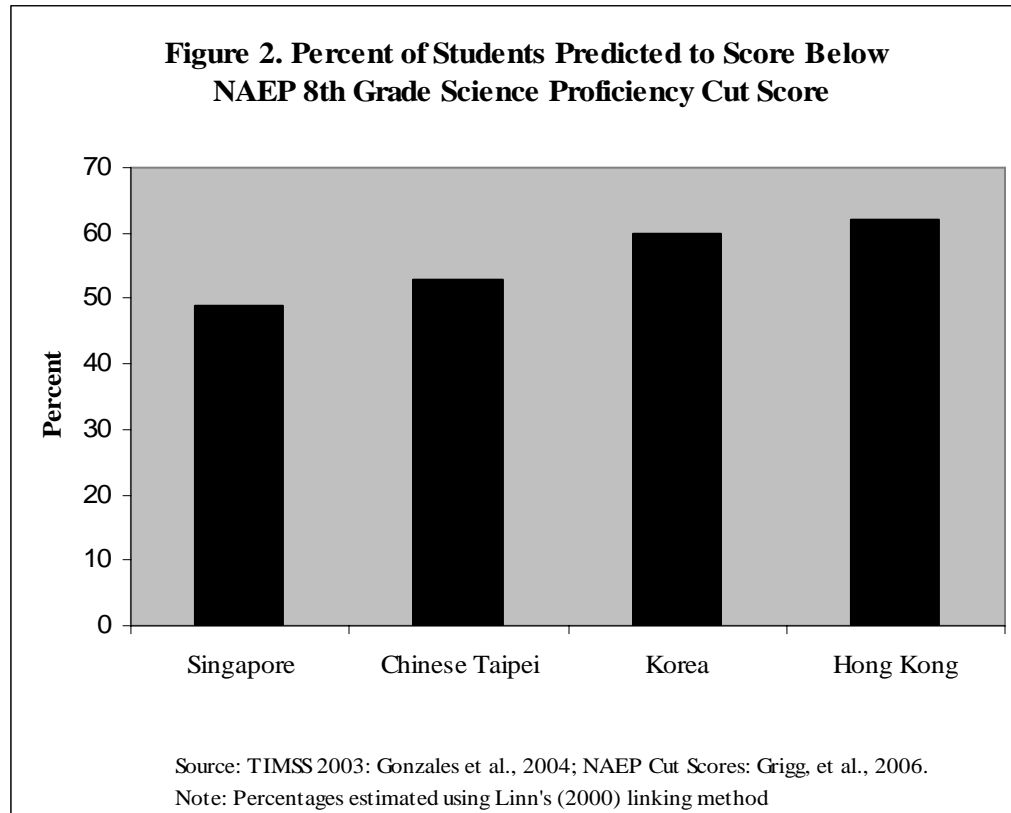
**Figure 1. Percent of Students Predicted to Score Below
NAEP 8th Grade Math Proficiency Cut Score**



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.



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.

