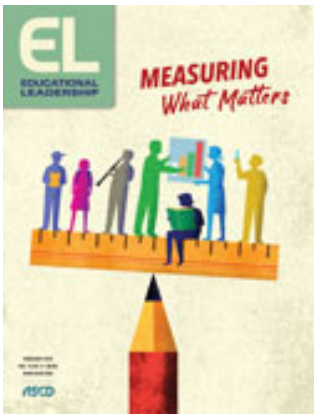
EDUCATIONAL LEADERSHIP

February 2018 | Volume **75** | Number **5**
Measuring What Matters Pages 64-69

Issue Table of Contents | Read Article Abstract



The Problem with "Proficient"

James Harvey

How NAEP's achievement-level benchmarks fail U.S. schools.

In 1996, the International Education Assessment (IEA) released one of the earliest examinations of how well 4th grade students all over the world could read. IEA is a highly credible international institution that monitors comparative school performance; it also administers the Trends in Mathematics and Science Study (TIMSS), a global assessment of 4th and 8th grade mathematics and science achievement. Its 1996 assessment (The IEA Reading Literacy Study, a predecessor to the Progress in International Reading Literacy Study, or PIRLS) demonstrated that out of 27 participating nations, U.S. 4th graders ranked number two in reading (National Center for Education Statistics, 1996). Only Finland ranked higher. To the extent these rankings mean very much, this second-place finish for the United States was an impressive accomplishment.

But around the same time, the National Assessment Governing Board of the National Assessment of Educational Progress (NAEP) reported that just one-third of American 4th graders were "proficient" in reading. To this day, the board of NAEP continues to release similarly bleak findings about American 4th graders' reading performance (National Center for Education Statistics, 2011). And IEA continues to release global findings indicating that the performance of U.S. 4th graders in reading remains world class (Mullis et al., 2012).

How could both these findings be accurate? Was it true, as NAEP results indicated, that U.S. 4th graders couldn't walk and chew gum at the same time? Or was IEA's conclusion—that the performance of American 4th graders in an international context was first class—more valid? A broader question arises here, one that has intrigued researchers for years: How would other nations perform if their students were held to the NAEP achievement-level benchmark for "proficient"? How might they perform on Common Core-aligned assess-ments with benchmarks that reflect those of NAEP?

How Would Other Nations Score on NAEP?

In 2015, statistician Emre Gönülates and I set out to explore these questions on behalf of the National Superintendents Roundtable (of which I am executive director) and the Horace Mann League (on whose board I serve). The results of our examination, recently released in a report titled *How High the Bar?* (Harvey & Gönülates, 2017), are eye-opening. In short, the vast majority of students in the vast majority of nations would not clear the NAEP bar for proficiency in reading, mathematics, or science. And the same is true of the "career and college-readiness" benchmarks in mathematics and English language arts that are used by the major Common Core-aligned assessments.

This finding matters because in recent years, communities all over the United States have seen bleak headlines about the performance of their students and schools. Many of these headlines rely on reports about student achievement from NAEP or the Common Core assessments. One particular concern is that only a minority of students in the United States meet the NAEP Proficient benchmark. Frequently, arguments in favor of maintaining this particular benchmark as the desired goal for American students and education institutions are couched in terms of establishing demanding standards so the United States becomes more competitive internationally.

But the reality is that communities around the world would face identical bleak headlines if their students sat down to take the NAEP assessments. So, when U.S. citizens read that "only one-third" or "less than half" of the students in their local schools are proficient in mathematics, science, or reading (or other subjects), they can rest assured that the same judgments could be applied to national education systems throughout the world if students in those nations participated in NAEP or Common Core-related assessments. (This is true despite the widespread perception that average student performance in some other nations exceeds average student performance in the United States. The metric applied in our study is not a rank ordering of mean scores by nation but the percentage of students in each nation likely to exceed the NAEP Proficient benchmark.)

Our findings may not even be surprising when we consider questions that have arisen from previous research on NAEP. For example:

> Why would anyone expect 4th grade students to obtain a "proficient" score in reading on a NAEP test within which they are asked to respond to reading passages that two separate reading analyses defined as appropriate for students in Grade 7 (Binkley & Kelly, 2003)?[1]
>
> Is it reasonable to believe that 30 percent of U.S. math students who have completed calculus aren't proficient in mathematics, as Pellegrino and colleagues (1999) found?
>
> Should we believe that 69 percent of precalculus students and 92 percent of those who completed trigonometry and Algebra II are similar "failures" who couldn't reach NAEP's benchmark for proficiency (Loveless, 2016)?

International Comparisons

The principle international assessments that can be reliably linked to NAEP are those that test reading in grade 4 (PIRLS) and mathematics and science in grade 8 (TIMSS).[2] The linking that Emre Gönülates

and I did in our research "maps" NAEP scores to comparable scores on TIMSS and PIRLS and to other assessments, such as those de-veloped by the Partnership for Assessment of Readiness for College and Careers and the Smarter Balanced Assessment Consortium.

The mapping procedures, extensively described in *How High the Bar?*, rely on methodologies known as *statistical moderation* and *equipercentile ranking*, both widely employed by testing experts and the U.S. Department of Education. Statistical moderation uses the mean and standard deviation of different assessments to put the scores of one assessment (such as NAEP 8th grade math) on the same distribution as the second assessment (such as TIMSS 8th grade math). *How High the Bar?* draws heavily on analyses completed by Gary W. Phillips, chief scientist at the American Institutes for Research and a former Acting Commissioner of the National Center for Education Statistics, although Phillips had no role in developing the report.[3]

When the NAEP benchmark for proficiency is applied to the results of these international assessments in reading (Grade 4) and math and science (Grade 8), it's the rare nation—even among advanced economies—in which 50 percent or more of students would reach this target. Not one of the nations participating in these assessments can demonstrate that a majority of its 4th graders can clear the equivalent of the NAEP proficiency bar in reading in 2011. (The 2016 PIRLS results, released as this article was going to press, do not appear to alter this finding.) Applying the NAEP benchmark to 8th graders in countries that participated in TIMSS, a majority of students in just three nations (Japan, Singapore, and South Korea) can be thought of as proficient in mathematics. That number falls to *one* (Singapore) in science.

The results for 4th grade reading are especially telling. By our analysis, in no country would a majority of students be considered proficient by NAEP's standard (whether these students speak English or a different language). When we apply the NAEP benchmark for proficient, in fact, the performance of 4th grade students on the 2011 PIRLS confirms the International Education Assessment's findings since 1996: The reading scores of students in most assessed nations fall far short of American students' performance. And data show that, using scores that align with the NAEP Proficient benchmark as a standard, not a single nation among the 40 that participated in PIRLs can demonstrate that even 40 percent of its 4th graders clear the NAEP proficiency bar. The U.S. ranks fifth in this 40-nation comparison, performing comparably with England among English-speaking nations.

What About Common Core Assessments?

When the Partnership for Assessment of Readiness for College and Careers and the Smarter Balanced Assessment Consortium developed their Common Core-based assessments, test developers faced considerable pressure to align the career and college readiness benchmarks with NAEP's Proficient benchmark. The pressure appears to have been successful; as shown in Figure 1, the benchmarks defining college and career readiness in these assessments at Grades 4 and 8 either match the NAEP proficiency benchmark or are very close to it. Several states that abandoned the assess-ments these consortiums created and developed their own assessments also adopted these benchmarks. Because the benchmarks for being "career and college ready" on Common Core assessments align with or approach NAEP's Proficient benchmark, we can expect that students around the world would also be found wanting if they took these assessments.

FIGURE 1. Relationship of "Career and College Ready" Benchmarks of Common Core-related Assessments to NAEP Proficient Benchmark

Grade and Subject	Assessment	NAEP Equivalent of "Career and College Ready"
Grade 4 English Language Arts	PARCC	Approaches Proficient
	SBAC	Basic
	Florida	Proficient
	New York	Proficient
Grade 4 Mathematics	PARCC	Approaches Proficient
	SBAC	Basic
	Florida	Proficient
	New York	Proficient
Grade 8 English Language Arts	Florida	Proficient
	New York	Proficient
Grade 8 Mathematics	PARCC	Proficient
	SBAC	Approaches Proficient
	Florida	Proficient
	New York	Proficient

Note: PARCC = Partnership for Assessment of Readiness for College and Careers. SBAC = Smarter Balanced Assessment Consortium. State names listed indicate a standardized test aligned to the Common Core standards developed by that state. Source of data: National Benchmarks for State Achievement Standards by Gary Phillips, 2016 American Institutes for Research.

The fault here lies not in the students. Not in the schools. Not in the Common Core standards or even in the assessments themselves. The fault lies in the flawed benchmarks defining acceptable performance that are attached to NAEP and the Common Core assessments. No matter how well-meaning advocates are, when they push for school improve-ment on the grounds of highly questionable assessment benchmarks, they aren't strengthening schools and building a better America. By painting a false picture of student achievement, they compromise public confidence in the nation's schools, in the process undermining education and weakening the United States.

Mis-setting the Bar

The history of NAEP's benchmarks partially explains how policymakers and the public have become convinced of the catastrophic failure of American public education. It's a tale of an audacious coup in public policy that combined a determination to demonstrate that large numbers of students were failing, rejection of expert advice about how to set benchmarks, and misuse of the word *proficient*.

In May 1990, NAEP's governing board voted to approve three achievement levels for NAEP tests: Basic, Proficient, and Advanced. The National Assessment Governing Board (NAGB) Chair Chester E. Finn, Jr., praised the decision, saying, "NAEP will report ... how good is good enough. What has been a descriptive process will become a normative process" (Rothman, 1990/1998). In a breathtakingly fast process, an advisory panel appointed by NAGB in June 1990 reached agreement by November on three achievement levels (Basic, Proficient, and Advanced) and the proportion of students at each level who should answer each question correctly. The NAGB adopted the recommendation in May 1991 by a vote of 19–1. In doing so, the board accepted the percentage of correct answers for each level recommended by the panel and decided to use this new methodology in the reading assessment to be mounted later that year (Vinovskis, 1998, p. 45).

Challenged about the speed and quality of this process, Finn responded that he was unwilling to sacrifice the "sense of urgency for national improvement," according one account (Vinovskis, 1998, p. 46). The board subsequently attempted to fire a team of contractors hired to assess the standards-setting process after it issued a critical report (Vinovskis, 1998, p. 47). In a later interview, Finn (2004) dismissed the value of technical expertise: "I get fed up with technical experts [who] take an adversarial stance toward some of the things that are most important in the views of those operating NAEP, such as setting standards" (Finn, 2004, p. 261).

At the time, many independent analysts rejected the process used to develop the NAEP benchmarks. Subsequent to NAGB's rejection of expert advice on the standard-setting, a number of institutions that were asked by Congress or the U.S. Department of Education to review this standard-setting process issued blistering critiques. The criticism from the U.S. General Accounting Office (1993), the National Academy of Sciences (Pellegrino, Jones, & Mitchell, 1999), contractors to the U.S. Department of Education (Scott, Ingels, & Owings, 2007), and analysts at the Brookings Institution (Loveless, 2016) and the National Academy of Sciences, Engineering, and Medicine (Edley & Koenig, 2016) was withering. The procedures for setting the benchmarks were "flawed," "defied reason and common sense," and produced results of "doubtful validity," concluded the critics. In the face of this fusillade, the NAGB made minor adjustments around the margins—while maintaining that it had to preserve its current benchmarks in the absence of anything better.

The most puzzling aspect of NAEP's standard-setting exercise was the use of the term *proficient*. Although observers and high-level government officials often misinterpret this term in regard to the assessments, NAEP and the NAGB have been clear over the years that, as they use the term, *proficient* doesn't mean performance at grade level (Loomis & Bourque, 2001, p. 2). For purposes of NAEP, *proficient* doesn't even mean proficient in the usual sense of the word. As NAGB officials have explained: "Students who may be proficient in a subject, given the common usage of the term, might not satisfy the requirements for performance at the NAEP achievement level" (Loomis & Bourque, 2001, p. 2).

Little wonder that public officials and citizens are confused. A term commonly understood to mean one thing has been redefined to be accurate only in a narrow technical sense. That term is then used in messaging tied to prominent national and state assessments, messaging delivered to the public and policymakers—who can only be expected to understand the term in its common usage. It's hard to avoid concluding that the word was consciously chosen to confuse policymakers and the public.

What's To Be Done?

It's time to take seriously the possibility that the NAEP bar for proficiency has been set so mistakenly high that it defeats NAEP's stated purpose of providing valuable insights into the performance of American students. So how to fix this situation? *How High the Bar?* makes no recommendation to lower standards. Instead it suggests a couple of common-sense improvements, starting with redefining NAEP's basic terminology to get rid of normative terms like *proficient*. *How High the Bar?* recommends replacing the terminology NAEP currently applies to its performance levels (Below Basic, Basic, Proficient, and Advanced) with the performance levels employed in international assessments: Low, Intermediate, Advanced, and Superior. This simple change in terminology would go a long way toward reducing the confusion the term *proficient* has introduced into the national discussion of school performance. And we should educate the public about the flaws embedded in these benchmarks and emphasize to everyone the caution that Congress has always assigned to them. It would also be highly desirable if the views of independent psychometricians and assessment experts guided NAEP's thinking about other technical judgments that could improve NAEP.

I'd also recommend that states revisit their decisions to tie their Common Core assessments to NAEP's benchmarks. To this last point, we should ask ourselves whether it makes sense to align benchmarks on Common Core assessments (potential gatekeepers for high school graduation or college enrollment) with NAEP's Proficient benchmark when fully 50 percent of students judged merely "Basic" by NAEP's metrics go on to obtain a four-year degree (Scott, Ingels, & Owings, 2007).

Going beyond these initial ideas, I believe the NAGB members, public officials, and American citizens need to learn from the wisdom of George Orwell. Orwell (1968) warned about the harm that can be done when words and terms are misused. The careless use of language was a mortal sin in his mind. He wrote: "[The English language] becomes ugly and inaccurate because our thoughts are foolish, but the slovenliness of our language makes it easier for us to have foolish thoughts."

I maintain that slovenly use of the term *proficient* has made it easy for American citizens and policymakers to have foolish beliefs about the failure of schools in the United States. No one denies that we have many problems in our education system. But the problems cannot be addressed if they are defined by trying to meet inappropriate benchmarks that were developed as much to prove a political point as to enlighten public understanding about the nature of our educational challenges.

References

Binkley, M., & Kelly, D. (2003). A content comparison of the NAEP and PIRLS fourth-grade reading assessments. Washington, DC: National Center for Education Statistics.

Edley, C., Jr., & Koenig, J. A. (Eds). (2016). Evaluation of the achievement levels for mathematics and reading on the National Assessment of Educational Progress. Washington, DC: National Academies Press.

Finn, C. E., Jr. (2004). "An interview with Chester E. Finn, Jr." In L. V. Jones and I. Olkins (Eds.), *The Nation's Report Card: Evolution and Perspectives*. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Harvey, J., & Gönülates, E. (2017). *How high the bar: NAEP and Common-Core benchmarks in an international context*. Seattle: National Superintendents Roundtable.

Loomis, S. C., & Bourque, M. L. (Eds.). (2001). *National Assessment of Educational Progress achievement levels, 1992–1998 for reading*. Washington, DC: National Assessment Governing Board.

Loveless, T. (2016, June 13). The NAEP proficiency myth [blog post]. Retrieved from www.brookings.edu/blog/brown-center-chalkboard

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). PIRLS 2011 international results in reading. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

National Center for Education Statistics. (1996). Reading literacy in the United States: Findings from the IEA reading literacy study. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (2011). The nation's report card: Reading 2011 (NCES 2012-457). Washington, DC: U.S. Department of Education.

Orwell, G.(1968). "Politics and the English language." in *The collected essays, journalism and letters of George Orwell* (vol. 4). Sonia Orwell and Ian Angos (Eds), pp. 127–140. New York: Harcourt, Brace, Javanovich.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). Grading the nation's report card: Evaluating NAEP and transofmring the assessment of educational progress. Washington, DC: National Academy Press.

PIRLS 2016 International Results in Reading: IEA, TIMSS, & PIRLS International Study Center, Lynch School of Education, Boston COllege. Retrieved from http://tinyurl.com/yascxnux

Rothman, R. (1990, May 23). "NAEP to create three standards for performance." *Education Week*. Cited in Vinovskis, M. A. (1998). Overseeing the nation's report card: The creation and evolution of the National Assessment Governing Board (NAGB). Ann Arbor, MI: University of Michi-gan

Scott, L. A., Ingels, S. J., & Owings, J. A. (2007). Interpreting 12th-graders' NAEP-scaled mathematics performance and postsecondary outcomes from the National Education Longitudinal Study of 1988 (NELS:88). U.S. Department of Education, Institute for Education Sciences, National Center for Education Statistics, (NCES 2007-328).

U.S. General Accounting Office. (1993). Educational achievement standards: NAGB's approach yields misleading interpretations. GAO/PEMD–993–12. Washington, DC: Author.

Vinovskis, M. A. (1998). Overseeing the nation's report card: The creation and evolution of the National Assessment Governing Board (NAGB). Ann Arbor, MI: University of Michigan.

Endnotes

[1] Binkley and Kelly of the assessment division of the National Center for Education Statistics applied two separate readability formulas to the 2002 NAEP 4th grade reading assessment. The formulas develop a metric for English texts that provides reading difficulty level by grade, based on sentence length and word complexity. Both analyses concluded that NAEP 4th grade reading passages were written at a level appropriate for 7th grade readers.

[2] It is not possible to link results from the Program on International Student Assessment (PISA) reliably to NAEP or Common Core-related assessment benchmarks. PISA assessments are administered to a sample of 15-year-old students who are found, in different nations and to different degrees, in grades ranging between Grade 7 and Grade 12. Given the small sample sizes per nation in international assessments, it's unlikely that a valid comparison could be drawn between the limited number of Grade 8 students assessed per nation in PISA and the nationally representative samples of U.S. Grade 8 students assessed in NAEP and TIMSS.

[3] See these reports by Gary Phillips, all published by the American Institutes for Research in Washington, D.C. Linking NAEP Achievement Levels to TIMSS (2007); Linking the 2011 National Assessment of Educational Progress (NAEP) in Reading to the 2011 Progress in International Reading Literacy Study (PIRLS) (2014), and *National Benchmarks for State Achievement Standards* (2016).

James Harvey (www.superintendentsforum.org) is executive director of the National Superintendents Roundtable and a member of the board of the Horace Mann League. Follow him on Twitter.

KEYWORDS

Click on keywords to see similar products:
assessment and grading, education policy, standards, common core state standards, whole child: supported, whole child: challenged, audience: administrators, audience: district-based-administrators, audience: higher-education, audience: new-principals, audience: new-teachers, audience: principals, audience: teacher-leaders, audience: teachers, audience: building-level-specialist, audience: instructional-coaches, audience: superintendents, audience: students, level: k-12