Embargoed: Not for release or discussion before 10:00 am (EST) Wednesday, January 17, 2018

HOW HIGH THE BAR?

How would other nations perform if their students were judged by Common Core or NAEP benchmarks?

THE NATIONAL SUPERINTENDENTS ROUNDTABLE AND THE HORACE MANN LEAGUE

JANUARY 2018

TABLE OF CONTENTS

INTRODUCTION

In 1996, the U.S. Department of Education released an analysis of a global assessment of Grade 4 reading administered earlier by the International Association for the Evaluation of Educational Achievement (IEA). The assessment demonstrated that among 27 nations, as measured by average reading performance, American fourth-graders ranked number two. Only Finland ranked higher. To the extent these rankings mean very much, for the United States this second-place finish was impressive.

Nevertheless, at about the same time, the National Assessment Governing Board (NAGB) of the National Assessment of Educational Progress (NAEP) was reporting that just one-third of American fourth-graders were "proficient" in reading. To this day, NAGB continues to release similar bleak findings about Grade 4 reading for American students. And IEA continues to release global findings on Grade 4 reading indicating that the performance of American students in reading at the fourth-grade level remains world-class.

How could both of these findings be accurate? Could they be reconciled? More broadly, a question that has intrigued researchers for 20 years arises: How would other nations perform if their students were held to the NAEP benchmark of Proficient? Similar questions can be anticipated about the Common Core assessments if these state tests are aligned with the NAEP proficiency benchmark. These are the issues this report sets out to explore.

Fortunately, several high-quality international assessments – the Progress in International Reading Literacy Study of Grade 4 reading (PIRLS) and the Trends in International Mathematics and Science Study in Grade 8 mathematics and science (TIMSS) – enable us to map the NAEP and Common Core benchmarks onto PIRLS and TIMSS results.*

The National Superintendents Roundtable and the Horace Mann League support high standards. The members of these associations are all educators. They have no interest in undermining their own profession. They believe the pursuit of excellence requires rigorous standards. They also believe in assessment. The value of large-scale assessments (national or international) is that, properly administered and reported, they provide a window into the world of schools along with solid estimates of student performance. The Roundtable and the League understand that. Each association is especially committed to the sort of assessment practices that help states, districts, schools, and teachers determine areas in which students are performing well and those where students need additional support. Several aspects of the new Common Core tests promise that sort of information.

But educators and policymakers must be confident that benchmarks defining acceptable performance on domestic assessments are valid guides to action. Without such confidence, conclusions about student performance in U.S. schools may be flawed. Responses based on flawed conclusions can only lead to distorted policies.

In discussions about assessment, the temptation to get into complex psychometric issues is well-nigh irresistible. This report sets out to do three things: (1) It aims to demystify assessment terminology and methodology so that front-line educators can understand what lies behind pronouncements about the performance of American students. (2) It brings together and examines two different strands describing the performance of our students – domestic and international assessments – to shed some light on how valid, in the broadest sense, these domestic benchmarks are. And (3) it provides a critical examination of the validity of NAEP benchmarks, defined broadly not technically, by asking how students in other nations measure up to them.

The central finding of this report is that the NAEP benchmark of Proficient is a defective and a misleading guide to action that is frequently inappropriately linked to Common Core assessments about "career and college readiness."

*The Grade 4 reading results reported here are based on a comparison of two 2011 reading assessments: NAEP's domestic assessment and PIRLS' international assessment involving dozens of nations. As this report went to press, IEA released international results from a 2016 administration of PIRLS. The 2016 PIRLS' results do not alter, in any appreciable way, the major conclusions of this report.

HOW HIGH THE BAR?

In recent years, communities all over the United States have been faced with bleak headlines about the performance of their students and schools. Many of these headlines rely on national and state results about performance on the National Assessment of Educational Progress (NAEP) or on the new Common Core assessments aligned with NAEP. A particular concern is that just a minority of students in the United States meets a NAEP benchmark of "Proficient" or Common Core benchmarks of "career and college readiness." Frequently, the arguments in favor of establishing these benchmarks as the desired goals for students and schools are couched in terms of making the United States more competitive internationally.

This report does not endorse an anti-testing agenda. Nor is it opposed to rigorous standards, high-quality assessment, or demanding accountability. The report hopes to inform the agenda for assessment-based accountability and to promote standards that are both rigorous and reasonable.

The analysis included here maps the performance of students abroad against the NAEP benchmark of Proficient. This promises to be a useful exploration because the performance of American students and American schools is frequently criticized on the basis of two different but apparently related pieces of information. On our national assessment, just a minority of students is deemed to be Proficient. And internationally, assessment experts report that the average performance of students in many other nations in reading, mathematics, and science exceeds the average performance of American students. Bringing these two strands of evidence together to ask how the students in other nations would perform if held to NAEP's Proficient benchmark (or comparable benchmarks in the Common Core assessments) should shed some light on how valid, in the broadest sense, these domestic benchmarks are.

This report sets out to do several things:

- First, it describes NAEP's structure and benchmarks and compares them with those of two major international assessments: TIMSS (Trends in Mathematics and Science Study, which assesses mathematics and science achievement in grade 8) and PIRLS (Progress in International Reading Literacy Study, which assesses fourth-grade reading).
- Then it reviews existing research linking the NAEP standards to international assessments in mathematics and science in Grade 8.
- Third, it explores more recent research linking the NAEP proficiency benchmark in Grade 4 reading to an international assessment of fourth-grade reading, before providing a new analysis identifying nations in which a majority of students clear the NAEP proficiency bar for Grade 4 reading.
- Next, it examines the benchmarks for "college readiness" established by the two major Common Core assessments – the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC).
- Finally, it concludes with a discussion of the findings before moving on to conclusions and recommendations.*

*Readers may wonder where PISA fits into this discussion. PISA (Program on International Student Assessment) is a test administered by the Organization for Economic Collaboration and Development in Paris. It purports to judge national school system performance based on the assessed achievement of 15-year-olds enrolled in school (not 15-year-olds in the general population).

It is not possible to link PISA results reliably to NAEP's 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. Most are in Grades 9 and 10. Given the comparatively small sample sizes per nation in international assessments, it is highly 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, PIRLS and TIMSS.

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

NAEP is the largest, continuing, nationally representative assessment of what American students know and can do in various subject areas. Since 1969, it has conducted periodic assessments of student competence in reading, mathematics, science, writing, U.S. history, civics, geography, and other subjects. This report is concerned almost solely with NAEP assessments governing reading in fourth grade and mathematics and science in grade 8. The National Center for Education Statistics (NCES), which oversees these assessments, considers them to be the "Nation's Report Card" and an integral part of the nation's ability to evaluate the condition and progress of education in America.

Administration. NAEP is administered by a tri-partite structure: NCES is the federal contracting agency that provides funds to a policymaking National Assessment Governing Board (NAGB), which in turn contracts much of the work involved with developing, administering and evaluating the assessments to experts at institutions such as the Educational Testing Service (ETS) and the American Institutes for Research. NAEP policy is established by NAGB, a politically appointed body, which developed the benchmarks discussed in this report.

Reporting. NAEP does not produce results for individual students or schools. Nor does it assess every student in the nation. Even students who take the NAEP assessment do not take the entire assessment. Instead NAEP tests representative samples of students and, through complex psychometric procedures, provides estimates of performance for the nation and selected demographic groups. NAEP's great value lies in producing results against a common yardstick for the entire population and for demographic groups such as all males, all females, or all African-American or Hispanic students.

Participation: Voluntary but Required. Participation by students, schools, districts and states is voluntary. However, federal law requires states in the NAEP sample that are receiving Title I funds to participate in NAEP reading and mathematics assessments in fourth and eighth grades.

Sample size. NAEP samples are very large. For example, the 2015 mathematics assessment involved 139,900 fourth-graders from 7,810 schools and 136,900 eighth-graders from 6,150 schools. The samples also include students with disabilities and English language learners. Since 1996, NAEP has made special efforts to include students challenged with disabilities.

NAEP & INTERNATIONAL ASSESSMENTS EXAMINED IN THIS REPORT

This report examines several assessments, including NAEP, PIRLS, and TIMSS. The U.S. Department of Education administers NAEP through the National Center for Education Statistics; both PIRLS and TIMSS are administered by the International Association for the Evaluation of Educational Achievement (IEA).

National Assessment of Educational Progress

NAEP (See sidebar) is the largest nationally representative and continuing assessment of what American students know and can do in various subject areas. It is often referred to as the "nation's report card" and is considered the "gold standard" of large-scale assessments. It includes a number of different assessments administered over the years, including a long-term assessment administered to students aged 9, 13, and 17 and the "main NAEP" assessment, administered every two years in reading and mathematics at grades 4 and 8, and more recently grade 12. This report focuses exclusively on "main NAEP."

NAEP Benchmarks. NAEP subject-area scales typically range from 0 - 500. In 1990, the National Assessment Governing Board (a politically appointed body that sets policy for NAEP) developed achievement levels to describe performance at certain standards or benchmarks. The achievement levels define Basic, Proficient, and Advanced performance. These three levels can be understood to describe, respectively, "partial mastery" of knowledge and skills, "solid academic performance…over challenging subject matter," and "superior performance." Each of these benchmarks is defined by "cut scores" established by grade, as outlined in Table 1 and Table 2.

As Table 1 makes clear, a state or demographic group of Grade 4 students that produced an average score of 237 by these metrics would be deemed to be performing at NAEP's Basic level. A one-point increase in that average score to 238, on the other hand, would denote Proficient performance. At the same time, it is clear that the range of scores deemed to be Basic or Proficient in reading and mathematics is quite wide. The Grade 4 reading range for Proficient covers 29 points (Table 1); the comparable range in mathematics (Table 2) covers 32 points. Given that the spread between Basic and Advanced in Grade 4

reading is just 60 points in reading (from 208 to 268), and 68 points in mathematics (214-282), 29- and 32-point spreads cover a significant amount of ground.

On one hand, a single point can make the difference between a finding of Basic or Proficient. On the other, the 29-or 32-point range of scores in each achievement level is significant.

Progress in International Reading and Literacy Survey (PIRLS)

PIRLS is also a highly regarded assessment. It is an international examination of reading and literacy skills in the fourth grade. It has been monitoring international trends in reading achievement in fourth-grade every five years since 2001. It is coordinated by the International Association for the Evaluation of Educational Achievement (IEA), the same organization that administers the TIMSS assessments in mathematics and science. IEA is a complex international organization. It maintains a headquarters in Amsterdam; an International Study Center at Boston College's Lynch School of Education; and a major data processing and research center in Hamburg, Germany.

PIRLS Benchmarks. Three aspects of the benchmarks associated with PIRLS are worth noting when compared with the NAEP achievement levels. First, the PIRLS scale runs from 0 to 1,000 (instead of NAEP's 0-500). This is not to imply the PIRLS scale is more precise, it is simply to point out it is different. Then too, the PIRLS scale has to accommodate only one grade level, while the more compact NAEP scale has to accommodate three. Finally, the PIRLS benchmark levels – Low, Intermediate, High, and Advanced – can be thought of as descriptive. They define where student performance fits on the scale. The NAEP benchmarks of Basic, Proficient, and Advanced seem more judgmental, especially in relation to the term "Proficient." They make a judgment about where student performance should be; clearly the intent is to define preferable student performance as "Proficient" or better, not merely "Basic."

Again, the significance of the cut scores is worth noting. A difference of just one point separates judgments about whether results are Low or Intermediate, or High or Advanced. Meanwhile each standard accommodates about 75 points, so that a nation whose students produced a mean score of 474 would be judged to be low performing, while one producing an average score of 475 would be judged to be intermediate. What the general public does not understand is that each of these scores is accompanied by estimates of standard error, perhaps as much as 10 points on the PIRLS and TIMSS assessments. So a score of 475 with a standard error of ten covers a range of approximately 465 to 475. Technically there is no significant difference between a score of 474 and 475.* But in the public mind there is a huge difference. Indeed, there may be no practically significant difference between a score of 471 and 479.

Apart from those issues, how is one to know how to compare a reading score of 401 at the 4th-grade level on NAEP with an identical score on PIRLS? Such a score would denote exceptionally "low" performance on PIRLS but very "advanced" performance on NAEP. Aligning and linking these scales lies at the heart of the research described in this report.

TABLE 1: READING
NAEP Cut Scores and Range of Scores, by Achievement Level and Grade

	Basic	Proficient	Advanced
Grade 4	208-237	238-267	268-500
Grade 8	243-280	281-322	323-500

TABLE 2: MATHEMATICS
NAEP Cut Scores and Range of Scores, by Achievement Level and Grade

	Basic	Proficient	Advanced
Grade 4	214-248	249-281	282-500
Grade 8	262-298	299-332	333-500

TABLE 3: PIRLS Cut Scores and Range of Scores, by Benchmark Level

	Low	Intermediate	High	Advanced
Grade 4	400-474	475-549	550-624	625+

* A score of 475 on PIRLS might be interpreted this way: We are 95% confident that the score is between 465 and 485. That always leaves a possibility of course that the true score lies below 465 or above 485.

TABLE 4: TIMSS Mathematics Benchmarks Cut Scores and Range of Scores, by Benchmark Level

	Low	Intermediate	High	Advanced
Grade 8	400-474	475-549	550-624	625+

TABLE 5: TIMSS Science Benchmarks Cut Scores and Range of Scores, by Benchmark Level

	Low	Intermediate	High	Advanced
Grade 8	400-474	475-549	550-624	625+

Trends in International Mathematics and Science Study (TIMSS)

Since 1995, TIMSS has monitored trends in mathematics and science achievement every four years, in fourth and eighth grade. TIMSS 2015 was the sixth such assessment. In 1995, 2008, and 2015, TIMSS also administered an assessment to advanced mathematics and physics students completing their final year of secondary school. The analysis in this report addresses only the linkages between the NAEP benchmarks and eighth-grade TIMSS mathematics and science assessments.

TIMSS Benchmarks. Like the PIRLS scale, the TIMSS scales (see Tables 4 & 5) run from 0 to 1,000. Again, the TIMSS benchmarks – Low, Intermediate, High or Advanced – are statistically descriptive, not normative. Although originally defined as percentiles – the 25th percentile; the 50th percentile; the 75th percentile, and the 90th percentile – the TIMSS benchmarks are now defined by scale scores as low, intermediate, high, and advanced. Finally, these benchmarks cover a range of approximately 75 points.

LINKING DIFFERENT ASSESSMENTS

Educators' heads can begin to spin in the effort to keep track of different assessments at different grade levels, testing different curricular areas. Table 6 below is a summary displaying the salient characteristics of NAEP, PIRLS, and TIMMS that are of interest in this analysis.

Analysts face two challenges in linking international assessments to NAEP benchmarks. The first is how to express the results of an international assessment with a scale of 0-1,000 in terms of a domestic assessment (NAEP) with a scale of 0-500. Having succeeded in that task, the second is identifying the nations by name and number in which a significant proportion of their students (say a simple majority) clear the NAEP proficiency bar. If the proportion of students who meet the NAEP standard of Proficient in many foreign nations dramatically exceeds the proportion in the United States, the argument that too few American students are meeting an appropriate achievement standard can be maintained. If, on the other hand, very few nations can demonstrate that the majority of their

TABLE 6: Key Characteristics of NAEP, PIRLS & TIMSS

Assessment	Scale	Grade	Subject	Nations	Benchmarks
NAEP	0-500	4	Reading	1	Below Basic, Basic, Proficient, Advanced
PIRLS	0-1000	4	Reading	57*	Low, Intermediate, High, Advanced
NAEP-M	0-500	8	Mathematics	1	Below Basic, Basic, Proficient, Advanced
TIMSS-M	0-1000	8	Mathematics	38*	Low, Intermediate, High, Advanced
NAEP-S	0-500	8	Science	1	Below Basic, Basic, Proficient Advanced
TIMSS-S	0-1000	8	Science	38*	Low, Intermediate, High, Advanced

* The 57 nations in PIRLS include 13 sub-national jurisdictions such as French-speaking Belgium and four nations that tested their students in Grade 6. When these are eliminated, 40 nations or city-states remain, including the United States. With respect to TIMSS 1999, the 38 nations assessed include three sub-national jurisdictions such as Flemish-speaking Belgium. When these are eliminated, 35 nations or city-states remain.[1]

students can meet the NAEP standard of Proficient, the argument is more difficult to sustain.

The Challenge of Linking Assessments. How does one link two assessments that differ in their metrics? Although the challenge is statistically complex, conceptually it is similar to converting the temperature in Celsius to the temperature in Fahrenheit. Considerable progress has been made in recent years in responding to the psychometric challenge of linking different assessments.[2]

Gary W. Phillips, chief scientist at the American Institutes of Research and former Acting Commissioner of the National Center for Education Statistics, has expressed the purpose of linking different assessments together: Linking, he said, is designed to project the NAEP achievement levels on to the scales of the international assessments. The purpose is to answer the question: "How would other countries perform if their [international assessment results] could be expressed in terms of NAEP achievement levels?"[3]

Equipercentile Ranking

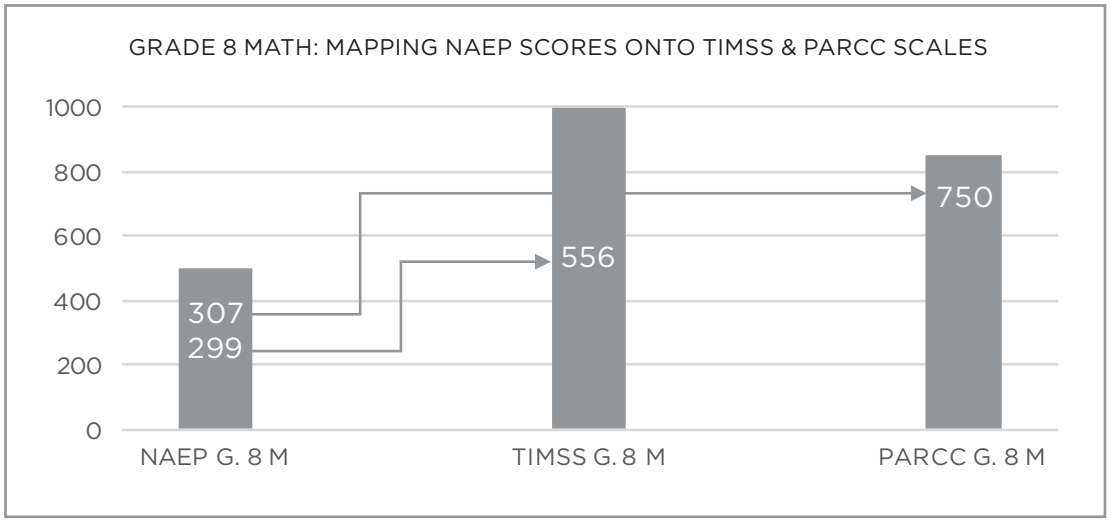
Assessments, national and international, have several things in common. Linking efforts take advantage of these commonalities. One is that assessments report results by percentile level. This makes it possible to "map" the percentile for a given score from one assessment on to the corresponding percentile on another, thereby identifying comparable scores on the two assessments. This equipercentile ranking procedure is, in fact, how the U.S. Department of Education compares the proficiency levels set by state assessments with NAEP's proficiency standard. Figure 1 below provides an example. Once the percentile at which NAEP's Proficient benchmark is determined (based on samples of U.S. students), it is a simple matter to find the equivalent score by percentile on companion assessments (based, it must be acknowledged on different samples).

Statistical Moderation

A second commonality is that every assessment reports major statistical features such as the mean (the arithmetic average), the median (the point at which half the respondents can be found above the line and half below), and standard deviation (points on either side of the mean that define where two-thirds of respondents lie). A process known as "statistical moderation," cited as early as 1992 by Mislevy draws on these features.[4] In a complex formula, this approach uses the mean and standard deviation of different tests to put the scores of one test (e.g., NAEP) on the same distribution as the second (e.g., TIMSS).[5] While the history of the term dates back to 1992, Johnson and his colleagues were the first to fully and successfully employ it to equate TIMSS assessments with NAEP.[6]

Phillips acknowledged that Johnson and his colleagues "did all the hard work" in developing this technique. Phillips used statistical moderation to link the 2000 NAEP achievement level in Grade 8 mathematics and science to comparable TIMSS assessments conducted in 1999. In 2014, he repeated

FIGURE 1: Example of Equipercentile Mapping



Sources: Phillips 2007, Table 3 and Phillips 2016. Table 10. **Read as:** A score of 299 on the NAEP 8th grade math scale converts into 556 on the comparable TIMSS scale. A NAEP score of 307 translates into a PARCC score of 750.

this approach for fourth-grade reading, using the 2011 NAEP and PIRLS administrations to compare achievement levels. This paper relies on these analyses as a foundation.

Although Phillips describes the work as "an extremely easy process" (because Johnson et al. "did all the hard work"), the Johnson-Phillips approach involves statistical formulas that appear so complex to the lay reader as to be intimidating (see Appendix B). Of necessity, a number of assumptions are built into these formulas. Potential error in each of the assessments being linked has to be estimated, including estimates of sampling and measurement error as well as errors in the parameters linking the two different assessments. These are not trivial issues. For example, as Phillips, noted in 2014, the linking parameters in the NAEP/PIRLS analysis were based on data collected in the United States, where students took both NAEP and PIRLS: "In all other countries, however, students only took PIRLS…. There is no guarantee that linking parameters estimated from… the United States will be the same as those in other nations."[7]

Still, with those assumptions acknowledged, fairly rough approximations are possible that (a) link the NAEP benchmark of Proficient to scales employed by PIRLS and TIMSS; and (b) provide estimates of the proportion of each nation's students who clear the NAEP Proficient bar.

NATIONS IN WHICH MOST STUDENTS CLEAR THE NAEP PROFICIENT BAR

There are 195 nations recognized by the United Nations in the world.[8] Most of them do not participate in international assessments. We may assume that many of the countries that do not participate in these assessments are developing, most performing in relatively modest ways on the international stage and in trade. As is true with many developing nations, large proportions of the populations in many of these nations leave school before entering high school.[9]

Among jurisdictions that tend to be larger and wealthier, 38 participated in the 1999 administration of TIMSS, which evaluated student performance in *mathematics and science in Grade 8.* In 2011, an international assessment *of reading in Grade 4* was administered by PIRLS in 57 jurisdictions. The 57 jurisdictions involved with PIRLS and the 38 involved with TIMSS include many sub-jurisdictions that are not national entities. These include jurisdictions such as Hong Kong and Taipei; Andalusia in Spain; several provinces in Canada; and the French-speaking population of Belgium. None of these jurisdictions is recognized as a nation by either the United Nations or the United States. When they are removed from the analysis, 40 nations or city-states remain for the PIRLS analysis and 35 remain for the analysis of TIMSS.

When the NAEP benchmark of Proficient is statistically applied to the results of these assessments in reading (Grade 4) and math and science (Grade 8), it is extremely rare to find any nation that can demonstrate that 50 percent or more of its students are "Proficient."

Grade 8 Mathematics and Science

Turning first to Grade 8 mathematics and science, Phillips (2007) benchmarked TIMSS results in Grade 8 against NAEP's standards. Table 7 summarizes the results. In Grade 8, when the question put by Phillips is raised around mathematics – How would other countries perform if their international assessment results could be expressed in terms of NAEP achievement levels? – just three nations can demonstrate that a majority of their students clear the NAEP proficiency bar. In science, just one city-state can do so.

Lim and Sireci also completed an equipercentile comparison of NAEP mathematics in 2017.[10] It produced considerably higher estimates of "NAEP Proficient" students in Singapore, the Republic of Korea, and Japan than did Phillips' earlier "statistical moderation" approach. It is noteworthy that the same nations were identified in each of these analyses. Lim and Sireci did not examine science.

It is by no means the case that nations in which a majority of students can be thought of as clearing the proficiency bar performed at the same high levels in both mathematics and science. Mathematics students meet the NAEP proficiency benchmark in impressive fashion in three nations, with 62 percent or more of tested students meeting or exceeding the standard, according to Phillips. But in science, Singapore, the exemplar jurisdiction, barely scrapes over the bar, with just 51 percent of its students at or above Proficient. With only trivial adjustments in the linking assumptions, there is little doubt that Singapore's performance might dramatically improve. On the other hand, it might just as easily sink below the 50 percent bar.

Grade 4 Reading

Phillips repeated the first part of the statistical moderation exercise for Grade 4 reading in 2014, comparing NAEP's 2011 reading assessment with the 2011 PIRLS assessment. He then determined how NAEP's benchmarks compared to PIRLS' international achievement levels. He did not apply the second formula (to determine the proportion of students in each nation meeting NAEP's benchmark of Proficient). Without that information, however, he concluded: "At each level, the linking shows that the NAEP Grade 4 reading achievement levels are higher than the PIRLS international benchmarks. This finding provides one piece of validity evidence that NAEP results are internationally competitive."

Without access to the complete NAEP data base, this current study built on Phillips' work and employed an equipercentile ranking approach to determine the number of nations that could demonstrate that a majority of their students met NAEP's standard of proficiency in Grade 4 reading. Table 9 presents the results.

The result is stark. When Phillips' question is put to the test around fourth-grade reading – how would other countries perform if the results of their fourth-grade reading assessment in PIRLS could be expressed in terms of NAEP achievement levels? – not a single nation among the 40 nations or city-states that participated in PIRLS can demonstrate that a simple majority of its students clear the NAEP proficiency bar.

TABLE 7: Nations in which a Majority of Grade 8 Students Clear the NAEP Proficiency Bar in Mathematics

Nation	Proportion of Students at or above NAEP Proficient
Singapore	76.8%
Republic of Korea	69.8%
Japan	61.7%

Source: Phillips, 2007*

TABLE 8: Nations in which a Majority of Grade 8 Students Clear the NAEP Proficiency Bar in Science

Nation	Proportion of Students at or above NAEP Proficient
Singapore	51%

Source: Phillips, 2007*

TABLE 9: Nations in which a Majority of Grade 4 Students Clear the NAEP Proficiency Bar in Reading

Grade/Subject	Number of Nations
Grade 4 Reading	0

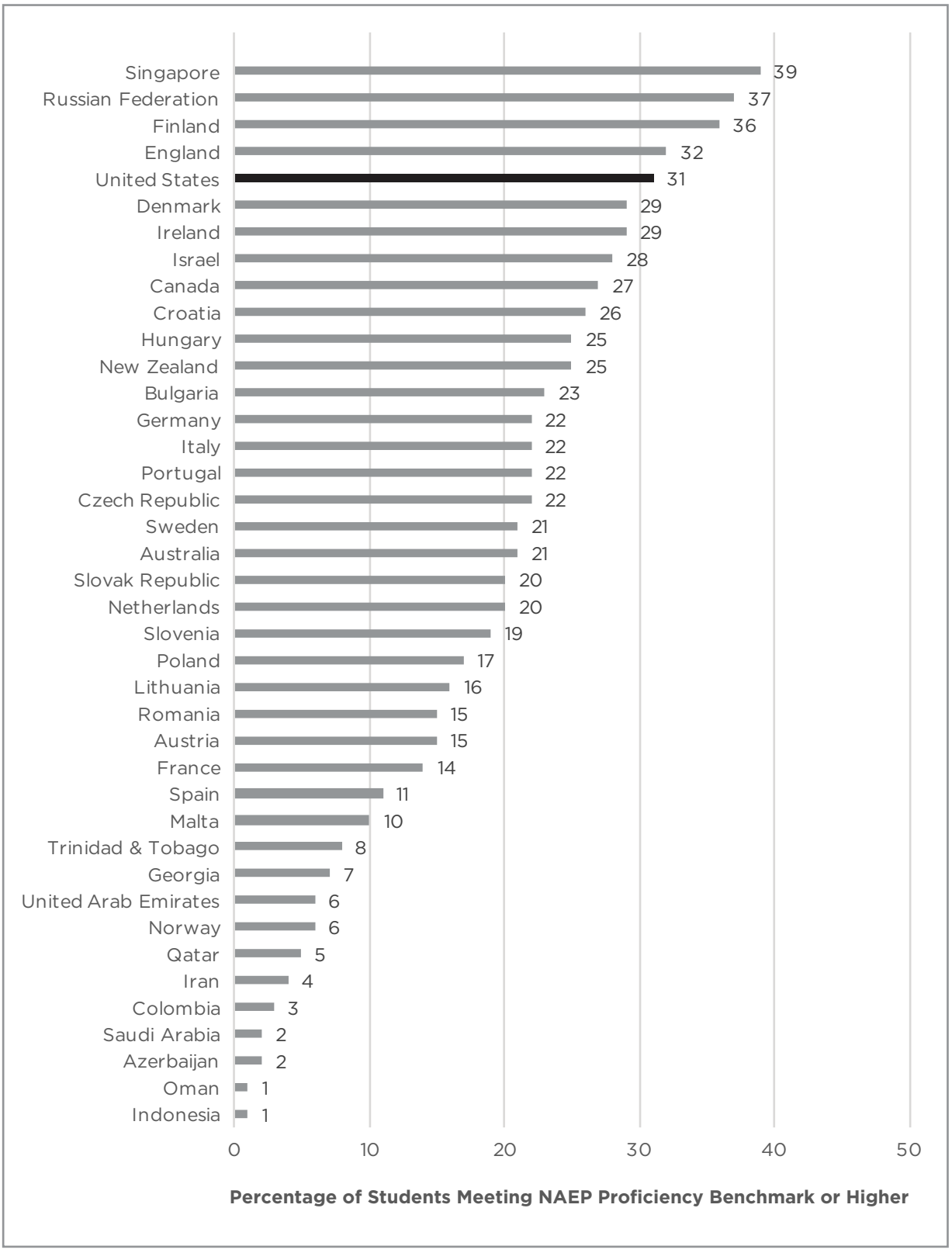
Source: Phillips, 2015 supplemented by Gönülates and Harvey, 2017

Assessing Reading in Different Languages. Anticipating that reading achievement results might differ by native language, this study initially analyzed PIRLS Grade 4 reading results separately for English-speaking nations and for non-English-speaking nations. (The results are displayed in Appendix C.) The distinction proved to be unnecessary. Whether students in different nations speak English as their native tongue or a different language, not a single nation can demonstrate that a majority of its students would be considered Proficient by NAEP's Grade 4 reading standard. By the NAEP standard of Proficient, in fact, the performance of Grade 4 students in reading in the vast majority of assessed nations falls very far short of the performance of students in the United States.

Figure 2 (next page) presents the results. With the NAEP standard of Proficient or better as the benchmark, American fourth-graders rank fifth among the 40 nations or city-states that participated in the PIRLS assessment. Among English-speaking nations in Grade 4, the United States and England have the highest proportion of students performing at NAEP's Proficient Level. As a practical and statistical matter, the one-point difference between these two English-speaking nations is insignificant.

*Phillips was the first to identify the outstanding performance of students from these three nations when their TIMSS mathematics results were aligned with NAEP's benchmarks. Subsequently, Hambleton, Sireci, and Smith (2009) and Lim and Sireci (2017) in separate analyses identified the stellar performance of students from these nations, although the proportions deemed to meet the NAEP Proficient benchmark varied somewhat. Lim and Sireci's analysis also reported that 53.4 percent of students in the Russian Federation met the NAEP mathematics standard.

FIGURE 2: Percentage of Grade 4 Students by Nation Who Meet the NAEP Benchmark of Proficient (or Higher) in Reading



Source: Gönülates and Harvey, See Appendix C

PRESSURE TO CONFORM STATE PROFICIENCY BENCHMARKS TO NAEP'S

It is against that background that the pressure of recent years to conform state standards to NAEP's proficiency benchmarks should be examined. To what extent are the "career and college readiness" benchmarks of the major Common Core assessments – PARCC and SBAC – aligned with NAEP's definition of proficiency?

A series of reports from the National Center for Education Statistics employed an equipercentile approach to link the definition of proficiency used in state assessments for grades 4 and 8 – typically performance at grade level – with NAEP's benchmark of Proficient. The 2015 analysis revealed that when adjusted to the NAEP metric, levels of difficulty across states differed dramatically, in both fourth and eighth grade. The NCES analyses have led advocacy groups to accuse states of deceiving their citizens with artificially low definitions of proficiency. Advocates then used these findings to justify policy proposals to align state assessment benchmarks with NAEP's definition of Proficient.[12]

Indeed, as states have moved their assessments closer to NAEP's proficiency standard, results state-by-state have dismayed educators. In Florida, just 39 percent of fourth graders and 30 percent of eighth graders were proficient in reading.[13] In Wisconsin, just 50 percent of fourth graders were deemed proficient on the state's version of SBAC.[14] Just 33 percent of students in California met or exceeded mathematics standards on the state's version of SBAC.[15]

Parents in some states, alarmed by these results, responded by having their children boycott the assessments. As many as 250,000 students "opted out" of the assessments in New York state in 2016.

PARCC & SBAC Assessments

Phillips' work shines some light on this issue. In 2016, he issued an analysis that aligned NAEP's benchmarks with the Common Core assessments.[16] His study examined achievement standards for PARCC, SBAC, ACT's Aspire, and statewide assessments in non-consortium states. He examined mathematics as well as English and Language Arts in both grades 4 and 8, but not science (since Common Core assessments in science do not exist). In the interests of parsimony, this paper restricts its examination to the subjects explored in the international comparisons – fourth-grade reading, and fourth- and eighth-grade mathematics.

Each of the consortium assessments has its own achievement standards tied to "career and college readiness." For SBAC, that standard is set at Level 3; for PARCC it is set at Level 4. In Grades 4 and 8, the standards are related to being "on track" to be college-ready by the time the student graduates.

Table 10 compares the performance standards by grade and subject of PARCC and SBAC, along with comparable benchmarks in selected state assessments.[†] One caveat offered by Phillips is that for both SBAC and PARCC the exercise involves mapping English/Language Arts (ELA) standards, which include writing, on to NAEP's reading standards, which do not.

Observations

Three observations can be made about these results. The first relates to how closely these assessments seem to be aligned with NAEP's national benchmark of Proficient. Of the 14 comparisons outlined above, nine are tightly aligned with NAEP's Proficient benchmark (Florida and New York in both mathematics and English and Language Arts both grades, along with PARCC in Grade 8 mathematics). Three others approach Proficient (PARCC in Grade 4 English and Language Arts and Grade 8 Mathematics). "Approaches Proficient" should be understood in terms of what was noted earlier in this report: Each of the NAEP benchmarks is accompanied by a range of approximately 30 or more points. While the equivalent NAEP score for SBAC's "college-ready" benchmark (222) places SBAC's Grade 4 ELA standard solidly in the middle of the NAEP *Basic* range, the same cannot be said of the other three benchmarks. PARCC's Grade 8 mathematics standard places it well within NAEP's *Proficient* range. Meanwhile, both PARCC's Grade 4 ELA standard and SBAC's Grade 8 mathematics standard are separated from NAEP's benchmark of *Proficient* by just a few points.

[†]State assessments in 2018 represent a moving target. New legislation enacted in 2015, the *Every Student Succeeds Act,* provides states with greater assessment flexibility than was available under the *No Child Left Behind Act*. Several states have formally abandoned PARCC and SBAC, but retained many of the features in their new assessments. If these new assessments retain benchmarks similar to those in PARCC or SBAC, it is highly unlikely that the results in terms of student success will differ greatly.

TABLE 10: Relationship of Common Core "Career and College Ready" Benchmarks 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 Mathematics	PARCC	Proficient
	SBAC	Approaches Proficient
	Florida	Proficient
	New York	Proficient
Grade 8 English/Language Arts	Florida	Proficient
	New York	Proficient

Source: Phillips, 2016

PARCC and SBAC

The development of the Common Core in recent years by the National Governors Association and the Council of Chief State School Officers has created a shared expectation of what students across all 50 states should know and be able to do. This also provided the opportunity for test consortia, backed up with $300 million from the U. S. Department of Education, to develop assessments grounded in the Common Core.

Two consortia developed assessments around the Common Core – the Partnership for Assessment of Readiness for College and Career (PARCC) and the Smarter Balanced Assessment Consortium (SBAC).

PARCC: Participants in 2016-17 included nine jurisdictions (six states, plus the District of Columbia, the Bureau of Indian Education, and the Department of Defense Schools). In addition, Massachusetts and Louisiana participate "at various levels," according to the PARCC website. The PARCC program offers a common set of K–12 assessments in English and math. When fully implemented, the four key components for Grades 3–11 will include:

1. Diagnostic assessment administered at beginning of each school year
2. Mid-year assessment predictive of a student's likely performance by end-of-year
3. Performance-based assessment in the last quarter of the school year
4. End-of-year summative assessment

Second, although technically it is possible to extend Phillips' analysis here to examine the central question raised in his 2007 and 2014 reports – How would other countries perform if the results of their reading and mathematics assessments in PIRLS and TIMSS were to be expressed in terms of PARCC or SBAC achievement levels? – the temptation to do so has been resisted. Extending the analysis in that way would easily double the margin of error, creating a situation in which casual readers might take seriously the very specific numbers produced without understanding just how unreliable and unstable the estimates are.[17]

Third, the confidence with which these assessments promise to predict "college readiness" is impressive but hardly convincing. While advocates have called for benchmarks related to college readiness, the predictive value of PARCC Grade 10 assessments in terms of college success leave somewhere between 84 and 99.5 percent of what accounts for first-year college success unaccounted for (see sidebar on college and career readiness). If the tenth-grade assessment is such a weak predictor, it is hard to put a lot of confidence in the accuracy of the "on track" assessments in Grade 4 and Grade 8.

DISCUSSION

This study, like Phillips' work, is oriented around assessment in the United States and grounded in American perspectives. Phillips' several *caveats* about these analyses should be kept in mind. Perhaps the most significant is that it is not clear the linking procedures are stable in other countries or that the "normal distribution" assumed in the United States is evident elsewhere. All of these caveats – ranging from assessments given in different years and at different times of the year, to the content differences between NAEP reading and PARCC and SBAC ELA assessments – should usefully be kept in mind. These cautions indicate that the international comparisons cannot be precise. Nevertheless, they do provide rough approximations offering insights into proficiency and college readiness benchmarks in an international context that otherwise would not be available.

Jurisdictions versus Nations

As noted earlier, the results from sub-national jurisdictions are ignored in this report. This decision revolves around a fundamental matter of definition. If U.S. performance is to be compared with other countries, entire nations should be the unit of comparison, not smaller and typically more advantaged sub-jurisdictions.

SBAC: Participants include 15 states, a territory, and the Bureau of Indian Education, according to SBAC's website. Its goal is to allow "all students to demonstrate what they know." Administered in Grades 3–8 and again in high school, the program's components include:

1. Computer-adaptive summative assessment that will be administered during the last 12 weeks of the school year
2. Interim assessments that can be used to predict student performance on the summative assessment while also providing feedback on student progress (mandatory)
3. Formative assessment resources to help teachers diagnose and respond to the needs of their students relative to CCSS,

PARCC and SBAC: The goal of both assessments is to add coherence and clarity to the testing process and assess higher-order thinking skills using performance tasks and innovative technology-enhanced items. Both are administered on computers.

Other: A 2015 map of state testing plans indicates that 21 states use other assessments, including state-specific assessments (many aligned with PARCC or SBAC) and ACT tests.

Sources: Presentations by PARCC and SBAC representatives to the National Superintendents Roundtable, July 2013; PARCC website (http://www.parcconline.org/about/states); SBAC website (http://www.smarterbalanced.org/about/members/); and *Education Week*, February 4, 2015 (http://tinyurl.com/m3ro87k).

COLLEGE AND CAREER READINESS

In 2004, two seminal assessment reports around 12th-grade exit standards were released. First, a national commission issued ***12th-Grade Student Achievement in America: A New Vision for NAEP.***[18] It called for expanding NAEP from grade 4 and 8 to grade 12. Then, three advocacy organizations released an influential report, *Ready or Not,* that called for creating a high school diploma signifying readiness for jobs and college.[19] Achieve, The Education Trust, and the Thomas B. Fordham Foundation called on higher education leaders, employers, and policymakers to tie admissions, college placement, and hiring decisions to more demanding high school exit standards. A decade later the "college and career readiness" movement was in full bloom.

Responding to a request from NAGB, ACHIEVE, Inc. provided two reports to the NAEP governing board **recommending that NAGB align 12th-grade NAEP with college and workforce expectations.** The first report, issued in 2005, governed reading; the second, issued in 2006, addressed mathematics.[20] The reports emphasized that 40% of college students require remediation.[21] The authors grounded their support for such alignment around the knowledge and skills defined in the American Diploma Project (developed in 2004 by a coalition including ACHIEVE). The writing report acknowledged that an assessment designed to focus on college and career preparedness "faces a daunting challenge of validation."

Meanwhile, **PARCC and SBAC** proceeded with developing Common Core-aligned assessments that incorporated benchmarks of "college and career readiness."

By 2014, **NAGB reported it was moving toward endorsing the concept of NAEP as an indicator of preparedness for college and career training** and offered some provisional estimates. **But it cautioned that inferences could be made only at the national level (and not for states or student subgroups),** that the plausibility of inferences of "preparedness" were more solid in mathematics than in English, and that "some proportion" of 12th-grade students would be judged to be falsely negative or falsely positive. (Since NAEP assesses a sample of students not all of them, the false-negative and false-positive challenge relates to population inferences, not findings on individual students.)

To compare PARCC's college readiness standard with its own assessment, the Massachusetts Comprehensive Assessment System (MCAS), the state asked Mathematica Policy Research to examine which test best measured college preparedness. In the summer of 2016, Mathematica concluded that the **PARCC and MCAS 10th-grade exams do equally well in predicting students' college success,** as a function of first year GPA in English (aligned with ELA).[22] That is to say, Mathematica concluded that both MCAS and PARCC exams are "at least as correlated with first-year college grades as are SAT scores." Positive news for both assessments in terms of college and career readiness.

However, it is not clear that either assessment does as good a job as the SAT in predicting total first-year grades. According to a 2015 publication from the College Board, the SAT's sponsor, there is a consistent correlation of about 0.55 between the composite SAT score (Language, Math, and Reading) and first-year GPA.[23] That means the **SAT score explains about 30 percent of first-year GPA.** High school GPA trumps the SAT as a predictor; the two together combine for a correlation between 0.625 and 0.65 from 2006 through 2010. The combination predicts between 39 and 42 percent of first-year GPA.

How well do PARCC and MCAS do in predicting first-year GPA? According to William Mathis, a former school superintendent now with the National Education Policy Center at the University of Colorado, **PARCC math tests predict 16 percent of first-year GPA at best, while it's possible the ELA PARCC assessment explains as little as one-half of one percent of first-year grades.**[24] As Mathis points out, that leaves somewhere between 99.5 percent (ELA) and 84 percent (math) of the variance in first-year grades unexplained. MCAS's ability to predict first-year grades is about the same as PARCC's in terms of ELA, and somewhat lower than PARCC's in mathematics.

It might be argued that similar logic should be applied to comparisons of student performance in the U.S. with student performance in smaller countries. That is a reasonable position, but each of smaller countries at least meets the threshold qualification of recognition as an independent nation or city-state, something that cannot be said of Shanghai, Hong Kong, Northern Ireland, individual Canadian provinces, or other similar sub-jurisdictions.

The Case of Singapore. Of note in the analysis above is the special distinction Singapore holds: the only nation (actually a city-state) in which a majority of its students seem to clear the NAEP proficiency bar in both mathematics and science in Grade 8. There may be aspects of the Singapore educational system that can inform American schools, but they should be understood in the context of the Singaporean culture that produced these results.

Singapore's remarkable rise to becoming an international financial hub was facilitated by central government control exercised for decades by a benevolent dictator, accompanied by what can only be thought of as a punitive legal system. This system, governing everything from littering on streets and eating on public transit to violence, drug addiction and murder, manages citizens' and visitors' behavior on a daily basis, accompanied by fines, punishments, and beatings unthinkable in Western democracies. This punitive culture extends to schools, where caning is common.

Analysts suggesting that Americans have much to learn from the top-down education system in Singapore[25] might acknowledge in passing the clash of cultures and values between this authoritarian city-state and the values undergirding free-market democracies in the West.

Japan and the Republic of Korea. Japan and the Republic of Korea join Singapore on the honor roll of nations in which a substantial proportion of eighth-grade students meet NAEP's proficiency benchmark in Grade 8 mathematics (but not science). Their accomplishment is impressive by any standard. Both nations are well known for obsessive and competitive attention to education. Japan is characterized by "examination hell" in the final year of secondary school, with students grinding to prepare for university admissions examinations from early morning to midnight. South Korea's "education fever" is part of a competitive system embedded in a culture in which education is considered central to national life and a source of family status. In both nations, the intense competitive pressure around school examinations is thought to contribute to high rates of suicide among teenagers.

Controversy Around the Term "Proficient"

To most people, the term "proficient" when applied to an individual is understood to mean that person is reasonably good at doing something. It might even imply advanced or expert skill. With respect to NAEP, the term Proficient is often confused with being at grade level.[26] However, the National Center on Education Statistics has repeatedly stressed that Proficient is NOT synonymous with being at grade level. For example, Loomis and Bourque, two officials associated with NAEP's National Assessment Governing Board (NAGB), said clearly in 2001: "[I]t is important to understand that the Proficient achievement level does not refer to 'at grade' performance."[27]

Indeed, a 2003 NCES comparison of the content and level of difficulty of PIRLS and NAEP's fourth-grade reading assessments concluded that the NAEP assessment asked students to work with reading segments that were twice as long and consisted of longer and more complex sentences than the reading material in PIRLS.[28] Drawing on two different "readability" formulas the study found that both formulas agreed on the average level of difficulty of the NAEP reading passages in fourth grade: these passages would be appropriate for students in Grade 7. One of the formulas suggested the PIRLS material was aimed at students in Grade 5; the other suggested PIRLS material was appropriate for grades 5-6. So the conflation of the term Proficient with performance at grade level dramatically understates the level of difficulty of NAEP material, certainly in the fourth-grade reading assessment.

"Proficient" Does Not Mean Proficient. More surprisingly, it turns out NAEP's definition of proficiency does not mean proficient as most people understand the term. As Loomis and Bourque wrote:

> 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.

How did such an unusual definition come to dominate educational discourse in the United States? In a breathtakingly fast process, an advisory panel appointed by the National Assessment Governing Board in June 1990 reached agreement by November on the three NAEP achievement levels and the proportion of students at each level who should answer each question correctly. NAGB adopted the recommendations the following May. In doing so, NAGB members rejected the advice of technical experts to go slow on the benchmarking process.[29] Challenged about the speed of the process, NAGB Chair Chester E. Finn, Jr. responded that he was unwilling to sacrifice the "sense of urgency for national improvement."[30] In a later interview, Finn 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."[31] Commenting on this history, an analyst from the Economic Policy Institute and former education reporter for the *New York Times*, Richard Rothstein, concluded that Finn believed the "realism of proficiency cut scores was unimportant…compared to the desirable impact on public psychology of demonstrating that large numbers of students were failing."[32]

It is understandable that the general public, educators and policymakers would confuse Proficient with grade-level performance. But it is difficult to understand why the government has for so long encouraged this careless use of language around a topic so fundamental to American social and economic well-being. It sows confusion about the performance of American students and the quality of American schools.

NAEP Standard-Setting Process. That this situation has persisted for so long is all the more puzzling in light of the major controversy associated with NAEP's standard-setting process and the resulting benchmarks the process produced (see sidebar). The controversy has persisted for the last quarter century. Judgments by independent analysts have ranged from conclusions that the standards are "procedurally flawed," producing results of "doubtful validity" (U.S. General Accounting Office, 1993), to comparisons with other data that indicate NAEP's definition of proficiency "defies reason" and "refutes common sense" (Loveless, 2016).

As the sidebar notes, the benchmarks do have their defenders. But when students taking pre-Calculus, Calculus, and Advanced Placement classes fail to clear the Proficient bar in 12th grade, while 50 percent of those judged to be merely Basic by NAEP's metrics later obtain a four-year college degree, it stretches credulity to propose that all students be held to a standard closely aligned with NAEP's Proficient benchmark before being permitted to graduate from high school or admitted to a two- or four-year college.

In light of this ongoing measurement controversy about the validity of the benchmarks NAGB established for NAEP, Congress has insisted since 2001 that NAEP continue to use the achievement levels on a "trial basis," noting that they should be interpreted "with caution." Caution about the NAEP benchmark of Proficient should not be thrown to the four winds. To the degree analysts can make these determinations, the vast majority of students in the vast majority of the nations of the world fail to measure up.

CONCLUSIONS

The results of the analyses in this report are clear, unambiguous, and broadly valuable for policymaking purposes.

- If the NAEP benchmark of Proficient was to be applied to the results of international assessments, the vast majority of students in the vast majority of nations in the world would fail to clear the bar in reading, mathematics, and science.
- With respect to Common Core assessments, it can be concluded with some confidence that to the extent the Common Core assessments align with NAEP's standard of Proficient, it is highly likely that most students in the United States and all over the world will be similarly frustrated if held to typical "college ready" benchmarks.
- It is time to take seriously the possibility that the NAEP bar for Proficient has been set so mistakenly high that it (a) defeats NAEP's purpose of providing valuable insights into the performance of American students; and (b) establishes a standard that defeats the best efforts of educational systems around the world.
- The term "Proficient" is judgmental, not evaluative and its use has misled the public and policymakers.

CONTROVERSY AROUND NAEP STANDARD-SETTING PROCESS

Although most public discussion of NAEP benchmarks assumes their development and validity are settled matters, the truth is that a scientific debate has raged for decades about both the definitions and how they were developed. This is by no means a settled question among psychometricians. For example:

- The U.S. General Accounting Office (1993) concluded that NAEP's standard-setting process was "procedurally flawed" and the results of "doubtful validity."[33]
- The National Academy of Sciences (1999) agreed that NAEP's achievement-level setting procedures were flawed. "The judgment tasks are difficult and confusing; raters' judgments of different item types are internally inconsistent; appropriate validity evidence for the cut scores is lacking; and the process has produced unreasonable results."[34]
- In a report to the Department of Education (2007), independent researchers noted that among seniors who completed calculus only 68 percent scored at the Proficient level or better.[35]
- In addition, the 2007 report noted that eight years after high school graduation, 50 percent of those who scored at Basic on NAEP mathematics in twelfth grade had obtained a bachelor's degree.[36]
- The Buros Institute (2009) argued that NAEP lacked a "transparent, organized validity framework, beginning with a clear definition of the intended and unintended uses of the NAEP assessment scores. We recommend that NAGB continue to explore achievement level methodologies."[37]
- A Brookings Institution researcher (2016) recently echoed the 2007 concern about calculus students. Fully 30 percent of 12th-grade students who completed calculus were deemed to be below Proficient, a figure that jumped to 69 percent for pre-calculus students and 92 percent for students who completed trigonometry and Algebra II. These data "defy reason" and "refute common sense," he concluded.[38]
- A detailed study from the National Academy of Sciences, Engineering, and Medicine (2016) took note of the "controversy and disagreement" around the achievement levels, and concluded that considerable variability existed among cut-score judgments, including inconstancy around different item formats and different levels of difficulty.[39]
- The National Academy (2016) also pointed to several other challenges, including: final achievement-level descriptors were not those used to set the cut scores; interpretive guidance to understand the NAEP achievement levels is inconsistent and piecemeal, leading to possible misuse; and the current achievement-level descriptors do not provide clear, accurate, and specific information about what students know and can do at each achievement level (2016).[40]

That is not to say the benchmarks do not have their defenders. Phillips, Hambleton et al, ACHIEVE, The Fordham Institute and NAGB among others cite additional evidence, including internal validity studies, 12th-grade NAEP results connected to college success, and procedural integrity in the development of the benchmarks as justification for them.[41]

But the doubts of researchers such as Loveless (2016) and Pellegrino and his colleagues (2007) persist. Like Loveless, Pellegrino found the results "not believable," in large part because too few students were judged Proficient when compared to other indicators of advanced work, including participation in Calculus classes and Advanced Placement courses.

It is hard to avoid the conclusion that the challenge of clearing the proficiency bar is not simply a challenge for the United States. It is a global issue when the NAEP standard of Proficient defines the benchmark for student performance. In light of those findings and conclusions, several recommendations for improvement are outlined below.

RECOMMENDATIONS

At the outset, this report noted that the purpose is not to promote an anti-assessment agenda or oppose accountability and standards. It argued that bringing together the two strands of evidence about American school performance (NAEP benchmarks and international assessments) should shed some light on how valid, in the broadest sense, the American benchmark of Proficient is. Despite the questions that have been raised over the years about the misuse of international assessments, this study does not argue they are not useful or dispute the conclusion that students in some nations perform at higher levels than students in the United States. This study's central point is that the NAEP benchmark of Proficient establishes a standard that is unreasonable and defies common sense. Common Core "college ready" standards set close to NAEP's benchmark of proficiency will also frustrate students, both here and abroad.

The analysis in this report supports the conclusion that communities all over the world would face bleak headlines if their students sat down to take the NAEP or Common Core assessments. When citizens of the United States read that "only one-third" or "less than half" of the students in their local schools are proficient in mathematics, science, or reading, they can rest assured that the same judgments can be applied to students throughout most of the world. The fault lies not in the students. Not in the schools. Not in the Common Core. Nor even in the assessments themselves.

The fault lies in the peculiar definition of proficiency embedded in NAEP, an activity otherwise widely recognized as setting the standard for state-of-the-art assessment.*

It is time to say that no matter how well-meaning, advocates who push for school improvement justified by faulty data and benchmarks are not strengthening schools and building a better America. They are undermining education and weakening the United States.

Against that backdrop, we offer five recommendations to point the way ahead.

*NAEP markets itself as the "gold standard" of assessment. It is widely understood to be so. But the appellation applies to the technical quality of the assessment – its pioneering sampling standards, questionnaire development, quality control, and the like – not to its benchmarks. As is clear in the literature review contained in this report, controversy has dogged the benchmark-setting procedures from the time NAGB, the politically appointed policy-making board, established them in 1990 to the present. Claims, such as that recently made in *The Atlantic*, that the benchmarks themselves are accepted as a gold standard are mistaken, as even a casual review of NAEP's history reveals. (Mikhail Zinshteny, "How Much Tougher is Common Core?" *The Atlantic*, July 10, 2015.)

I. REDEFINE NAEP'S BASIC TERMINOLOGY

WE RECOMMEND that the National Assessment Governing Board rename the NAEP benchmarks as Low, Intermediate, High, and Advanced.

NAGB should examine its achievement levels once again. The misuse of the term "Proficient" has misled policymakers and the American public. There is no reason also not to revisit the standard-setting process itself. Adjusting the standards might complicate long-range trend analysis, but the wisdom of an ancient Turkish adage rings true: "No matter how far you have gone down the wrong road, turn back."

If it is essential to maintain the broad framework of the standards set years ago, a simple change in terminology can go a long way toward fixing the damage: simply rename the benchmarks to make them more similar to international benchmarks: Low, Intermediate, High, and Advanced. Such terminology eliminates the judgmental nature of "Below Basic" and "Proficient" in the current jargon. It also permits analysts to continue long-term trend analyses without interruption.

We believe there is also a lot to be said for (1) readjusting the NAEP scale scores from 0 – 500 to 0 – 1,000; and (2) setting the mean for each grade level at 500, instead of forcing three different grade levels into the narrow 0-500 scale. With respect to point (1): changing the NAEP scale to resemble those associated with international assessments simplifies matters in the public mind. With regard to (2): the public, and many advocates, are confused into believing that average results for white students in fourth grade are higher than those for African-American students in eighth grade, because results for both groups are reported on the same constricted 0-500 scale, with the expectation that the general public will understand the differences. Nothing could be further from the truth. Different scales for different grades send a confusing message to the general public and policymakers.

II. EMPHASIZE CAUTION IN INTERPRETING THESE BENCHMARKS

WE RECOMMEND that the U.S. Department of Education emphasize in every NAEP publication the Congressional insistence that NAEP benchmarks be understood as acceptable only on a "trial basis" and that the results based on the benchmarks should be interpreted "with caution."

Congress has insisted since 2001 that NAEP use its achievement benchmarks on a "trial basis," noting that they should be interpreted "with caution." While NAGB has followed the letter of the law in that regard, it has violated the spirit. The acknowledgment of Congress's insistence tends to be buried in the middle of NAGB's reports, often as a sentence added out of context at the end of paragraphs describing the assessment. When NAGB issues reports comparing state benchmarks unfavorably with NAEP's standard of proficiency, it moves far beyond understanding proficiency as a standard to be used on a "trial basis" and interpreted "with caution." It instead encourages states to adopt what many observers consider to be a highly questionable benchmark.

In recommending that the Congressional insistence be emphasized in every NAEP publication, we suggest that instead of hiding this information in the middle of reports, it be featured prominently in all NAEP publications as a one-page, stand-alone epigraph that cannot be overlooked. If our first recommendation (replace the term Proficient with the term High) is accepted, this warning is still required – because standard-setting process described in the body of this report remains so controversial.

III. EDUCATE THE PUBLIC ABOUT THE ASSESSMENT FINDINGS OUTLINED IN THIS REPORT

WE RECOMMEND that local school leaders – state chiefs, superintendents, board members, and teachers – vigilantly educate their local communities about the flaws embedded in the term Proficient and how school systems abroad would perform if held to that standard.

In the broadest terms the findings of this report support the conclusions that (1) the NAEP standard of Proficient is set unreasonably high; (2) state assessment benchmarks aligned with the NAEP proficiency standard are also set unreasonably high; and (3) the vast majority of students in most nations throughout the world cannot meet these unreasonable benchmarks.

Despite the publication and distribution of this report, it is highly likely that misrepresentations around the term "Proficient" will continue to be common and reported frequently. Local educators should not stand idly by. Through newsletters, PTA outreach, board meetings, regional and state gatherings, and opinion pieces in local newspapers and letters to the educator, they should maintain a consistent line of thinking emphasizing that:

- Proficient does not mean performance at grade level.
- Proficient does not mean what most people understand the term Proficient to mean.
- The procedures under which the term was developed have long been subjects of controversy.
- Congress has long held that the term should be used on a trial basis and interpreted with caution.
- Statistical tests reveal that the vast majority of students in almost all countries all over the world fail to meet the NAEP Proficient standard.
- Common Core "college-readiness" benchmarks aligned with the NAEP standard of Proficient should be treated with the greatest skepticism.

The Value of Assessment

The value of large-scale assessments (national or international) is that properly understood they provide a window into the world of schools and student performance.

There is another form of assessment that is not for accountability but for learning. These are assessments that are diagnostic in nature (helping us understand what individual students know and where they need to improve), formative (designed to let teaching staff know how well they are doing), and summative (providing year-end judgments about what students have learned). Each of these is valuable in its own way and in fact Common Core assessments such as PARCC and SBAC include such assessments *for* learning, as opposed to assessments *of* performance. In this respect, they promise to be helpful.

The point is that assessments of learning should not have high stakes attached to them and they should not so overwhelm the assessment agenda that local diagnostic assessments for learning are put at risk.

IV. REVISIT THE DECISION TO TIE STATE ASSESSMENTS' "COLLEGE READINESS" STANDARDS TO NAEP'S PROFICIENT (ADVANCED) BENCHMARK

WE RECOMMEND extreme caution before acting on the assumption that state agencies (or psychometricians) understand who is "college ready" and who is not, especially in determining whether students in Grades 4 and 8 are "on track" to be "college ready."

For decades, college admissions officials and psychometricians have understood that college entrance examinations and the high school record, in combination, are the best predictors of first-year grades in a four-year institution. They predict very little beyond that. Of the two, the high school record, reflecting four years of student effort, is the superior indicator of potential success in the first year. College

entrance examinations lag behind.

It is therefore a surprise to find policymakers and advocates (who should know better) and psychometricians (who do know better) united behind a belief that new Common Core assessments have predictive validity in determining students' readiness for college. The "college readiness" standard rests on a very flimsy reed – that students meeting the standard are unlikely to require enrollment in remedial courses in the first college year and can hope to attain a "B" in related mathematics or literature courses. True, there is a correlation but, as noted in the "College and Career Readiness" sidebar earlier in this report, analysts report that the correlations are, to put it as charitably as possible, only modest. It is estimated that they predict as little as .07 percent of first year college GPA (an English/Language Arts "college readiness" standard) and as much as 16 percent (a mathematics "college readiness" standard). In the best case scenario, that means that 84 percent of variance in first-year grades is unaccounted for; in the worst case, what accounts for 99.5 percent of first-year grades is a mystery.

The idea that psychometricians or state agency officials can accurately predict how individual students will perform in the future should be treated with the greatest suspicion. Even parents don't have that foresight.

V. DEVELOP A NATIONAL K-12 CAPACITY FOR ASSESSMENT ANALYSIS

WE RECOMMEND that the major national organizations representing a variety of K-12 constituencies develop significant capacity to analyze and comment on developments in national and international assessments.

With some notable exceptions,[42] the K-12 community as a whole has tended to remain silent in the face of official or other apparently authoritative pronouncements about the proficiency of American students or U.S. educational standing in the larger world. Although academics and independent analysts have questioned the definition of proficiency or over-reliance on international comparisons of student performance,[43] the K-12 community does not speak with a common voice on these critical issues.

This silence risks leaving the impression that the K-12 community accepts these reports, typically issued with extremely well-funded public relations campaigns, uncritically. But the reality is that, behind the scenes, leading school administrators, board members, principals, and teachers complain in quite sophisticated fashion about how these judgments are reached and disseminated without advance notice to the community or adequate opportunities to respond.

We suggest that instead of accepting these reports without comment, leading organizations in the K-12 community should attempt to speak with one voice on these issues. The opinions of individual organizations representing teachers, principals, superintendents, state and local board members and even parent-teacher organizations can be dismissed as self-serving. But in combination, representing as they do the professionals associated with educating more than 50 million students, they speak with an authority on behalf of students that no think tank, foundation, or policy analyst can claim.

To that end, we recommend the creation of a significant, independent, analytical body, funded jointly by leading K-12 associations, to produce at a minimum an annual report commenting on developments in educational assessment.

A LARGER PURPOSE

No one can doubt that among the central purposes of schooling in the modern world is the obligation of educators to produce graduates who are competent in reading, writing, and mathematics -- and prepared to earn a living. Assessment and accountability are critical components of delivering on that promise. But it is not too much to say that schools have been overwhelmed by a species of assessment imperialism in which what is tested becomes what is important. NAEP and international assessments have been co-opted in support of this conception of what schools are all about.

Education is about more than testing. It is about more than earning a living. It is about living a life. Students are not just standardized test results. When curriculum is forced into a straitjacket of what will be tested and the purposes of schools become constrained by the economic utility of their graduates, the larger purposes of education in a democracy are at risk.

In today's complex modern world, this nation needs graduates with a well-grounded knowledge of literature, history, and science. They need to be skilled problem-solvers and critical thinkers, with an appreciation for the arts and, ideally, some experience in developing their artistic talents. Clearly they need to be emotionally and physically healthy, as well as good citizens with a well-developed ethic for work and public service. Those have always been among the central purposes of public schools.

That is why, despite their superficial appeal, phrases such as "Proficient" and "career and college ready" do not capture the essential nature of the public good that is the public school. Put simply, the larger purpose of public schools is to produce public-minded citizens – whatever their political preferences – capable of functioning in and contributing to a democratic society. Citizens with a commitment to the welfare of their nation and to the future of the Republic. Recent evidence suggests that in pursuit of "college and career readiness" our schools are failing at this larger purpose.[44]

Economic competitiveness is important but potentially at risk is something much more significant: the ideal and the dream of America. That dream is made up of opportunity, community, and security. In pursuit of it, public education has been this nation's greatest strength and most powerful force. Despite the challenges facing public schools, we must not lose sight of their importance in creating the America we all know and love.

It is that America that is at risk. And it is the values embedded in that America that represent the real standards around which educators, citizens, and the assessment community should rally.

APPENDIX A: ACKNOWLEDGMENTS

The National Superintendents Roundtable and the Horace Mann League want to acknowledge the contributions of several people who were critical to the completion of this work.

Our first acknowledgment goes to James Harvey (National Superintendents Roundtable) and Emre Gönülates (Michigan State University and Teachers College, Columbia University) for developing the research and framing the argument in this document. We want also to thank the National Superintendents Roundtable for supporting Dr. Harvey's time to complete this project and for covering the expenses associated with Dr. Gönülates' invaluable contributions.

We thank the Roundtable's Steering Committee and members of the board of the Horace Mann League who endorsed this work from the outset.

Several individuals reviewed an earlier draft of this report and provided helpful guidance. We are extremely grateful to the following for taking the time to give us the benefit of their views, many of which are incorporated here:

- David Berliner, Regents Professor of Education Emeritus, Arizona State University, former president of the American Educational Research Association;
- Eva Baker, Founding Director of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles;
- Henry Braun, Boisi Professor of Education, Boston College, and Director of the Center for the Study of Testing, Evaluation, and Education Policy;
- Charles Fowler, Lead Schools, Inc., New Hampshire;
- Tom Loveless, Brown Center, Brookings Institution; and
- David Rutkowski, Center for Educational Measurement, University of Oslo.

We want to acknowledge also the early support and encouragement of this research by John Chattin-McNichols and John Jacob Gardiner of Seattle University, along with the late Dean of the Seattle University College of Education, Sue Schmitt. William Schmidt and Richard Houang of Michigan State University were also sources of inspiration.

Finally, we thank Rhenda Meiser, of Meiser Communications, Kai Hiatt of the National Superintendents Roundtable, and Anne Paxton from ProForum for their careful reading of the manuscript. Kathy Mathes of Mathes Design designed the report. Ms. Meiser counseled on the release and public announcement of the report. We are deeply in their debt.

APPENDIX B: STATISTICAL MODERATION

In statistical moderation, the aim is to put different scales from two different assessments onto the same scale. We can map NAEP scores on PIRLS scale using the following equations:

B.1: $PIRLS_{level} = \hat{A} + \hat{B} \times NAEP_{level}$

where

B.2: $\hat{A} = \hat{\mu}_{PIRLS} - \hat{B}\hat{\mu}_{NAEP}$

$$\hat{B} = \frac{\hat{\sigma}_{PIRLS}}{\hat{\sigma}_{NAEP}}$$

In Equation B.1, $PIRLS_{level}$ is the estimated PIRLS score that is associated with the corresponding NAEP level. $\hat{A}$ and $\hat{B}$ are the estimated intercept and slope parameters, respectively, of the line that maps the NAEP scale onto PIRLS.

Equation B.2 shows the equations to calculate these $\hat{A}$ and $\hat{B}$ values. $\hat{\mu}_{PIRLS}$ and $\hat{\mu}_{NAEP}$ are the estimated national means for PIRLS and NAEP scores of the students in U.S, respectively. $\hat{\sigma}_{PIRLS}$ and $\hat{\sigma}_{NAEP}$ are the estimated national standard deviations for PIRLS and NAEP scores of the students in U.S, respectively.

Example Calculation

In both PIRLS and NAEP data, the calculation of these estimates are done for each plausible value and then averaged. For simplicity's sake, results for only one of the plausible values will be shown here. Complete calculations can be seen in Phillips (2014).

For plausible value 4:

$\hat{\mu}_{PIRLS} = 556.36$
$\hat{\mu}_{NAEP} = 220.07$
$\hat{\sigma}_{PIRLS} = 73.62$
$\hat{\sigma}_{NAEP} = 36.05$

Consequently, we can calculate $\hat{B}$ as: $\hat{B} = \frac{\hat{\sigma}_{PIRLS}}{\hat{\sigma}_{NAEP}} = \frac{73.62}{36.05} = 2.042$

With the result for in hand, we can calculate $\hat{A}$ as: $\hat{A} = 556.36 - (2.042 \times 220.07) = 106.977$

For plausible value 4, the equation B.1 becomes: $PIRLS_{level} = 106.977 + (2.042 \times NAEP_{level})$

We can use the equation above to calculate PIRLS equivalent of each NAEP level. For example, the NAEP reading achievement level for Advanced is 268. This corresponds to a PIRLS score of 654:

$$PIRLS_{Advanced} = 106.977 + (2.042 \times 268) = 654.23$$

The same logic can be applied for Proficient (238) and Basic (208):

$$PIRLS_{Proficient} = 106.977 + (2.042 \times 238) = 592.97$$
$$PIRLS_{Basic} = 106.977 + (2.042 \times 208) = 531.71$$

APPENDIX C: APPLYING NAEP BENCHMARKS TO PIRLS RESULTS

Emre Gönülates
James Harvey

April 29, 2017

This report aims to compare 4th-grade reading scores of the countries and jurisdictions that participated in the Progress in International Reading Literacy Study (PIRLS) 2011 against the performance benchmarks for the 4th-grade National Assessment of Educational Progress (NAEP) reading test. These benchmarks were defined by the National Assessment Governing Board, which establishes NAEP policy.

Calculation of comparable scores between these tests requires finding a linking function between NAEP and PIRLS 4th-grade reading assessments. Both of these tests aim to measure the same construct, i.e. reading ability of a 4th grader. If one desires to use the scores of NAEP and PIRLS tests interchangeably (as in comparing SAT scores obtained in May and October administrations), these tests should be equated. But equating requires satisfying stringent conditions (Holland, 2007). The differences between the two large-scale assessments of interest here makes strict equating impossible. The content of the tests is slightly different (Binkley & Kelly, 2003). In addition, the tests were constructed using different specifications, such as test length, item format, and length of reading passages.

However, although strict equating is not possible, the technique of "statistical moderation" can be used to link tests such as these (Linn, 1993). A linking equation* that finds the PIRLS scale equivalents of NAEP benchmark scores can be built using the score distributions of the examinees from each test. The resulting PIRLS scores that are equivalent to NAEP contain some error and the error margins have to be estimated for each point estimate. Two sources of error need to be taken into account: sampling and measurement error. Phillips (2014) used the statistical moderation approach to link PIRLS 2011 and the NAEP 4th-grade reading assessment in the same year. He reported that about 94% of the linking error was due to sampling and 6% was due to measurement error.

In this study, we used the PIRLS scale equivalents of NAEP benchmark scores calculated by Phillips (2014). Table 1 shows the 4th-grade 2011 PIRLS scale equivalents of 4th grade NAEP benchmark scores.

We then compared the score distributions of countries that participated in PIRLS 2011 to the PIRLS-equivalent scores in Table 1. For each participating country, the percentage of students at or above NAEP benchmarks were calculated by means of an equipercentile procedure. Each calculated percentage contains some margin of error due to sampling and measurement error. For each percentage, an accompanying standard error that reflects the

TABLE 1: Fourth Grade 2011 PIRLS Scale Equivalents of Fourth Grade NAEP Benchmark Scores

	NAEP Reading Benchmark Score	PIRLS Equivalent Score	Standard Error of PIRLS Equivalent Score
Basic	208	532	1.8
Proficient	238	593	1.9
Advanced	268	654	2.2

Source: Phillips (2014, p.13)

* The linking equation is a simple linear equation such as $y = A + B \cdot x$, where x represents the NAEP score and y represents the equivalent PIRLS score. Phillips (2014) reported that the equation that links 2011 NAEP to 2011 PIRLS is $y = 108.2 + 2.04 \cdot x$.

sampling and measurement error was calculated. These standard errors reflect only the errors in PIRLS. As Table 1 shows, there is also an error due to projecting NAEP benchmarks on to the PIRLS scale. The standard errors below do not reflect this linking error. A separate analysis showed that the linking error can increase the standard errors reported below up to 1.29 percentage points for English-speaking jurisdictions.

Results for English-Speaking Jurisdictions

Anticipating that international comparisons of reading ability might differ depending on whether English was the official language of various nations, we first calculated the proportion of students in English-speaking jurisdictions whose PIRLS results placed them within the various NAEP benchmarks (Table 2). For example, according to the fifth column in Table 2, an estimated 8% of the students in the United States who took PIRLS scored at or above 654 (the PIRLS-equivalent score to NAEP's Advanced benchmark).

We also calculated the percentage of students in English-speaking jurisdictions at or above each of NAEP's benchmarks (Basic, Proficient, and Advanced), and created "whisker bars" to represent the standard error associated with each of the estimates. These standard errors take into account complex sample design issues and measurement error, including that associated with calculating plausible values.

Figure 1 and Table 2 confirm each other. Students in England, Northern Ireland, and the United States clearly outperform students in the other six English-speaking nations. While their performance differs modestly in rank order, the proportions are well within the margin of error of each other.

Figure 2 arrays the information for English-speaking jurisdictions in a different light. It displays the score range of students between the 5th and the 95th percentile for each country. The points in the middle represent the median score of each jurisdiction. For example, for the United States, 90% of student scores were between 427.77 (5th percentile) and 670.6 (95th percentile). The dashed colored lines represent the mapped NAEP benchmarks. We can say that in Australia, more than 95% of 4th graders assessed in PIRLS would test below Advanced, according to NAEP's benchmarks. Trinidad and Tobago and South Africa present an even more difficult policy challenge with regard to 4th-grade reading.

We also calculated separately the range of students in English-speaking jurisdictions scoring between the first and the 99th percentile. For analytical purposes, it adds modestly to what can be gleaned from Figure 2, but for the purposes of this analysis it is put aside. The calculation is available on request.

Results for All Jurisdictions

At first blush, it seems that none of the English-speaking jurisdictions can demonstrate that a majority of their students meet the NAEP Proficient benchmark for 4th-grade reading when their scores on PIRLS are linked to NAEP. We now consider all jurisdictions, including those that speak English as the official language and those that do not.

Table 3 displays the percentages of students, by

TABLE 2: Percentages of Students in PIRLS at or Above NAEP Proficiency Benchmarks for English Speaking Countries

Country	Below Basic	Basic	Proficient	Advanced
Australia	49%	30%	17%	4%
Canada	39%	35%	21%	6%
Ireland	37%	34%	22%	7%
New Zealand	47%	28%	18%	7%
South Africa	82%	11%	5%	2%
Trinidad and Tobago	74%	18%	7%	1%
United States	35%	33%	23%	8%
England	38%	31%	23%	9%
Northern Ireland	33%	33%	25%	9%

FIGURE 1: Percentage of Students in PIRLS at or Above Proficient in English Speaking Countries

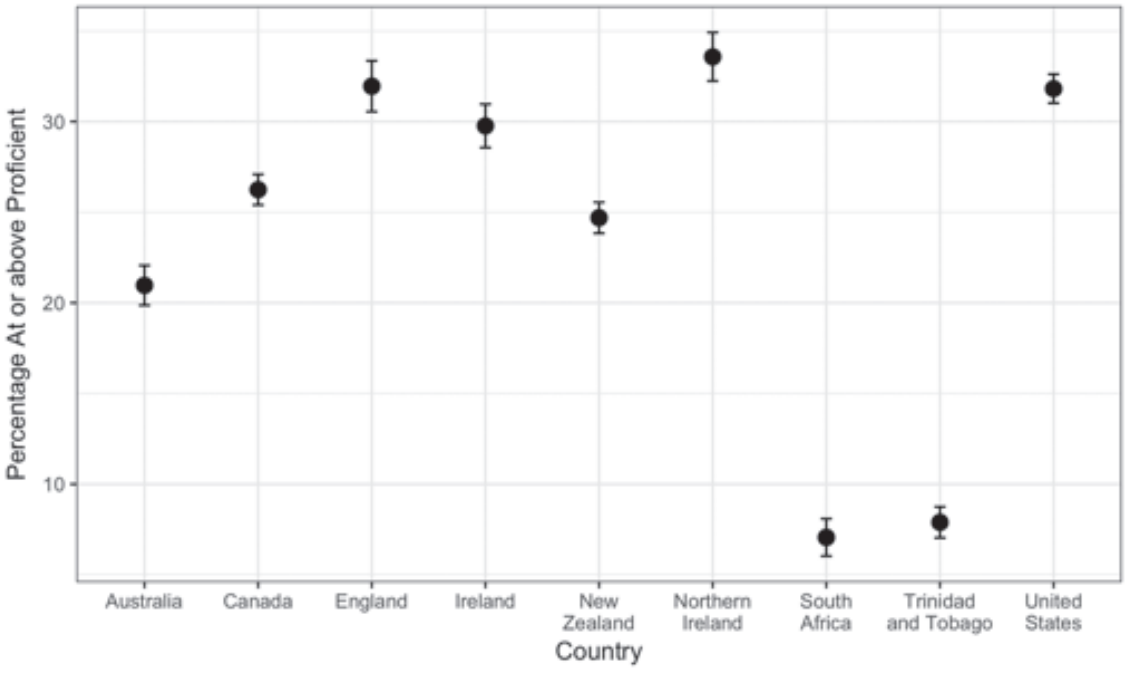


FIGURE 2: Score Range of Students between 5th Percentile and 95th Percentile in English Speaking Countries

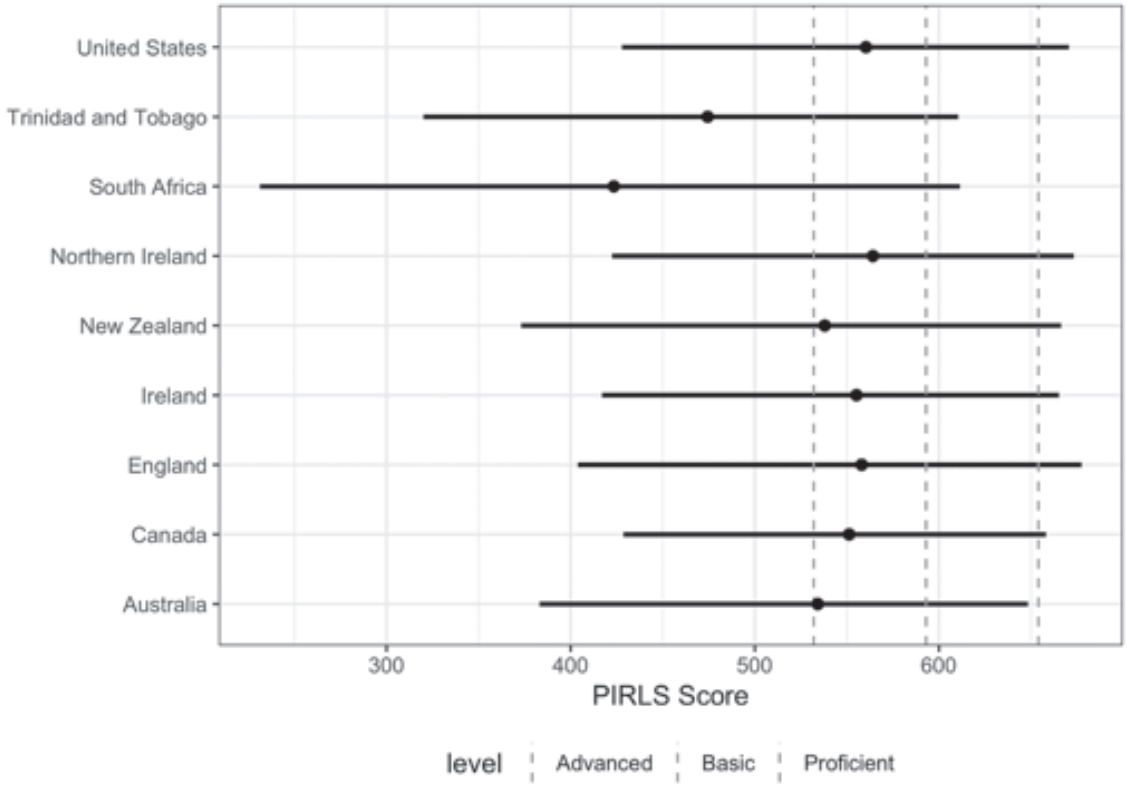


TABLE 3: Percentages of Students in PIRLS at or Above NAEP Proficiency Benchmarks for All Participating Countries

Country	Below Basic	Basic	Proficient	Advanced
Azerbaijan	85.57%	12.49%	1.80%	0.14%
Australia	48.80%	30.23%	16.76%	4.21%
Austria	49.51%	35.49%	13.40%	1.60%
Botswana	87.80%	8.61%	3.17%	0.42%
Bulgaria	45.51%	31.28%	18.46%	4.75%
Canada	38.99%	34.76%	20.53%	5.72%
Chinese Taipei	34.64%	36.76%	23.72%	4.88%
Colombia	85.53%	11.53%	2.62%	0.32%
Croatia	34.54%	39.78%	21.69%	3.99%
Czech Republic	37.80%	40.79%	18.60%	2.81%
Denmark	34.31%	37.10%	23.86%	4.73%
Finland	27.02%	37.15%	27.98%	7.85%
France	54.59%	31.49%	12.12%	1.80%
Georgia	70.29%	22.78%	6.19%	0.74%
Germany	43.13%	35.22%	17.90%	3.75%
Honduras	84.99%	11.83%	2.81%	0.37%
Hong Kong SAR	23.80%	38.70%	30.72%	6.78%
Hungary	43.03%	31.41%	20.34%	5.22%
Indonesia	92.41%	6.72%	0.80%	0.07%
Iran, Islamic Rep. of	80.58%	15.34%	3.80%	0.28%
Ireland	36.63%	33.60%	22.49%	7.28%
Israel	41.62%	30.23%	20.39%	7.76%
Italy	43.04%	35.18%	18.13%	3.65%
Kuwait	84.46%	11.47%	3.29%	0.78%
Lithuania	49.96%	34.37%	13.70%	1.97%
Malta	68.94%	20.86%	8.42%	1.78%
Morocco	98.36%	1.51%	0.13%	0.00%
Oman	92.71%	5.98%	1.19%	0.12%
Netherlands	38.90%	41.58%	17.74%	1.78%
New Zealand	47.30%	28.00%	17.62%	7.08%
Norway	63.77%	29.48%	6.25%	0.50%
Poland	50.82%	31.84%	14.45%	2.89%
Portugal	41.82%	36.43%	18.69%	3.06%
Qatar	83.88%	11.53%	3.64%	0.95%
Romania	59.00%	26.24%	12.08%	2.68%
Russian Federation	27.59%	35.73%	27.88%	8.80%
Saudi Arabia	87.70%	10.24%	1.81%	0.25%
Singapore	30.06%	30.24%	26.42%	13.28%
Slovak Republic	44.13%	36.47%	16.66%	2.74%
Slovenia	48.09%	33.64%	15.58%	2.69%
South Africa	82.21%	10.73%	4.99%	2.07%
Spain	58.97%	29.75%	10.09%	1.19%
Sweden	41.77%	36.97%	17.94%	3.32%
Trinidad and Tobago	74.14%	17.97%	6.65%	1.24%
United Arab Emirates	81.02%	13.04%	4.71%	1.23%
United States	34.76%	33.42%	23.49%	8.33%
England	37.51%	30.53%	22.62%	9.34%
Northern Ireland	33.06%	33.36%	24.62%	8.96%
Belgium (French)	64.52%	27.30%	7.57%	0.61%
Morocco 6	88.61%	9.49%	1.72%	0.18%
Dubai	67.01%	20.31%	9.77%	2.91%
Abu Dhabi, UAE	86.15%	10.13%	2.86%	0.86%
Canada, Ontario	36.64%	33.52%	23.03%	6.81%
Canada, Quebec	45.28%	36.70%	15.51%	2.51%
Canada, Alberta	38.76%	34.54%	21.04%	5.66%
Maltese-Malta	79.20%	16.35%	4.01%	0.44%
Andalusia, Spain	58.66%	29.97%	10.18%	1.19%

jurisdiction, whose PIRLS results place them within the corresponding NAEP benchmarks. The results are presented alphabetically to help guard against the natural human tendency to impose a rank order by achievement level in an effort to read more into the results than they warrant. It is, however, just a simple matter of combining the percentages of students who may be deemed to be Proficient and Advanced, by the NAEP metrics, to obtain an estimate of what proportion of students in each jurisdiction would be considered to be at or above the NAEP Proficient benchmark in 4th-grade reading, as indicated by their performance on the PIRLS 4th-grade reading assessment.

Across all 57 jurisdictions, fewer than fifty percent of the students assessed in 4th-grade reading would be deemed to be Proficient when judged by the NAEP 4th-grade standard.

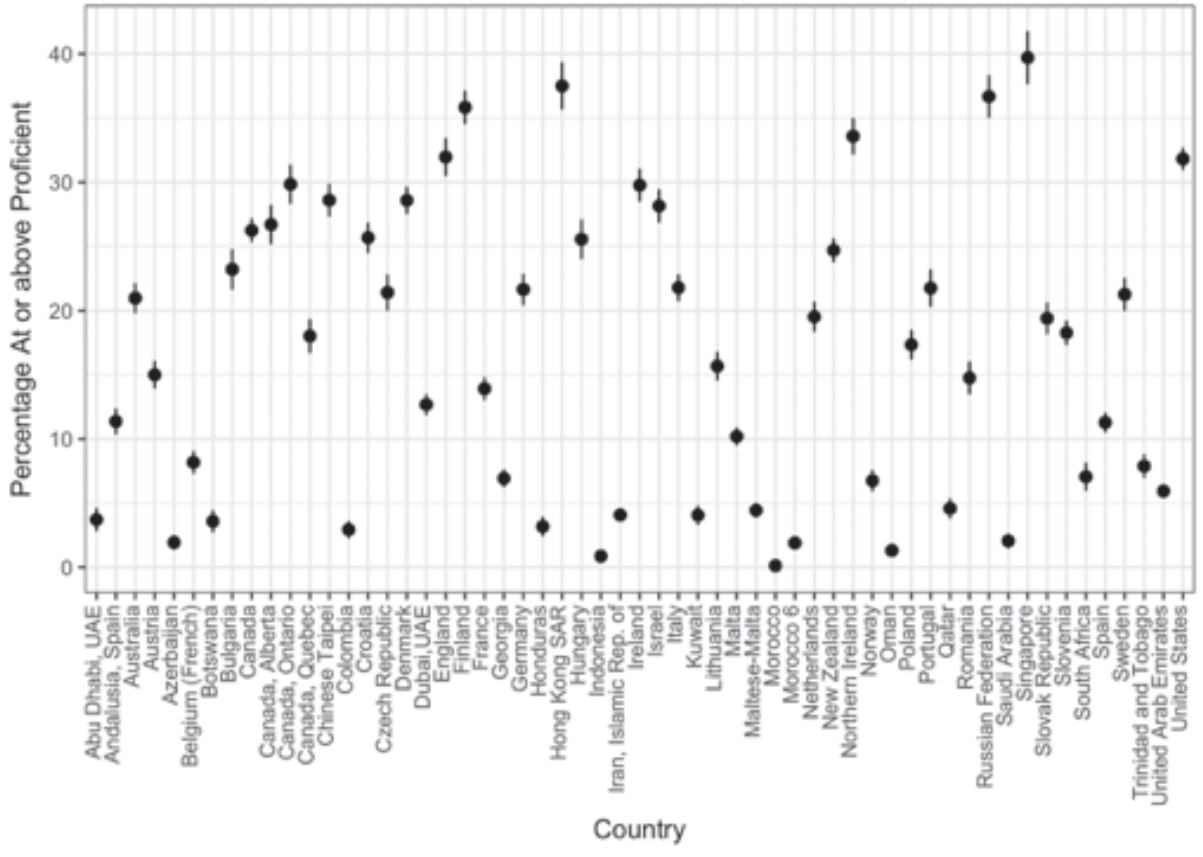
Table 3 shows the percentages of students in PIRLS who were within the corresponding NAEP benchmarks for all participating jurisdictions.

As with the English-speaking nations, we also calculated the percentage of students in all jurisdictions at or above each of NAEP's benchmarks (Basic, Proficient, and Advanced), and created "whisker bars" to represent the standard error associated with each of the estimates. Figure 3 presents the results for the proportion of students at or above the NAEP benchmark of Proficient. (As with the analysis of English-speaking nations, we present only the results for the Proficient analysis here. The results for Basic and Advanced are available on request.)

It seems clear from Figure 3 that although none of the 57 jurisdictions that participated in PIRLS 2011 can demonstrate that a majority of their students would meet the NAEP benchmark of Proficient, the results indicate that students in four jurisdictions – Singapore, the Russian Federation, Hong Kong, and Finland – outperform students in the United States on this metric. By rank order, Northern Ireland and England might be added to this list of four high performers, but, as noted earlier, their results are well within the margin of error.

Meanwhile, Figure 4 (next page) displays the score range for students in all jurisdictions between the 5th and 95th percentile. (As with the English-speaking

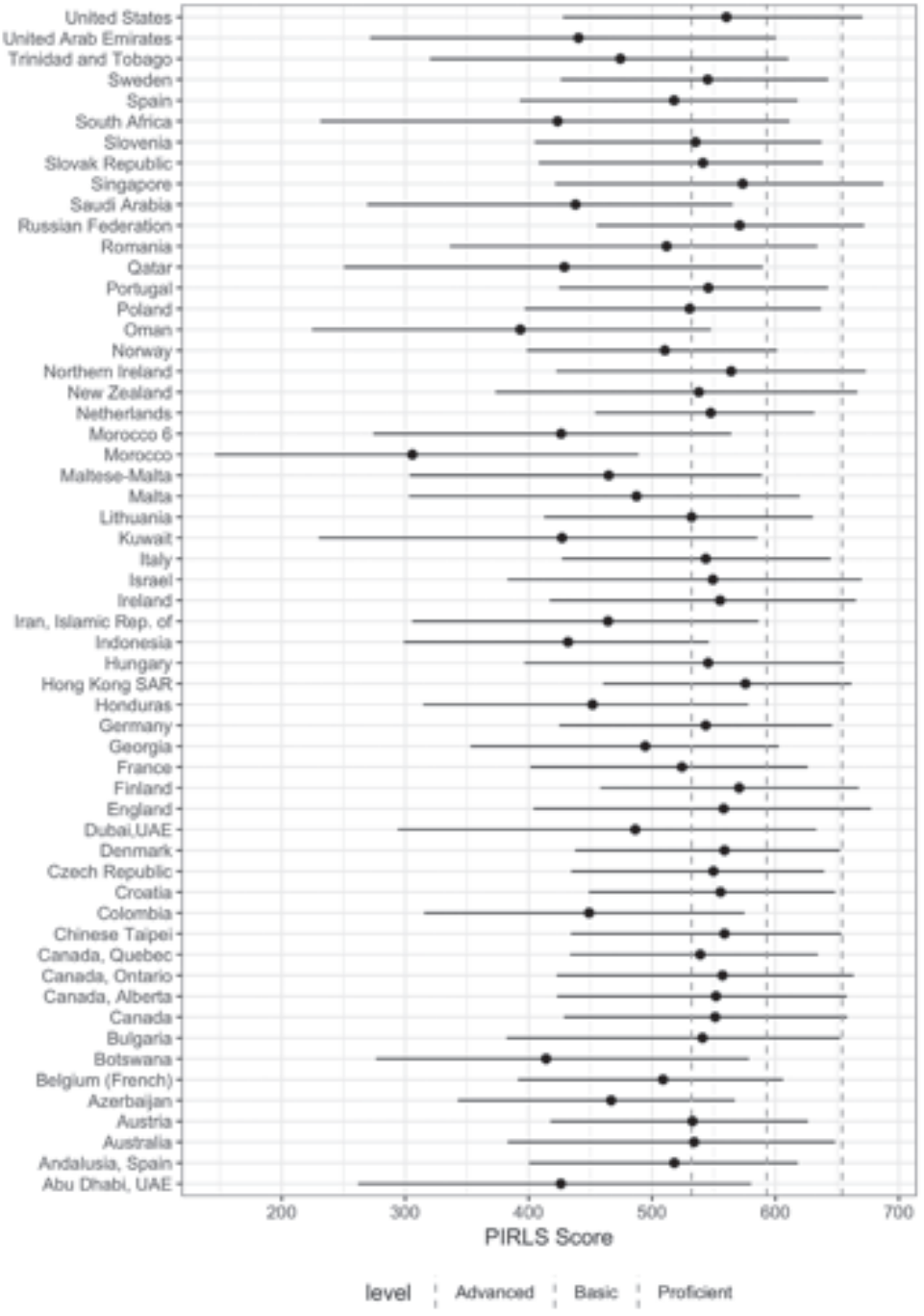
FIGURE 3: Percentages of Students at or above Proficient in All Participating Countries



nations, we also calculated the range of students between the first and 99th percentile and can provide those results on request.)

Once again, the point on each line represents the median, with the colored dotted lines indicating the relevant NAEP benchmarks or cut scores for each Basic, Proficient, and Advanced. We can say that students in a number of jurisdictions (the United States, Singapore, Russian Federation, Northern Ireland, Israel, Ireland, Finland, and England) are able to achieve at the NAEP Advanced level in reading at the 4th-grade level. It is also transparently clear that 4th-grade reading performance throughout much of the Middle East (UAR, Saudi Arabia, Qatar, Oman, Morocco, Iran, and Dubai) is disappointing in the extreme.

FIGURE 4: Score Range of Students between 5th Percentile and 95th Percentile for All Participating Countries



REFERENCES

Binkley, M. & Kelly, D. L. (2003). *A content comparison of the NAEP and PIRLS fourth-grade reading assessments* (No. NCES 200310). National Center for Education Statistics. Washington, DC.

Holland, P. (2007). A framework and history for score linking. In N. Dorans, M. Pommerich, & P. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30). New York, NY: Springer Verlag.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6(1), 83–102. doi:10.1207/s15324818ame0601 5

Phillips, G. W. (2014). *Linking the 2011 national assessment of educational progress (NAEP) in reading to the 2011 progress in international reading literacy study (PIRLS).* NAEP Validity Studies Panel. San Mateo, CA.

ENDNOTES

[1] Sub-jurisdictions eliminated from the 57 PIRLS participants: Abu Dhabi, UAE; Andalusia, Spain; Belgium (French); Canadian provinces of Alberta, Ontario, and Quebec; Dubai, UAE; Florida, United States; Hong Kong, China; Taipei, Taiwan; and Northern Ireland in the United Kingdom.

Sub-jurisdictions eliminated from the 38 TIMSS participants: Belgium (Flemish), Taipei, and Hong Kong.

[2] See for example, Hambleton, R.K., Sireci, S. G., and Smith, Z.R. (2009). "How Do Other Countries Measure Up to the Mathematics Achievement Levels on the National Assessment of Educational Progress?" *Applied Measurement in Education,* 22 (4), 376-393; Johnson, E., Cohen, J., Chen, W. H., Jiang, T., & Zhang, Y. (2005). *2000 NAEP – 1999 TIMSS Linking Report.* Washington, DC: National Center for Education Statistics. (NCES 2005–01); Lim H. and S. G. Sireci, "Linking TIMSS and NAEP Assessments to Evaluate International Trends in Achievement." *Education Policy Archives,* February 13, 2017 (Vol. 25, No. 11); and Phillips, G. W. (2007). *Expressing international education achievement in terms of U.S. performance standards: Linking NAEP achievement levels to TIMSS.* Washington, DC: American Institutes for Research.

[3] Phillips, G. W. (2007). *Expressing international education achievement in terms of U.S. performance standards: Linking NAEP achievement levels to TIMSS.* Washington, DC: American Institutes for Research.

[4] Statistical moderation has a significant history. Thissen in 2007 noted that Mislevy referred to this kind of statistical manipulation as statistical moderation as early as 1992. Holland in 2007 referred to it as anchor scaling. (See: Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales.* New York: Springer-Verlag. Meanwhile, Johnson and his colleagues developed the formulas adopted by Phillips in 2007. (See: Johnson, E., Cohen, J., Chen, W. H., Jiang, T., & Zhang, Y. (2005). *2000 NAEP – 1999 TIMSS Linking Report.* Washington, DC: U.S. Department of Education, Institute for Education Sciences, National Center for Education Statistics. (NCES 2005–01).

[5] Lim, H. & S. G. Sireci, 2017.

[6] See Johnson et al, 2005.

[7] Phillips, G. W. (2014). *Linking the 2011 National Assessment of Educational Progress (NAEP) in Reading to the 2011 Progress in International Literacy Study (PIRLS).* Washington, DC: American Institutes for Research.

[8] 193 members of the United Nations and two non-members holding observer status – The Holy See (the Vatican) and the State of Palestine.

[9] Laurie Burkitt, "China to Abandon One-Child Policy," *Wall Street Journal,* October 30, 2015.

[10] Lim H. and S. G. Sireci, "Linking TIMSS and NAEP Assessments to Evaluate International Trends in Achievement." *Education Policy Archives,* February 13, 2017 (Vol. 25, No. 11)

[11] See for example, *Mapping State Proficiency Standards on to NAEP Scales.* Washington: National Center for Education Statistics, 2015. (NCES 2015-046)

[12] Cronin, J., Dahlin, M., Adkins, D. and Kingsbury, G.G. (2007). *The Proficiency Illusion.* Washington: The Fordham Foundation.

[13] Denise Smith Amos, "Business leaders concerned over new Florida school grade formula," Thinkstock, January 8, 2016.

[14] News Release, "State Results Issued for Badger Exam," Wisconsin Department of Public Instruction, January 13, 2016. (DPI-NR 2016-02B)

[15] News Release, "State schools chief Torlaskon calls first year of CAASP results California's starting point toward goal of career and college readiness," California Department of Education, September 9, 2015. (Release # 15-69)

[16] Gary W. Phillips. (2016). *National Benchmarks for State Achievement Standards.* Washington: American Institutes for Research.

[17] Among Phillips' qualifiers: "The estimates provided . . . should be considered rough, ballpark estimates and should be used only for broad policy understandings. . . In the United States, students took both NAEP and TIMSS; in all other countries…students only took TIMSS. Whether the linking parameters are stable in other countries is an empirical question [which] no international linking study has been designed to answer…

[T]he percentage at or above basic, proficient, and advanced levels in the tables is based on the assumption of a "normal distribution" of performance within each country…

[T]he linking parameters are assumed to be stable across years. …Finally, the achievement levels developed for the NAEP were based on the content of the NAEP.…At best, these concordance tables should be used for rough approximations and should not be used for less granular inferences." (2007).

Phillips repeated many of these caveats in his 2014 analysis of PIRLS and added notes to the effect that the NAEP assessment was administered between January and March, although PIRLS was administered from April to June, and that PIRLS excluded 7.2 percent of the population, nearly twice the NAEP exclusion rate of 4 percent. He also acknowledged that NCES felt the validity evidence was insufficient to use PIRLS results for one state to predict state PIRLS results for all 50. In the 2016 report on PARCC and SBAC, he noted that the comparison was between a NAEP assessment of reading with Common Core assessments that include writing.

[18] National Commission on NAEP 12th-Grade Assessment and Reporting. (2004). *12th-Grade Student Achievement in America: A New Vision for NAEP.* Washington: National Commission.

[19] American Diploma Project (2004). *Ready or Not: Creating a High School Diploma that Counts.* Washington: Achieve, Inc.

[20] See Achieve, Inc. (2005). *Recommendations to the National Assessment Governing Board on Aligning 12th Grade NAEP with College and Workplace Expectations:* Reading; and (2006) Mathematics.

[21] Figures on college remediation rates are notoriously difficult to tie down. They include rates at public and private four- and two-year colleges, including open-door community colleges, which admit applicants regardless of possession of a high school diploma. Education Reform Now provided a valuable service in 2016 by analyzing remediation data solely for recent high school graduates. ERN estimated that 25 percent of first-year students require remediation, still a very high figure, with more than 50 percent of that total enrolled in second-chance community colleges. See: Mary N. Barry and Michael Dannenberg, Out of Pocket: The High Cost of Inadequate High Schools. Washington: Education Reform Now, April 2016.

[22] Nichols-Barrer, I. and Brian Gill (2016). "Testing College Readiness: Massachusetts Compares the Validity of Two Standardized Tests." *Education Next* (Vol. 16, No. 3).

[23] Shaw, E.S. (2015). A SAT Validity Primer. New York: College Board.

[24] William J. Mathis (2016). "Alice in PARCCland: Does 'Validity Study' Really Prove the Common Core Test is Valid?" Washington Post, June 1, 2016.

[25] See, for example, OECD, *Lessons from PISA for the United States.* Paris: OECD (2011); Sclafani, Susan, and E. Lim, *Rethinking Human Capital in Education: Singapore as a Model for Teacher Development,* Washington: The Aspen Institute (2008); and Tucker, Marc S., *Standing on the Shoulders of Giants: An American Agenda for Education Reform.* Washington: National Center on Education and the Economy (2011);

[26] See, for example, a comment by Campbell Brown, May 16, 2016: "Two out of three eighth graders in this country cannot read or do math at grade level." Ms. Brown, a former CNN anchor, co-founded *The 74*, an education news site. The only conceivable source of Ms. Brown's claim is NAEP's eighth-grade mathematics assessment. Available: "Campbell Brown's Advice for the Next President," *Slate*, May 16, 2016. Downloaded February 25, 2017 at: http://tinyurl.com/hzpa9wk

[27] Loomis, S.C, and Bourque, M.L, eds. (2001). *National Assessment of Educational Progress Achievement Levels, 1992-1998 for Reading.* Washington: National Assessment Governing Board.

[28] Binkley, M. (2003). *A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments.* Washington: National Center for Education Statistics.

[29] 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

[30] Vinovskis, M.A., (1998).

[31] Chester E. Finn, Jr. (2004). "An Interview with Chester E. Finn, Jr.," in Lyle V. Jones and Ingram Olkins (eds.), The Nation's Report Card. Evolution and Perspectives. Bloomington, Indiana: Phi Delta Kappa Educational Foundation.

[32] Rothstein, R., and R. Jacobsen and T. Wilder. (2008). *Trading Education: Getting Accountability Right.* New York: Teachers College Press.

[33] U.S. General Accounting Office. (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations.* GAO/PEMD–993–12. Washington, DC: Author. Retrieved from ERIC database. (ED359268).

[34] Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the nation's report card.* Washington, DC: National Academy Press.

[35] Scott, L. A., Ingels, S. J., and 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). Retrieved from ERIC database. (ED498359).

[36] Scott, Ingels, and Owings, 2007.

[37] Buckendahl, C.W. et al, (2009). *Evaluation of the National Assessment of Educational Progress.* Lincoln, NE: Buros Center for Testing, University of Nebraska.

[38] Loveless, T., "The NAEP Proficiency Myth," (2016). Brookings Institution: Brown Center Chalkboard, June 13, 2016.

[39] Edley, C., Jr. and J. A. Koenig (eds). 2017. *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress.* Washington: National Academies Press.

[40] Edley and Koenig, 2016.

[41] See for example: Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W. J., & Zwick, R. (2000). A response to "setting reasonable and useful performance standards" in the National Academy of Sciences "Grading the Nation's Report Card." *Educational Measurement: Issues and Practice, 19*(2), 5–14.

[42] See for example: Riddile, M. (2014). "PISA: It's Still 'Poverty not Stupid'," blog published by the National Association of Secondary School Principals, February 12, 2014 at http://tinyurl.com/h84456b; and also: *School Performance in Context* (2015). Seattle, Washington: Horace Mann League and National Superintendents Roundtable; and statements from the Learning First Alliance about the need for a moratorium on new high-stakes tests (June 2013) and the importance of understanding international assessment results in context (December 2013).

[43] See for example, the work of several researchers at the National Education Policy Center, University of Colorado at Boulder; Tom Loveless, Brown Center, Brookings Institution, and David Rutkowski and Leslie Rutkowski, Center for Educational Management, the University of Oslo, Norway.

[44] Kahlenberg, R. D and C. Janey. (2016. *Putting Democracy Back into Public Education.* Washington: The Century Foundation.

The Horace Mann League

FOR MORE INFORMATION:

National Superintendents Roundtable
9425 35th Avenue NE, Suite E | Seattle, WA 98115 | 206-526-5336
www.superintendentsforum.org | jamesharvey@superintendentsforum.org