AIR®

AMERICAN INSTITUTES FOR RESEARCH®

INTERNATIONAL BENCHMARKING

State and National Education Performance Standards

September 2014

Gary W. Phillips
Vice President and Institute Fellow

Contents

Executive Summary

State performance standards represent how much the state expects the student to learn in order to be considered proficient in reading, mathematics, and science. In the past, these performance standards have been used by each state to report adequate yearly progress (AYP) under No Child Left Behind federal legislation, and they are currently being used for federal reporting under the Department of Education's flexibility waivers. These standards are also used by the state to monitor progress from year to year, and to report on the success of each classroom, school, and district to parents and the public.

This report uses international benchmarking as a common metric to examine and compare what students are expected to learn in some states with what students are expected to learn in other states.[1] The performance standards in each state were compared with the international benchmarks used in two international assessments, and it was assumed that each state's expectations are embodied in the stringency of the performance standards (also called achievement standards) it uses on its own state accountability tests. The international assessments were the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). The data were obtained through a statistical linking study tying the National Assessment of Educational Progress (NAEP) to TIMSS and PIRLS (see Appendix A). The international benchmarking not only provided a mechanism for calibrating the difficulty and gauging the global competitiveness of each state standard but also yielded an international common metric with which to compare state expectations.

The overall finding in the study is that there is considerable variance in state performance standards, exposing a large gap in expectations between the states with the highest standards and the states with the lowest standards. Although this gap in expectations is large, many policymakers may not be aware of just how large it is. In general,

- The difference between the standards in the states with the highest standards and the states with the lowest standards is about 2 standard deviations. In many testing programs, a gap this large represents three to four grade levels.
- This "expectations gap" is so large that it is more than twice the size of the national black–white *achievement gap*. Closing the achievement gap is important, but so is closing the larger expectation gap. Reducing the expectation gap will require consistently high expectations from *all* states.

[1] This report is a follow-up to a previous AIR report in which 2007 NAEP was linked to 2007 TIMSS (Phillips, 2010). The data in the current report link 2011 NAEP to 2011 TIMSS. Subsequent to 2011, some states may have raised performance standards and some may have lowered them. For example, since 2011, Kentucky, New York, Utah and Wisconsin have substantially raised their performance standards, to a level that is consistent with a "B" in this report.

The 2011 percent proficient for each state test was obtained from Ed*Facts*, U.S. Department of Education, http://www2.ed.gov/admins/lead/account/consolidated/index.html. The state NAEP Coordinators in 22 states were contacted to confirm or correct state results reported in Ed*Facts*. Some states were excluded in some tables because AIR was unable to reliably confirm the state's percent proficient. The author would like to thank Jonathan Phelan and Steven Hummel at AIR for conducting this review of the data.

- The report also found that success under No Child Left Behind is largely related to using low performance standards. The states reporting the highest numbers of proficient students have the lowest performance standards. More than two-thirds of the variation in state success reported by No Child Left Behind is related to how high or low the states set their performance standards.

These results help explain why the United States does poorly in international comparisons. Many states think they have high standards and are doing well, and feel no urgency to improve because almost all their students are proficient.

The report estimated how the 2011 state results reported to No Child Left Behind would have looked had all the states used a common metric. When the data were reanalyzed using a common metric, higher achievement was correlated with a higher performance standard. With a different metric used in each state, as encouraged by NCLB, higher achievement is obtained by setting low standards. When a common metric is used in each state, such as the state NAEP assessment, higher achievement is associated with setting higher standards.

The data show that the No Child Left Behind paradigm of encouraging each state to establish its own unique performance standard is fundamentally flawed and misleading. The big policy problem associated with the current No Child Left Behind state testing paradigm is that the lack of a common metric results in a lack of transparency. Because test results across the 50 states are not comparable, any inference about national progress is impossible; we cannot even determine if progress in one state is greater than progress in another state. Clearly, 50 states going in 50 different directions cannot lead to national success that is globally competitive. Transparency in measurement (through use of a common metric) is the most fundamental requirement for scientific measurement and the first step in determining if our educational programs are succeeding. The lack of transparency among state performance standards leads to a kind of policy jabberwocky: the word *proficiency* means whatever one wants it to mean. This misleads the public, because low standards can be used to artificially rack up high numbers of "proficient" students. This looks good for federal reporting requirements, but it denies students the opportunity to learn college and career readiness skills. If we believe almost all students are already proficient, what is the motivation to teach them higher-level skills? This may be the main reason why less than 40 percent of 12th grade students are academically prepared for college.[2] Furthermore, over a third of students enrolled in college need remedial help. They thought that they were college ready because they passed their high school graduation test, but they were not.

[2] On May 14, 2014, the National Assessment Governing Board (NAGB) released a study in which it estimated that 39 percent of 12th graders are prepared for college in mathematics and 38 percent are prepared in reading. This was done by establishing a college preparedness predictive score of 163 on a 300-point scale in mathematics and 302 on a 500-point scale in reading.

To reduce the expectations gap, this report recommends re-engineering the current standard-setting paradigm used by the states. Almost all states use test content-based standard setting methods such as the bookmark method (Mitzel, Lewis, Patz, & Green, 2001). The problem with this approach is that it uses an inward focus on internal state content standards and does not focus on how state expectations stack up against the expectations of other states, the nation, and other countries. Rather than deriving performance standards exclusively from internal state content considerations, this report recommends a different method of evidence-based standard setting that incorporates more empirical data. An example of this is the *Benchmark Method* (Phillips, 2011) of standard setting, which argues that performance standards are fundamentally a policy-judgment decision (not just a content decision) and that these standards need to be guided by knowledge of the real world around us and the requirements that our students will face as they compete in a global economic and technological world.

Introduction

For the past quarter century, we as a country have believed that our underachieving educational system has put our nation at risk (National Council for Excellence in Education, 1983). National policymakers have responded to this crisis with aspirational, far-reaching goals, such as "being the first in the world in mathematics and science achievement by 2000" (National Education Goals Panel, 1999), "all students will be proficient in reading and mathematics by 2014" (No Child Left Behind Act, 2001), or "every student should graduate from high school ready for college and a career, regardless of their income, race, ethnic or language background, or disability status" (U.S. Department of Education, 2010).

Each of these national goals recognizes that in the 21st century, students must compete in a global economy, not just a local economy. The need for states to set high, internationally competitive standards has recently been emphasized by a number of policymakers. A recent report by the NGA, CCSSO, and Achieve (2008) concludes:

> *Governors recognize that new economic realities mean it no longer matters how one U.S. state compares to another on a national test; what matters is how a state's students compare to those in countries around the globe. America must seize this moment to ensure that we have workers whose knowledge, skills, and talents are competitive with the best in the world.* (p. 1)

Andreas Schleicher (2006), director of the Organization for Economic Co-operation and Development (OECD) Program for International Student Assessment (PISA), stated,

> *It is only through such benchmarking that countries can understand relative strengths and weaknesses of their education system and identify best practices and ways forward. The world is indifferent to tradition and past reputations, unforgiving of frailty and ignorant of custom or practice. Success will go to those individuals and countries which are swift to adapt, slow to complain, and open to change.* (p. 16)

The President of the United States (Obama, 2009), in a speech to the U.S. Hispanic Chamber of Commerce, recognized the need for high and consistent standards. He stated,

> *Let's challenge our states to adopt world-class standards that will bring our curriculums into the 21st century. Today's system of 50 different sets of benchmarks for academic success means fourth-grade readers in Mississippi are scoring nearly 70 points lower than students in Wyoming—and getting the same grade.*

Over the last decade within the United States, many states have been busy developing new content standards and new criterion-referenced tests that measure success on those content standards. Much of this frenetic activity is related to the federal No Child Left Behind legislation that requires states to report annually on whether they are making adequate yearly progress (AYP) toward meeting state standards. When states set performance standards, however, they generally have little knowledge of how those state performance standards compare with international performance standards, such as those used on TIMSS, PIRLS, and PISA. Yet, states

should care about how their students compare internationally and how their performance standards compare internationally. States compete with international companies, and their students will need to compete in an international market place.

International Benchmarking

International benchmarking is a way to calibrate the difficulty level of state performance standards to international standards. This type of benchmarking is similar to benchmarking in business and industry. For example, the fuel efficiency and quality of American-built cars are often benchmarked against those built in Japan and South Korea. Such benchmarking is important in education if we are to expect our students to compete in a global economy.

International Benchmarking Using TIMSS, PIRLS, and PISA

The international data already collected for three assessments could provide the data needed for international benchmarks. Two of these are used in this study: TIMSS and PIRLS. Both surveys are sponsored by the International Association for the Evaluation of Educational Achievement (IEA), currently located in the Netherlands. TIMSS is an assessment of Grade 4 and Grade 8 students in mathematics and science, and PIRLS is an assessment of Grade 4 students in reading. The third survey is PISA, sponsored by the Paris-headquartered OECD. PISA is an assessment of 15-year-old students in mathematics, science, and reading literacy. Statistical techniques for international benchmarking using PISA can be found in Phillips and Jiang (2014).

Expressing International Benchmarks as Grades

International benchmarks using TIMSS and PIRLS can be obtained by states by statistically linking their state tests to the state NAEP, then linking national NAEP to national TIMSS or PIRLS. This process of *chain linking* places the state's own performance standards on the TIMSS or PIRLS scale. States can then determine how their own state performance standards compare with the international benchmarks on TIMSS and PIRLS. One of the primary ways TIMSS and PIRLS report their results is in terms of international benchmarks. The labels and cut-points on the TIMSS and PIRLS scales for the international benchmarks are Advanced (625), High (550), Intermediate (475), and Low (400). These performance standards apply to both the Grade 4 and Grade 8 mathematics assessment in TIMSS and the Grade 4 reading assessment in PIRLS. Full descriptions of the TIMSS and PIRLS international benchmarks are contained in Appendix D.

To facilitate discussion, this report will relabel the international benchmarks as grades, with Advanced assigned an A, High assigned a B, Intermediate a C, and Low a D. These grades are indicated in Table 1.

Table 1: Determining Benchmark Grades

International benchmark on TIMSS and PIRLS	Cut-score on TIMSS and PIRLS	Grade for international benchmark
Advanced	650	A+
	625	A
	600	A-
High	575	B+
	550	B
	525	B-
Intermediate	500	C+
	475	C
	450	C-
Low	425	D+
	400	D
	375	D-

International Benchmarks for State Performance Standards

After each state performance standard is expressed on the common scale of TIMSS or PIRLS, comparing them and gauging their international competitiveness is possible. To see how, compare Figures 1 through 4 with Figures 5 through 8. Figures 1 through 4 show the percentage of students reported proficient by the states in 2011 in Grade 4 mathematics and reading and in Grade 8 mathematics and science. The percent proficient is the state result for spring 2011 under the federal reporting requirements of No Child Left Behind. The 2011 percent proficient results were reported on the U.S. Department of Education website at http://www2.ed.gov/admins/lead/account/consolidated/index.html. Using Grade 8 mathematics as an example, as shown in Figure 3, we see that the state with the greatest percentage of students reported proficient under No Child Left Behind is Georgia, whereas the percentage of proficient students in Massachusetts is among the lowest across the states. If parents used No Child Left Behind data to choose a state in which to live so their children could attend the best schools, they might choose Georgia. But there is something wrong with this picture. We know that NAEP reports exactly the opposite, with Massachusetts the highest-achieving state and Georgia among the lowest-achieving states. If we look deeper into the state performance standards, we can begin to explain this contradiction.

In each state, the number of proficient students is influenced by how high or low the state sets the performance standard for proficiency. The only way to compare the stringency or difficulty level of the performance standards across states is to express them in a common metric. This is done in Figures 5 through 8 by converting the state performance standards to the metric of TIMSS (i.e., the TIMSS equivalent of the state performance standard in mathematics and science) and converting the state performance standards to the metric of PIRLS (i.e., the PIRLS equivalent of the state performance standard in reading). The TIMSS equivalents and PIRLS

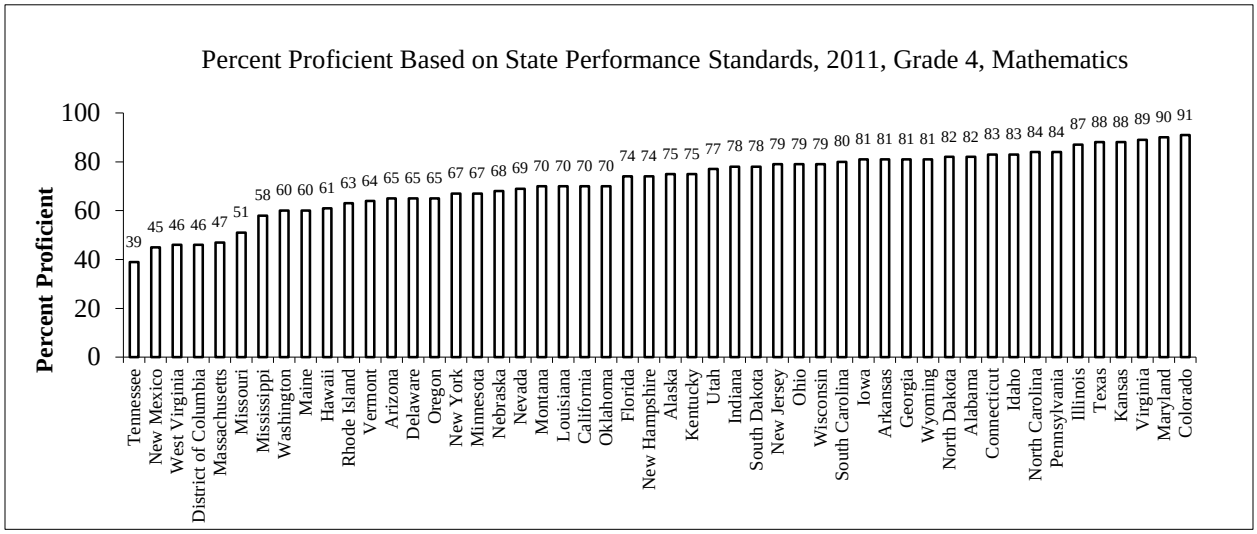
equivalents are then expressed as a grade (see Table 1, above). These grades represent the international benchmark for the state performance standards. A state performance standard that is mapped to a TIMSS equivalent in the D range of the TIMSS scale (i.e., a Low international benchmark) is requiring only a minimal level of mathematics. On the other hand, a state performance standard that is mapped to a TIMSS equivalent in the B range of the TIMSS scale (i.e., a High international benchmark) is requiring a level of mathematics similar to that needed to perform at the TIMSS and PIRLS level of the typical student in the highest-performing countries.

In Figures 5 through 8, the states have been ordered by their reported rates of percent proficient. The states with the lowest percent proficient are on the left and the states with the highest percent proficient on the right. The negatively sloping line shows that the performance standards drop as percent proficient increases.

Comparing the international benchmarks in Figures 5 through 8 to the percent proficient in Figures 1 through 4 shows why so many states can claim so many proficient students for federal reporting requirements. These states are using low standards to define proficiency. For example, in Grade 8 mathematics, seven states require only the equivalent of a D or D+ to be considered proficient. Massachusetts, on the other hand, has the highest performance standard in the country, a B−, which is why that state has fewer proficient students. The correlation between the difficulty of the state performance standard and the percent proficient is equal to −.83 for Grade 4 mathematics, −.83 for Grade 4 reading, −.79 for Grade 8 mathematics, and −.88 for Grade 8 science. This means that about two-thirds of the variance in No Child Left Behind reporting is due to how high—or low—the state sets the performance standard. In other words, high state performance as reported by No Child Left Behind is largely determined by how low a state sets its performance standards.

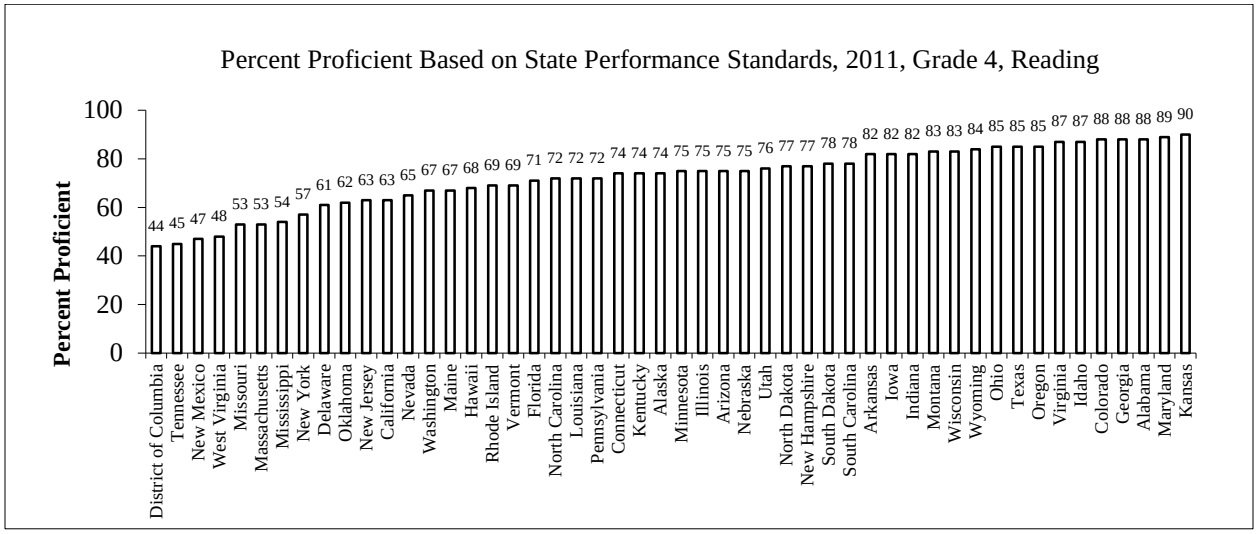
Another important observation emerging from Figures 5 through 8 is that the difference between the highest and lowest performance standards represents a difference in expectations. The states with the highest standards are expecting more than the states with the lowest standards, and this expectation gap is huge. We can get a solid understanding of this expectation gap if we express it in terms of TIMSS and PIRLS standard deviation units. Expressed as units of the U.S. national standard deviations of TIMSS or PIRLS, the standard deviation differences between the highest and lowest performance standard are 2.0, 1.6, 1.6, and 2.1 for Grade 4 mathematics, Grade 4 reading, Grade 8 mathematics, and Grade 8 science, respectively. To get a feel for the magnitude of these differences, note that a difference of two standard deviations equals about a three- to four-grade-level difference in student proficiency. Also, two standard deviations is about twice the size of the black–white achievement gap, which is often characterized as about one standard deviation. For example, the average national scores on the 2013 Grade 8 NAEP mathematics assessment were 263 and 295, for blacks and white, respectively with a standard deviation equal to 33. Expressed as a standard deviation unit, the black-white achievement gap was (263-295) / 33 = -.97. This means black students scored almost one standard deviation below white students.

Figure 1: Percent Proficient Based on State Performance Standard, Mathematics, Grade 4



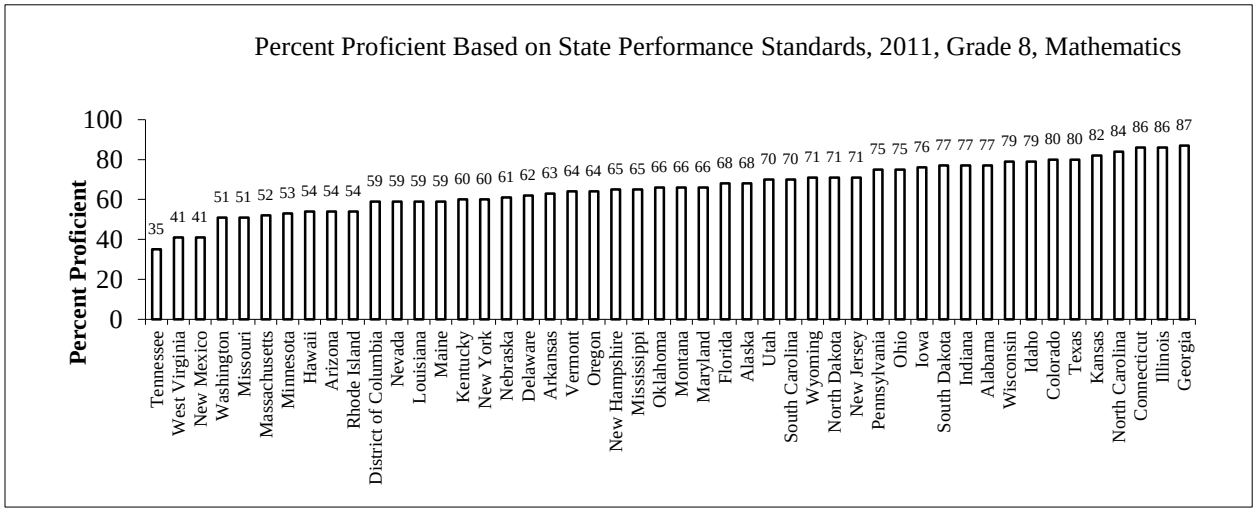
Source: Ed*Facts*, U.S. Department of Education, http://www2.ed.gov/admins/lead/account/consolidated/index.html.

Figure 2: Percent Proficient Based on State Performance Standard, Reading, Grade 4



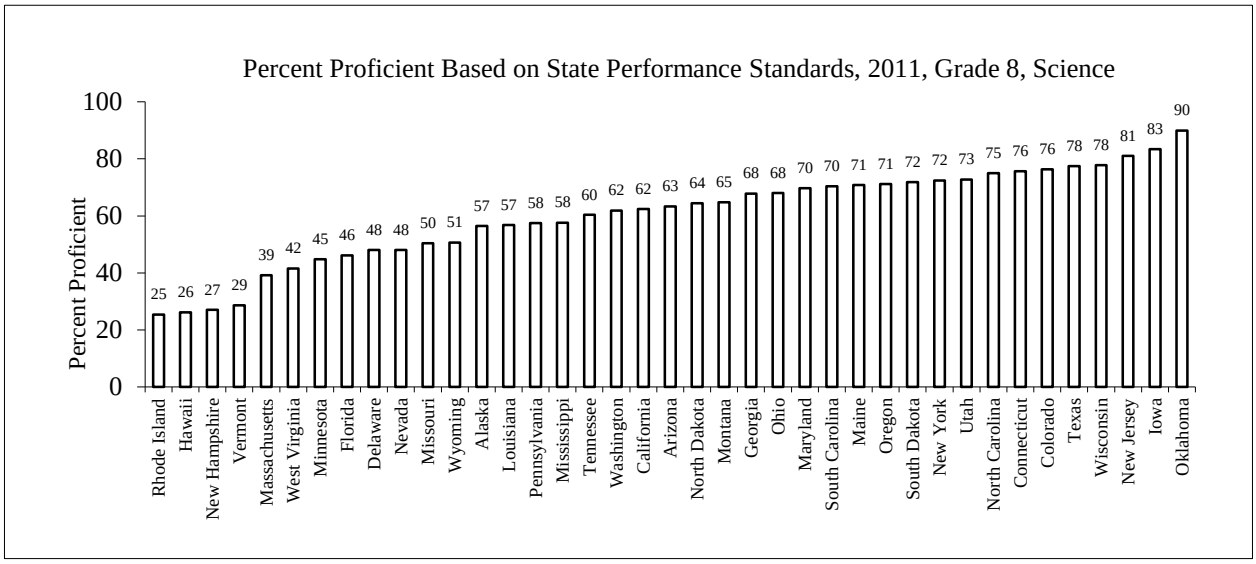
Source: Ed*Facts*, U.S. Department of Education, http://www2.ed.gov/admins/lead/account/consolidated/index.html.

Figure 3: Percent Proficient Based on State Performance Standard, Mathematics, Grade 8



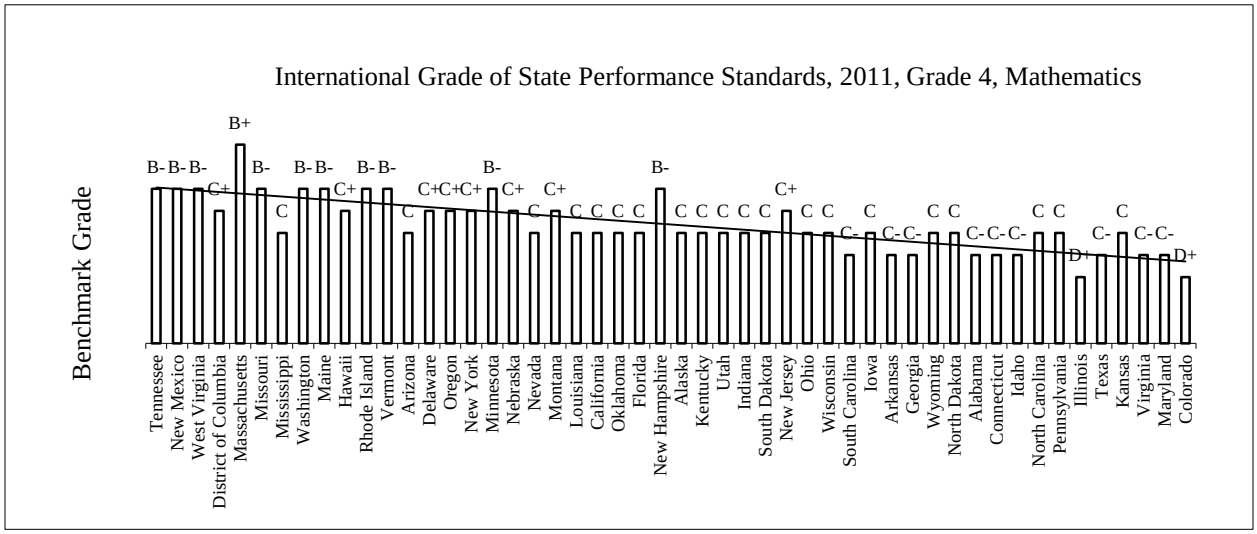
Source: Ed*Facts*, U.S. Department of Education, http://www2.ed.gov/admins/lead/account/consolidated/index.html.

Figure 4: Percent Proficient Based on State Performance Standard, Science, Grade 8



Source: Ed*Facts*, U.S. Department of Education, http://www2.ed.gov/admins/lead/account/consolidated/index.html.

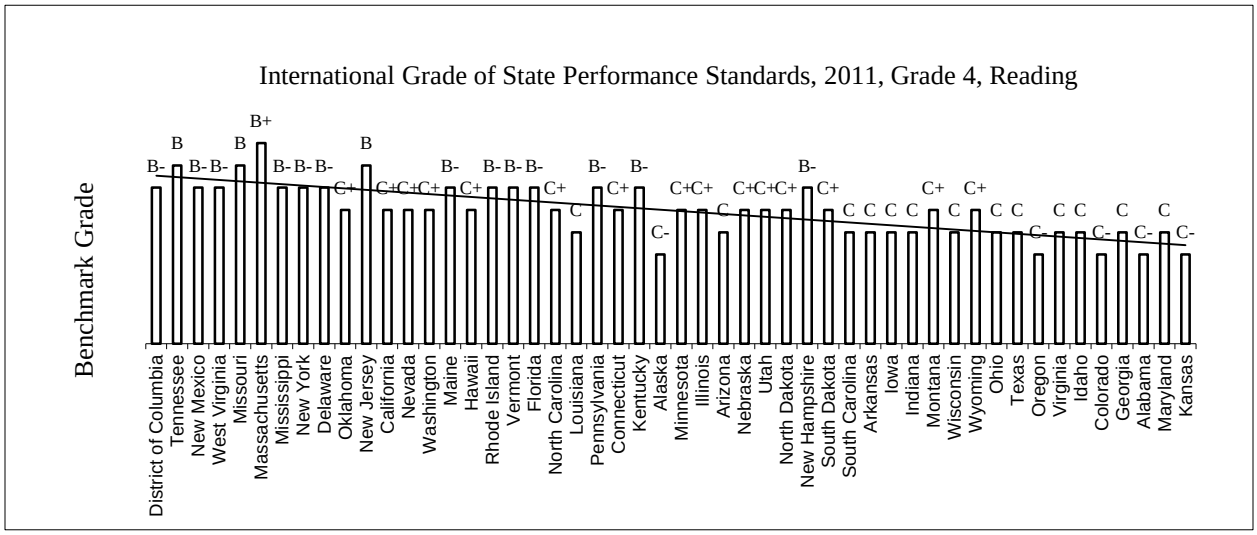
Figure 5: International Benchmarks for Mathematics, Grade 4



Note: The negatively sloping line of best fit represents the negative linear relationship between the state performance standard and the state percent proficient.

Source: Phillips, G. (2014). *International benchmarking: State and national education performance standards.* Washington, DC: American Institutes for Research.

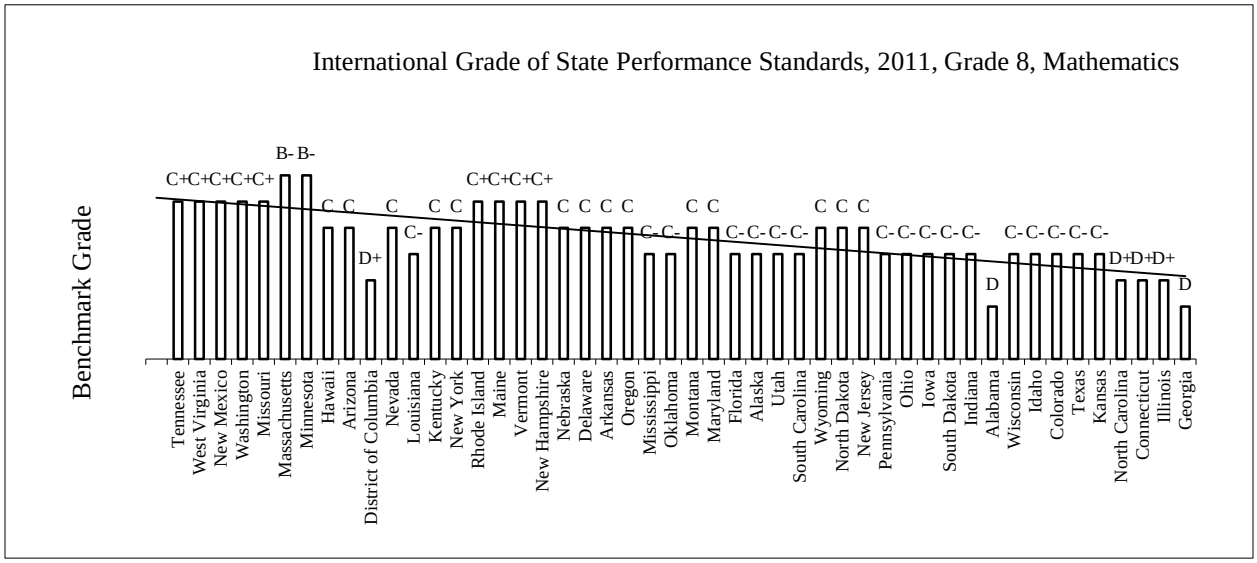
Figure 6: International Benchmarks for Reading, Grade 4



Note: The negatively sloping line of best fit represents the negative linear relationship between the state performance standard and the state percent proficient.

Source: Phillips, G. (2014). *International benchmarking: State and national education performance standards.* Washington, DC: American Institutes for Research.

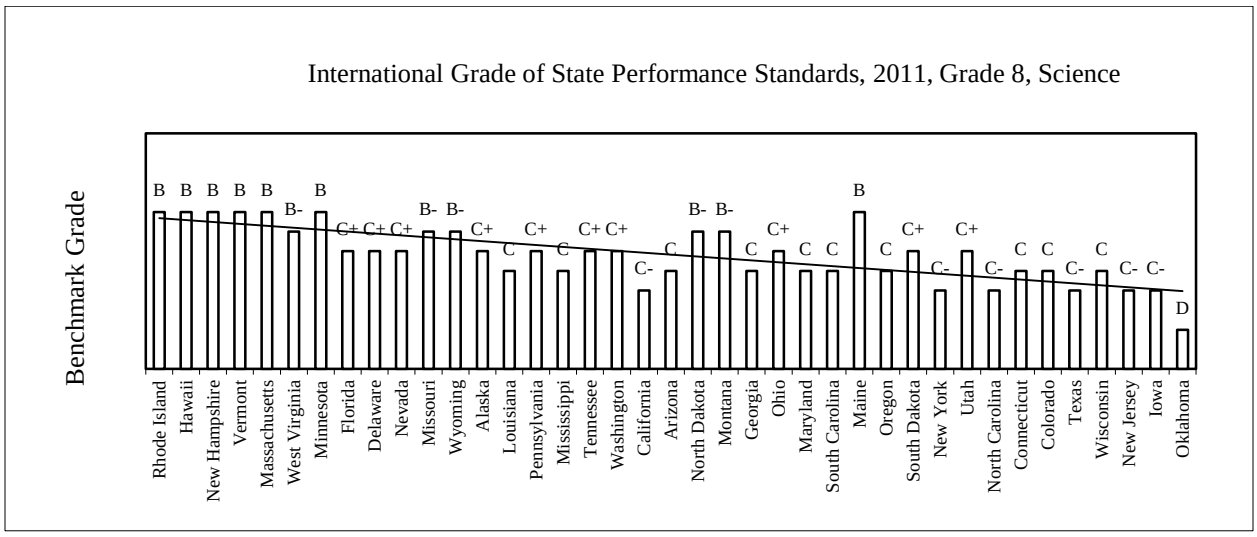
Figure 7: International Benchmarks for Mathematics, Grade 8



Note: The negatively sloping line of best fit represents the negative linear relationship between the state performance standard and the state percent proficient.

Source: Phillips, G. (2014). *International benchmarking: State and national education performance standards.* Washington, DC: American Institutes for Research.

Figure 8: International Benchmarks for Science, Grade 8



Note: The negatively sloping line of best fit represents the negative linear relationship between the state performance standard and the state percent proficient.

Source: Phillips, G. (2014). *International benchmarking: State and national education performance standards.* Washington, DC: American Institutes for Research.

Expressing State Performance Standards with a Common Metric

As indicated above, there is a large negative correlation between the stringency of the state standards and the percent proficient reported to the federal government as required by NCLB. This implies that, based on NCLB reporting, setting higher state standards is associated with lower levels of student performance. So does that mean there is no benefit to setting higher expectations? Actually, there is a benefit, but we have to use a common metric to see it. For example, Tables 12 and 13 (in Appendix B) show the TIMSS equivalent of the state performance standards (column 2) and the estimates of how many students would be expected to reach the High international benchmark on TIMSS (column 6). The latter estimates are available for Grade 8 from the NCES 2011 NAEP–TIMSS linking study (NCES, 2013). The question becomes this: Are higher state performance standards associated with higher percentages of students estimated to achieve the High level of performance on TIMSS? The correlations are +.37 for mathematics and +.41 for science. The correlations are positive and statistically significant. If we compare the TIMSS equivalent of the state performance standards with the percent proficient on state NAEP, we find similar results. The correlations are +.35 for mathematics and +.39 for science. The correlations are again positive and statistically significant.

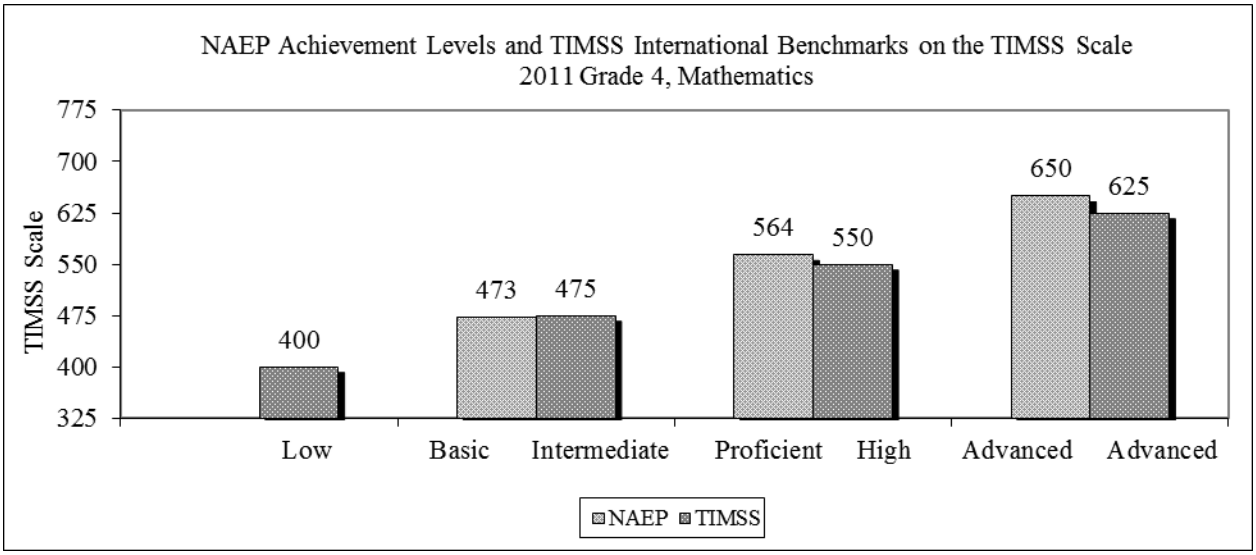
These results show a positive association between raising state performance standards and improved student performance. Such results are visible only when the performance standards are expressed with a common metric across the states.

International Benchmarks for National Performance Standards

In addition to benchmarking state performance standards, TIMSS and PIRLS can be used to internationally benchmark NAEP national achievement levels. This can be seen in Figures 9 through 12. The general conclusion from the linking results is that the NAEP Proficient achievement level is higher than the TIMSS High benchmark and the NAEP Advanced level is higher than the TIMSS Advanced benchmarks for Grade 4 mathematics and Grade 8 science. Furthermore, the NAEP Proficient and Advanced standards are higher than the PIRLS international benchmarks in Grade 4 reading. However, the NAEP Proficient and Advanced achievement levels are lower than the TIMSS High and Advanced international benchmarks for Grade 8 mathematics. These results are graphed in Figures 9 through 12, where the NAEP performance standards are expressed in the TIMSS and PIRLS metric. The same results are displayed again in Figures 13 through 16, but with the TIMSS and PIRLS international benchmarks expressed in the NAEP metric.

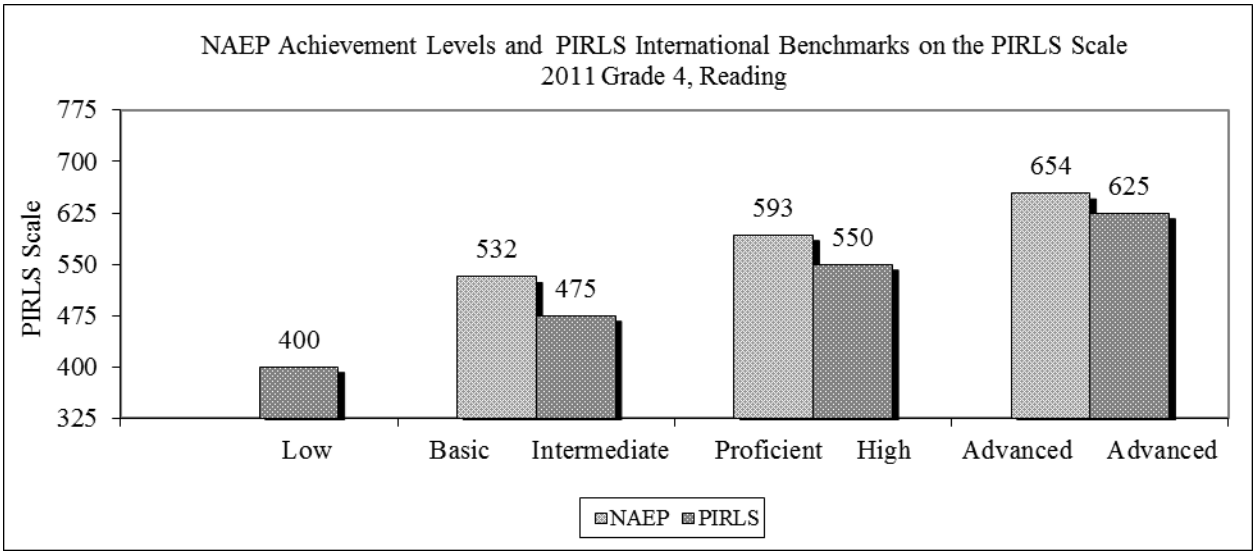
These findings may help explain several anomalies when comparing NAEP results with TIMSS and PIRLS results. For example, it is often reported that the United States does very well on international reading comparisons but has a low level of proficiency based on NAEP. For example, the 2011 PIRLS shows that 56 percent of U.S. students were reading at the High level and ranked sixth among the participating countries, implying that the United States produces students who are world-class readers. However, only 34 percent of students were reported proficient on the 2011 NAEP, suggesting that very few students in the United States are proficient readers. The reason for this discrepancy is that the NAEP Proficient standard is substantially higher than the PIRLS High international benchmark, as indicated in Figure 10 and Figure 14.

Figure 9: International Benchmarks for NAEP Performance Standards in Mathematics, Grade 4, Using the TIMSS Metric



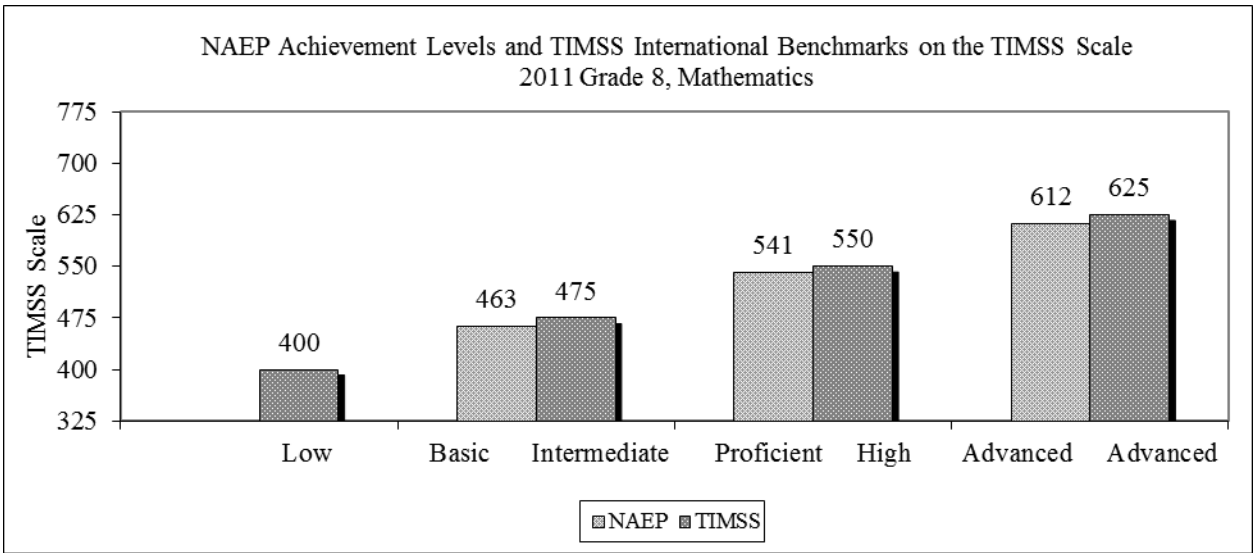
Note: NAEP Proficient and Advanced are significantly higher than TIMSS High and Advanced, respectively.

Figure 10: International Benchmarks for NAEP Performance Standards in Reading, Grade 4, Using the PIRLS Metric



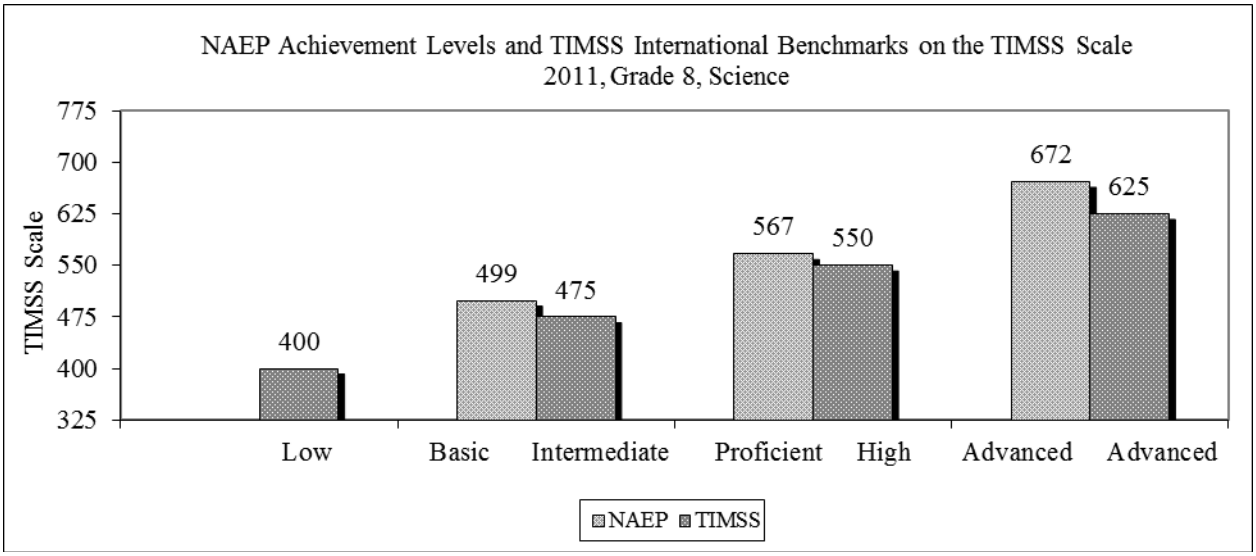
Note: NAEP achievement levels are significantly higher than PIRLS international benchmarks.

Figure 11: International Benchmarks for NAEP Performance Standards in Mathematics, Grade 8, Using the TIMSS Metric



Note: NAEP achievement levels are significantly lower than TIMSS international benchmarks.

Figure 12: International Benchmarks for NAEP Performance Standards in Science, Grade 8, Using the TIMSS Metric



Note: NAEP achievement levels are significantly higher than TIMSS international benchmarks.

Figure 13: International Benchmarks for NAEP Performance Standards in Mathematics, Grade 4, Using the NAEP Metric

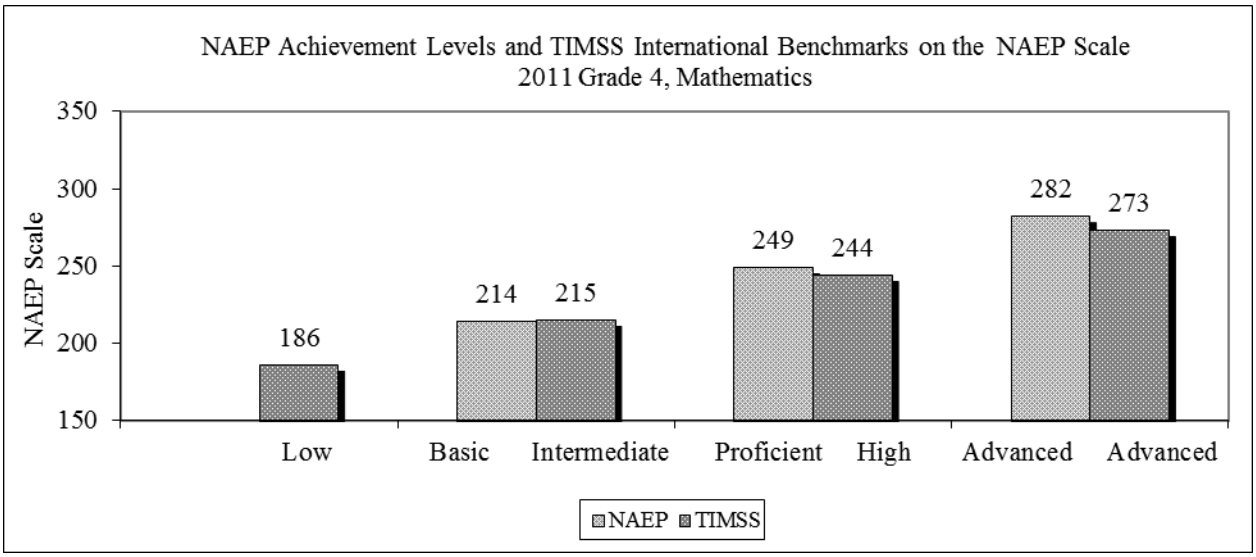


Figure 14: International Benchmarks for NAEP Performance Standards in Reading, Grade 4, Using the NAEP Metric

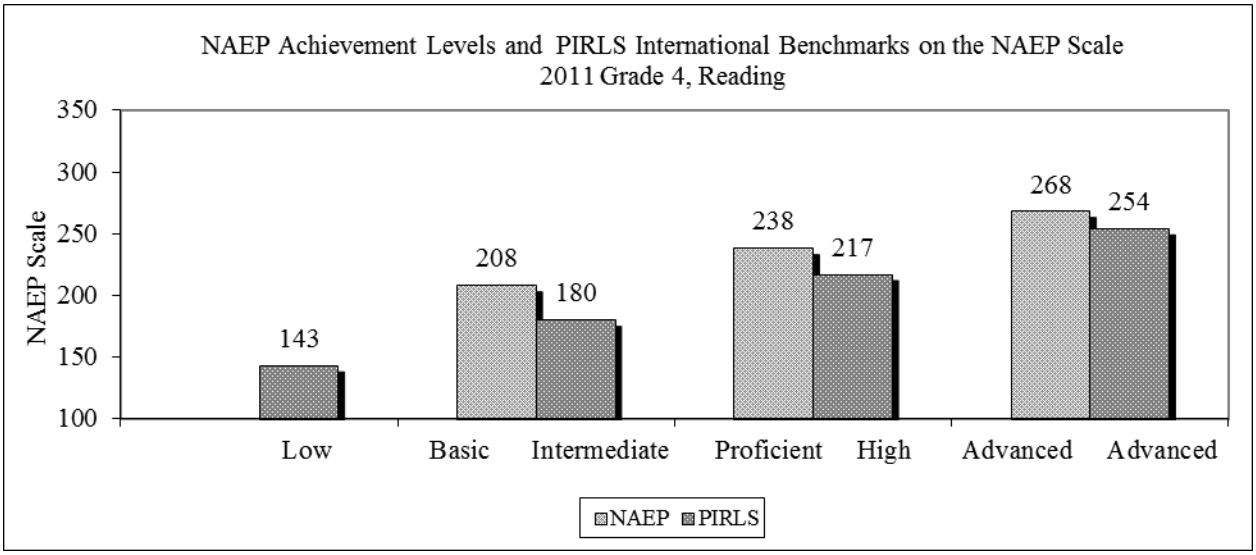


Figure 15: International Benchmarks for NAEP Performance Standards in Mathematics, Grade 8, Using the NAEP Metric

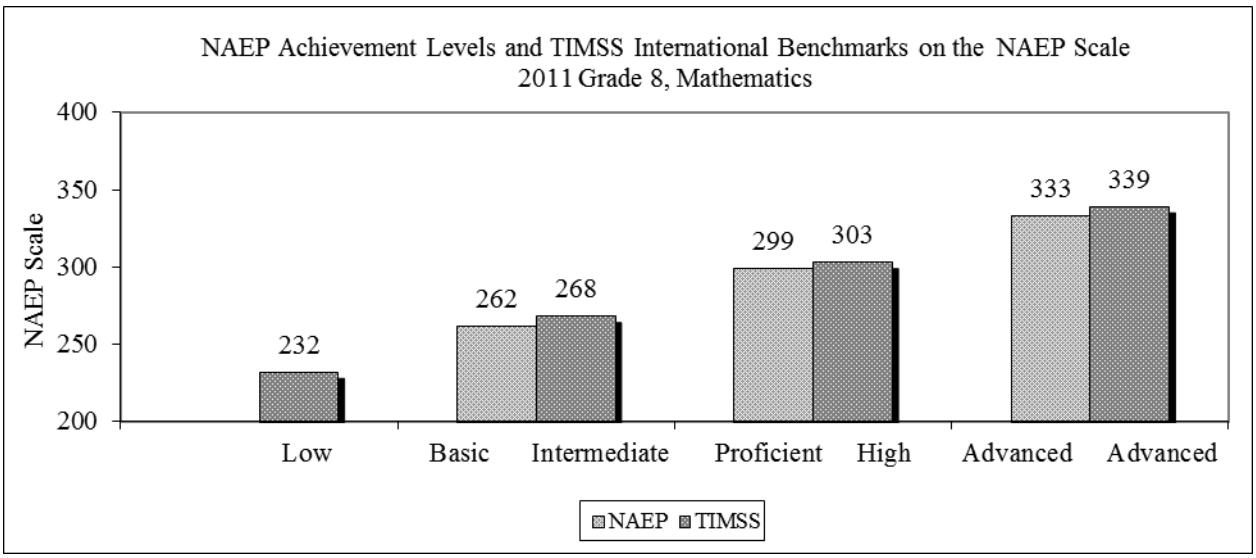
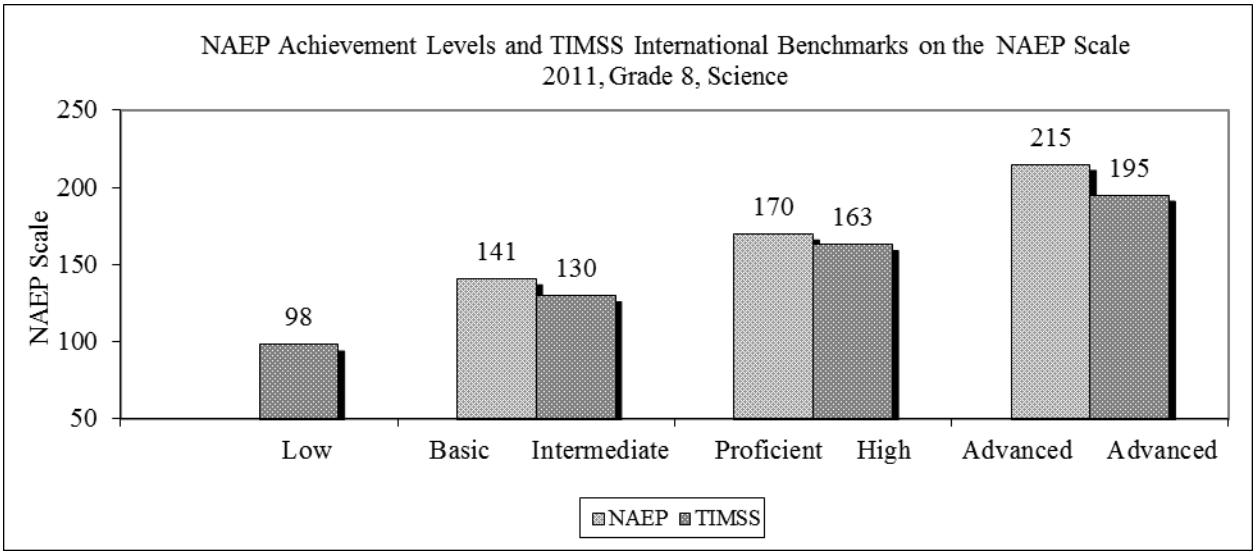


Figure 16: International Benchmarks for NAEP Performance Standards in Science, Grade 8, Using the NAEP Metric



How to Get Higher and More Consistent Standards

The lack of transparency among state performance standards is beginning to dawn on national policymakers. Recent calls for *fewer*, *clearer*, and *higher* standards by Secretary of Education Arne Duncan are recognition of the need for transparency. The Common Core project by CCSSO and NGA is partly motivated by the nation's lack of progress toward internationally competitive educational excellence if the 50 states are going in 50 different directions.

Both the Secretary of Education and the CCSSO–NGA project are primarily talking about fewer, clearer, and higher *content* standards. Content standards are statements about the scope and sequence of what students should learn in each grade and subject in school. Their concern is whether the state content standards are challenging and at least comparable to what is taught to students in the highest-performing countries in the world. This is an important first step, but it does not address the expectations gap discussed in this report. Many states already have highly challenging 21st-century content standards, but use low *performance* standards to increase the number of proficient students for federal reporting. States need a way to set consistently high performance standards. This can only happen if the current standard-setting paradigm in the testing industry is changed.

One of the main reasons states set low performance standards is related to the methodology currently in vogue in state testing programs to establish performance standards. Frequently used techniques like the Bookmark Method (Mitzel, Lewis, Patz, & Green, 2001) set standards based primarily (and in some cases exclusively) on test content. Teachers and other stakeholders set standards by reviewing test items and relating them to descriptions of performance levels and state content standards. The use of external empirical data is usually relegated to secondary importance in the standard-setting process. The standard-setting process is content-based not evidence-based.

The problem with narrowly focused content-based standard-setting methods is that nothing in the standard-setting process ensures that the performance standards are nationally or internationally challenging. The panelists usually believe that they are setting rigorous standards, basing their belief on the personal classroom experiences of the teachers and the anecdotal experiences of other stakeholders on the panel. Unfortunately, the panelists are flying without radar and have no clue as to whether they are setting standards that will help their students compete outside their state. Across the country, the strict emphasis on internal state content in setting performance standards has had the net effect of creating wide variations in rigor across all the states, and dumbed-down performance standards in many. These wide variations and low standards bespeak a lack of credibility and lack of transparency in state and federal education reporting, confuse policymakers, and mislead the public in some states into believing that their students are proficient when they are not. To correct this problem, this report recommends a more evidence-based approach to standard setting, such as the Benchmark Method (Phillips, 2013), in which panelists are guided by external data from other educational systems.

In the near future, many states are likely to function as a consortium and adopt the Common Core standards developed by CCSSO and NGA. Eventually, the Common Core content standards will need to establish Common Core performance standards. The Benchmark Method of establishing performance standards represents a departure from the narrow focus on internal

content standards currently used in most states. The Benchmark Method recognizes that performance standards are policy decisions, and that they need to be consistent and be set high enough to prepare students to compete for college and careers beyond the state borders. If the Benchmark Method were to be used in the future by individual states (or a consortium of states), state performance standards would be consistent and more on par with the high standards used by national and international assessments such as NAEP, TIMSS, PIRLS, and PISA.

Conclusion

The overall finding in the study is that the difference in the stringency of the performance standards used across the states is huge and probably far greater than most policymakers realize. The difference between the state with the highest standards and the state with the lowest standards was about 2 standard deviations. This difference is so great that it is more than twice the size of the national black–white achievement gap. In many state testing programs, a difference this great represents three to four grade levels.

These large differences among states clearly indicate why we need more common assessments and the Common Core State Standards. It is not that each state should teach the same thing at the same time in every grade every year—instead, we need to reduce the extreme variability that we now have, whereby some low-achieving states have low expectations and higher-achieving states have higher expectations. These huge differences in expectations deny students in states with low performance standards the opportunity to learn from a challenging curriculum.

Unfortunately, at the time of this report, much of the support for the Common Core State Standards has eroded. Initially, 46 states (including the District of Columbia) planned to conduct common assessments on the Common Core State Standards either through the Smarted Balanced Assessment Consortium (SBAC) or the Partnership for Assessment of Readiness for College and Careers (PARCC). Based on a recent tally, that number has now dropped to 27 states (including the District of Columbia), with 17 participating in the SBAC and 10 participating in PARCC (Gewertz & Ujifusa, 2014). In addition, recent 2014 polls by Education Next and the 46th annual PDK/Gallup Poll have shown a drop in public support for the Common Core based on the public's misperception that that the Common Core was a federal initiative (Camera, 2014).

Our analysis found that success under No Child Left Behind is largely related to using low performance standards. The stringency of state performance standards had a high negative correlation with the percentage of proficient students reported by the states. The states reporting the highest numbers of proficient students had the lowest performance standards. Another way of saying this is that high state performance reported by No Child Left Behind is significantly correlated with low state performance standards. About two-thirds of the variation in states' success reported by No Child Left Behind reflects differences in how individual states set their performance standards.

This report also estimated how the 2011 state results reported to No Child Left Behind would have looked had all the states used performance standards expressed in a common metric. When the data were reanalyzed on this basis, higher expectations reported by states were correlated with higher achievement.

This report argues that the No Child Left Behind paradigm of encouraging each state to set a different performance standard is fundamentally flawed, misleading, and lacking in transparency. Test results across the 50 states are not comparable, inferences about national progress are impossible, and we cannot even determine if progress in one state is greater than progress in another state. The lack of transparency among state performance standards misleads the public, because low standards can be used to artificially inflate the numbers of proficient students. This practice denies the nation's students the opportunity to learn college and career readiness skills.

References

Camera, L. (2014). Polls capture publics sour view of common core state standards, *Education Week*, 34(2).

Gewertz, C., & Ujifusa, A. (2014). State plans for testing fragmented. *Education Week*, 33(32).

Johnson, E. G., Cohen, J., Chen, W., Jiang, T., & Zhang, Y. (2005). *2000 NAEP–1999 TIMSS linking report.* Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Chestnut Hill, MA: Lynch School of Education, Boston College.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Erlbaum.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: Lynch School of Education, Boston College.

National Center for Education Statistics. (2011a). *The nation's report card: Mathematics 2011* (NCES 2012-458). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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

National Center for Education Statistics. (2012). *The nation's report card: Science 2011* (NCES 2012-465). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

National Center for Education Statistics. (2013). *U.S. states in a global context: Results from the 2011 NAEP–TIMSS linking study* (NCES 2013-460). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

National Council for Excellence in Education. (1983). *A nation at risk.* Washington, DC: U.S. Department of Education.

National Education Goals Panel. (1999). *Building a nation of learners*. Washington, DC: Author.

National Governors Association, Council of Chief State School Officers, Achieve, Inc. (2008). *Benchmarking for success: Ensuring U.S. students receive a world-class education*. Washington, DC: National Governors Association.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Obama, B. (2009). *Remarks to the Hispanic Chamber of Commerce*, the New York Times, March 10, 2009.

Phillips, G. W. (2010). *International benchmarking: State education performance standards.* Washington, DC: American Institutes for Research.

Phillips, G. W. (2013). The Benchmark Method of standard setting. In G. Cizek (Ed.), *Setting performance standards* (2nd ed.). New York, NY: Routledge.

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).* A publication of the NAEP Validity Studies Panel. San Mateo, CA: American Institutes for Research.

Phillips, G. W., & Jiang, T. (2014). *Using PISA as an international benchmark in standard setting.* Manuscript accepted for 2015 publication in the *Journal of Applied Measurement*.

Schleicher, A. (2006). The economics of knowledge: Why education is key to Europe's success. *Lisbon Council Policy Brief 1*(1).

United States Department of Education. (2010). *A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act*. Washington, DC, page 3.

Wolter, K. (1985). *Introduction to variance estimation*. New York, NY: Springer-Verlag.

Appendix A: Statistically Linking NAEP to TIMSS and PIRLS

This report uses the statistical linking procedures outlined in Johnson, Cohen, Chen, Jiang, and Zhang (2005). One major difference is that this report uses reported statistics from the NAEP 2011, TIMSS 2011, and PIRLS 2011 published reports, and the 2011 NAEP reports in mathematics and reading, rather than recalculating them from the public-use data files and plausible values available for the NAEP, TIMSS, and PIRLS assessments. The international benchmarking in this study is based on data obtained from several publically available reports. Data on mathematics, reading, and science NAEP were obtained from 2011 NAEP reports (National Center for Education Statistics, 2011a, 2011b, 2010). Data on TIMSS mathematics and science results were obtained from 2011 TIMSS reports (Mullis, Martin, Foy, & Arora, 2011; Martin, Mullis, Foy, & Stanco, 2011). Data on PIRLS were obtained from the 2011 PIRLS report (Mullis, Martin, Kennedy, Foy, & Drucker, 2012).

In the following discussion, *Y* denotes TIMSS (or PIRLS) and *X* denotes NAEP. In statistical moderation, the estimated z score is a transformed x score expressed in the y metric

$$\begin{aligned} z &= \hat{A} + \hat{B}(x) \\ &= \left(\hat{\mu}_Y - \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \hat{\mu}_X \right) + \left(\frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \right) x \end{aligned} \quad (1)$$

The *z* is the TIMSS equivalent (or PIRLS equivalent) of the NAEP score *x* associated with the state performance standard. The NAEP score *x* is obtained from determining the scaled score on NAEP that is the equipercentile equivalent of the performance standard on the local state accountability test (that is used for federal reporting required by No Child Left Behind).

In equation (1), $\hat{A}$ is an estimate of the intercept of a straight line, and $\hat{B}$ is an estimate of the slope of the linear transformation of NAEP to TIMSS or PIRLS defined by

$$\hat{A} = \hat{\mu}_Y - \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \hat{\mu}_X \quad (2)$$

$$\hat{B} = \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \quad (3)$$

In the above equations, $\hat{\mu}_X$ and $\hat{\mu}_Y$ are the national means of the U.S. NAEP and U.S. TIMSS (or PIRLS), respectively, while $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are the national standard deviations of the assessments.

Linking Error Variance

The linking error variance in the TIMSS equivalents and PIRLS equivalents can be determined through the following equation:

$$\hat{\sigma}_z^2 = \hat{B}^2\hat{\sigma}_x^2 + \hat{\sigma}_A^2 + 2(x)\hat{\sigma}_{AB} + (x)^2\hat{\sigma}_B^2 \quad (4)$$

According to Johnson et al. (2005), the error variances in this equation, $\hat{\sigma}_A^2$, $2\hat{\sigma}_{AB}$, and $\hat{\sigma}_B^2$, can be approximated by Taylor-series linearization (Wolter, 1985).

$$\hat{\sigma}_A^2 = \hat{B}^2\hat{\sigma}_x^2 + \hat{\sigma}_y^2 + x^2\hat{B}^2\left[\frac{Var(\hat{\sigma}_Y)}{\hat{\sigma}_Y^2} + \frac{Var(\hat{\sigma}_X)}{\hat{\sigma}_X^2}\right]$$

$$2\hat{\sigma}_{AB} = -2x\hat{B}^2\left[\frac{Var(\hat{\sigma}_Y)}{\hat{\sigma}_Y^2} + \frac{Var(\hat{\sigma}_X)}{\hat{\sigma}_X^2}\right] \quad (5)$$

$$\hat{\sigma}_B^2 = \hat{B}^2\left[\frac{Var(\hat{\sigma}_Y)}{\hat{\sigma}_Y^2} + \frac{Var(\hat{\sigma}_X)}{\hat{\sigma}_X^2}\right].$$

Equations (4) and (5) were used with data in the U.S. linking sample to derive the estimates of linking error variance in this paper.

The statistics needed to use equations (1) through (3) are contained in the tables below.

Table 2: Means and Standard Deviations for National Samples of Grade 4 TIMSS 2011 and NAEP 2011 in Mathematics

	Mean	Standard error of mean	Standard deviation	Standard error of standard deviation
TIMSS 2011, Math, Grade 4	540.65	1.81	75.58	1.11
NAEP 2011, Math, Grade 4	240.11	0.22	29.08	0.33

Table 3: Means and Standard Deviations for National Samples of Grade 4 PIRLS 2006 and NAEP 2011 in Reading

	Mean	Standard error of mean	Standard deviation	Standard error of standard deviation
PIRLS 2006, Reading, Grade 4	556.37	1.54	73.43	0.95
NAEP 2011 Reading, Grade 4	220.03	0.31	36.05	0.16

Table 4: Means and Standard Deviations for National Samples of Grade 8 TIMSS 2011 and NAEP 2011 in Mathematics

	Mean	Standard error of mean	Standard deviation	Standard error of standard deviation
TIMSS 2011, Math, Grade 8	506.89	2.63	76.04	1.59
NAEP 2011, Math, Grade 8	282.73	0.20	36.25	0.17

Table 5: Means and Standard Deviations for National Samples of Grade 8 TIMSS 2011 and NAEP 2011 in Science

	Mean	Standard error of mean	Standard deviation	Standard error of standard deviation
TIMSS 2011, Science, Grade 8	522.19	2.53	80.42	1.43
NAEP 2011, Science, Grade 8	150.74	0.23	34.50	0.17

The parameter estimates $\hat{A}$ and $\hat{B}$ are indicated in Tables 6 through 9. These are the intercepts and slopes, respectively, needed to re-express NAEP results on the TIMSS or PIRLS scale.

Table 6: Estimating TIMSS 2011 Mathematics From NAEP 2011, Mathematics, Grade 4

Estimates of linking parameters *A* and *B*		
	A	*B*
Parameter	-83.45	2.60
Standard error	11.74	0.05
Covariance	-0.56	

Table 7: Estimating PIRLS 2006 Reading From NAEP 2011, Reading, Grade 4

Estimates of linking parameters *A* and *B*		
	A	*B*
Parameter	108.20	2.04
Standard error	6.33	0.03
Covariance	-0.17	

Table 8: Estimating TIMSS 2011 Mathematics from NAEP 2011, Mathematics, Grade 8

Estimates of linking parameters *A* and *B*		
	A	*B*
Parameter	-86.14	2.10
Standard error	12.98	0.04
Covariance	-0.57	

Table 9: Estimating TIMSS 2011 Mathematics from NAEP 2011, Science, Grade 8

Estimates of linking parameters *A* and *B*		
	A	*B*
Parameter	170.78	2.33
Standard error	6.97	0.04
Covariance	-0.28	

Appendix B: State Proficient Standards Expressed in the Metric of TIMSS or PIRLS

This appendix provides the TIMSS equivalents and PIRLS equivalents of the state proficient performance standards used for reporting to No Child Left Behind in 2011. For example, in Table 10, the TIMSS equivalent of the Massachusetts proficient standard in Grade 4 mathematics was 580. In other words, the Massachusetts proficient standard is comparable in difficulty to the TIMSS score of 580. A score of 580 on TIMSS is at the High international benchmark and is comparable to a B+, based on the grading system in Table 1 of this report (B+ is assigned if the TIMSS equivalent or PIRLS equivalent of the state proficient standard is between 575 and 599 on the TIMSS or PIRLS scale).

Table 10: International Benchmarks Based on the TIMSS Equivalents of State Proficient Standards, Mathematics, Grade 4, 2011

State	TIMSS equivalent of state proficient standard	Standard error of TIMSS equivalent	International benchmark level of state proficient standard	International benchmark grade	Percent of students estimated to reach High TIMSS benchmark	Standard error of percent of students
Massachusetts	580	2.3	High	B+	64	2.3
Tennessee	543	3.0	Intermediate	B-	35	2.2
Missouri	540	3.5	Intermediate	B-	46	2.3
West Virginia	534	4.0	Intermediate	B-	37	2.1
New Mexico	531	2.5	Intermediate	B-	35	2.0
Minnesota	530	2.6	Intermediate	B-	57	2.2
Washington	529	3.5	Intermediate	B-	49	2.3
New Hampshire	528	2.2	Intermediate	B-	63	2.2
Vermont	528	2.2	Intermediate	B-	54	2.1
Rhode Island	528	2.2	Intermediate	B-	47	2.0
Maine	528	2.2	Intermediate	B-	51	2.2
Hawaii	514	3.2	Intermediate	C+	44	1.9
Delaware	514	2.8	Intermediate	C+	45	2.0
Montana	514	3.2	Intermediate	C+	50	2.2
Nebraska	506	3.9	Intermediate	C+	44	2.5
New Jersey	503	3.7	Intermediate	C+	56	2.5
District of Columbia	502	2.8	Intermediate	C+	27	1.8
Oregon	502	3.5	Intermediate	C+	41	2.1
New York	501	3.3	Intermediate	C+	42	2.1

State	TIMSS equivalent of state proficient standard	Standard error of TIMSS equivalent	International benchmark level of state proficient standard	International benchmark grade	Percent of students estimated to reach High TIMSS benchmark	Standard error of percent of students
Oklahoma	500	2.9	Intermediate	C	40	2.4
Mississippi	500	3.1	Intermediate	C	31	2.3
Arizona	498	3.1	Intermediate	C	39	2.5
Kentucky	496	3.2	Intermediate	C	46	2.4
Nevada	496	2.8	Intermediate	C	41	2.1
Indiana	496	3.5	Intermediate	C	50	2.6
North Dakota	495	3.2	Intermediate	C	52	2.2
Florida	495	3.3	Intermediate	C	44	2.3
Ohio	494	3.3	Intermediate	C	50	2.3
Wisconsin	493	3.1	Intermediate	C	51	2.2
Wyoming	493	2.9	Intermediate	C	50	2.1
Utah	493	3.3	Intermediate	C	48	2.2
South Dakota	491	4.0	Intermediate	C	46	2.2
Iowa	486	3.6	Intermediate	C	48	2.3
North Carolina	485	3.3	Intermediate	C	51	2.2
Pennsylvania	482	3.3	Intermediate	C	53	2.7
California	481	4.3	Intermediate	C	38	2.7
Kansas	480	3.5	Intermediate	C	54	2.7
Louisiana	479	3.9	Intermediate	C	32	2.5
Alaska	476	3.2	Intermediate	C	41	2.1
Idaho	473	3.3	Low	C-	45	2.1
Connecticut	470	3.8	Low	C-	48	2.7
South Carolina	470	3.1	Low	C-	41	2.4
Georgia	469	3.8	Low	C-	43	2.0
Arkansas	469	3.1	Low	C-	42	2.1
Virginia	466	5.7	Low	C-	52	2.3
Texas	465	4.2	Low	C-	46	3.0
Maryland	459	3.5	Low	C-	55	2.2
Alabama	450	6.2	Low	C-	33	2.4
Illinois	448	3.7	Low	D+	44	2.4
Colorado	446	5.8	Low	D+	51	2.2

Table 11: International Benchmarks Based on the PIRLS Equivalents of State Proficient Standards, Reading, Grade 4, 2011

State	PIRLS equivalent of state proficient standard	Standard error of PIRLS equivalent	International benchmark level of state proficient standard	International benchmark grade	Percent of students estimated to reach High TIMSS benchmark	Standard error of percent of students
Massachusetts	586	3.2	High	B+	74	2.1
New Jersey	558	2.6	High	B	67	2.5
Tennessee	554	2.4	High	B	47	2.3
Missouri	551	3.1	High	B	54	1.9
Delaware	549	2.4	Intermediate	B-	60	2.0
West Virginia	549	3.1	Intermediate	B-	47	1.9
New York	548	3.0	Intermediate	B-	56	2.1
New Mexico	537	2.8	Intermediate	B-	40	2.1
New Hampshire	530	1.9	Intermediate	B-	67	2.0
Vermont	530	1.9	Intermediate	B-	61	1.8
Rhode Island	530	1.9	Intermediate	B-	56	1.9
Maine	530	1.9	Intermediate	B-	56	2.0
District of Columbia	530	2.6	Intermediate	B-	35	1.8
Florida	529	2.3	Intermediate	B-	59	2.4
Pennsylvania	529	3.2	Intermediate	B-	61	2.3
Mississippi	527	2.8	Intermediate	B-	42	2.3
Kentucky	526	2.4	Intermediate	B-	60	2.4
Oklahoma	525	2.8	Intermediate	C+	48	2.3
Washington	524	2.9	Intermediate	C+	54	2.2
Connecticut	523	3.7	Intermediate	C+	61	2.3
North Dakota	523	2.3	Intermediate	C+	61	2.0
North Carolina	516	2.9	Intermediate	C+	55	2.2
Nebraska	515	3.2	Intermediate	C+	57	2.1
California	513	3.5	Intermediate	C+	44	3.0
Nevada	512	2.5	Intermediate	C+	45	2.0
Minnesota	511	2.8	Intermediate	C+	56	2.3
Hawaii	507	2.9	Intermediate	C+	47	1.9
Utah	506	4.0	Intermediate	C+	54	2.1
Montana	505	2.5	Intermediate	C+	60	2.0

State	PIRLS equivalent of state proficient standard	Standard error of PIRLS equivalent	International benchmark level of state proficient standard	International benchmark grade	Percent of students estimated to reach High TIMSS benchmark	Standard error of percent of students
Illinois	502	2.7	Intermediate	C+	53	2.0
South Dakota	502	3.4	Intermediate	C+	53	2.0
Wyoming	500	4.0	Intermediate	C+	59	2.1
Ohio	499	3.5	Intermediate	C	59	2.3
Louisiana	494	3.2	Intermediate	C	43	2.6
Maryland	493	4.1	Intermediate	C	66	1.9
Indiana	493	2.8	Intermediate	C	54	2.0
Iowa	492	3.2	Intermediate	C	54	1.9
Wisconsin	490	3.0	Intermediate	C	55	1.9
Virginia	488	3.6	Intermediate	C	61	2.2
Arizona	488	3.0	Intermediate	C	45	2.2
South Carolina	486	3.0	Intermediate	C	48	2.2
Texas	484	3.3	Intermediate	C	52	3.0
Arkansas	481	3.7	Intermediate	C	50	2.1
Idaho	478	3.8	Intermediate	C	54	1.9
Georgia	475	4.8	Intermediate	C	55	2.2
Alabama	475	4.9	Low	C-	54	2.6
Alaska	474	3.2	Low	C-	42	1.9
Colorado	472	3.2	Low	C-	57	2.2
Kansas	471	4.8	Low	C-	57	2.2
Oregon	468	4.6	Low	C-	50	2.1

Table 12: International Benchmarks Based on the TIMSS Equivalents of State Proficient Standards, Mathematics, Grade 8, 2011

State	TIMSS equivalent of state proficient standard	Standard error of TIMSS equivalent	International benchmark level of state proficient standard	International benchmark grade	Percent of students estimated to reach High TIMSS benchmark	Standard error of percent of students
Massachusetts	536	3.1	Intermediate	B-	57	3.2
Minnesota	527	3.1	Intermediate	B-	49	2.8
Tennessee	517	3.4	Intermediate	C+	21	2.3
Washington	516	3.4	Intermediate	C+	38	2.5
New Mexico	505	3.2	Intermediate	C+	22	2.1
Missouri	503	3.8	Intermediate	C+	31	2.6
West Virginia	503	3.3	Intermediate	C+	21	2.1
New Hampshire	501	2.9	Intermediate	C+	47	2.6
Vermont	501	2.9	Intermediate	C+	51	2.7
Maine	501	2.9	Intermediate	C+	44	2.5
Rhode Island	501	2.9	Intermediate	C+	37	2.3
Montana	498	3.3	Intermediate	C	41	2.6
Arizona	491	3.3	Intermediate	C	29	2.4
North Dakota	491	3.6	Intermediate	C	38	2.8
Hawaii	489	3.3	Intermediate	C	29	2.1
New Jersey	489	4.3	Intermediate	C	50	2.8
Nebraska	488	3.4	Intermediate	C	30	2.5
Kentucky	487	4.3	Intermediate	C	27	2.4
Delaware	485	3.1	Intermediate	C	32	2.4
Maryland	485	3.9	Intermediate	C	34	2.6
New York	483	3.7	Intermediate	C	40	2.3
Wyoming	480	2.9	Intermediate	C	36	2.7
Nevada	480	3.5	Intermediate	C	26	2.2
Oregon	480	4.1	Intermediate	C	32	2.5
Arkansas	475	3.3	Intermediate	C	29	2.4
South Dakota	473	3.8	Low	C-	35	2.6
Ohio	472	3.2	Low	C-	36	2.8
Oklahoma	472	4.7	Low	C-	20	2.5
Alaska	471	4.0	Low	C-	39	2.4
Louisiana	470	4.0	Low	C-	24	2.3
Utah	469	3.6	Low	C-	30	2.4
Texas	466	3.3	Low	C-	31	2.8

State	TIMSS equivalent of state proficient standard	Standard error of TIMSS equivalent	International benchmark level of state proficient standard	International benchmark grade	Percent of students estimated to reach High TIMSS benchmark	Standard error of percent of students
South Carolina	464	3.8	Low	C-	27	2.5
Colorado	462	3.4	Low	C-	35	2.7
Florida	462	3.1	Low	C-	31	3.2
Pennsylvania	461	4.5	Low	C-	40	2.6
Iowa	461	3.6	Low	C-	39	2.5
Indiana	461	4.1	Low	C-	35	3.3
Wisconsin	460	3.4	Low	C-	43	2.6
Kansas	459	3.6	Low	C-	36	2.7
Idaho	457	3.8	Low	C-	31	2.5
Mississippi	451	4.1	Low	C-	15	2.3
District of Columbia	441	4.0	Low	D+	21	1.6
North Carolina	440	4.7	Low	D+	44	3.6
Connecticut	436	4.3	Low	D+	37	2.9
Illinois	427	4.7	Low	D+	34	2.5
Alabama	423	4.3	Low	D	15	2.5
Georgia	415	4.7	Low	D	24	2.4

Note: For nine states the estimates of the percentage reaching the High TIMSS benchmark are based on actual TIMSS results. The states are Alabama, California, Colorado, Connecticut, Florida, Indiana, Massachusetts, Minnesota, and North Carolina. For the remaining states the estimates are based on the NCES 2011 NAEP–TIMSS linking study (NCES, 2013).

Table 13: International Benchmarks Based on the TIMSS Equivalents of State Proficient Standards, Science, Grade 8, 2011

State	TIMSS equivalent of state proficient standard	Standard error of TIMSS equivalent	International benchmark level of state proficient standard	International benchmark grade	Percent of students estimated to reach High TIMSS benchmark	Standard error of percent of students
Massachusetts	569	3.4	High	B	61	2.8
Maine	562	3.0	High	B	55	3.1
Vermont	562	3.0	High	B	60	3.0
New Hampshire	562	3.0	High	B	57	3.0
Rhode Island	562	3.0	High	B	43	2.3
Minnesota	556	3.6	High	B	54	2.6
Hawaii	554	3.4	High	B	31	2.2
Wyoming	543	3.3	Intermediate	B-	52	3.0
Missouri	534	3.9	Intermediate	B-	45	3.1
West Virginia	534	3.3	Intermediate	B-	34	2.8
North Dakota	530	3.4	Intermediate	B-	56	3.2
Montana	526	3.5	Intermediate	B-	53	3.0
Florida	525	3.3	Intermediate	C+	42	3.5
Delaware	524	3.9	Intermediate	C+	40	2.5
Alaska	516	3.5	Intermediate	C+	50	2.6
Washington	513	3.4	Intermediate	C+	45	2.8
Nevada	511	4.7	Intermediate	C+	33	2.4
South Dakota	511	3.0	Intermediate	C+	50	2.9
Pennsylvania	508	3.0	Intermediate	C+	46	2.8
Ohio	506	3.5	Intermediate	C+	51	2.9
Utah	504	3.2	Intermediate	C+	51	2.8
Tennessee	500	3.5	Intermediate	C+	39	2.7
Colorado	491	4.9	Intermediate	C	51	3.2
Oregon	490	5.0	Intermediate	C	45	2.7
Louisiana	490	4.3	Intermediate	C	34	3.2
Georgia	485	4.8	Intermediate	C	38	3.0
Wisconsin	484	4.3	Intermediate	C	53	2.9
Maryland	481	4.7	Intermediate	C	42	2.7
Michigan	479	3.6	Intermediate	C	45	2.8
Arizona	479	3.2	Intermediate	C	31	2.7
Connecticut	477	4.0	Intermediate	C	45	2.5
Mississippi	476	3.7	Intermediate	C	22	2.6
South Carolina	475	4.5	Intermediate	C	36	2.6

State	TIMSS equivalent of state proficient standard	Standard error of TIMSS equivalent	International benchmark level of state proficient standard	International benchmark grade	Percent of students estimated to reach High TIMSS benchmark	Standard error of percent of students
Iowa	472	3.9	Low	C-	52	2.9
California	470	4.5	Low	C-	28	1.9
Texas	469	4.1	Low	C-	39	2.7
New York	468	4.2	Low	C-	46	2.4
North Carolina	462	3.9	Low	C-	42	3.2
New Jersey	462	6.0	Low	C-	52	2.9
Oklahoma	421	8.0	Low	D	35	2.8

Note: For nine states the estimates of the percentage reaching the High TIMSS benchmark are based on actual TIMSS results. The states are Alabama, California, Colorado, Connecticut, Florida, Indiana, Massachusetts, Minnesota, and North Carolina. For the remaining states the estimates are based on the NCES 2011 NAEP–TIMSS linking study (NCES, 2013).

Appendix C: Validity of International Benchmarking

The international benchmarking in this report depends on several statistical assumptions. The first assumption is that the 2011 state percent proficient data are accurate. The data were obtained from the federal Ed*Facts* website (http://www2.ed.gov/about/inits/ed/edfacts/index.html). The percent proficient include the 1 percent of students who take the state alternate assessment. A separate analysis (based on 2009 data that had the alternate assessment excluded) indicated that including the 1 percent did not make any material difference in the overall aggregate percent proficient. The correlation between the percent proficient with the 1 percent alternate assessment included and the percent correct with the 1 percent alternate assessment excluded was about .99. In addition, the state-NAEP Coordinators in 22 states were contacted to confirm or correct state results reported in Ed*Facts*.

The international benchmarking in this report for mathematics and science uses a chain-linking approach, in which the state test is first linked to state NAEP through equipercentile linking. Then, the national NAEP is linked to national TIMSS or PIRLS through statistical moderation. For this approach to be valid, the TIMSS equivalents based on the chain linking need to be comparable to the actual TIMSS results for the state. Fortunately, in 2011 there were nine states that took the TIMSS assessment at Grade 8 in mathematics and science. NCES conducted a NAEP–TIMSS linking study that used the nine states to validate the TIMSS predictions for all states (National Center for Education Statistics, 2013). This study can also use data from the nine states for validation. This permits us to compare the percentage we estimate would reach the High level of performance on TIMSS with the actual percentage that reached the High level in nine states. If the two estimates are comparable, then that would indicate the statistical linking is reasonably valid. Below are the comparisons for mathematics and science.

The international benchmarking for reading uses the results of a previous linking study between NAEP and PIRLS (Phillips, 2014).

Table 14: Comparisons Between Percentage Reaching High Level on TIMSS Equivalents and the Actual Percentage for Mathematics, Grade 8, 2011

State	TIMSS equivalent state percentage	Standard error linking	Actual TIMSS state percentage	Standard error state TIMSS	Overall standard error	z-Test	Significant difference
Alabama	17	2.2	15	2.5	3.4	0.63	NS
California	22	2.1	24	2.5	3.2	-0.76	NS
Colorado	37	2.7	35	2.7	3.8	0.59	NS
Connecticut	32	2.6	37	2.9	3.9	-1.09	NS
Florida	23	2.1	31	3.2	3.8	-2.04	Significant
Indiana	29	2.6	35	3.3	4.2	-1.53	NS
Massachusetts	44	2.6	57	3.2	4.1	-3.11	Significant
Minnesota	41	2.7	49	2.8	3.9	-2.13	Significant
North Carolina	32	2.4	44	3.6	4.4	-2.87	Significant

Two-tailed z-test, with alpha = .05. NS = not significant.

Table 15: Comparisons Between Percent Reaching High Level on TIMSS Equivalents and the Actual Percentage for Science, Grade 8, 2011

State	TIMSS equivalent state percentage	Standard error linking	Actual TIMSS state percentage	Standard error state TIMSS	Overall standard error	z-Test	Significant difference
Alabama	26	2.6	24	2.7	3.7	0.60	NS
California	28	2.4	28	1.9	3.1	-0.13	NS
Colorado	47	3.0	48	2.6	4.0	-0.11	NS
Connecticut	41	2.7	45	2.5	3.6	-1.19	NS
Florida	34	2.5	42	3.5	4.3	-1.76	NS
Indiana	38	2.6	43	2.9	3.9	-1.37	NS
Massachusetts	48	2.7	61	2.8	3.9	-3.40	Significant
Minnesota	48	2.8	54	2.6	3.8	-1.55	NS
North Carolina	33	2.5	42	3.2	4.1	-2.31	Significant

Two-tailed z-test, with alpha = .05. NS = not significant.

The results in Tables 14 and 15 indicate that the linking results were adequate for Grade 8 mathematics. In Grade 8 mathematics, the percent reaching the High TIMSS equivalents were statistically comparable to the actual percent reaching the High benchmark in five out of nine comparisons. The linking primarily underestimated the percent reaching High TIMSS benchmarks for states with very high performance standards. In Grade 8 science, the linking results were especially good and were comparable in seven out of nine comparisons. As can be seen from these tables, the estimates based on linking were not perfect, but they were adequate in most cases.

The validity evidence for the NAEP-PIRLS linking was also encouraging (see Phillips, 2014). In 2011, Florida administered a statewide assessment in PIRLS. In general, PIRLS equivalents are not statistical significantly different from the actual PIRLS benchmarks. For example, the mean difference between the PIRLS equivalent and the actual PIRLS mean is not significant (see **Error! Reference source not found.**). The only significant difference between the PIRLS equivalent and the actual PIRLS result is for the percentage of advanced students (see Table 17). Even though there is only a 1 percent difference between the predicted and the actual percentage, the difference is statistically significant because the standard errors are so small.

Table 16: Comparing Mean for the State PIRLS Equivalents With the Actual State PIRLS, Reading, Grade 4, 2011

Florida	Equivalent state mean	Standard error	PIRLS state mean	Error state PIRLS	Standard error	z-Test	Significant difference
Mean	566	2.8	569	2.9	4.0	-0.83	NS

Table 17: Comparing Percentages Above Benchmarks for the State PIRLS Equivalents With the Actual State PIRLS, Reading, Grade 4, 2011

Florida	Equivalent of state percentage	Error PIRLS equivalent	PIRLS state percentage	Error state percentage	Standard error	z-Test	Significant difference
Advanced	18	1.9	22	1.7	2.5	-1.40	NS
High	59	2.4	61	1.7	2.9	-0.59	NS
Intermediate	91	1.2	91	1.1	1.7	0.24	NS
Low	99	0.2	98	0.4	0.5	3.04	Significant

Two-tailed z-test, with alpha = .05.

Appendix D: International Benchmarks for TIMSS and PIRLS[3]

International Benchmarks for Grade 4 TIMSS Mathematics

Advanced

Students can apply their understanding and knowledge in a variety of relatively complex situations and explain their reasoning. They can solve a variety of multi-step word problems involving whole numbers, including proportions. Students at this level show an increasing understanding of fractions and decimals. Students can apply geometric knowledge of a range of two- and three-dimensional shapes in a variety of situations. They can draw a conclusion from data in a table and justify their conclusion.

High

Students can apply their knowledge and understanding to solve problems. Students can solve word problems involving operations with whole numbers. They can use division in a variety of problem situations. They can use their understanding of place value to solve problems. Students can extend patterns to find a later specified term. Students demonstrate understanding of line symmetry and geometric properties. Students can interpret and use data in tables and graphs to solve problems. They can use information in pictographs and tally charts to complete bar graphs.

Intermediate

Students can apply basic mathematical knowledge in straightforward situations. Students at this level demonstrate an understanding of whole numbers and some understanding of fractions. Students can visualize three-dimensional shapes from two-dimensional representations. They can interpret bar graphs, pictographs, and tables to solve simple problems.

Low

Students have some basic mathematical knowledge. Students can add and subtract whole numbers. They have some recognition of parallel and perpendicular lines, familiar geometric shapes, and coordinate maps. They can read and complete simple bar graphs and tables.

[3] The text is this appendix is taken from the 2011 TIMSS international mathematics report (Mullis, Martin, Foy & Arora, 2012), the 2011 TIMSS international science report (Martin, Mullis, Foy & Stanco, 2012), and the 2011 PIRLS international reading report (Mullis, Martin, Kennedy, Foy & Drucker, 2012).

International Benchmarks for Grade 4 PIRLS Reading

Advanced

When reading Literary Texts, students can:

- Integrate ideas and evidence across a text to appreciate overall themes
- Interpret story events and character actions to provide reasons, motivations, feelings, and character traits with full text-based support

When reading Informational Texts, students can:

- Distinguish and interpret complex information from different parts of text, and provide full text-based support
- Integrate information across a text to provide explanations, interpret significance, and sequence activities
- Evaluate visual and textual features to explain their function

High

When reading Literary Texts, students can:

- Locate and distinguish significant actions and details embedded across the text
- Make inferences to explain relationships between intentions, actions, events, and feelings, and give text-based support
- Interpret and integrate story events and character actions and traits from different parts of the text
- Evaluate the significance of events and actions across the entire story
- Recognize the use of some language features (e.g., metaphor, tone, imagery)

When reading Informational Texts, students can:

- Locate and distinguish relevant information within a dense text or a complex table
- Make inferences about logical connections to provide explanations and reasons
- Integrate textual and visual information to interpret the relationship between ideas
- Evaluate content and textual elements to make a generalization

International Benchmarks for Grade 4 PIRLS Reading

Intermediate

When reading Literary Texts, students can:

- Retrieve and reproduce explicitly stated actions, events, and feelings
- Make straightforward inferences about the attributes, feelings, and motivations of main characters
- Interpret obvious reasons and causes and give simple explanations
- Begin to recognize language features and style

When reading Informational Texts, students can:

- Locate and reproduce two or three pieces of information from within the text
- Use subheadings, text boxes, and illustrations to locate parts of the text

Low

When reading Literary Texts, students can:

- Locate and retrieve an explicitly stated detail

When reading Informational Texts, students can:

- Locate and reproduce explicitly stated information that is at the beginning of the text

International Benchmarks for
Grade 8 TIMSS Mathematics

Advanced

Students can reason with information, draw conclusions, make generalizations, and solve linear equations. Students can solve a variety of fraction, proportion, and percent problems and justify their conclusions. Students can express generalizations algebraically and model situations. They can solve a variety of problems involving equations, formulas, and functions. Students can reason with geometric figures to solve problems. Students can reason with data from several sources or unfamiliar representations to solve multi-step problems.

High

Students can apply their understanding and knowledge in a variety of relatively complex situations. Students can use information from several sources to solve problems involving different types of numbers and operations. Students can relate fractions, decimals, and percents to each other. Students at this level show basic procedural knowledge related to algebraic expressions. They can use properties of lines, angles, triangles, rectangles, and rectangular prisms to solve problems. They can analyze data in a variety of graphs.

Intermediate

Students can apply basic mathematical knowledge in a variety of situations. Students can solve problems involving decimals, fractions, proportions, and percentages. They understand simple algebraic relationships. Students can relate a two-dimensional drawing to a three-dimensional object. They can read, interpret, and construct graphs and tables. They recognize basic notions of likelihood.

Low

Students have some knowledge of whole numbers and decimals, operations, and basic graphs.

International Benchmarks for Grade 8 TIMSS Science

Advanced

Students communicate an understanding of complex and abstract concepts in biology, chemistry, physics, and earth science. Students demonstrate some conceptual knowledge about cells and the characteristics, classification, and life processes of organisms. They communicate an understanding of the complexity of ecosystems and adaptations of organisms, and apply an understanding of life cycles and heredity. Students also communicate an understanding of the structure of matter and physical and chemical properties and changes and apply knowledge of forces, pressure, motion, sound, and light. They reason about electrical circuits and properties of magnets. Students apply knowledge and communicate understanding of the solar system and Earth's processes, structures, and physical features. They understand basic features of scientific investigation. They also combine information from several sources to solve problems and draw conclusions, and they provide written explanations to communicate scientific knowledge.

High

Students demonstrate understanding of concepts related to science cycles, systems, and principles. They demonstrate understanding of aspects of human biology, and of the characteristics, classification, and life processes of organisms. Students communicate understanding of processes and relationships in ecosystems. They show an understanding of the classification and compositions of matter and chemical and physical properties and changes. They apply knowledge to situations related to light and sound and demonstrate basic knowledge of heat and temperature, forces and motion, and electrical circuits and magnets. Students demonstrate an understanding of the solar system and of Earth's processes, physical features, and resources. They demonstrate some scientific inquiry skills. They also combine and interpret information from various types of diagrams, contour maps, graphs, and tables; select relevant information, analyze, and draw conclusions; and provide short explanations conveying scientific knowledge.

Intermediate

Students recognize and apply their understanding of basic scientific knowledge in various contexts. Students apply knowledge and communicate an understanding of human health, life cycles, adaptation, and heredity, and analyze information about ecosystems. They have some knowledge of chemistry in everyday life and elementary knowledge of properties of solutions and the concept of concentration. They are acquainted with some aspects of force, motion, and energy. They demonstrate an understanding of Earth's processes and physical features, including the water cycle and atmosphere. Students interpret information from tables, graphs, and pictorial diagrams and draw conclusions. They apply knowledge to practical situations and communicate their understanding through brief descriptive responses.

Low

Students can recognize some basic facts from the life and physical sciences. They have some knowledge of biology, and demonstrate some familiarity with physical phenomena. Students interpret simple pictorial diagrams, complete simple tables, and apply basic knowledge to practical situations.

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000 | TTY: 877.334.3499

www.air.org

Making Research Relevant

3178_09/14