AMERICAN INSTITUTES FOR RESEARCH®

International Benchmarking:
State Education Performance Standards

Gary W. Phillips, Ph.D.
Vice President and Chief Scientist
AIR®

October 2010

www.air.org

About This Report

The American Institutes for Research (AIR) has funded and conducted this report as part of our effort to make research relevant to policymakers and practitioners in the field of education. Our mission at AIR is to conduct and apply behavioral and social science research to improve people's lives and well-being, with a special emphasis on the disadvantaged. This report helps meet this goal by providing policymakers international benchmarks against which they can compare and monitor the educational performance of students.

In a highly interconnected world, U.S. students will require strong mathematic skills to compete against their peers around the globe. Reports such as International Benchmarking: State Education Performance Standards help policymakers and educators to know how well they are doing in meeting this challenge and to track progress over time.

About AIR®

Established in 1946, with headquarters in Washington, DC, and with nearly 30 offices in the United States and around the world, AIR is a nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of health, education, and workforce productivity.



Contents

List of Tables

List of Figures

Copies of this paper can be downloaded by searching http://www.air.org, and questions can be addressed to the author at gwphillips@air.org. Proper citation is as follows: Phillips, G. W. (2010), International Benchmarking: State Education Performance Standards. Washington, DC: American Institutes for Research.

Executive Summary

National policymakers have recently encouraged states to adopt world-class education standards as a way for the nation to compete in the 21st century. However, high national expectations can never be realized if expectations across the states are wildly inconsistent and are extremely low in some states. By setting low performance standards, states commit the educational equivalent of short selling. Rather than betting on student success, the educators sell the student short by lowering standards. What the educator gets out of this practice is the illusion of high rates of proficiency, which have a palliative effect on public opinion and meet the requirements of federal reporting. What the student gets out of it is a dumbed-down education, with little opportunity to learn college-ready and career-ready skills.

This report uses international benchmarking to examine the *expectations gap* between what students are expected to learn in some states and what students are expected to learn in others. This report assumes that each state's expectations are embodied in the stringency of the performance standards it uses on its own state accountability tests. The state performance standards represent how much the state expects the student to learn in order to be considered proficient in reading and mathematics. Performance standards are used by each state to report adequate yearly progress (AYP) under current No Child Left Behind federal legislation. These standards are also used by the state to monitor progress from year to year, and to report to parents and the public on the success of the each classroom, school, and district.

In the examination of this issue, the proficiency standards in each state were compared with the international benchmarks used in two international assessments. These were the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). The international benchmarking not only provided a mechanism for calibrating the difficulty and gauging the global competitiveness of each state standard but yielded an international common metric with which to compare state expectations.

The overall finding in the study is that the differences in the stringency of the performance standards used across the states are huge.[1] Although this gap in expectations is large, few policymakers are aware of it. For this reason, it is important that the reader get a feel for how large it is. As an example, we will use the gap between what is expected in Massachusetts and in the states with the lowest standards.

- The difference between the standards in Massachusetts and the standards of the states with the lowest standards is about 2 standard deviations.[2] In many testing programs, a gap this large may represent as much as four grade levels.
- This expectations gap is so large that it is more than twice the size of the national black–white *achievement gap*. Before the nation can close the achievement gap, it must close the bigger expectations gap. Reducing the national achievement gap will require high expectations from *all* states.
- What if Massachusetts used a performance standard comparable to the one in Tennessee? The Massachusetts Grade 8 mathematics Proficient standard, for example, is at the 55th percentile. If Massachusetts used a Proficient standard comparable in difficulty to the Tennessee proficient standard, it would be at the 4th percentile. This is a dramatic illustration of how far apart the performance standards are among the states.

The report also found that success under No Child Left Behind is largely related to using low performance standards. For example, in Grade 8 mathematics, the stringency of the state performance standards had a negative correlation of about -.81 with the number of proficient students reported by the state. The states reporting the highest numbers of proficient students have the lowest performance standards. More than 60% of 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 are doing well and feel no urgency to improve because almost all their students are proficient. They have a type of Lake Woebegone delusion where they have no idea how they stack up when compared with peers outside their own state.

The report also estimated how the 2007 state results reported to No Child Left Behind would have looked had all the states used an internationally benchmarked *common performance standard*. Under this approach, all the states would have reported their percent Proficient based on a level playing field. When the data were reanalyzed on the basis of a level playing field, there was a dramatic drop among the states reporting the highest levels of proficiency. For example, in Grade 8 mathematics, Tennessee dropped from 88% to 21% and Massachusetts went from being one of the lowest performing states to the highest achieving state in the nation.

The report shows that the No Child Left Behind paradigm of encouraging each state to set a different performance standard is fundamentally flawed and misleading. The big policy problem associated with the current No Child Left Behind state testing paradigm is lack of transparency. Test results across the 50 states are not comparable, any inference about national progress is impossible, and we cannot even determine if progress in one state is greater than progress in another state. Transparency in measurement is the most fundamental requirement for determining success in educational programs. 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

[1] The data in this report are from 2007. In subsequent years, some states may have raised performance standards and some may have lowered them.

[2] The standard deviation is a measure of how far apart the performance standards are or how large the expectations gap is. Using Massachusetts as a reference state, and Grade 8 as an example, the largest expectations gap is between Massachusetts and Tennessee.

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-ready and career-ready skills. If almost all students are proficient, what is the motivation to teach them higher level skills? This may be the main reason why almost 40% of students entering college need remedial courses. They thought they were college ready because they passed their high school graduation test, but they were not.

In order to reduce the expectations gap, this report recommends that the current standard-setting paradigm used by the states be reengineered. Rather than deriving performance standards exclusively from internal state content considerations, the report recommends a new method for setting standards that is influenced more by empirical data. The *Benchmark Method* (Phillips, 2011) of standard setting starts with empirical data rather than ending with it. The Benchmark Method acknowledges 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 national and global economic and technological world. Content considerations are used to describe the performance standard, but content is not the primary driver of how high or low the standard should be. Instead, the benchmark is the primary driver in determining whether the performance standard is high enough to allow students to compete in a national and international context. The report recommends that the Benchmark Method of standard setting be used in the future if states function as a consortium with funding from the federal Race to the Top assessment program. After states adopt and implement the Common Core State Standards Initiative developed by the Council of Chief State School Officers (CCSSO) and the National Governors Association (NGA), they will need to establish common performance standards. At this stage, the Benchmark Method could help guarantee consistently high, internationally competitive, performance standards.

International Benchmarking and National Education Policy

The need for high internationally competitive education standards has recently been emphasized by national policymakers. A recent report by the National Governors Association (NGA), the Council of Chief State School Officers (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)."*

The President of the United States (Barack Obama, in a speech to the U.S. Hispanic Chamber of Commerce, 2009) recognizes the need for high and consistent standards. He has 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 8 years 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 AYP toward meeting state standards. When states set performance standards, however, they generally have little knowledge of how those state performance standards compare with national standards, such as those used on the National Assessment of Educational Progress (NAEP). Even more important, they have no understanding of how their state performance standards compare with international standards, such as those used on the Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), and Program for International Student Assessment (PISA).

International Benchmarking

International benchmarking is one way to calibrate the difficulty level of state performance standards. What do we mean by *international benchmarking state performance standards?* Understanding international benchmarking requires first understanding national benchmarking. When states establish performance standards (e.g., the Proficient level), they need to know how the state standards compare with national standards. This provides a national benchmark for the state performance standard. NAEP has recently provided national benchmarks through the 2007 state mapping study (Bandeira de Mello, Blankenship, and McLaughlin, 2009). These benchmarks were obtained for states by linking their state tests to state NAEP and thereby placing their state performance standards on the NAEP scale. States can then determine how their own state performance standards compare with NAEP national performance standards (e.g., Basic, Proficient, and Advanced). The linking procedure provides the states with NAEP scores that are equivalent to the performance standards on their state tests (referred to as *NAEP-equivalent scores*).

The international benchmarking in this report piggybacked on the National Center for Education Statistics (NCES) study by taking the linking one step further and linking the state test to TIMSS or PIRLS.[3] 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

Three assessments collect international data, and therefore could provide the data needed for international benchmarks. Two of these are 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.

[3] See Appendix A for details of the statistical linking of NAEP to TIMSS and PIRLS. Appendix B reports the *TIMSS equivalents* and *PIRLS equivalents* for state proficient performance standards. Appendix C provides evidence of the validity of the linking, using data from the states of Massachusetts and Minnesota.

The third survey is PISA, sponsored by the Organization for Economic Cooperation and Development (OECD), located in Paris. 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 (2010).

Expressing International Benchmarks as Grades

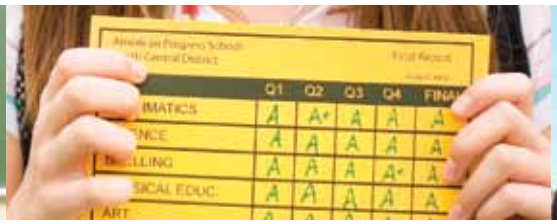
International benchmarks using TIMSS and PIRLS can be obtained by states by statistically linking their state tests to the state NAEP, then linking NAEP to TIMSS or PIRLS. This process of *chain linking* places the state's own performance standards on the TIMSS or PIRLS scale. States can then determine 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 Grade 4 reading in PIRLS.

To facilitate communication, this report will re-label 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[4]

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

[4] The grade designations in this report are slightly different from those in a previous report by the author (Phillips, 2009). In the previous report, some of the grades were determined by statistical criteria. In this report, all the grades represent equal 25-point intervals on the TIMSS scale.


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 we can do this, we need to compare Figures 1 through 3 with Figures 4 through 6. Figures 1 through 3 display the percent of proficient students reported by the states in 2007 in Grades 4 and 8 mathematics and Grade 4 reading. The percent proficient is the state results for spring 2007 under the federal reporting requirements of No Child Left Behind. The 2007 percent proficient results were first reported in the NCES 2007 state mapping study (Bandeira de Mello, Blankenship, and McLaughlin, 2009) and can be found at the U.S. Department of Education Web site at **http://www.ed.gov/admins/lead/account/consolidated/sy06-07part1/index.html.** Using Grade 8 mathematics as an example, as shown in Figure 2, we see that the state with the greatest number of proficient students reported under No Child Left Behind is Tennessee, whereas the number 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 Tennessee. But there is something wrong with this picture. We know that NAEP reports exactly the opposite, with Massachusetts the highest achieving state and Tennessee one of 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 Proficient performance standard. 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 4 and 5 by converting the state performance standards to the metric of TIMSS (i.e., the TIMSS equivalent of the state performance standard in mathematics) and in Figure 6 by 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 the TIMSS and PIRLS achievement of the typical student in the highest performing countries.

Once the state performance standards are expressed on a common metric (i.e., the TIMSS or PIRLS scale), the range in difficulty from the lowest to the highest performance standard is incredible. Using Grade 8 mathematics as an example, the lowest TIMSS equivalent of the Proficient performance standard was in Tennessee (408) and the highest was in Massachusetts (557).[5] The Massachusetts proficient standard was 149 units higher than the Tennessee proficient standard. This gap in expectations is about 2 standard deviation units on the TIMSS scale. In many states, a difference this large represents more than four grade levels.

The four grade level difference can be demonstrated if we look at the differences in performance standards between Massachusetts and Tennessee, using the NAEP metric in mathematics (these data are reported in the 2007 NCES State Mapping Study, 2009). The Tennessee 8th-grade NAEP-equivalent performance standard (234) is substantially below the Massachusetts 4th-grade NAEP-equivalent performance standard (254). This is further reinforced by the fact that the average NAEP scores of 4th-grade students in Massachusetts (252) are above the Tennessee 8th-grade NAEP-equivalent performance standard (234).

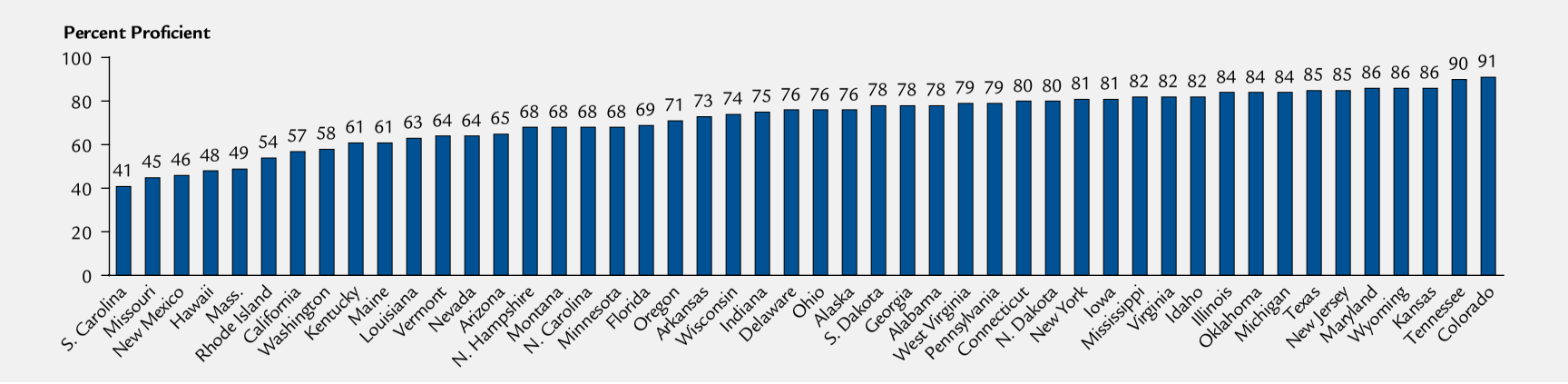
Comparing the international benchmarks in Figures 4 through 6 to the percent proficient in Figures 1 through 3 shows why so many states can claim so many proficient students for federal reporting requirements. These states are using low standards to define *proficiency*. For example, in Grade 8 mathematics, seven states only require a D or D+ to be considered Proficient. Massachusetts, on the other hand, has the highest performance standard in the county, 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 -.77, -.81, and -.78 for Grades 4 and 8 in mathematics, and Grade 4 in reading, respectively. 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 reported by No Child Left Behind is largely determined by how low a state sets its performance standards.

We should note that not all states are achieving high rates of proficiency by lowering their standards. For example, Hawaii is a small and relatively poor state that has made the right policy decision, which is in the best interest of its children, by requiring high standards in Grade 8 mathematics (slightly lower than those in Massachusetts). Over the past several years, Hawaii's leadership has maintained the high standards and the student performance in Hawaii has gradually improved (as indicated by their NAEP scores).

[5] See Appendix A for the TIMSS equivalents for Grades 4 and 8 in mathematics and Grade 4 in reading. In Grade 8 mathematics, the TIMSS equivalent of 408 in Tennessee is equal to a grade of D and the TIMSS equivalent of 557 in Massachusetts is equal to a grade of B.

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

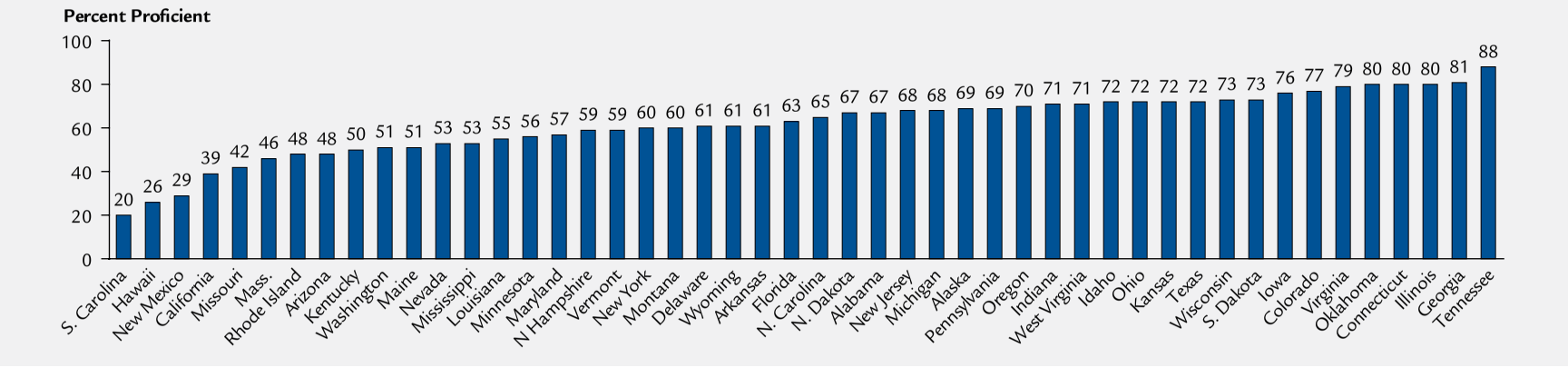
Percent Proficient Based on State Performance Standards, 2007, Grade 4, Mathematics



Source: Phillips, 2010, *International Benchmarking State Education Performance Standards,* AIR, Washington, DC.

Figure 2: Percent Proficient Based on State Performance Standard, Mathematics, Grade 8

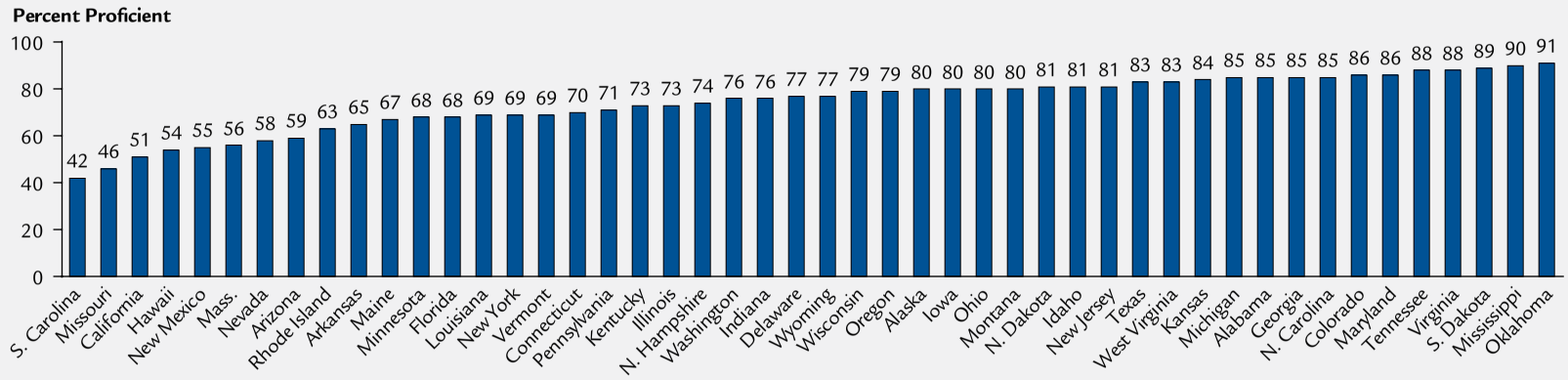
Percent Proficient Based on State Performance Standards, 2007, Grade 8, Mathematics



Source: Phillips, 2010, *International Benchmarking State Education Performance Standards,* AIR, Washington, DC.

Figure 3: Percent Proficient Based on State Performance Standard, Reading, Grade 4

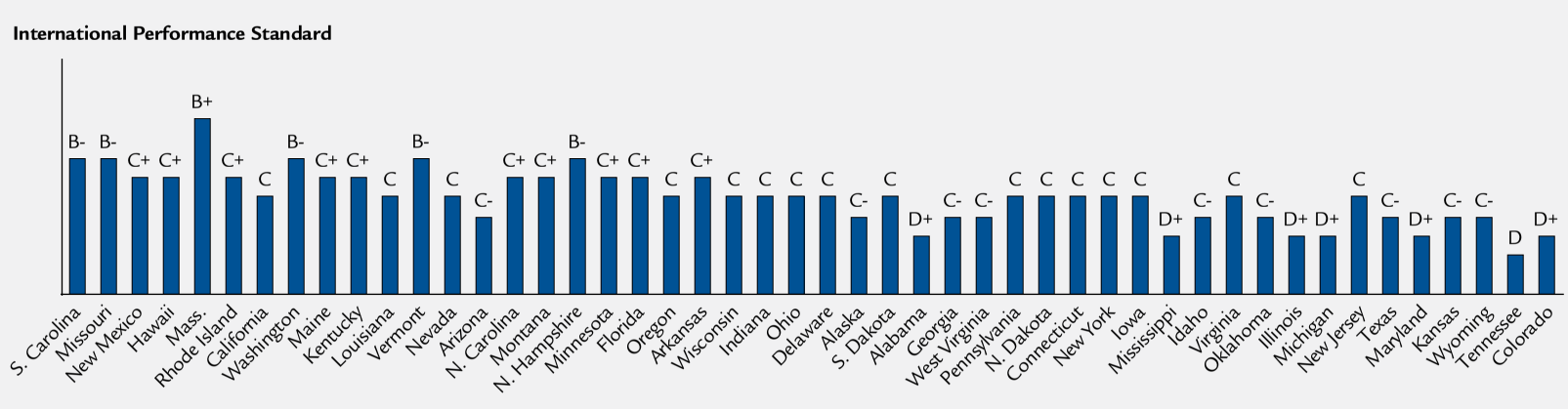
Percent Proficient Based on State Performance Standards, 2007, Grade 4, Reading



Source: Phillips, 2010, *International Benchmarking State Education Performance Standards,* AIR, Washington, DC.

Figure 4: International Benchmarks for Mathematics, Grade 4

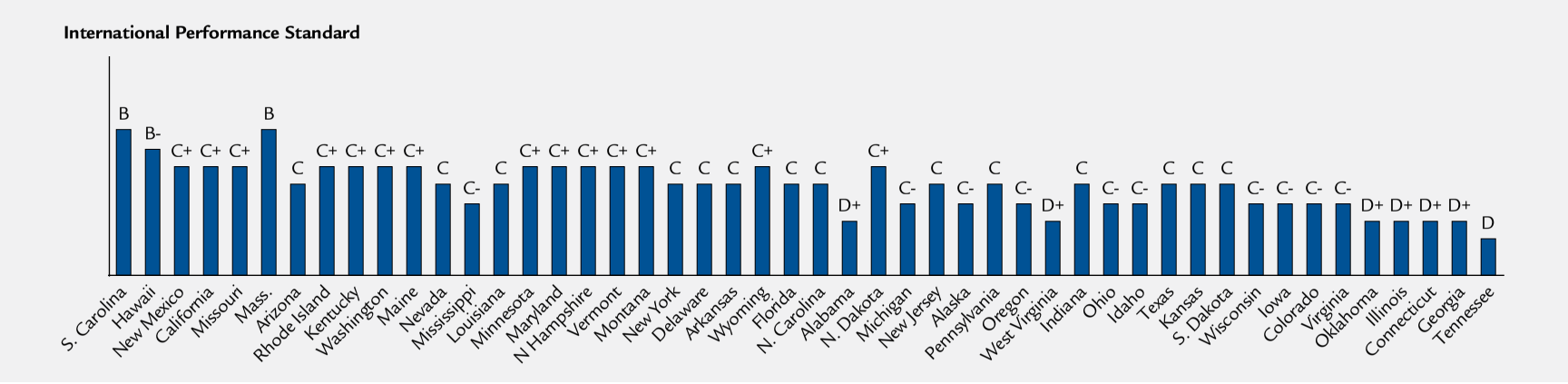
International Performance Standards Comparable to State Performance Standards, 2007, Grade 4, Mathematics



Source: Phillips, 2010, *International Benchmarking State Education Performance Standards,* AIR, Washington, DC.

Figure 5: International Benchmarks for Mathematics, Grade 8

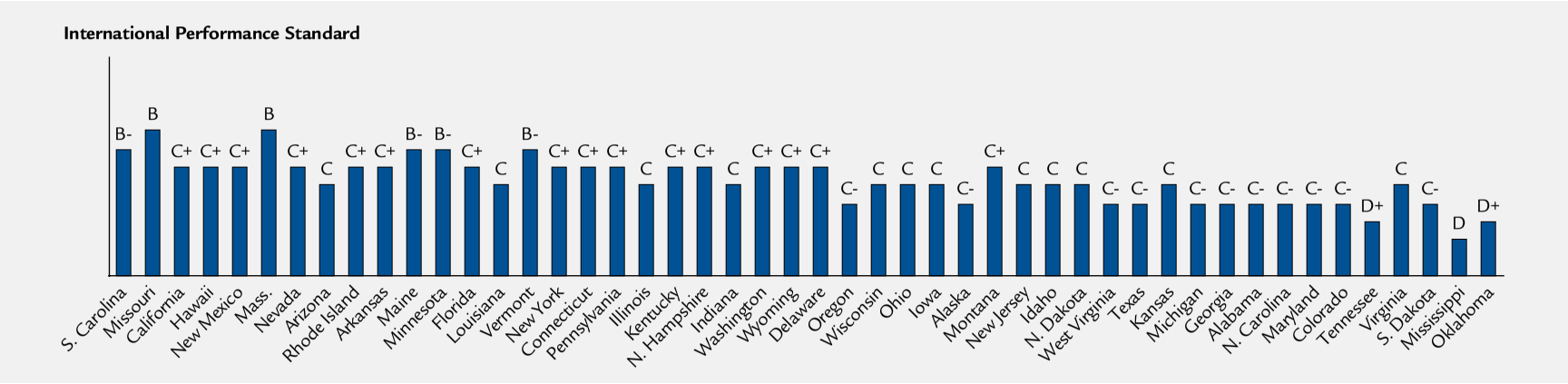
International Performance Standards Comparable to State Performance Standards, 2007, Grade 8, Mathematics



Source: Phillips, 2010, *International Benchmarking State Education Performance Standards,* AIR, Washington, DC.

Figure 6: International Benchmarks for Reading, Grade 4

International Performance Standards Comparable to State Performance Standards, 2007, Grade 4, Reading



Source: Phillips, 2010, *International Benchmarking State Education Performance Standards,* AIR, Washington, DC.

Which States Have World-Class Standards?

Using a standard of B to represent world class, we see that for Grade 4 mathematics, only Massachusetts had world-class mathematics standards. These standards are comparable to the mathematical skill and knowledge of the typical (or average) 4th-grade student in Japan, Taiwan, Singapore, and Hong Kong—the highest achieving countries on the 2007 TIMSS.

For Grade 8 mathematics, Massachusetts and South Carolina were the only states with world-class standards. These standards are comparable to the mathematical skill and knowledge of the typical (or average) 8th-grade student in South Korea, Japan, Taiwan, Singapore, and Hong Kong—the highest achieving countries on the 2007 TIMSS.

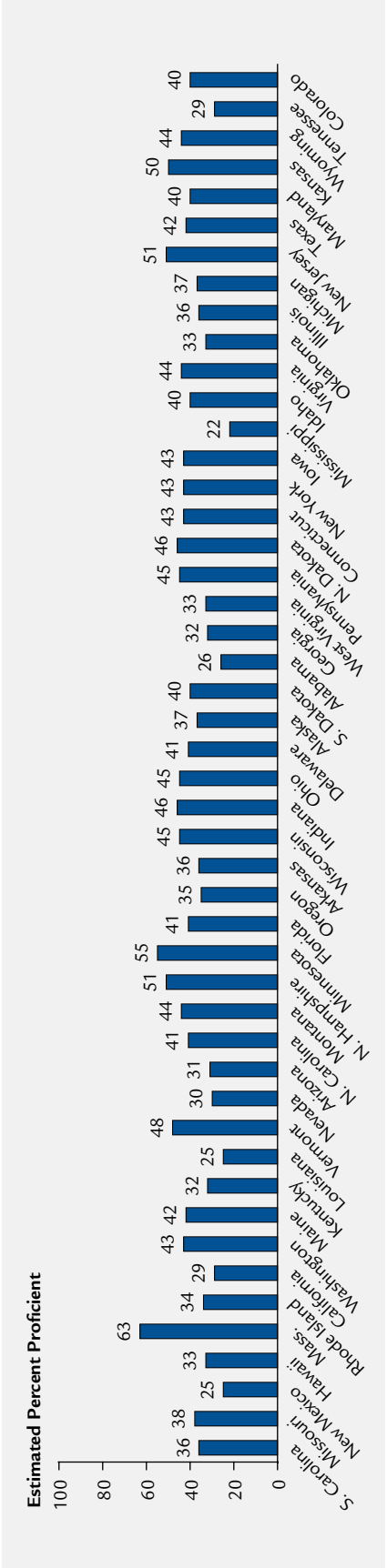
For Grade 4 reading, Massachusetts and South Carolina were the only states with world-class standards. These standards are comparable to the reading skills of the typical (or average) 4th-grade student in Hungary, Italy, Luxembourg, Singapore, Hong Kong, the Russian Federation, and Canada (Ontario, Alberta, and British Columbia). These were the highest achieving countries on the 2006 PIRLS.

Estimating State Performance With a Common Performance Standard

How would the 2007 state results reported to No Child Left Behind have looked had all the states used a common performance standard that had been internationally benchmarked to TIMSS or PIRLS

(e.g., the High international benchmark, or B)? Had the states used a common standard, then all would have reported their percent proficient on the basis of performance standards of comparable difficulty, using a level playing field. Figures 7 through 9 shows what this might have looked like. A common standard gives a dramatically different picture of state performance. High levels of percent proficient are no longer related to low levels of performance standards. Instead, high levels of proficiency are now related to high levels of academic achievement. For example, in Grade 8 mathematics the percent proficient in Tennessee drops to 21%, with Massachusetts now the highest achieving state in the nation. If parents were using the information shown in Figures 7 through 9 to choose a state in which to live so their children can attend the best schools, they might choose Massachusetts. The estimates of percent proficient have similar patterns for Grade 4 mathematics and Grade 4 reading. In each case, Massachusetts outperforms all other states.

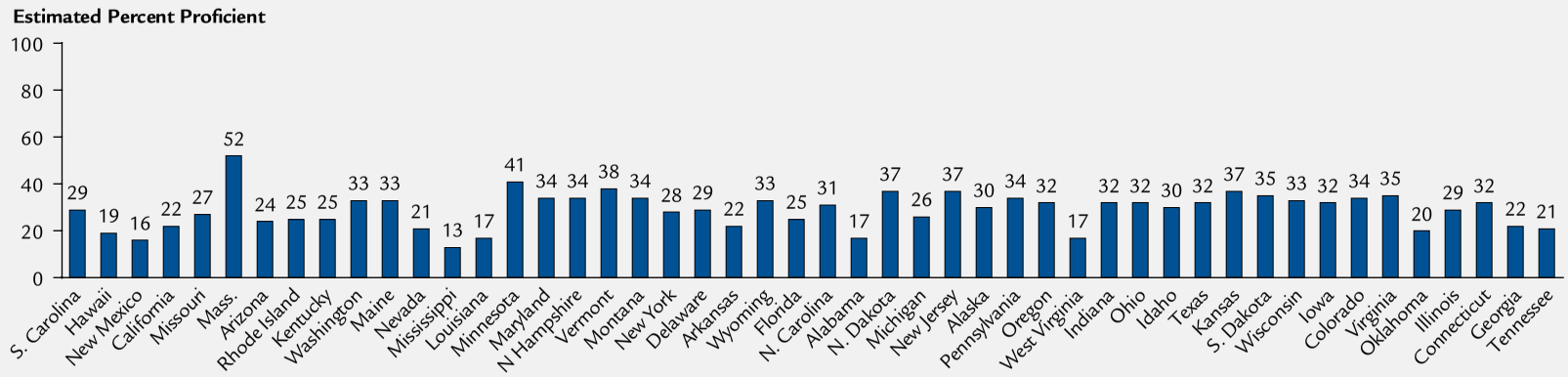
Figure 7: Estimate of Percent Proficient If All States Had Used an Internationally Benchmarked Common Performance Standard of B, Mathematics, Grade 4



Source: Phillips, 2010, *International Benchmarking State Education Performance Standards*, AIR, Washington, DC.

Figure 8: Estimate of Percent Proficient If All States Had Used an Internationally Benchmarked Common Performance Standard of B, Mathematics, Grade 8

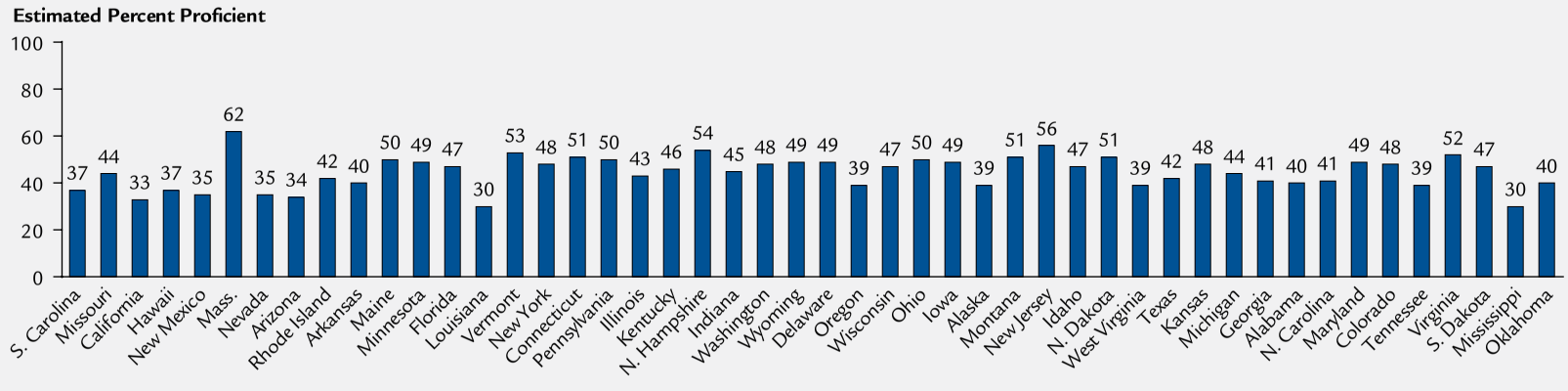
Estimated % Proficient if Each State Used an Internationally Benchmarked Common Standard of B, 2007, Grade 8, Math



Source: Phillips, 2010, *International Benchmarking State Education Performance Standards,* AIR, Washington, DC.

Figure 9: Estimate of Percent Proficient If All States Had Used an Internationally Benchmarked Common Performance Standard of B, Reading, Grade 4

Estimated % Proficient if Each State Used an Internationally Benchmarked Common Standard of B, 2007, Grade 4, Reading



Source: Phillips, 2010, *International Benchmarking State Education Performance Standards,* AIR, Washington, DC.

How To Fix This Problem: Reengineer Standard Setting

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 is recognition of the need for transparency. The Common Core project by CCSSO and NGA acknowledges that the nation cannot make 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 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 then they use low performance standards to increase the number of proficient students making AYP for No Child Left Behind. States need a way to set consistently high *performance* standards. This can only happen if the current standard-setting paradigm used in the testing industry is reengineered.

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. In state testing programs, the sequence of events for setting performance standards is pretty much routine. First, the state typically develops content standards (statements about the range of what students should learn, e.g., in reading and mathematics). Then the state develops *performance-level descriptors,* or statements about how much of the content standards students should learn. Finally, the state establishes performance standards (cut-scores on the test scale) that represent degrees of proficiency (e.g., Basic, Proficient, and Advanced). The performance standards are usually recommended by a broadly representative group of educators, business leaders, and other stakeholders. Throughout the testing industry, it is almost a religious mantra that the performance standards must be based on the content standards and performance-level descriptors and not be influenced by normative data.

Frequently used techniques like the Bookmark Method (Mitzel, Lewis, Patz, and Green, 2001) set the standards over two or three rounds, seeking convergence on a final standard. The use of empirical impact data (what percentage of students in the state would reach the chosen standard) is usually relegated to secondary importance in the standard-setting process. Impact data are often presented to the standard-setting panelists after round 1 or round 2, after they have already made up their minds about how high the standard should be. The research literature indicates that this practice of introducing impact data late in the process has almost no influence on the panelists' decisions. Rarely do the panelists change their minds as a result of impact data.

The problem with narrowly focused content-based standard-setting methods is that there is nothing in the standard-setting process that ensures that the performance standards are challenging. The panelists will usually believe that they are setting rigorous standards, basing their belief on the personal classroom experiences of the teachers and the anecdotal experiences of parents, business leaders, and other stakeholders on the panel. However, the content-based standard-setting methods are relatively impervious to the influence of empirical data. Internal state impact data are introduced too late in the process to make any real difference in the standard setters' deliberations. But even more important is the fact that there are almost never any national or international data used to help set nationally or internationally competitive standards. Instead, 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 have created a lack of credibility and lack of transparency in state and federal education reporting, have confused policymakers, and have misled the public in some states into believing that their students are proficient when they are not. In order to correct this problem, this report recommends a Benchmark Method (Phillips, 2011, in press) of setting standards that increases the chances that state standards will be consistently high and nationally and internationally competitive.

The Benchmark Method of Standard Setting

The Benchmark Method of standard setting starts with empirical data rather than ending with them. This method acknowledges that performance standards are fundamentally a policy-judgment decision (not just a content decision) that needs to be guided by knowledge of the real world and the requirements U.S. students will face as they compete in a national and global economic and technological world. Content considerations are used to guide and describe the standard that is set, but content is not the primary driver of how high or low the standard should be. In a nutshell, the Benchmark Method of standard setting would use the following steps:

1. ***Content standards:*** A broad consensus on content standards (e.g., statewide content standards) is established and helps guide the scope and sequence of curriculum and teaching strategies and test development blueprints. A large pool of items that is representative of the content standards is developed and field-tested.
2. ***Test development:*** Items are assembled into a test form that covers the content standards and matches the test blueprint. For standard-setting purposes, the items in the test are often ordered from easy to hard in a document referred to as the ordered-item booklet.
3. ***Benchmarking through statistical linkages:*** The state test scale is statistically linked to national and/or international scales (see Phillips and Jiang, 2010). The statistical linkage is used to determine national or international benchmarks on the state test scale (these are the performance-standard equivalents on the state test that are comparable in difficulty to the performance standards on the national or international test). The statistical linkage allows the benchmarks to be expressed as page numbers in the ordered-item booklet.
4. ***Internationally benchmarked performance-level descriptors:*** Content specialists use the state content standards and the items in the ordered-item booklet, mapped to the performance levels on the national or international test, to develop a performance-level descriptor that describes what students know and can do on the state test. The key concept in the Benchmark Method of standard setting is that the performance-level descriptors represent a performance standard on the state test that is comparable to the rigor of the performance

standard on the national or international test. The state then makes a policy decision on whether the benchmarked performance-level descriptor (on the state test) represents what the state wants their students to know and be able to do (e.g., in order to be considered proficient in mathematics). The performance levels should be challenging but achievable for most of the students in that grade. Once this process is complete, the performance-level descriptors will represent the policy vision of the state as to how high the performance standard should be and how much students need to know and be able to do in order to reach that standard. The policy vision will be informed by external referents that help the state know whether the expectations are reasonable, achievable, and nationally and internationally competitive.

5. ***Standard-setting panel:*** Once the performance-level descriptors are drafted, the next step is to find the specific cut-score on the state scale that best represents each performance level. Panelists review the content standards and the performance-level descriptors, and make recommendations on where the cut-score should be on the test. On the basis of content and other considerations, the panelists can lower or raise the cut-score, but now they do this with the full knowledge that they are going below or above the benchmark.

The $350 million from the Race to the Top assessment program and the reauthorization of the Elementary and Secondary Education Act (ESEA) could provide an unprecedented opportunity for states to improve their testing programs. 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 high enough for 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), then state performance standards would be consistent and more on a par with the high standards used by national and international surveys such as NAEP, TIMSS, PIRLS, and PISA.

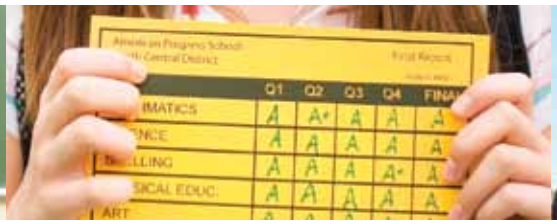
Conclusion

The overall finding in the study is that the difference in the stringency of the performance standards used across the states is far greater than most policymakers realize. The difference between the state with the highest standards and the state with the lowest standards was more than 2 standard deviations. This difference is so great that it is more than twice the size of the national black–white achievement gap (which is on the order of 1 standard deviation). In many state testing programs, a gap this large may represent as much as four grade levels.

The report also 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 percent 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. More than 60% of variation in state success reported by No Child Left Behind is because of the way in which the states set their performance standards.

This report also estimated how the 2007 state results reported to NCLB would have looked had all the states used an internationally benchmarked common performance standard. Had this been the case, then all the states would have reported their percent proficient on the basis of a level playing field. When the data were reanalyzed on this basis, there was a dramatic drop in percent proficient among the states reporting the highest levels of proficiency.

This paper 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, inference about national progress is not possible, 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 students the opportunity to learn college-ready and career-ready skills. If almost all students are proficient, what is the motivation to teach them higher level skills?


References

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. (2009). *Mapping state proficiency standards onto NAEP scales: 2005–2007* (NCES 2010-456). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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.

Lee, J., Grigg, W., & Dion, G. (2007). *The nation's report card: Mathematics 2007* (NCES 2007-494). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Lee, J., Grigg, W., & Donahue, P. (2007). *The nation's report card: Reading 2007* (NCES 2007-496). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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., & Foy, P. (2008). *TIMSS 2007 international mathematics report.* Boston: Lynch School of Education–Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries.* Chestnut Hill, MA: Boston College.

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.

Phillips, G. W. (2009). *The second derivative: International benchmarks in mathematics for American states and school districts.* Washington, DC: American Institutes for Research.

Phillips, G. W., & Jiang, T. (2010). Statistical methods for international benchmarking performance standards (under journal review).

Phillips, G. W. (2011, in press). The Benchmark Method of standard setting. In G. Cizek (Ed.), *Setting performance standards* (2nd ed.). New York: Routledge.

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

Appendix A

Statistically Linking NAEP to TIMSS and PIRLS

This report uses the statistical-linking procedures outlined in Johnson and colleagues (2005). One major difference is that this report uses extant statistics from the NAEP 2007, TIMSS 2007, and PIRLS 2006 published reports, and the 2007 NAEP reports in mathematics and reading, rather than recalculating them from the public-use data files and plausible values available from the NAEP, TIMSS, and PIRLS assessments.

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

$$z = \hat{A} + \hat{B}(x)$$
$$= \left(\hat{\mu}_Y - \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \hat{\mu}_X \right) + \left(\frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \right) x \qquad (0.1)$$

The z is the TIMSS equivalent (or PIRLS equivalent) of The NAEP equivalent (x). The NAEP equivalent is obtained from the NCES 2007 Mapping Study (Bandeira de Mello, Blankenship, and McLaughlin, 2009). In equation (0.1) $\hat{A}$ is an estimate of the intercept of a straight line, and $\hat{B}$ is an estimate of the slope defined by

$$\hat{A} = \hat{\mu}_Y - \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \hat{\mu}_X \qquad (0.2)$$

$$\hat{B} = \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \qquad (0.3)$$

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 of the state proficient standard for each state can be determined through the following equation:

$$\hat{\sigma}_z^2 = \hat{B}^2 \hat{\sigma}_x^2 + \hat{\sigma}_A^2 + 2(x)\hat{\sigma}_{AB} + (x)^2 \hat{\sigma}_B^2 \qquad (0.4)$$

The error variance term $\hat{\sigma}_x^2$ in equation (0.4) is the linking error variance from the NCES 2007 State Mapping Study. According to Johnson and colleagues (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).

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

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

$$\hat{\sigma}_B^2 = \hat{B}^2\left[\frac{Var(\hat{\sigma}_Y)}{\hat{\sigma}_Y^2} + \frac{Var(\hat{\sigma}_X)}{\hat{\sigma}_X^2}\right] \qquad (0.5)$$

Equations (0.4) and (0.5) were used with data in the United States linking sample to derive the estimates of linking error variance in this paper.

The statistics needed to use equations (0.1) through (0.5) are contained in the tables below.

Table 2: Means and Standard Deviations for National Samples of Grade 4 TIMSS 2007 and NAEP 2007 in Mathematics

	Mean	Error of mean	Standard deviation	Error of standard deviation
TIMSS 2007, Math, Grade 4	529.00	2.45	75.33	1.76
NAEP 2007, Math, Grade 4	239.72	0.17	28.63	0.10

Sources: Mullis, Martin, and Foy, 2008; Lee, Grigg, and Dion, 2007.

Table 3: Means and Standard Deviations for National Samples of Grade 8 TIMSS 2007 and NAEP 2007 in Mathematics

	Mean	Error of mean	Standard deviation	Error of standard deviation
TIMSS 2007, Math, Grade 8	508.45	2.83	76.74	2.04
NAEP 2007, Math, Grade 8	281.35	0.27	36.07	0.13

Sources: Mullis, Martin, and Foy, 2008; Lee, Grigg, and Dion, 2007.

Table 4: Means and Standard Deviations for National Samples of Grade 4 PIRLS 2006 and NAEP 2007 in Reading

	Mean	Error of mean	Standard deviation	Error of standard deviation
PIRLS 2006, Reading, Grade 4	539.93	3.55	74.06	2.56
NAEP 2007 Reading, Grade 4	220.99	0.26	35.73	0.15

Sources: Mullis, Martin, Kennedy, and Foy, 2007; Lee, Grigg, and Donahue, 2007.

The parameter estimates $\hat{A}$ and $\hat{B}$ are indicated in Table 5 through Table 7. These are the intercepts and slopes, respectively, needed to re-express NAEP results on the TIMSS or PIRLS scale.

Table 5: Estimating TIMSS 2007 Mathematics From NAEP 2007, Mathematics, Grade 4

	Estimates of Linking Parameters A and B		
	A		B
Parameter	-101.79		2.63
Standard error	15.13		0.06
Covariance		-0.93	

Table 6: Estimating TIMSS 2007 Mathematics From NAEP 2007, Mathematics, Grade 8

	Estimates of Linking Parameters A and B		
	A		B
Parameter	-90.13		2.13
Standard error	16.29		0.06
Covariance		-0.91	

Table 7: Estimating PIRLS 2006 Reading From NAEP 2007, Reading, Grade 4

	Estimates of Linking Parameters A and B		
	A		B
Parameter	81.79		2.07
Standard error	16.33		0.07
Covariance		-1.15	

Appendix B

State Proficient Standards Expressed in the Metric of TIMSS

This appendix provides the TIMSS equivalents and PIRLS equivalents of the state proficient performance standards used for reporting to NCLB in 2007. For example, in Table 8, 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). The standard error of the TIMSS equivalents and PIRLS equivalents includes both the linking error from the equipercentile linking of the state tests to state NAEP in the NCES 2007 mapping report (Bandeira de Mello, Blankenship, and McLaughlin, 2009) and the linking error from the statistical moderation linking of NAEP to TIMSS and PIRLS.

Table 8: International Benchmarks Based on the TIMSS Equivalents of State Proficient Standards, Mathematics, Grade 4, 2007

State	TIMSS equivalent of state proficient standard	Standard error of TIMSS equivalent	International benchmark level of state proficient standard	Phillips International Benchmark Grade
Massachusetts	580	5	High	B+
Missouri	543	4	Intermediate	B-
New Hampshire	527	4	Intermediate	B-
South Carolina	543	4	Intermediate	B-
Vermont	527	4	Intermediate	B-
Washington	529	4	Intermediate	B-
Arkansas	500	4	Intermediate	C+
Florida	503	4	Intermediate	C+
Hawaii	524	3	Intermediate	C+
Kentucky	502	4	Intermediate	C+
Maine	519	4	Intermediate	C+
Minnesota	523	4	Intermediate	C+
Montana	513	4	Intermediate	C+
North Carolina	506	3	Intermediate	C+
New Mexico	511	4	Intermediate	C+
Rhode Island	519	4	Intermediate	C+
California	492	4	Intermediate	C
Connecticut	478	4	Intermediate	C
Delaware	491	4	Intermediate	C
Indiana	498	4	Intermediate	C
Iowa	476	5	Intermediate	C
Louisiana	485	5	Intermediate	C
North Dakota	492	4	Intermediate	C
Nevada	487	4	Intermediate	C
New Jersey	476	4	Intermediate	C
New York	475	4	Intermediate	C
Ohio	490	5	Intermediate	C
Oregon	477	4	Intermediate	C
Pennsylvania	485	4	Intermediate	C
South Dakota	488	4	Intermediate	C
Virginia	475	4	Intermediate	C
Wisconsin	483	7	Intermediate	C
Alaska	467	5	Low	C-
Arizona	460	5	Low	C-
Georgia	460	4	Low	C-
Idaho	470	4	Low	C-
Kansas	474	5	Low	C-
Oklahoma	460	5	Low	C-
Texas	469	4	Low	C-
West Virginia	469	5	Low	C-
Wyoming	467	4	Low	C-
Alabama	438	6	Low	D+
Colorado	426	6	Low	D+
Illinois	445	4	Low	D+
Maryland	441	5	Low	D+
Michigan	434	6	Low	D+
Mississippi	436	4	Low	D+
Tennessee	419	5	Low	D

Table 9: International Benchmarks Based on the TIMSS Equivalents of State Proficient Standards, Mathematics, Grade 8, 2007

State	TIMSS equivalent of state proficient standard	Standard error of TIMSS equivalent	International benchmark level of state proficient standard	Phillips International Benchmark Grade
Massachusetts	557	7	High	B
South Carolina	574	5	High	B
Hawaii	536	4	Intermediate	B-
North Dakota	503	4	Intermediate	C+
Wyoming	504	4	Intermediate	C+
Montana	508	5	Intermediate	C+
Vermont	514	4	Intermediate	C+
New Hampshire	511	4	Intermediate	C+
Maryland	501	5	Intermediate	C+
Minnesota	518	4	Intermediate	C+
Maine	518	4	Intermediate	C+
Washington	518	4	Intermediate	C+
Kentucky	504	4	Intermediate	C+
Rhode Island	504	4	Intermediate	C+
Missouri	524	4	Intermediate	C+
California	508	4	Intermediate	C+
New Mexico	517	4	Intermediate	C+
South Dakota	486	4	Intermediate	C
Kansas	485	5	Intermediate	C
Texas	481	4	Intermediate	C
Indiana	476	5	Intermediate	C
Pennsylvania	487	4	Intermediate	C
New Jersey	489	4	Intermediate	C
North Carolina	484	4	Intermediate	C
Florida	476	4	Intermediate	C
Arkansas	498	5	Intermediate	C
Delaware	489	4	Intermediate	C
New York	490	4	Intermediate	C
Louisiana	478	4	Intermediate	C
Nevada	477	4	Intermediate	C
Arizona	480	4	Intermediate	C
Virginia	460	5	Low	C-
Colorado	461	5	Low	C-
Iowa	472	5	Low	C-
Wisconsin	467	5	Low	C-
Idaho	473	5	Low	C-
Ohio	474	4	Low	C-
Oregon	467	5	Low	C-
Alaska	474	4	Low	C-
Michigan	464	5	Low	C-
Mississippi	468	4	Low	C-
Georgia	426	6	Low	D+
Connecticut	446	6	Low	D+
Illinois	443	4	Low	D+
Oklahoma	439	5	Low	D+
West Virginia	449	4	Low	D+
Alabama	448	6	Low	D+
Tennessee	408	7	Low	D

Table 10: International Benchmarks Based on the PIRLS Equivalents of State Proficient Standards, Reading, Grade 4, 2007

State	PIRLS equivalent of state proficient standard	Standard error of PIRLS equivalent	International benchmark level of state proficient standard	Phillips International Benchmark Grade
Massachusetts	563	5	High	B
Missouri	552	5	High	B
Vermont	525	5	Intermediate	B-
Minnesota	527	5	Intermediate	B-
Maine	525	5	Intermediate	B-
South Carolina	543	5	Intermediate	B-
Montana	502	5	Intermediate	C+
Delaware	501	5	Intermediate	C+
Wyoming	505	5	Intermediate	C+
Washington	502	6	Intermediate	C+
New Hampshire	517	4	Intermediate	C+
Kentucky	506	6	Intermediate	C+
Pennsylvania	520	5	Intermediate	C+
Connecticut	524	5	Intermediate	C+
New York	515	5	Intermediate	C+
Florida	515	5	Intermediate	C+
Arkansas	523	5	Intermediate	C+
Rhode Island	518	5	Intermediate	C+
Nevada	510	5	Intermediate	C+
New Mexico	517	4	Intermediate	C+
Hawaii	521	5	Intermediate	C+
California	518	5	Intermediate	C+
Virginia	477	6	Intermediate	C
Kansas	479	6	Intermediate	C
North Dakota	499	5	Intermediate	C
Idaho	489	5	Intermediate	C
New Jersey	499	6	Intermediate	C
Iowa	493	6	Intermediate	C
Ohio	492	7	Intermediate	C
Wisconsin	482	6	Intermediate	C
Indiana	495	5	Intermediate	C
Illinois	496	5	Intermediate	C
Louisiana	482	7	Intermediate	C
Arizona	492	5	Intermediate	C
South Dakota	466	6	Low	C-
Colorado	469	6	Low	C-
Maryland	467	6	Low	C-
North Carolina	460	5	Low	C-
Alabama	453	6	Low	C-
Georgia	466	6	Low	C-
Michigan	451	7	Low	C-
Texas	471	6	Low	C-
West Virginia	459	6	Low	C-
Alaska	462	5	Low	C-
Oregon	467	7	Low	C-
Oklahoma	438	10	Low	D+
Tennessee	444	7	Low	D+
Mississippi	420	7	Low	D

Appendix C

Validity of International Benchmarking

The international benchmarking in this report depends on several statistical-linking studies. The first is the state test to state NAEP equipercentile linking study in the NCES report *Mapping State Proficiency Standards Onto NAEP Scales* (Bandeira de Mello, Blankenship, and McLaughlin, 2009). The second is the national NAEP to national TIMSS statistical-moderation-linking study reported in *The Second Derivative: International Benchmarks in Mathematics for American States and School Districts* (Phillips, 2009).

The international benchmarking in this report uses a chain-linking approach, in which the state test is linked to state NAEP and then the national NAEP is linked to national TIMSS or PIRLS. 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, there were two states (Massachusetts and Minnesota) in 2007 that provided some data to validate the linking. In 2007 both Massachusetts and Minnesota administered TIMSS to state-representative samples. Tables 11 through 14, below, show the differences between the estimates for Massachusetts and Minnesota from the national linking study (Phillips, 2009) versus the actual TIMSS estimates from the state TIMSS survey in 2007. If the linking done by Phillips (2009) is valid for states, then the estimates for Massachusetts and Minnesota, based on the national linking study, should yield aggregate statistics comparable to those provided by the actual TIMSS survey. As can be seen from these tables, the estimates based on linking were not perfect, but they were adequate in most cases. Among the comparisons contained in the tables below, the state estimates from the linking study were consistent with the actual TIMSS study in 17 of 20 comparisons.

Table 11: Accuracy of Linking Validated in Massachusetts, Grade 4, Mathematics

	Estimates from 2007 linking study	Standard error	Actual 2007 state TIMSS	Standard error	95% confidence interval	Statistically significant difference
Mean	562	3.4	573	3.5	4.03	Yes
% above A	17	2.8	18	1.7	0.59	No
% above B	58	3.6	63	2.5	2.45	No
% above C	91	1.8	92	2.2	0.77	No
% above D	99	0.3	99	1.2	-0.54	No

Note. A difference is significant if its associated confidence interval is less than -2.58 or greater than +2.58 (includes a Bonferoni adjustment to alpha based on the number of comparisons in the table, $\alpha = .025/5$).

Table 12: Accuracy of Linking Validated in Minnesota, Grade 4, Mathematics

	Estimates from 2007 linking study	Standard error	Actual 2007 state TIMSS	Standard error	95% confidence interval	Statistically significant difference
Mean	548	3.7	554	5.9	1.68	No
% above A	15	2.4	18	1.4	2.04	No
% above B	49	3.4	55	2.8	2.64	Yes
% above C	84	2.4	85	2.0	0.79	No
% above D	98	0.7	99	1.0	2.08	No

Note. A difference is significant if its associated confidence interval is less than -2.58 or greater than +2.58 (includes a Bonferoni adjustment to alpha based on the number of comparisons in the table, $\alpha = .025/5$).

Headquarters ■ 1000 Thomas Jefferson Street, NW ■ Washington, DC 20007-3835 ■ www.air.org

Table 13: Accuracy of Linking Validated in Massachusetts, Grade 8, Mathematics

	Estimates from 2007 linking study	Standard error	Actual 2007 state TIMSS	Standard error	95% confidence interval	Statistically significant difference
Mean	544	4.0	547	4.6	1.10	No
% above A	13	2.6	16	1.7	1.57	No
% above B	47	3.9	52	2.5	2.29	No
% above C	82	2.7	82	2.2	-0.26	No
% above D	97	0.8	95	1.2	-3.29	Yes

Note. A difference is significant if its associated confidence interval is less than -2.58 or greater than +2.58 (includes a Bonferoni adjustment to alpha based on the number of comparisons in the table, $\alpha = .025/5$).

Table 14: Accuracy of Linking Validated in Minnesota, Grade 8, Mathematics

	Estimates from 2007 linking study	Standard error	Actual 2007 state TIMSS	Standard error	95% confidence interval	Statistically significant difference
Mean	531	3.7	532	4.4	0.56	No
% above A	10	2.1	8	1.4	1.75	No
% above B	40	3.6	41	2.8	0.51	No
% above C	77	2.9	81	2.0	2.04	No
% above D	96	1.1	97	1.0	1.24	No

Note. A difference is significant if its associated confidence interval is less than -2.58 or greater than +2.58 (includes a Bonferoni adjustment to alpha based on the number of comparisons in the table, $\alpha = .025/5$).