Testimony for the
U.S. Senate Committee on Health, Education, Labor and Pensions

Dirksen Senate Office Building
First and C Streets NE, Washington, DC 20510, Room 430
Washington, DC 20002

Dr. Gary W. Phillips
Vice President and Chief Scientist
American Institutes for Research

April 28, 2010

Thank you Chairman Harkin and members of the committee for the invitation to be here today. My name is Gary W. Phillips, and I am a Vice President and Chief Scientist at the American Institutes for Research (AIR). AIR is a 65-year-old, not-for-profit, nonpartisan organization whose mission is to conduct behavioral and social science research to improve people's lives and well-being, with a special emphasis on the disadvantaged. Previously, I was the Acting Commissioner at the National Center for Education Statistics (NCES). My career has been devoted to providing policymakers with better data to help them improve American education.

Today I would like to make two central points about No Child Left behind and the reauthorization of the ESEA.

1. I will demonstrate that No Child Left Behind (NCLB) has a large loop hole that has misled the public and I encourage Congress to close this loop hole in the reauthorization of the ESEA. Other people will be providing you testimony on whether this legislative act will improve education. I will focus on whether this legislative act provides enough information to know if education has been improved.

2. I will propose that Congress encourage states to abandon their outmoded 20th century paper/pencil-based testing paradigm for a new generation of 21st century technology-based tests that are more accurate, less burdensome, faster, and cheaper.

What Is Wrong With No Child Left Behind?

The most significant thing wrong with NCLB is a lack of transparency. Contributing to this lack of transparency is the fact that the NCLB results represent state efforts to reach unattainable national goals. For the last quarter century, education reform professionals have known that our underachieving educational system has put our nation at risk (*A Nation at Risk: The Imperative for Education Reform,* April 1983). National policymakers have responded to this crisis with slogans and unattainable utopian goals, such as *"being the first in the world in mathematics and science achievement by 2000"* (1990 National Education Goals Panel); or *"all students will be proficient in reading and mathematics by 2014"* (No Child Left Behind Act of 2001); or "by 2020 . . . *ensure that every student graduates from high school well prepared for college and a career*" (A Blueprint for Reform: The Reauthorization of the Elementary and Secondary Education Act, 2010). A national goal should be high but reachable. A good example of a challenging but achievable national goal is the *Proficient* standard used by the National Assessment of Educational progress (NAEP) and the National Assessment Governing Board (NAGB). The *Proficient* standard is challenging but achievable by most (although not all) students. The new ESEA should contain career and college-ready national goals that are internationally competitive but not so high that they are unattainable by states and schools.

The greatest contributor to the lack of transparency in NCLB, however, is the misleading data used by policymakers to monitor progress toward the goals (referred to as Adequate Yearly Progress). Both the federal government and the states have an unfortunate history of presenting flawed state testing data to the public.

From 1984 to 1989, the U.S. Department of Education compared state performance using the Wall Chart that showed average state aggregates of SAT and ACT scores. The Wall Chart was used even though it was widely criticized because it measured only the self-selected college-bound population. The larger the percentage of the population taking the SAT or ACT tests, the lower the state's ranking on the Wall Chart. The states with the least number of students heading for college tended to have the highest ranking. In fact, the 1986 correlation between the SAT and the proportion of college-bound students was –0.86 (College Board, 1986). The fact that it was a misleading indicator due to self-selection did not deter the department from using the system for 6 years under two Secretaries of Education, Terrell H. Bell and William J. Bennett.

In 1987, a West Virginia physician produced a report in which he stated that he had found that on norm-referenced tests, all 50 states were claiming they were above the national average (Cannell, 1987). This so-called Lake Woebegone report sparked much interest in Washington because it was hoped that norm-referenced tests might overcome some of the problems of the SAT and ACT in the Wall Chart as indicators of state-by-state performance. Although this was a black eye for educators, the practice continues today. States are still asked to explain how they can be above the national average on their norm-referenced test when they are below the national average on the National Assessment of Educational progress (NAEP).

The biggest flaw in state testing data, however, is in use today in all states, sanctioned and encouraged by the No Child Left Behind Act of 2001. NCLB provides a new type of Wall Chart where again state aggregates are not comparable and are misleading. The most significant thing wrong with NCLB is a lack of transparency. The severe consequences of failing to meet AYP had the unintended consequence of encouraging states to lower, rather than raise, their own standards. The law inadvertently encouraged the states to dumb down their performance standards to get high rates of proficiency. The fact that states dumb down their performance standards can be seen in Figures 1 and 2 in this document. The “percent proficient” in these tables represent what was reported by NCLB in Grades 4 and 8 in mathematics in 2007. In my remaining remarks I will use Grade 8 to illustrate my points. In Grade 8 we see that Tennessee is the highest achieving state in the nation while Massachusetts is one of the lowest. If parents were looking to raise a family in a state with an excellent track record of success based on NCLB data, they should move their family to Tennessee. However, there is something wrong with this picture. We know that NAEP reports exactly the opposite with Massachusetts the highest achieving state and Tennessee being one of the lowest achieving states.

However, if we look deeper into state performance standards, we can begin to explain this contradiction. The grades imposed on the chart are from an upcoming AIR report titled “The Expectation Gap” that internationally benchmarked state proficient standards to the Trends in International Mathematics and Science Study (TIMSS) (Phillips, 2010). The report then expressed the international benchmarks as international grades. To do this I statistically linked the test in each state to the Trends in International Mathematics and Science Study (TIMSS) and expressed the state standards as international grades on a comparable scale. (*A* = Advanced, *B* =

High, *C* = Intermediate, *D* = Low). This gives policymakers an international benchmarked common metric by which to compare state performance standards. Returning to Grade 8 we see that many states obtain high levels of proficiency by lowering their standards. The states with the highest levels of proficiency require only a *D*, which is comparable to being below the *Basic* standard on NAEP and the lowest level of mathematics knowledge and skills on TIMSS. On the other hand, the states with the lowest levels of proficiency require the highest standards (where a *B* is comparable to the *Proficient* standard on NAEP and equal to the *High* level on TIMSS). In fact, the correlation between the percent proficient reported by the state under NCLB and the difficulty of their standards is -.81.

The gap in expectations in the state performance standards is not just a minor accounting irregularity. It has real equity consequences for a student's opportunity to learn. If my child attends school in a state where almost everyone is proficient, what leverage do I have as a parent to ask the state to provide a more challenging education? The gap in expectations has major educational consequences. The difference between the standards in Massachusetts and the standards of the states with the lowest standards is about 2 standard deviations. This gap in expectations is so large that I would like to take a minute impress on you just how large it is.

1. This *expectation gap* is so large that it is more than twice the size of the national black–white *achievement gap*. The nation will never be able to close the achievement gap until it closes the bigger problem of the expectations gap.

2. The gap in expectations represents two-to-three grade-level differences between what the states are expecting their students to know and be able to do. What the low standard- states are expecting in middle school is comparable in difficulty to what Massachusetts expected back in elementary school.

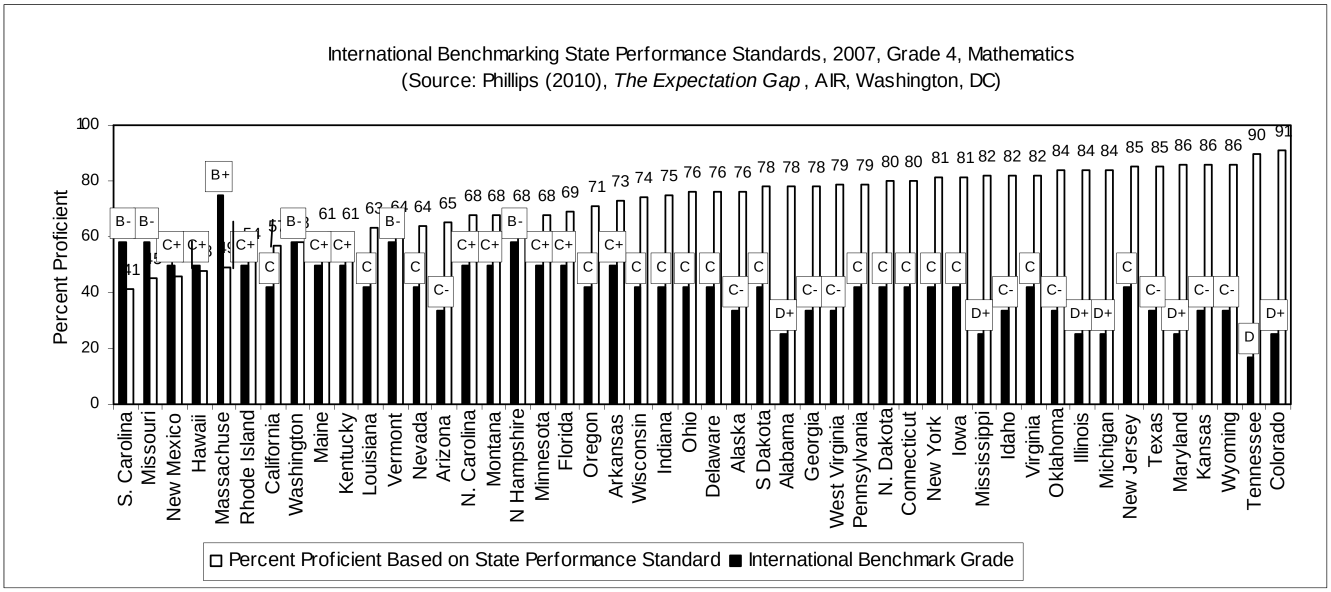
3. The Massachusetts proficient standard is at the 54th percentile. If Massachusetts used the Tennessee proficient standard in Massachusetts it would be at the 4th percentile.

This helps explain why the United States does poorly on international comparisons. Many States think they are doing well and feel no urgency to improve because almost all their students are proficient. They have no idea how they stack up when compared to peers outside their own

Lake Woebegone. This also helps explain 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.

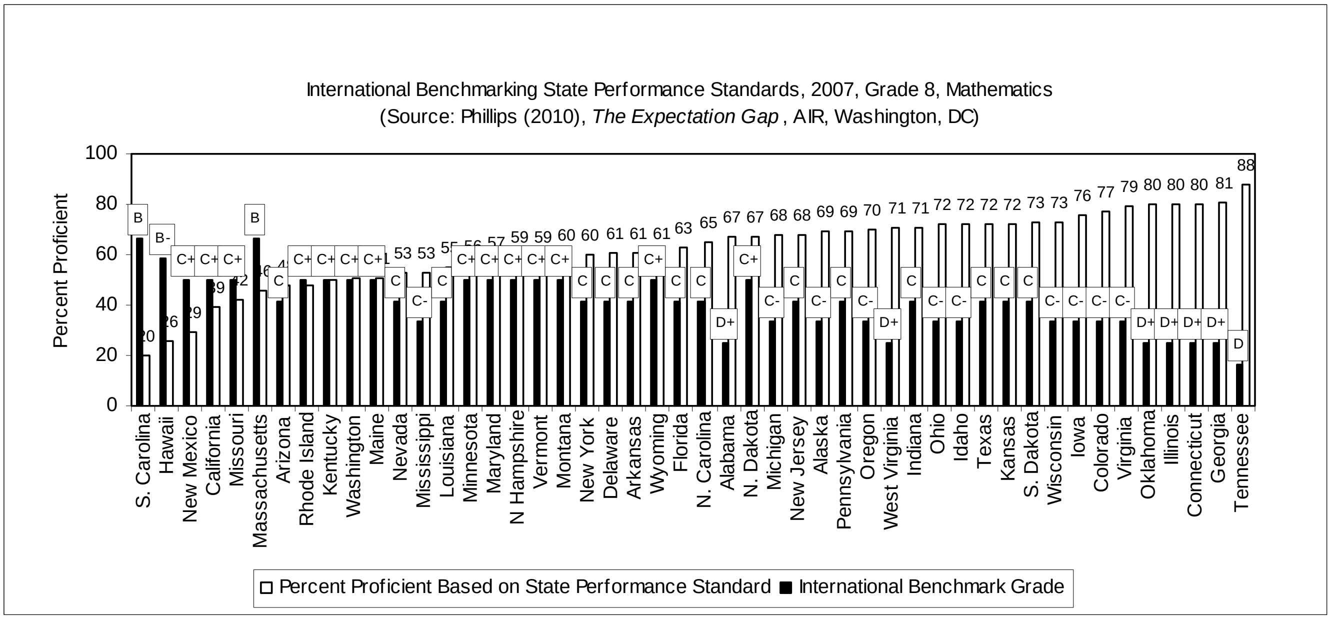
We should note that not all states are achieving high rates of proficiency by lowering their standards. For example, Hawaii is a small and relatively poor state that has made the right policy decision that is in the best interest of its children by requiring high standards (just under those in Massachusetts), although student performance is relatively low. Even though the state has been internally criticized for having too high standards, the state leadership has maintained the high standards and the student's performance in Hawaii have gradually improved (as indicated by their NAEP scores) over the years.

Figure 1: How High Were the State Proficient Standards in 2007 in Grade 4 Mathematics?



International Benchmarking State Performance Standards, 2007, Grade 4, Mathematics
(Source: Phillips (2010), *The Expectation Gap* , AIR, Washington, DC)

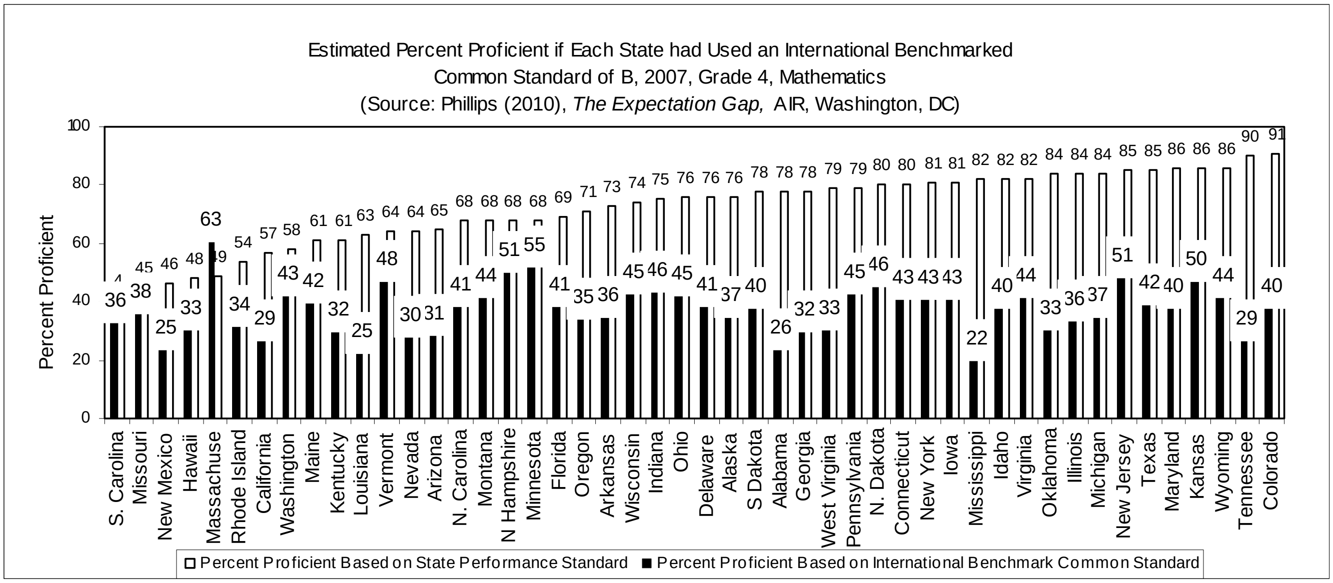
Figure 2: How High Were the State Proficient Standards in 2007 in Grade 8 Mathematics?



International Benchmarking State Performance Standards, 2007, Grade 8, Mathematics
(Source: Phillips (2010), *The Expectation Gap* , AIR, Washington, DC)

How would the 2007 state results reported to NCLB have looked had all the states used a *common performance standard* that was comparable in difficulty to the High International Benchmark on TIMSS? Had this been done, then all of the states would have reported their percent proficient based on performance standards of comparable difficulty using a level playing field. Figure 4 gives an example of what this might have looked like for Grade 8 mathematics—a dramatically different picture of state performance. We see that when all the states use an internationally competitive *common performance standard*, the performance in Tennessee drops from 88% to 21%. Now Massachusetts is the highest achieving state. If the parents mentioned above were using the information shown in Figure 4 to choose a state in which to live, where their children would attend schools with the highest educational expectations and achievement, they might choose Massachusetts.

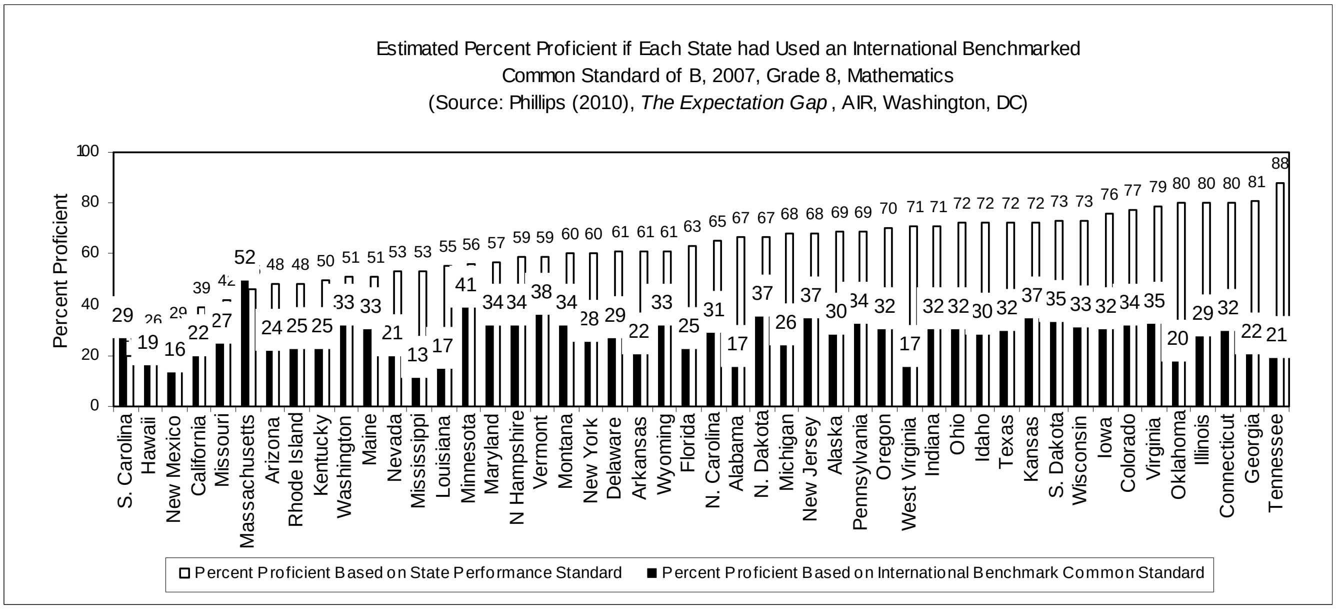
Figure 3: How Many Students Would Have Been Proficient if Each State Had Used a Level Playing Field in 2007 Grade 4 in Mathematics?



Estimated Percent Proficient if Each State had Used an International Benchmarked Common Standard of B, 2007, Grade 4, Mathematics

(Source: Phillips (2010), *The Expectation Gap,* AIR, Washington, DC)

Figure 4: How Many Students Would Have Been Proficient if Each State Had Used a Level Playing Field in 2007 Grade 8 in Mathematics?



Estimated Percent Proficient if Each State had Used an International Benchmarked Common Standard of B, 2007, Grade 8, Mathematics

(Source: Phillips (2010), *The Expectation Gap* , AIR, Washington, DC)

The Need for a New Generation of Technology-based State Testing

NCLB requires that states develop their own tests but does not provide funding for doing so. Therefore, states suffering budget cutbacks have no incentive to try new and better approaches to testing. The outdated pencil/paper tests are used in most states require costly and time-consuming administration, followed by costly and time-consuming scoring, followed by costly and time-consuming reporting. With spring testing, getting test results back to teachers and parents before the summer recess is nearly impossible. States like to claim they teach 21st century skills but they measure learning with 20th century tests. The only way state testing will move into the 21st century and take advantage of high-speed modern technology is with federal funding. Furthermore, the current model of one-size-fits-all, paper/pencil test provides poor measurement for much of the student population. The tests are too easy for high-achieving students and too hard for low-achieving students, students with disabilities, and English language learners.

The 350 million dollars from the Race to the Top Assessment Program and the reauthorization of the ESEA could provide an unprecedented opportunity for states to upgrade their testing capacity. In the near future, many states are likely to function as consortia and adopt the Common Core Standards developed by the Council of Chief State School Officers (CCSSO) and the National Governors Association (NGA). I would recommend that the ESEA encourage the consortia of states to use *Computer-Adaptive Testing* as their standard modus operandi.

Computer-adaptive tests are already in partial use in many states. However, in three states—Delaware, Hawaii, and Oregon—the entire state testing program are already computer-adaptive. Since AIR is the vendor for these three states I can speak with some authority on how their computer-adaptive tests operate. In all three of these states, the test consists of multiple-choice items and challenging constructed-response items that are both administered and scored by computer (no booklet printing cost and no scoring cost). The total cost of the computer-adaptive test is half that of a paper/pencil test. In each of these three states, the computer-adaptive test is developed based on universal design principles, and each test administered to a student covers all of the content standards. The technology platform provides three opportunities to take the summative test each year (used for accountability and federal reporting). In addition, the computer-adaptive test administers formative assessments (developed and used by teachers for diagnostic purposes) and interim assessments (used by teachers to get an early fix on how much students are progressing during the year) all on the same scale as the summative test. The results are available for each student within 15 seconds. Not only are these assessments faster and cheaper, but computer-adaptive testing yields more accurate measurement for high- and low-achieving students and better measurement for students with disabilities and English language learners.

What Should Be Included in the Reauthorization of ESEA?

Common content standards and *common performance standards* should be included in the reauthorization of ESEA. The CCSSO and the NGA are currently developing common content standards. Content standards represent the scope and sequence of content that should be taught in the schools. This is an important first step in creating transparency and accountability in ESEA. However, this needs to be followed by an equally important second step—establishing common performance standards. Performance standards represent how much of what is taught students are expected to learn. Because every student cannot learn everything that is taught in every grade and every subject, educators need a realistic performance goal. This performance standard needs to be common to all the states (or consortium of states) so that all the states have a level playing field. Each state does not get to set its own bar. The United States cannot be internationally competitive in our educational achievement if states are going in 50 different directions (different content standards) and have 50 different expectations of what their students should learn (different performance standards).

Computer-adaptive testing and the use of the best available modern technology should be included in the reauthorization of the ESEA. The reauthorized ESEA should encourage and fund states to use modern technology to administer, score, and report results. The best of all options is computer-adaptive testing that provides a more reliable measurement of student achievement involving less time, fewer items, and less cost. Computer-adaptive testing also provides better measurement for both high-achieving students and low-achieving students such as students with disabilities and English language learners.

In conclusion, I would like to thank you for the opportunity to give you my views on the next generation of state assessments. Setting internationally competitive education standards is a critical national priority. Students tomorrow will not be competing with the best students in their school. They will be competing with the best students in the world. In order to get states to establish high standards you must close the expectations loop hole in NCLB and reward states that set high internationally benchmarked standards. States also need federal funding in order to embrace the next generation of technology driven assessments. The technology for better, faster and cheaper testing already exists. National leadership is needed to move the states in this direction.

References

Bandeira de Mello, V. Blankenship, C, and McLaughlin, D (2009). *Mapping State Proficiency Standards onto NAEP Scales: 2005–2007* (NCES 2010-456). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Cannell, J. J., (1987), Nationally Normed Elementary Achievement Testing in America's Public Schools: How All 50 States Are Above the National Average, Friends for Education, Daniels, West Virginia.

College Board, (1986), *Press statement for release of 1986 SAT scores*. New York: The College Board.

Phillips, G. W., (2010), The Expectation Gap: Internationally Benchmarking State Performance Standards, American Institutes for Research, Washington, DC.