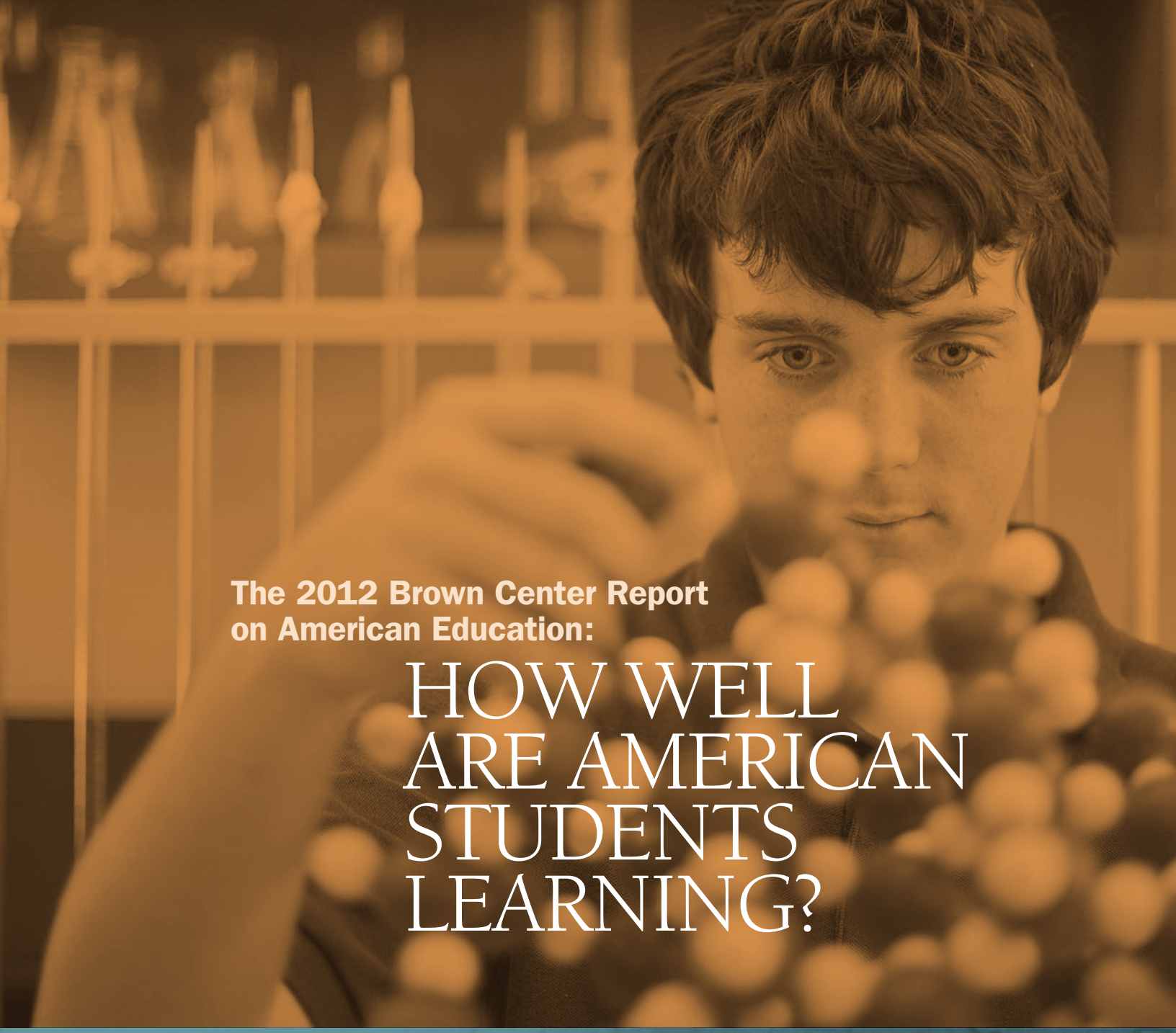


The 2012 Brown Center Report on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

With sections on predicting the effect of the Common Core State Standards, achievement gaps on the two NAEP tests, and misinterpreting international test scores.

BROWN CENTER on Education Policy
at BROOKINGS

ABOUT BROOKINGS

The Brookings Institution is a private nonprofit organization devoted to independent research and innovative policy solutions. For more than 90 years, Brookings has analyzed current and emerging issues and produced new ideas that matter—for the nation and the world.

ABOUT THE BROWN CENTER ON EDUCATION POLICY

Raising the quality of education in the United States for more people is imperative for society's well-being. With that goal in mind, the purpose of the Brown Center on Education Policy at Brookings is to examine the problems of the American education system and to help delineate practical solutions. For more information, see our website, www.brookings.edu/brown.

This report was made possible by the generous financial support of The Brown Foundation, Inc., Houston.

The 2012 Brown Center Report on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

With sections on predicting the effect of the Common Core State Standards, achievement gaps on the two NAEP tests, and misinterpreting international test scores.

February 2012
Volume III, Number 1

by:
TOM LOVELESS
Senior Fellow, The Brown Center on Education Policy

TABLE OF CONTENTS

Research assistance by:

MICHELLE CROFT

Copyright ©2012 by
THE BROOKINGS INSTITUTION
1775 Massachusetts Avenue, NW
Washington, D.C. 20036
www.brookings.edu

All rights reserved

THE 2012 BROWN CENTER REPORT ON AMERICAN EDUCATION

This edition of the Brown Center Report on American Education marks the first issue of volume three—and eleventh issue over all. The first installment was published in 2000, just as the Presidential campaigns of George W. Bush and Al Gore were winding down. Education was an important issue in that campaign. It has not been thus far in the current campaign for the Republican nomination (as of February 2012). And it is unlikely to be a prominent issue in the fall general election. Despite that, the three studies in this Brown Center Report investigate questions that the victor in the 2012 campaign, and the team assembled to lead the U.S. Department of Education, will face in the years ahead.

The first section is on the Common Core State Standards, a project that President Obama has backed enthusiastically. Forty-six states and the District of Columbia have signed on to the Common Core; detailed standards have been written in English language arts and mathematics; and assessments are being developed to be ready by the 2014–2015 school year. The first section attempts to predict the effect of the Common Core on student achievement.

Despite all the money and effort devoted to developing the Common Core State Standards—not to mention the simmering controversy over their adoption in several states—the study foresees little to no impact on student learning. That conclusion is based on analyzing states' past experience with standards and examining several years of scores on the National Assessment of Educational Progress (NAEP).

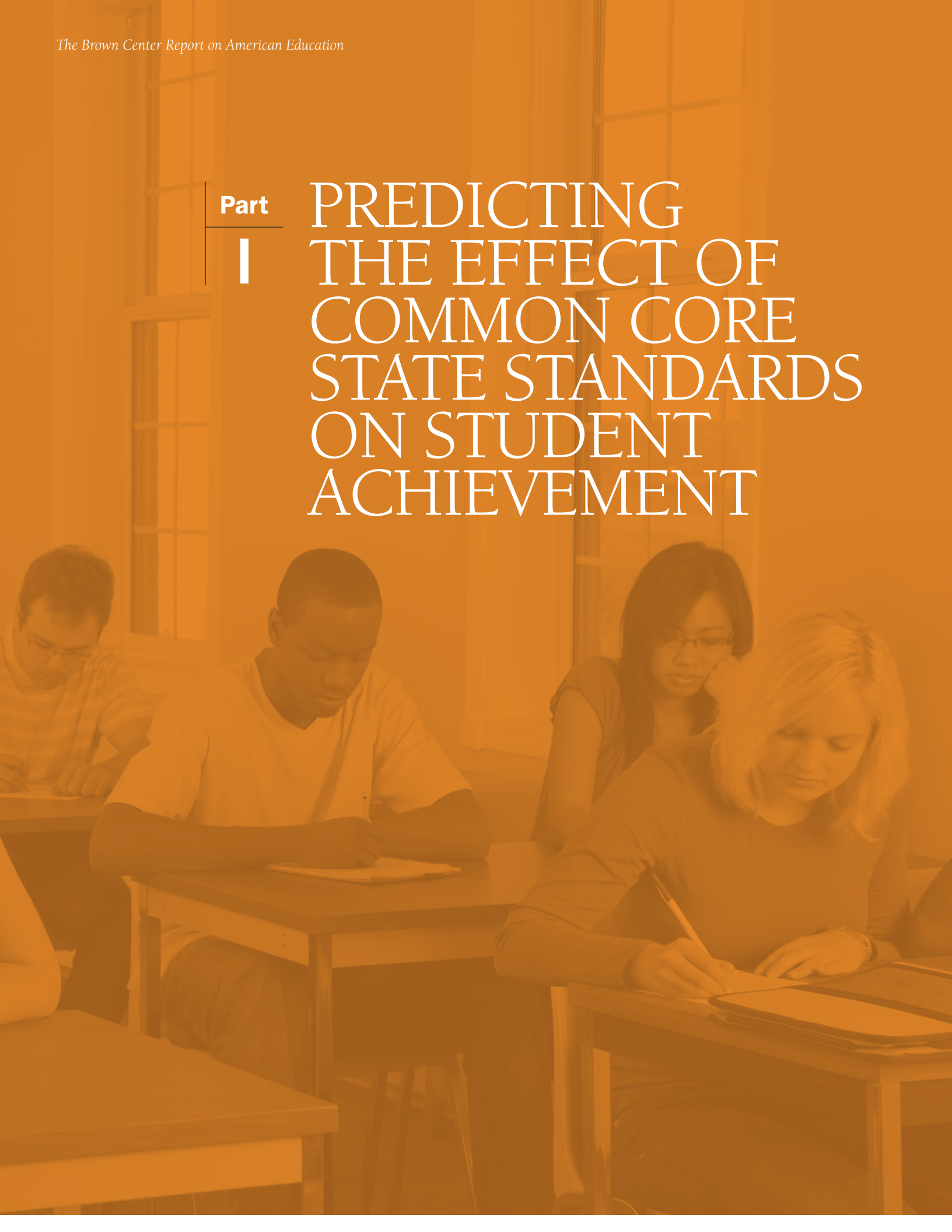
States have had curricular standards for schools within their own borders for many years. Data on the effects of those standards are analyzed to produce three findings. 1) The quality of state standards, as indicated by the well-known ratings from the Fordham Foundation, is not related to state achievement. 2) The rigor of state standards, as measured by how high states place the cut point for students to be deemed proficient, is also unrelated to achievement. Raising or lowering the cut point is related to achievement in fourth grade, but the effect is small, and the direction of causality (whether a change in cut point produces a change in test score or vice versa) is difficult to determine. 3) The ability of standards to reduce variation in achievement, in other words to reduce differences in achievement, is also weak.

Common standards will only affect variation between and among states (analysts use the grammatically suspect "between-state" as shorthand for this kind of variation). Achievement variation existing within states is already influenced, to the extent that standards can exert influence, by the states standards under which schools currently operate. Within state variation is four to five times larger than the variation between states. Put another way, anyone who follows NAEP scores knows that the difference between Massachusetts and Mississippi is quite large. What is often overlooked is that every state has a mini-Massachusetts and Mississippi contrast within its own borders. Common state standards only target the differences between states, not within them, sharply limiting common state standards' potential impact on achievement differences.

The second section of the Report investigates achievement gaps on NAEP. The NAEP has two different tests: the Long-Term Trend NAEP, which began in 1969, and the Main NAEP, which began in 1990. The two tests differ in several respects, but they both carry the NAEP label and both are integral components of "The Nation's Report Card."

Achievement gaps are the test score differences between groups of students with different socioeconomic (SES) characteristics: for example, racial or ethnic background, family income, or language status. The second section poses the question: Do the two NAEP tests report similar achievement gaps? Researchers and policy makers are well aware that significant test score gaps exist between SES groups. Researchers try to study them, policy makers try to close them. What NAEP has to say about the magnitude of such gaps plays an important role in the policy arena. The analysis presented in section two indicates that the two NAEPs do in fact differ. The Main NAEP consistently reports larger SES gaps. This is only a preliminary study, a first cut at the data that reveals a general pattern, so the findings must be viewed cautiously. And explanations for the phenomenon are necessarily speculative. More work needs to be done on this topic.

The third section of the report is on international assessments. Interpretations of international test scores are characterized by three common mistakes. The first occurs when a nation's scores go up or down dramatically and analysts explain the test score change by pointing to a particular policy. The case of Poland's gains in reading is offered as an excellent example of dubious causality attributed to a single policy. The second mistake stems from relying on rankings to gauge a country's academic standing. National rankings have statistical properties that can mislead observers into thinking that large differences are small or small differences are large. They can also make growth appear larger or smaller than it really is. Several examples are provided of misinterpretations of rankings and suggestions on how to avoid them. The third mistake is pointing to a small group of high-performing nations, often called "A+ countries," and recommending, with no additional analysis, that their policies should be adopted. The same policies may be embraced by the lowest performing nations or nations in the middle of the distribution. On any test, the entire distribution must be considered, not just scores at the top.



Part I PREDICTING THE EFFECT OF COMMON CORE STATE STANDARDS ON STUDENT ACHIEVEMENT

FORTY-SIX STATES AND THE DISTRICT OF COLUMBIA HAVE signed on to the Common Core State Standards Initiative, a project sponsored by the Council of Chief State School Officers (CCSSO) and the National Governors Association (NGA). The Common Core spells out what students should learn in mathematics and English-language arts from kindergarten to the end of high school. The standards were written by teams of curriculum specialists and vetted by panels of academics, teachers, and other experts.[1] In 2010, the federal government funded two consortia to develop assessments aligned with the Common Core. The new tests are to be ready in 2014.

The push for common education standards argues that all American students should study a common curriculum, take comparable tests to measure their learning, and have the results interpreted on a common scale, with the scale divided into performance levels to indicate whether students are excelling, learning an adequate amount, or falling short. Past experience with standards suggests that each part of this apparatus—a common curriculum, comparable tests, and standardized performance levels—is necessary. No one or two of them can stand alone for the project to succeed.

Proponents point to the intuitive appeal of a common curriculum. "It's ludicrous," Bill Gates told the *Wall Street Journal*, "to think that multiplication in Alabama and multiplication in New York are really different."[2] In a report called *The Proficiency Illusion*, The Fordham Institute made a similar point regarding state efforts to evaluate schools using fifty different assessments and fifty different definitions of what constitutes acceptable performance.[3] How can a school in one state be labeled a failure while a school in another state and with almost exactly the same test scores can be considered a success?

The authority to operate school systems is constitutionally vested in states. But states have undermined their own credibility when it comes to measuring student learning. Accounts of dumbed-down and

poorly-written state tests, manipulation of cut scores to artificially boost the number of students in higher performance levels, and assessments on which students can get fewer than 50% of items correct and yet score "proficient" fuel the belief that states individually cannot be trusted to give the public an accurate estimate of how American education is doing.[4]

Three Theorized Effects

The Common Core State Standards are theorized to improve education in three ways. First, proponents argue that the Common Core is superior to most current state standards. In a recent study, The Fordham Institute concluded that Common Core standards are better than 37 states' standards in English-language arts and 39 states in mathematics.[5] It follows, proponents believe, that the Common Core will raise the quality of education nationally by defining a higher-quality curriculum in English-language arts and mathematics than is currently taught. Let's call this the "quality theory." Achievement will increase because students will study a better curriculum.

The second idea is that the Common Core sets higher expectations than current state standards, the assumption being that cut points on the new assessments will be set at a higher level than states currently set on their own tests. Comparisons with the National Assessment of Educational Progress (NAEP) lead many analysts to conclude that states set proficiency standards far too low. States routinely report more students attaining proficiency than NAEP indicates, often 30–40 percentage points more.[6] The *No Child Left Behind Act* left it up to the states to design their own tests and to set performance levels wherever they want, but the pattern of states reporting significantly higher percentages of proficient students preceded NCLB.[7] A new Common Core test will presumably end such discrepancies by evaluating proficiency using the same standards for every state, and these standards are to be more rigorous than those currently used. Schools and students will respond by reaching for these loftier goals. Let's call this the "rigorous performance standards" theory.

The third hypothesis is that standardization yields its own efficiencies. In the same *Wall Street Journal* interview cited above, Bill Gates referred to this idea by complaining about the time and money wasted on the many different versions of textbooks that are published to conform to individual states' curricular tastes.[8] In a reverse spin on the same argument, others argue that textbooks are bloated with redundant content as publishers attempt to incorporate numerous idiosyncratic state curricular mandates into one book.[9] The assumption of both arguments is that one, high-quality textbook—or perhaps a few that are aligned with the same content standards—used by all American students attending the same grade would be an improvement over the status quo. Other proponents point to the potential gaps in learning that occur as students move from state to state. Especially when students move mid-year, important concepts might be missed while other concepts are studied unnecessarily a second time. Teachers who move from state to state experience similar difficulties in terms of lesson planning. Let's call this the "standardization" theory.

Opposing Arguments

Some analysts question the theories behind the Common Core. Writing in *Education Week* in the summer of 2011, Andrew Porter compared the Common Core to existing state standards and international standards from other countries and concluded that

The Common Core State Standards are theorized to improve education in three ways.

…data exist that can help predict the magnitude of effects from the Common Core.

the Common Core does not represent much improvement.[10] Opponents of the Common Core, including Sandra Stotsky, James Milgram, Ze'ev Wurman, and Williamson Evers, criticize the quality of the proposed standards for English-language arts and mathematics. They conclude that the math standards, in particular, are inferior to existing standards in Massachusetts and California.[11]

Critics of the Common Core issued a "counter-manifesto" arguing that the proposed common standards would undermine the decentralized, federalist principles on which education has been governed since America's founding. Declaring that a "one-size-fits-all, centrally controlled curriculum" does not make sense, the counter-manifesto states that only weak evidence supports the push for national standards. International test data are not helpful since most countries have national standards and the few that do not, including Canada and Germany, have both impressive and non-impressive scores. Concern for interstate student mobility is overblown, the counter-manifesto claims, because very few students move between states. Most mobility is within state, which is already addressed by the *No Child Left Behind Act*'s requirement that every state establish standards. Since 2003, every state has state curriculum standards that delineate the curriculum for public schools within its borders.[12]

Can empirical evidence shed light on the main points of contention in this debate? Not entirely. Much of the argument is philosophical. Those who believe that the Common Core enumerates what schools should be teaching and students should be learning support the proposed standards. And those who believe a greater degree of standardization would produce more common educational outcomes—and that common outcomes are desirable—also support the proposed standards. Those holding to the opposite beliefs, and believing that local school governance is preferable to governance by larger entities, are critics of the standards.

Despite the philosophical disagreements, there are empirical questions on which evidence exists. The nation has had several years of experience with education standards—since the 1980s in many states and since 2003 in all states—and data exist that can help predict the magnitude of effects from the Common Core. How much does raising the quality of standards matter in boosting student achievement? Will raising the bar for attaining proficiency—in other words, increasing the rigor of performance standards—also raise achievement? And how much variance will be reduced—or how much "sameness" in achievement will be attained—by having students across the country studying a common curriculum?

Quality and Achievement

Let's start with the theory that high-quality standards promote achievement gains. In October 2009, a colleague at Brookings, Grover "Russ" Whitehurst, investigated whether quality ratings for state standards, as judged by the two most cited ratings (from the American Federation of Teachers and Fordham Foundation), are correlated with state NAEP scores. Whitehurst found that they are not. States with weak content standards score about the same on NAEP as those with strong standards. The finding of no relationship held up whether NAEP scores from 2000, 2003, 2005, 2007, or the gains from 2000–2007 were used in the analysis. And it held up for the scores of both white and black students.[13]

The current study extends that inquiry by looking at NAEP data from 2003–2009. Gain scores on NAEP reading and math tests from 2003 and 2009 are combined to form a composite gain score. The scores

are adjusted to control for demographic characteristics of each state—the percent of students qualifying for free or reduced lunch, special education, or English language learner status. More precisely, scores are adjusted to control for changes that occurred in those demographic characteristics from 2003–2009. That prevents swings in states' demographic characteristics from skewing the results. Ratings of state curricular standards conducted by the Fordham Foundation in 2000 and 2006 are used to model the quality of state standards. It is particularly apt to model the quality of state standards with the Fordham ratings considering Fordham's high opinion of the Common Core.

The results are shown in Table 1-1. Three questions are answered by the data. The first row addresses the question: Do the Fordham ratings in 2000 successfully predict the NAEP gains that states made in reading and math from 2003–2009? One could imagine, since there is undoubtedly some lag time before standards are implemented in classrooms and realized in student learning, that the curriculum standards of 2000 would influence achievement gains made three, six, or even nine years down the road. The correlation coefficient of –0.06 indicates that they do not.

The second row examines whether the ratings of 2006 are statistically related to 2003–2009 NAEP gains. In other words, was the quality of standards in the middle of the gain period related to test score gains? Again, the answer is no, with a correlation coefficient of 0.01. The final row looks at the change in ratings from 2000 and 2006. According to Fordham, some states improved their standards in 2006 while others adopted weaker standards in 2006 than they had back in 2000. Are changes in the quality of standards related to changes in achievement? Again, the answer is that they are not (correlation coefficient of 0.08).

Relationship of Fordham's Ratings of State Content Standards with State NAEP Gains (2003–2009)

Table 1-1

Standards Rating	Correlation Coefficient
Fordham 2000	–0.06
Fordham 2006	0.01
Change in Fordham 2000–2006	0.08

Relationship of State Proficiency Level with NAEP Achievement (Correlation Coefficients)

Table 1-2

	2005 NAEP	2009 NAEP	Change 2005–2009
4th Grade Reading	–0.22	–0.08	0.35*
4th Grade Math	–0.12	0.01	0.34*
8th Grade Reading	–0.11	–0.09	0.06
8th Grade Math	0.00	0.01	0.02

* p < .05

Rigorous Performance Standards and Achievement

The second theory of improvement is based on performance standards. A 2006 NCES report found that the difficulty of state performance standards is uncorrelated with achievement.[14] Performance levels (or "cut points") for student proficiency were mapped onto the 2005 NAEP scale. States with higher, more rigorous cut points did not have stronger NAEP scores than states with less rigorous cut points. A new NCES report was released in 2011 with updated measures using 2009 NAEP data.[15]

Table 1-2 summarizes the correlations between the rigor of state performance levels and achievement. In a replication of the earlier NCES study, we also find that the states'

A 2006 NCES report found that the difficulty of performance standards is uncorrelated with achievement.

…the absolute level of performance standards does not seem to matter but raising or lowering levels does exhibit a relationship with fourth grade changes in achievement…

2005 NAEP scores are unrelated to where the states drew the line for proficiency in 2005. Fourth-grade reading and math have slightly negative correlations (–0.22 and –0.12, respectively), as does eighth-grade reading (–0.11). The correlation coefficient for eighth-grade math is 0.00. State achievement is unrelated to the level at which states define proficiency. The same is true for 2009 NAEP scores and the level at which proficiency was placed that year (see the second column of the table).

The final column of Table 1-2 investigates whether changes in state NAEP scores from 2005–2009 are related to changes in proficiency level. Did states that raised the bar also perform better? And did states that lowered the bar perform worse? Correlation coefficients for 8th grade are near zero. Positive and statistically significant correlations were found for fourth-grade reading (0.35) and fourth-grade math (0.34). It is interesting that the absolute level of performance standards does not seem to matter but raising or lowering levels does exhibit a relationship with fourth grade changes in achievement, explaining about 12% of the variation in the change in state NAEP scores.

Whether one phenomenon is causing the other is difficult to tell. Changes in proficiency cut points are probably endogenous to trends in test scores. In other words, states with rising scores may feel emboldened to raise their proficiency cut points and those with declining scores may feel compelled to lower theirs. That is quite a different story than the raising or lowering of cut points producing changes in test scores. Unfortunately, simple correlations cannot determine the direction of causality, or if causality exists at all, only whether these two variables are statistically related. In the current analysis, change in level is related to change in fourth-grade scores.

How Common Will Achievement Become?

The third theory concerns standardization. For the Common Core movement, attaining greater standardization of educational outcomes is an important goal. If standards do not reduce variation, then even if they boost performance, simply raising average scores will still leave many states—and the districts, schools, and students within states—far behind and far below acceptable levels of performance. The two previous analyses indicate that it is unlikely that common standards will boost performance; however, it is possible for the national average on NAEP to remain stable while variation is reduced—for instance, if top states decline a little while states at the bottom rise by the same amount. Another way would be for high flying schools within states to decline a little while poorly performing schools increase their performance by a commensurate amount.

In terms of state NAEP scores, variation comes in two forms: variation between states and variation within states. We would expect common standards to reduce variation between states, so that the NAEP score difference between states at the top and bottom of the rankings would be reduced. States that currently offer vastly different curricula, assessments, and performance standards will harmonize those elements of their educational systems. One would expect test score differences to shrink. That is the essence of common standards. Within-state variation, on the other hand, remains unaffected by common standards. Every state already has standards placing all districts and schools within its borders under a common regime. And despite that, every state has tremendous within-state variation in achievement. Schools that score at the top of the world on

international assessments are within a short car trip, sometimes even within a short subway ride, from schools that score at the level of the world's lowest achieving nations.

Let's compare these two forms of variation. Table 1-3 displays data on NAEP standard deviations between and within states. Standard deviation is a measure of variation, the amount of spread in a group of data. On any particular test, about two-thirds of observations are within one standard deviation (above and below) of the average score. "Between-State SD" is the standard deviation of NAEP scores for the fifty states and the District of Columbia—how much they differ from each other. "Within-State SD" is the average of the standard deviations for the fifty states and the District of Columbia—how much the students within each state, on average, differ from each other.

The findings are clear. Most variation on NAEP occurs within states not between them. The variation within states is four to five times larger than the variation between states. Much of the similarity of state scores comes from aggregating individual student scores, which differ greatly, to the state level. The variation in student performance within states washes out to produce means that are alike across states. Consider this: fourth-grade NAEP scores in math range from Massachusetts at the top with 252 down to the District of Columbia with 219. That 33 point difference is not too much larger than the average standard deviation within states (27.8). What does that mean? Consider Massachusetts and Mississippi, a state with low scores but not at the very bottom. Their NAEP means differ by 25 points. Every state, including Massachusetts and Mississippi, has a mini-Massachusetts and Mississippi contrast within its own borders. That variation will go untouched by common state standards.

Relationship of State Proficiency Level with NAEP Achievement (Correlation Coefficients)

Table 1-3

	Average State NAEP Score	Between-State SD	Within-State SD	Multiple (Within/Between)
4th Grade Reading	220.1	6.6	34.7	5.3
4th Grade Math	239.5	6.3	27.8	4.4
8th Grade Reading	263.3	6.5	32.9	5.1
8th Grade Math	282.4	8.5	34.8	4.1

Discussion

What effect will the Common Core have on national achievement? The analysis presented here suggests very little impact. The quality of the Common Core standards is currently being hotly debated, but the quality of past curriculum standards has been unrelated to achievement. The rigor of performance standards—how high the bar is set for proficiency—has also been unrelated to achievement. Only a change in performance levels has been related to an increase in achievement, and that could just as easily be due to test score changes driving changes in policy, not the other way around. The Common Core may reduce variation in achievement between states, but as a source of achievement disparities, that is not where the action is. Within-state variation is four to five times greater.

The sources of variation in educational outcomes are not only of statistical importance but also bear on the question of how much state policy can be expected to change schools. Whatever reduction in variation between, say, Naperville and Chicago that can be ameliorated by common standards has already been accomplished by Illinois's state efforts. State standards have already had a crack at it. Other states provide even more deeply rooted historical examples. California has had state curriculum frame-

The Common Core may reduce variation in achievement between states, but as a source of achievement disparities, that is not where the action is.

Standards in education are best understood as aspirational.

works since at least 1962, statewide testing with scores for every school published publically since 1971 (except for a brief timeout in the early 1990s), state textbook adoption for K–8 since the nineteenth century, and a court-ordered equalized spending system since the late 1970s. Any effect that these laws have on reducing achievement variation within the state has already occurred. The Common Core must go beyond these efforts to reduce variation in California's achievement. That is highly unlikely.

Two lessons can be drawn from the analysis above. First, do not expect much from the Common Core. Education leaders often talk about standards as if they are a system of weights and measures—the word "benchmarks" is used promiscuously as a synonym for standards. But the term is misleading by inferring that there is a real, known standard of measurement. Standards in education are best understood as aspirational, and like a strict diet or prudent plan to save money for the future, they represent good intentions that are not often realized.

Why don't aspirational standards make much of a difference? Researchers from the International Association for the Evaluation of Educational Achievement (IEA) first sketched the concept of opportunity to learn using international test score data in the 1970s.[16] Distinctions were drawn among the intended, implemented, and achieved curriculums. The intended curriculum is embodied by standards; it is what governments want students to learn. The differences articulated by state governments in this regard are frequently trivial. Bill Gates is right that multiplication is the same in Alabama and New York, but he would have a difficult time showing how those two states—or any other two states—treat multiplication of whole numbers in significantly different ways in their standards documents.

What is crucial is the distance between the intended curriculum and the two curriculums below. The implemented curriculum is what teachers teach. Whether that differs from state to state is largely unknown; what is more telling is that it may differ dramatically from classroom to classroom in the same school.[17] Two fourth-grade teachers in classrooms next door to each other may teach multiplication in vastly different ways and with different degrees of effectiveness. State policies rarely touch such differences. The attained curriculum is what students learn. Two students in the same classroom and instructed by the same teacher may acquire completely different skills and knowledge. One student understands and moves on; another struggles and is stuck. And that even happens in classrooms with outstanding teachers.

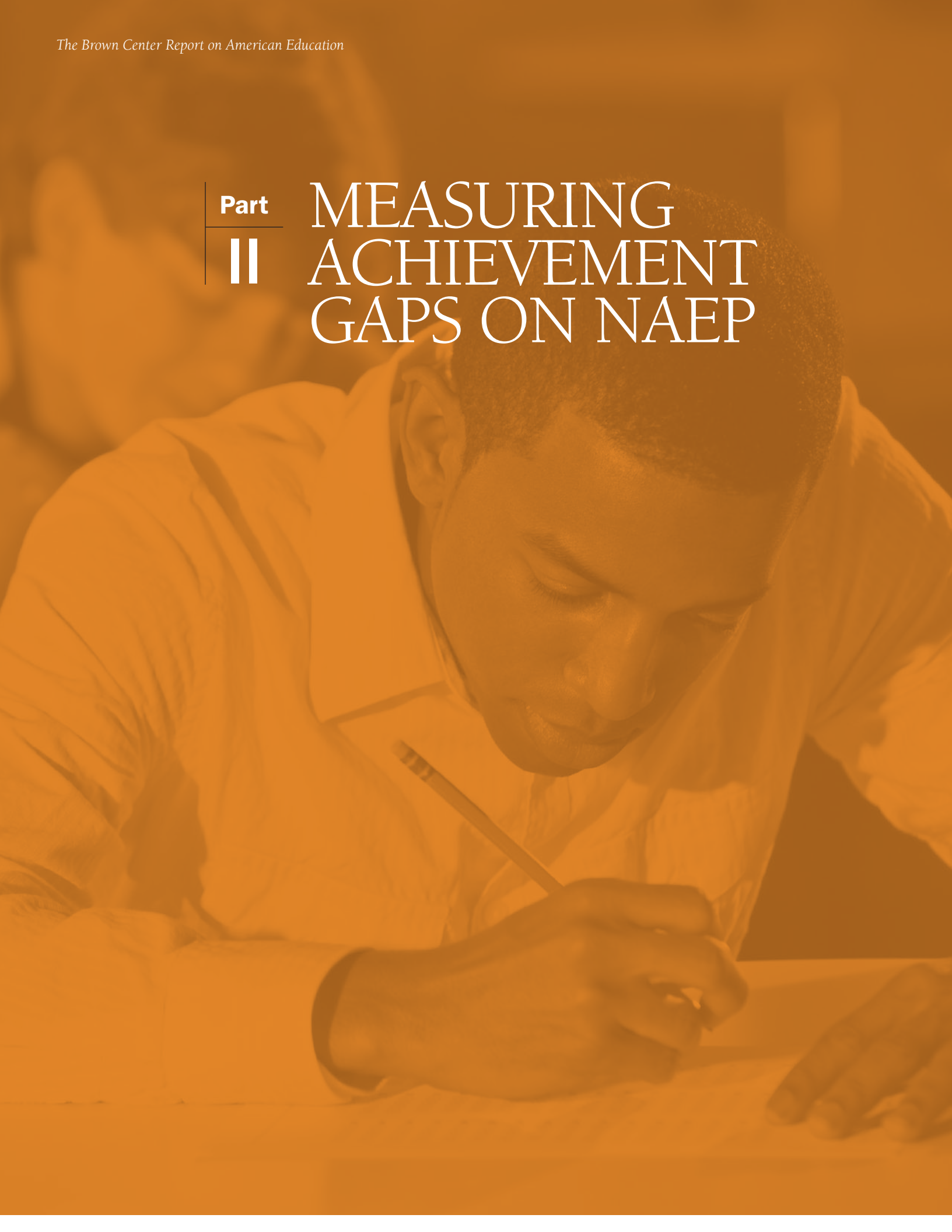
The whole system is teeming with variation. Policies at national, state, district, and school levels sit on top of these internal differences, but they rarely succeed in ameliorating them. The Common Core will sit on top of the implemented and attained curriculums, and notwithstanding future efforts to beef up the standards' power to penetrate to the core of schooling, they will probably fail to dramatically affect what goes on in the thousands of districts and tens of thousands of schools that they seek to influence.

A final word on what to expect in the next few years as the development of assessments tied to the Common Core unfolds. The debate is sure to grow in intensity. It is about big ideas—curriculum and federalism. Heated controversies about the best approaches to teaching reading and math have sprung up repeatedly over the past century.[18] The proper role of the federal government, states, local districts, and schools in deciding key educational questions, especially in deciding what should be taught,

remains a longstanding point of dispute. In addition, as NCLB illustrates, standards with real consequences are most popular when they are first proposed. Their popularity steadily declines from there, reaching a nadir when tests are given and consequences kick in. Just as the glow of consensus surrounding NCLB faded after a few years, cracks are now appearing in the wall of support for the Common Core.

Don't let the ferocity of the oncoming debate fool you. The empirical evidence suggests that the Common Core will have little effect on American students' achievement. The nation will have to look elsewhere for ways to improve its schools.

The empirical evidence suggests that the Common Core will have little effect on American students' achievement.



Part II MEASURING ACHIEVEMENT GAPS ON NAEP

THIS SECTION OF THE BROWN CENTER REPORT IS ABOUT TEST score gaps among groups of students identified by race, ethnicity, family income, or native language—in other words, characteristics related to socioeconomic status (SES). Much has been written about such gaps, and the National Assessment of Educational Progress (NAEP) frequently serves as the source of data to measure them.

There are two NAEP tests: the Main and the Long-Term Trend (LTT). Do the two NAEPs provide similar estimates of SES achievement gaps? As the analysis below shows, it appears that they do not. The discrepancy has implications for national education policy, especially since closing achievement gaps related to SES is a prominent goal of federal policy, included in the *No Child Left Behind Act* (NCLB), Race to the Top, and the Blue Ribbon Schools selection criteria. NAEP serves as the most widely-recognized instrument for determining whether the nation is making progress towards meeting the goal.

Background

Gaps between socioeconomic groups as measured by prominent tests have long interested observers of U.S. education. The Coleman Report of 1966 and subsequent analyses of the same data set, in particular, the Harvard-based *Inequality* studies, revealed large gaps between advantaged and disadvantaged youngsters on tests of academic achievement.[19] Gaps between blacks and whites on the SAT, American College Test (ACT), and Armed Forces Qualifying Test (AFQT) have been analyzed to estimate differences in preparation for college and military service and to document unequal opportunities for learning in American society. Racial differences on IQ tests (e.g., Stanford–Binet Intelligence Scales, Wechsler Intelligence Scale for Children) have been persistent for several decades and the explanations for them exceedingly controversial.[20]

In the 1970s, tests came under fire for containing racial bias. A district court judge in a California court case, *Larry P. v. Riles*,[21] found intelligence tests culturally biased against blacks and banned using them in that state for placing students into special education programs. Analysts often draw a distinction between achievement tests, designed to measure the material that students have learned in the past, and aptitude tests—of which IQ tests are a species—designed to measure students' capacity

for learning in the future. Aptitude tests are believed to be more prone to cultural bias because they depart from the school curriculum, meaning that much of the skill they measure may be picked up in families or among peers, social units that embody cultural differences. An achievement test, on the other hand—an Advanced Placement Chemistry Exam, for example—is explicitly designed to measure whether students have learned curricular material that is taught in school.[22]

These distinctions are fuzzy. What students have learned in the past undoubtedly influences their capacity for future learning. So there must be some overlap. A telling sign of the confusion is apparent in the Scholastic *Aptitude* Test (or SAT), which was renamed as the Scholastic *Assessment* Test in the 1990s, and then renamed again in 2003 to just the SAT, an initialism signifying no words.[23] The College Board was scrambling to get on the right side of public views toward the word "aptitude." Aptitude has a connotation of being innate and immutable. Achievement has a connotation of referring to the knowledge and skills one acquires by being a good student and working hard at school. As a statistical matter, the SAT and its more achievement-oriented competitor, the ACT, are highly correlated (with a correlation coefficient of about 0.92).[24]

Test items that are sensitive to cultural differences are described as "culturally loaded" or having a pronounced "cultural load factor." Test developers have strict protocols for screening items for cultural bias. Despite its name change, the SAT has received scrutiny. A study by Roy Freedle, a cognitive psychologist who worked for over thirty years at the Educational Testing Service, caused a furor in 2003 by reporting statistically significant black–white differences on many SAT items.[25]

Freedle examined data based on differential item functioning (DIF) of SAT items, a technique that compares the responses of two subgroups of examinees matched on proficiency. The technique controls for overall ability to answer a test's items correctly, thereby isolating whether other characteristics, typically race or gender, are associated with differential performance on particular items. Freedle found that easy items (those that most examinees got correct) on the verbal section of the SAT favored whites. Freedle hypothesized that easy items contain words and concepts that are often vague. Examinees flesh out the items' meanings by calling on their own life experiences, creating different interpretations across cultural groups. Hard items contain words and concepts that are specialized, more technical, more precise, learned through education, and with meanings that do not vary across cultural groups.

The study was replicated in 2010 by Maria Veronica Santelices and Mark Wilson employing a later SAT data set, more robust statistical controls, and a different methodology that addressed critiques of Freedle's study. They confirmed the relationship between item difficulty and differences by race on verbal items, but found no such differences with Hispanic students or among any subgroups on math items.[26] Again, black–white differences favoring whites emerged on the SAT's easier items.

In addition to the psychometric issues raised by these studies, policy concerns about the consequences of emphasizing either aptitude or achievement have been raised by other observers of American education. Writing in *Education Week* in 1993, Michael Kirst and Henry Rowen urged colleges and universities to base admissions on achievement tests linked to high school curriculum, not on aptitude-oriented tests such as the SAT. Such a move would enhance

Aptitude has a connotation of being innate and immutable.

NAEP Main and LTT Gap Differences
Free and Reduced Lunch

Table 2-1

	Reading			Math		
	LTT	Main	Main–LTT Diff*	LTT	Main	Main–LTT Diff*
Age 9/Grade 4	0.70	0.75	0.05	0.70	0.79	0.09
Age 13/Grade 8	0.61	0.69	0.07	0.72	0.76	0.04
Age 17/Grade 12	0.56	0.53	–0.03	0.66	0.68	0.01

Average Main gap = 0.70, average LTT gap = 0.66, Difference = 0.04 (Main larger).
*May not correspond to column difference due to rounding.

NAEP Main and LTT Gap Differences
Black–White

Table 2-2

	Reading			Math		
	LTT	Main	Main–LTT Diff*	LTT	Main	Main–LTT Diff*
Age 9/Grade 4	0.67	0.76	0.08	0.72	0.91	0.19
Age 13/Grade 8	0.60	0.79	0.19	0.84	0.91	0.08
Age 17/Grade 12	0.62	0.69	0.07	0.89	0.88	–0.01

Average Main gap = 0.83, average LTT gap = 0.72, Difference = 0.11 (Main larger).
*May not correspond to column difference due to rounding.

minority access to higher education, Kirst and Rowen argued, citing research by John Bishop of Cornell University showing that minority–majority differentials tend to be smaller (in standard deviation units) on achievement tests (including the LTT NAEP) than on aptitude tests (including the SAT).[27]

What We Did

The LTT and Main NAEPs are given in different years and to slightly different populations. The LTT is age-based; whereas the Main NAEP is grade-based. We first paired LTT NAEP data with Main NAEP data. The 2004 LTT was paired with the 2005 Main NAEP and the 2008 LTT with the 2009 Main. Age 9 on the LTT was paired with 4th grade on the Main, age 13 with 8th grade, and age 17 with 12th grade. These pairings are conventional in NAEP analysis, but they may introduce bias or influence the findings reported below in unknown ways.

We then calculated the achievement gaps for both tests on the groups defined by four SES variables: students qualifying for free and reduced lunch vs. those who do not, black vs. white, Hispanic vs. white, and students who are English language learners vs. those who are not. The gaps were standardized by dividing the point gap by the test's standard deviation. The standardized gaps are reported in the tables below, along with the differences between the LTT and Main.

Results

Table 2-1 reports NAEP gaps of students who do and do not qualify for free and reduced lunch. Qualifying for the free and reduced lunch program is often used as a proxy for poverty in studies of student achievement. The gaps range from 0.53 (about one-half standard deviation) on the Main's test of reading for 17-year-olds to 0.79 (about three-quarters of a standard deviation) on the Main's test of mathematics for fourth graders. The gaps on both tests tell us, not surprisingly, that students from wealthier families score significantly higher on NAEP than students from poorer families. The gaps are generally larger on the Main than the LTT—the one exception being reading at age 17/grade 12, in which the gap on the LTT is slightly larger. The six cells in the table report an average gap of 0.70 on the Main and 0.66 on the LTT, resulting in a difference of .04. Put another way, achievement gaps related to poverty status are about 0.04 standard deviations larger on the Main NAEP than on the LTT NAEP.

Table 2-2 shows data for black–white differences on the two tests. Again, gaps are larger on the Main than the LTT, with the exception at age 17/grade 12 in math. Two

of the differences are 0.19 standard deviation units: the gaps for reading at age 13/grade 8 and for math at age 9/grade 4. These are modest effect sizes at best, but as the difference between two tests, they are noteworthy. Take the age 13/grade 8 gaps in reading, for example. The LTT estimates the black–white gap as 0.60 and the Main as 0.79. That is a difference of about 32%. If policymakers were to use the LTT NAEP to establish a baseline and then use the Main NAEP as a later measure, they would be misled to conclude that the black–white gap had grown by 32%—even though this is an artifact of differences between the two NAEP tests and not a real trend. Conversely, using the Main NAEP as a baseline and the LTT as a later measure would give the false impression of the gap shrinking by 32%. The two NAEP tests are not interchangeable when it comes to measuring black–white test score gaps.

The Hispanic–white gap differences are reported in Table 2-3. The Main NAEP reports larger gaps than the LTT for all six subject/age/grade combinations. In age 9/grade 4 math the Hispanic–white gap is 0.72 standard deviations, about 57% wider than the 0.46 Hispanic–white gap on the LTT NAEP. The average gap difference is 0.10 between the two tests, but that is largely driven by the large age 9/grade 4 math difference. Omitting that figure leaves an average test difference of 0.06 for the remaining pairings.

The gaps for English language learners (ELL) are presented in Table 2-4. Again, the Main NAEP reports larger gaps than the LTT NAEP. These are the largest test differences for the four SES variables analyzed. Both differences at age 9/grade 4 are large: 0.19 in reading and 0.24 in math. They correspond to gaps that are 23% larger in reading and 41% larger in math when recorded by the Main NAEP compared to the LTT NAEP.

NAEP Main and LTT Gap Differences
Hispanic–White

Table 2-3

	Reading			Math		
	LTT	Main	Main–LTT Diff*	LTT	Main	Main–LTT Diff*
Age 9/Grade 4	0.61	0.72	0.11	0.46	0.72	0.26
Age 13/Grade 8	0.64	0.70	0.06	0.65	0.74	0.09
Age 17/Grade 12	0.53	0.56	0.03	0.67	0.68	0.01

Average Main gap = 0.69, average LTT gap = 0.59, Difference = 0.10 (Main larger).
*May not correspond to column difference due to rounding.

NAEP Main and LTT Gap Differences
English Language Learners

Table 2-4

	Reading			Math		
	LTT	Main	Main–LTT Diff*	LTT	Main	Main–LTT Diff*
Age 9/Grade 4	0.81	1.00	0.19	0.59	0.83	0.24
Age 13/Grade 8	1.12	1.26	0.14	0.96	1.09	0.13
Age 17/Grade 12	1.06	1.19	0.13	0.90	1.00	0.10

Average Main gap = 1.06, average LTT gap = 0.90, Difference = 0.16 (Main larger).
*May not correspond to column difference due to rounding.

The final two tables provide summary data. Table 2-5 reports the average gaps for each SES variable and the gap difference between the Main and LTT NAEPs. Keep in mind that there is undoubtedly some overlap of the SES categories; for example, the ELL gap differences are surely related to the Hispanic–white differences. In addition, black and Hispanic students are both more likely than white students to be poor, so a portion of the black–white and Hispanic–white gaps is shared with the gap pertaining to free and reduced lunch.

In terms of groups of students based on SES characteristics, the largest differences between the LTT and Main NAEP are with ELL students (see Table 2–5). The role that language plays in the two NAEP

The two NAEP tests are not interchangeable when it comes to measuring black-white test score gaps.

Summary of NAEP Main and LTT Gap Differences Average by SES Variables

Table 2-5

	Average Main Gap	Average LTT	Increase of Gap on Main Compared to LTT
Free and Reduced Lunch	.70	.66	5.79%
Black–White	.83	.72	13.92%
Hispanic–White	.69	.59	15.78%
ELL	1.06	.90	17.26%

Summary of NAEP Main and LTT Gap Differences Average By Subject and Grade

Table 2-6

	Average Main Gap	Average LTT	Increase of Gap on Main Compared to LTT
Reading	0.80	0.71	12.91%
Math	0.83	0.73	14.06%
Age 9/Grade 4	0.81	0.66	23.07%
Age 13/Grade 8	0.87	0.77	13.16%
Age 17/Grade 12	0.78	0.74	5.29%

The two tests are different instruments for measuring student learning.

tests could be producing different estimates of the ELL and non-ELL achievement gap. Table 2-6 takes another cut at the data by showing the tests' differences by subject and grade levels. Gaps in math and reading look similar, but the age 9/grade 4 subjects stand out with the largest differences. The divergence of the two NAEPs along the dimensions of language and age is an intriguing finding. That, along with the headline finding that the Main NAEP consistently reports larger SES gaps than the LTT NAEP's are important considerations for researchers who use NAEP data to investigate achievement gaps. They are also important factors for NAEP policymakers to think about when deciding the future of the Nation's Report Card.

Discussion

Let's tackle three questions provoked by the study's findings: 1) Which NAEP is right?, 2)Why do the two NAEP's differ?, and 3) Does it really matter?

Which NAEP is right?

They may both be right. The two tests are different instruments for measuring student learning, and although they share the NAEP label and would no doubt produce highly-correlated results if given to the same sample of students, they measure different constructs. The reading skills assessed on the LTT NAEP are not the same as the reading skills assessed on the Main NAEP—nor are the skills measured by the math tests comparable. In the future, investigations that dig down to make comparisons on an item by item basis may discover that the Main NAEP produces inflated estimates of achievement gaps or that the LTT understates those gaps, but this preliminary investigation only makes a start by comparing the relative performance gaps of subgroups on each test.

Why do the two NAEPs differ in measuring SES achievement gaps?

Content differences may play a role. As just mentioned, the Main NAEP was designed to assess different skills and concepts than the LTT NAEP, which had a nearly twenty-year track record when the Main NAEP was first launched in 1990. In math, for example, the LTT NAEP focuses more on computing with whole numbers and fractions; the Main NAEP on how students apply mathematics to solve problems. In reading, the LTT NAEP presents shorter passages, more vocabulary words in isolation, and more items asking students to identify the main idea of a passage. The Main NAEP has a broader selection of literary forms and asks students to compare multiple texts.[28]

Neither test is absolute on these dimensions; for instance, the Main NAEP in math includes some computation items, and the LTT includes problem solving items. The difference is one of emphasis. To better measure how well students apply knowledge, to assess a broader array of topics, and to "modernize" the national assessment, the Main NAEP was created. But the upshot is this: the contemporary skills and knowledge measured by the Main NAEP, compared to the more traditional skills assessed by the Long-Term Trend, may be more influenced by the background experiences of students. That would widen the measured gaps between groups defined by socioeconomic characteristics. If attending school works to reduce such differences, that would also explain why the largest discrepancies between the two NAEPs appear with the youngest pairing of students (9-year-olds/fourth graders) rather than the older ones.

Another possibility is that disadvantaged students are less likely to be exposed to the knowledge and skills on the Main NAEP as compared to the LTT NAEP. Fewer opportunities to learn the Main NAEP's content in schools serving a preponderance of poor children, for example, would widen gaps between students qualifying for free and reduced lunch and those from wealthier households. The Main NAEP also poses more open-ended, constructed-response items, as opposed to the multiple choice or short answer items that are more prevalent on the LTT. Research shows that different item formats can affect the performance of ELL students. A 1999 study by the American Institutes for Research investigated why students do not answer many NAEP items. The researchers found that constructed-response items were much more likely to be skipped than multiple choice items, and that students with limited English proficiency were especially prone to non-responses on constructed-response items.[29]

Does it really matter?

Achievement gaps command the nation's attention. The *No Child Left Behind Act* established a national goal of closing achievement gaps between groups based on race, ethnicity, economic status, and disability. An outpouring of scholarship has documented the persistence of gaps, explored their causes, and evaluated interventions to reduce them.[30] Analysis of trends in achievement gaps on NAEP is a regular feature of reports from the U.S. Department of Education.[31] Measuring the magnitude of achievement gaps is an important function of NAEP, and it is important that the estimates be as precise as possible. As noted above, the fact that the two NAEP tests generate different estimates is not necessarily a problem, but it does deserve investigation and explanation.

The mere existence of two NAEP tests confuses many followers of test scores. The confusion surely increases when historical accounts of gap trends are published. Because the Main NAEP was started in 1990, studies of achievement gaps before then must rely on the LTT.[32] But only national trends can be tracked before 1990 because the LTT does not produce state scores. Studies of state gap trends must rely on the Main NAEP.[33] In addition, studies examining gaps in the percentage of students attaining specific achievement levels (for example, the percentage of blacks and whites scoring at advanced, proficient, basic, or below basic) are confined to the Main NAEP. The LTT NAEP does not have such levels.[34]

The upshot is this: studies conducted with one NAEP may have yielded different findings if they had been conducted with the other NAEP. State gaps would be narrower if measured with an LTT NAEP. Pre-1990 gaps

Studies conducted with one NAEP may have yielded different findings if they had been conducted with the other NAEP.

The goal of assessing higher-order skills is laudable but must be implemented cautiously.

would be wider if measured with the Main NAEP. The widening of the black–white gap from 24 to 27 scale score points on the age 9 LTT math test from 1988–1990, a reversal of several years of steady narrowing, falls well within the LTT-Main difference found by the current study (about 26%) on measuring gaps in math at that age and grade. The point extends to other databases based on NAEP tests. The reading and math assessments of the Early Childhood Longitudinal Study (ECLS), which reveal large achievement gaps among kindergartners in their first days of school, are based on the Main NAEP.[35] They even include some publicly-released Main NAEP items.[36] The current study suggests that an ECLS with a test based on the LTT NAEP would find narrower gaps.

A final word regarding higher-order items. The goal of assessing higher-order skills is laudable but must be implemented cautiously. Higher-order test items are more prone to cultural bias than items assessing basic skills. Two reasons. First, basic skills are universally agreed upon, transcend culture, and even constitute a world curriculum.[37] Abstract skills—application, analysis, synthesis, and evaluation—are varied in interpretation and more strongly influenced by the cultural lens through which they are interpreted and expressed.[38] Second, higher-order items often ask for constructed responses from examinees as opposed to simpler response formats (e.g., multiple choice). In mathematics, such items typically go beyond "showing one's work" and require students to explain their reasoning or to communicate something about the mathematics of the problems. Items involving only mathematical symbols do not rely on language skills. As noted above, ELL students are more likely than native English speakers to skip constructed response items. The reliance on expressive language is unavoidable on English language arts tests, of course, but may introduce cultural bias into math tests that should be avoided on our national assessment.

Part III MISINTERPRETING INTERNATIONAL TEST SCORES

INTERNATIONAL TEST SCORES RECEIVE A LOT OF ATTENTION, especially when first released. The press scrambles to find pundits offering instant analysis. Policy makers pour over the results to glean lessons for governments. Advocates look for evidence to bolster pre-existing arguments. This section of the Brown Center Report is about errors that arise from such interpretations, focusing on the three most common and potentially most misleading mistakes that are made when interpreting international tests scores.

Approximately fifty years ago, the founders of international assessments believed that comparing nations on tests of academic achievement would allow the world to serve as a laboratory of innovation and experimentation, that international tests could illuminate the different approaches that countries take to solve education's problems and the relative effectiveness of these efforts. The promise of international assessments is not advanced when data are misused.[39]

Dubious Causality

Let's start with a mystery: what explains the huge leap in Poland's reading scores on the Programme for International Student Assessment (PISA)? The test is given to 15-year-olds every three years. In 2000, Poland scored 480, then 497 in 2003, and 508 in 2006, a truly remarkable gain of 28 points in six years. To place these gains in perspective, the international average on the PISA was 492 in 2006. In only six years, Poland went from below average to above average, leapfrogging such countries as Japan, Germany, and the United Kingdom.

What explains Poland's great success? Almost immediately, explanations converged on one reform: tracking. In 1999, Poland changed its tracking system. In the old system, compulsory education (primary school) ended after eight years of schooling—at about age 15. Of those students matriculating to secondary schools, about half went to vocational schools focused on preparation for industrial sectors, one-third went to technical vocational schools, and about one-fifth attended academic schools (*lyceum*) that prepared students for college. After the reforms, primary education ended after six years of schooling, with the next three years devoted to a new compulsory, comprehensive lower secondary level (*gymnasium*). This postponed separation into vocational and academic schools and extended compulsory

education to about age 16, giving most 15-year-olds an extra year of academic instruction that they did not experience under the old system.

The Polish system, whether pre- or post-reform, should not be confused with tracking in the U.S. All tracking systems separate students for instruction, but tracking differs as much across countries as health care systems or the side of the street on which cars are driven. In the Polish system, tracking begins when students are divided into vocational or academic streams that attend separate schools. In the U.S., 15-year-olds are typically sophomores attending a comprehensive high school, one offering both vocational and academic courses. There are no formal vocational or academic streams. All sophomores must take several academic courses, with the specific requirements mandated by states. Within academic subjects, students may be tracked into classes distinguished by skill level (e.g., Honors English, English 10, or Remedial Reading) or curricular content (e.g., Geometry, Algebra, or General Math). These groupings are decided subject by subject. It is possible for students to be enrolled in an above-grade-level class in one subject and an at-grade or even below-grade-level class in another subject. In addition, students are not locked into a single track and may move up or down a level depending on their performance the previous year.[40]

The two key elements of the Polish tracking reform were the delay in separating students by ability for one year and the extra year of exposure to academic learning that vocational students now received. The two initiatives are independent. Vocational programs, if policy makers so decide, may include one or more years of intensive academic study. Many analysts decided that the one year delay in tracking was the reason for Poland's jump in PISA scores. The first declaration came from the authors of the 2006 PISA score report, as evident in the 2006 PISA Executive Summary: "A long-term trend in OECD countries has been to reduce the amount of separation and tracking in secondary education. The most recent major example of this is Poland, whose reading results before and after this education reform are reported in PISA."[41] The Executive Summary goes on to cast the reform as cost-free, producing gains without affecting other students: "Here [in Poland], an improvement in results among lower ability students immediately after the reform was not at the expense of higher ability students, whose results also rose in the subsequent period."[42]

Soon after, a World Bank study pressed harder on the theme of causality, "Poland's reading score was below the OECD average in 2000, at the OECD average in 2003, and above the OECD average in 2006, ranking 9th among all countries in the world…. With regard to the factors responsible for the improvement, the delayed tracking into vocational streams appears to be the most critical factor."[43] The study also mentioned the extra language and reading instruction that prospective vocational students were receiving as a contributing factor.

The hypothesis that delayed tracking lead to gains is certainly reasonable. Hanushek and Woessmann show that nations with later tracking (ironically, the U.S. is coded as a late tracker in the study)[44] have higher PISA scores than nations with earlier tracking. Also, as the World Bank study and OECD analyses indicated, improvement in Poland's scores was most pronounced at the bottom of the achievement distribution. Low achievers made larger gains than high achievers. Several studies of tracking, whether the American or European style, have found that tracking can depress the achievement

Tracking differs as much among countries as health care systems or the side of the street on which cars are driven.

Nations that Gained in PISA Reading (2000–2009)
Sorted by Change in Score

Table 3-1

	PISA Reading 2009	Change in Score (2000–2009)	Share of low performers	Share of top performers
Peru	370	43	–14.8	0.4
Chile	449	40	–17.6	0.8
Albania	385	36	–13.7	0.1
Indonesia	402	31	–15.2	0.0
Latvia	484	26	–12.5	–1.2
Israel	474	22	–6.7	3.3
Poland	500	21	–8.2	1.3
Portugal	489	19	–8.6	0.6
Liechtenstein	499	17	–6.4	–0.4
Brazil	412	16	–6.2	0.8
Korea	539	15	0	7.2
Hungary	494	14	–5.1	1
Germany	497	13	–4.2	–1.2

Source: *PISA 2009 Results: Learning Trends: Changes in Student Performance Since 2000 (Volume V)* (OECD, 2010).

of low achievers. In Poland, these are the students who would have been assigned to the vocational track. It makes sense that their achievement would rise after gaining an extra year of academic instruction.

So what's the mystery? Isn't the evidence persuasive that tracking reform was responsible for Poland's gains. Well, no, it's not. Delaying tracking by a year may have contributed to Poland's gains, but that simply cannot be concluded from PISA data. Hypothesized, yes, but concluded, no.

Don't forget that 2000 was the first year of PISA. We don't know Poland's trend before 2000. Poland's reading gains, and indeed the growth at the bottom of the distribution, may have started before the tracking reform in 1999. The 2000 score may have merely detected a trend already underway. Also, as noted above, it was the 2006 PISA scores that led to tracking reform being identified as influencing Poland's gains. Interestingly, Poland's reading score dipped 8 points in the very next release of scores, in 2009. None of the analysts who ascribed the 2000–2006 gain to the new tracking policy suspected that the 2006–2009 decline was related to the reform.

It turns out that by 2009 analysts could see that rising scores at the bottom of the achievement distribution were not only happening in Poland. As an OECD publication published after the 2009 scores explains, "In nearly all the countries that showed improved performance during the period, [2000–2009] the percentage of low performers dropped, meaning that the number of students who scored below the PISA baseline reading proficiency Level 2 was significantly smaller in 2009 than in 2000."[45]

None of the other twelve nations in Table 3-1 implemented tracking reform. And all registered gains. Indeed, six nations managed larger reading gains than Poland! And six evidenced larger gains among students performing at the bottom of the distribution. Tracking reform was not the key to these successes. And it may not be the key to Poland's.

Poland's 1999 education reforms were not limited to tracking. Instead, they involved a complete overhaul of the Polish school system, including such important elements as decentralization of authority and greater autonomy for schools, an increase in teacher salaries, a new system of national assessment, adoption of a core curriculum and national standards, reform of teacher education at the university level, and a new system of teacher promotion.[46] Any one of these policies—or several in combination—may have produced Poland's gains on PISA. Some may have even produced negative effects, dragging down achievement, while others offset the losses with larger gains. The point is this: no single reform can be plucked from several reforms adopted simultaneously and declared to have had the

greatest positive impact. Not based on PISA data. The data do not allow it.

Indeed, it is also possible that Poland's gains were not the result of policy choices at all. In a 2011 address on his nation's PISA gains, Miroslaw Sielatycki, Under-Secretary of State, Ministry of National Education in Poland, shared polling data indicating that the attitudes of the Polish people shifted dramatically around the time of the reforms. In 1993, less than half (42%) believed it was "definitely worth it" to get an education. In 2002, the percentage had jumped to 66% and in 2009 reached 68%.[47]

The public's changing view of the importance of education may have produced the gains on PISA. Or perhaps the 1999 policy changes produced a positive effect, but only conditional on shifting public attitudes, with tracking reform contributing only a tiny bit to this dynamic. No one knows for sure. To attribute Poland's 2000–2006 gains in PISA reading scores to tracking reform is a clear case of dubious causality, unsupported by the evidence.

The Problem With Rankings

Everyone loves rankings. International test results are frequently summarized by national rankings, introducing all kinds of mistakes into the interpretation of results. When the 2009 PISA scores were released, for example, the Associated Press reported that "Out of 34 countries assessed, the U.S. ranked 14th in reading, 17th in science, and 25th in math."[48] The rankings are correct, but actually 65 national and sub-national participants took the test. The 34 nations referred to in the Associated Press article are the economically developed nations belonging to the Organisation for Economic Co-operation and Development (OECD), omitting the dozens of developing countries that took the test and mostly scored lower than the U.S. The Huffington Post's headline, "U.S. Falls in World Education Rankings, Rated 'Average'" misleads readers into thinking that American performance on PISA declined from the previous time American students took the test.[49] In fact, in all three subjects, U.S. scores improved: from 495 to 500 in reading (the previous score is from 2003), from 474 to 487 in math, and from 489 to 502 in science. The "rated average" conclusion is accurate. But the U.S. has never been rated above average on PISA, and the 2009 scores show improvement for the U.S., not decline.

Beyond the misleading press accounts, focusing on rankings has several pitfalls. First, the confidence that two close rankings are truly different may not be established in terms of statistical significance. This is fundamentally important in interpreting international scores. Because national scores are derived from a randomly selected group of students, tests like PISA and TIMSS (and NAEP, for that matter) contain sampling error, noise that necessitates placing some "wiggle room" around each estimate.[50] The word "error" is not pejorative in this usage but simply refers to a statistical property that must be considered in making an estimate from sampling. Fortunately, sampling error can be calculated. We refer to scores outside this wiggle room (based on what is known as the "standard error") as statistically significantly different.

The authors of PISA and TIMSS go to great lengths preparing tables of results that incorporate sampling error into the estimates of national averages. Unfortunately, the tables are usually ignored. Let's examine a small portion of the relevant table from TIMSS 2007 4th Grade Math to see how rankings can mislead if they are interpreted incorrectly.

Figure 3-1 displays the relative rankings of the top fifteen nations in fourth-grade mathematics, rank-ordered by their average scale score (second column). Here is a basic guide to how to read the table. The buttons to the right of the average scale scores point up

To attribute Poland's 2000–2006 gains in PISA reading scores to tracking reform is a clear case of dubious causality.

Sample Excerpt From *TIMSS 2007 International Mathematics Report*

Figure 3-1

Exhibit 1.2 **TIMSS 2007 Multiple Comparisons of Average Mathematics Achievement**

Instructions: Read across the row for a country to compare performance with the countries listed along the
the average achievement of the country in the row is significantly lower than that of the comparison countr
country, or if there is no statistically significant difference between the average achievement of the two cou

Country	Average Scale Score	Hong Kong SAR	Singapore	Chinese Taipei	Japan	Kazakhstan	Russian Federation	England	Latvia	Netherlands	Lithuania	United States	Germany	Denmark	Australia	Hungary
Hong Kong SAR	607 (3.6)			▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Singapore	599 (3.7)			▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Chinese Taipei	576 (1.7)	▼	▼		▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Japan	568 (2.1)	▼	▼	▼		▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Kazakhstan	549 (7.1)	▼	▼	▼	▼						▲	▲	▲	▲	▲	▲
Russian Federation	544 (4.9)	▼	▼	▼	▼						▲	▲	▲	▲	▲	▲
England	541 (2.9)	▼	▼	▼	▼						▲	▲	▲	▲	▲	▲
Latvia	537 (2.3)	▼	▼	▼	▼						▲	▲	▲	▲	▲	▲
Netherlands	535 (2.1)	▼	▼	▼	▼								▲	▲	▲	▲
Lithuania	530 (2.4)	▼	▼	▼	▼	▼	▼	▼	▼					▲	▲	▲
United States	529 (2.4)	▼	▼	▼	▼	▼	▼	▼	▼						▲	▲
Germany	525 (2.3)	▼	▼	▼	▼	▼	▼	▼	▼	▼					▲	▲
Denmark	523 (2.4)	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼					▲
Australia	516 (3.5)	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼			
Hungary	510 (3.5)	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼			

Source: Ina V.S. Mullis et al., *TIMSS 2007 International Mathematics Report* (Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College), page 36.

Rankings must be interpreted cautiously.

or down or are blank, comparing a country on a single row with another country in a single column. Now look at each cell at the intersection of rows and columns. Down buttons mean the row country scored below the column country to a degree that is statistically significant. Up buttons mean the row country scored higher than the column country by a statistically significant amount. And cells that are blank—again, these are at the intersection of row and column countries—indicate that the two countries' scores cannot be distinguished statistically. Those countries' scores should be considered statistically equivalent; sampling error does not allow the conclusion that one country outscored the other.

Let's look at the U.S. as an example. It ranks 11th among the fourth-grade TIMSS nations. Reading across the row, the average score was 529 with a standard error (in parentheses) of 2.4 points. In the next eight cells, the down buttons indicate that the U.S. scored below eight countries to an extent that is statistically significant. We can be pretty confident, at the 95% level of confidence, that U.S. math achievement on TIMSS 2007 was lower than those eight nations. Then there are five blank cells, indicating that the U.S. performance is indistinguishable statistically from the column countries—Netherlands, Lithuania, the U.S. itself, Germany, and Denmark. Then two countries, Austria and Hungary, have up buttons. The U.S. outscored them to a statistically significant degree, along with another 21 countries out of view to the right in the original table. A statement that we can be confident in making, then, is that U.S. scored below eight countries, the same as four countries, and above twenty-three countries in mathematics.

Note how statistical significance affects any changes in ranking. If the U.S. had slipped to 12th place in this table (with a score of 525, where Germany is) or had risen to 9th (with a score of 535, the place held by Netherlands), it would make no difference in terms of reporting a statistically indistinguishable score from its actual score of 529. None. This table is from a single test, the 2007 TIMSS. Both TIMSS and PISA also produce tables that factor in statistical significance in change scores over time. That involves a different calculation than illustrated here (the participants are different, and each nation's standard error is different because each test produces a different set of test scores) but this example demonstrates the basic point. Rankings must be interpreted cautiously.

The potential for misinterpretation is exacerbated by the fact that rankings are not equal interval. That is, the distance between two rankings may be larger or smaller somewhere else in the distribution.

Look at Figure 3-1 again. If Kazakhstan's score of 549 were boosted 18 points, to 567, it would still rank the same, in 5th place behind Japan's 568. But an increase in Germany's score of 18 points, from 525 to 543, would elevate it from 12th to 7th place. These two hypothetical gains are identical in terms of scale score points, but they result in very different changes in rankings. Careless observers might be impressed by Germany's increase of 18 points but consider Kazakhstan's exact same improvement inconsequential, and they would be seriously misled as to how the two nations are performing. It cannot be stressed enough: interpret rankings cautiously.

The A+ Country Fallacy

The third misinterpretation can be dealt with quickly because it is related to the previous two. It is misleading, as shown with the case of Poland, to pull one policy from a country's entire reform agenda and proclaim that it alone produced a change in test scores. It is also misleading to focus on rankings. Add to those two errors the practice of focusing on a single policy from a single country at the top of the rankings—and a triple mistake has been committed. That's what some people do (mostly pundits, but, sadly, a few researchers who should know better) when they say "Country X is doing something I favor, and Country X scores at the top on TIMSS and PISA; therefore what I favor is a good idea."

So two maxims right up front. First, in order to determine whether a policy is good or bad, there must be variation in the policy across observable units. Put another way, if everyone is doing the same thing, you cannot tell whether that thing is good or bad. There must be variation so that the nations, states, district, or schools embracing the policy can be compared to those without it. Second, the entire range of the distribution must be examined. Focusing on the top end of any distribution is bound to mislead, especially if causality is construed from mere observables (characteristics that can be seen and recorded). Low-performing and average-performing countries have as much to offer in providing knowledge about how policies work (or don't work) as do high-performing countries.

In the U.S., advocates of a national curriculum have for years pointed to nations at the top of TIMSS and PISA rankings and argued that because those countries have national curriculums, a national curriculum must be good. The argument is without merit. What the advocates neglect to observe is that countries at the bottom of the international rankings also have a national curriculum. In fact, almost all nations have national curriculums! And the most notable countries featuring federal systems and historically having no national curriculum—Australia, Germany, the United States—have been taking steps towards either adopting a national curriculum or reducing differences among state and provincial standards.

Another example of this mistake is the recent outbreak of Finland-worship in the United States. In the 1990s, it was Japan. In addition to dominating several industries, Japan scored at the top of international tests in the 1980s and 1990s. Americans rushed to copy various aspects of Japanese education and business (e.g., lesson study, the tenets of W. Edward Deming). Now that PISA is the most prominent international test of high school students and Finland scores very high on all three subjects (reading, math, and science), it is Finland that many believe deserves emulation. Writing in *The New Republic* in 2011, Samuel E. Abrams cited small class sizes, highly-paid teachers, providing ample play time to students, social promotion, detracking, and the spare

Focusing on the top of any distribution is bound to mislead.

…the prudence of policies, if being evaluated based on international evidence, should never be judged by a single nation's reaction to them.

use of standardized tests as policy positions on which the U.S. should follow Finland.[51] Pasi Sahlberg argues that the American trend towards using value-added measures for evaluating teacher performance would be met with opposition in Finland:

> "It's very difficult to use this data to say anything about the effectiveness of teachers. If you tried to do this in my country, Finnish teachers would probably go on strike and wouldn't return until this crazy idea went away. Finns don't believe you can reliably measure the essence of learning."[52]

The irony, of course, is that Finland's exalted status in education circles comes from what Finns apparently do not believe in—measurements of learning. The PISA data do not clearly confirm or deny Sahlberg's claims about value-added evaluations of teachers, but that's not the real lesson here.[53] The lesson is that the prudence of policies, if being evaluated based on international evidence, should never be judged by a single nation's reaction to them. Nor by the experience of a few nations. Case studies can be helpful in generating hypotheses, but not in testing for causal effects that can be generalized beyond the case study nations' borders. To simply select one or two favorite policies out of everything a few nations do, even if those nations are high performing, and declare that other countries should follow their example is not very informative and certainly not good policy analysis.

Conclusion

International test scores are a valuable resource for policy makers and of great interest to the public. As this section of the Brown Center Report has illustrated, the results are also vulnerable to misinterpretation, especially when cited as evidence in political battles over the wisdom of adopting particular policies. Three misinterpretations are common. First, dubious claims of causality. Arguments have been made that Poland's tracking reforms spurred achievement gains on the PISA reading test from 2000–2009. It is plausible that tracking reform contributed to Poland's success, but the evidence is weak. Other countries accomplished gains just as large as Poland's without engaging in tracking reform. Many of them also boosted the scores of low achievers as much as Poland did. Moreover, Poland adopted several important reforms at the same time that tracking reform took place, and it is impossible to disentangle the effects of one reform from the others. Polish attitudes towards education shifted dramatically during this period and may have provided cultural support for achievement.

A second common mistake is the misuse of national rankings. The test scores underlying two adjacent rankings, or even several close rankings, may not be statistically significantly different. Rankings are not equal interval—they differ in various parts of the distribution—so a nation may jump several rankings with a gain that is actually smaller than that of a country whose ranking stays the same. Rankings must be interpreted with great caution.

Finally, the A+ country fallacy is a common mistake. Pointing to a single, high-scoring country, or a group of them, and declaring that one or more of their policies should be adopted by other countries is misguided. It combines the errors of making dubious causal claims and misusing rankings, and by ignoring evidence from low or average scoring nations on the same policy question, produces a triple error, a true whopper in misinterpreting international test scores.

NOTES

1 See the Common Core State Standards Initiative website on About the Standards, http://www.corestandards.org/about-the-standards

2 Jason L. Riley, "Was the $5 Billion Worth It?" *Wall Street Journal*, July 23, 2011.

3 John Cronin, Michael Dahlin, Deborah Adkins, and G. Gage Kingsbury, *The Proficiency Illusion* (Washington, DC: The Thomas B. Fordham Institute, 2007).

4 See Jennifer Medina, "Standards Raised, More Students Fail Tests," *New York Times*, July 29, 2010, p A1. Paul E. Peterson and Frederick Hess, "Keeping an Eye on State Standards," *Education Next*, Vol 6, No. 3, Summer 2006. Michigan Department of Education, "State Board Gives Nod to Improve Standards for State Assessment Scores," September 13, 2011, http://www.michigan.gov/mde/0,4615,7–140--262249--,00.html.

5 Sheila Byrd Carmichael, Gabrielle Martino, Kathleen Porter-Magee, and W. Stephen Wilson, *The State of State Standards and the Common Core—in 2010* (Washington, DC: The Thomas B. Fordham Institute, 2010).

6 Paul E. Peterson and Frederick Hess, "Keeping an Eye on State Standards," *Education Next*, Vol 6, No. 3, Summer 2006.

7 Tom Loveless, "Are States Honestly Reporting Test Scores?" *The 2006 Brown Center Report on American Education: How Well Are American Students Learning?* (Washington, DC: The Brookings Institution, 2006), pp. 21–29.

8 Jason L. Riley, "Was the $5 Billion Worth It?" *Wall Street Journal*, July 23, 2011.

9 U.S. Department of Education, *Chapter 8: Instructional Materials, The Final Report of the National Mathematics Advisory Panel 2008* (Washington, DC: United States Department of Education, 2008).

10 Andrew Porter, "In Common Core, Little to Cheer About," *Education Week*, August 10, 2011, pp. 24–25.

11 R. James Milgram and Sandra Stotsky, *Fair to Middling: A National Standards Progress Report* (Boston, MA: Pioneer Institute, 2010); Sandra Stotsky, and Ze'ev Wurman, *Common Core's Standards Still Don't Make the Grade: Why Massachusetts and California Must Regain Control Over Their Academic Destinies* (Boston, MA: Pioneer Institute, 2010); Ze'ev Wurman and Bill Evers, "National Standards Would Harm Math Curriculum," *San Francisco Chronicle*, July 30, 2010, p. A14.

12 Closing the Door on Innovation, http://www.k12innovation.com/Manifesto/_V2_Home.html

13 Russ Whitehurst, "Don't Forget the Curriculum," *Brown Center Letters on Education # 3* (Washington, DC: Brookings Institution, October 2009), http://www.brookings.edu/papers/2009/1014_curriculum_whitehurst.aspx

14 National Center for Education Statistics, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales* (NCES 2007–482). (Washington, DC: U.S. Department of Education, 2007).

15 V. Bandeira de Mello, *Mapping State Proficiency Standards Onto the NAEP Scales: Variation and Change in State Standards for Reading and Mathematics*, 2005–2009 (NCES 2011–458). (Washington, DC: U.S. Department of Education, 2011).

16 Curtis C. McKnight, "Model for the Second International Mathematics Study," *SIMS Bulletin*, 4 (1979), pp. 6–39.

17 A recent study of math topics taught by 7,000 teachers in 27 states shows a huge redundancy in topics. More than 70% of the math topics taught in K–8 are repeated from the previous grade. State standards are just as repetitive, but in support of the current study's findings, there is no correlation between the redundancy of state standards and the redundancy of instruction (Morgan Polikoff, in press, "The Redundancy of Mathematics Instruction in U.S. Elementary and Middle Schools," *Elementary School Journal*).

18 See *The Great Curriculum Debate: How Should We Teach Reading and Math?* edited by Tom Loveless (Washington, DC: The Brookings Institution, 2001).

19 James S. Coleman, et al., *Equality of Educational Opportunity* (Washington, DC: U.S. Dept. of Health, Education, and Welfare, Office of Education, 1966); Christopher Jencks, et al., *Inequality: A Reassessment of the Effect of Family and Schooling in America* (New York: Basic Books, Inc., 1972).

20 For contrasting positions on the inheritability of intelligence, see Richard J. Hernstein and Charles Murray, *Bell Curve: Intelligence and Class Structure in America* (New York: Free Press, 1994) and *Intelligence, Genes, and Success: Scientists Respond to The Bell Curve*, edited by Bernie Devil, Stephen E. Fienberg, Daniel P. Resnick, and Kathryn Roeder (New York: Springer-Verlag, 1997).

21 495 F. Supp. 926 (N.D.Cal, 1979).

22 Christopher Jencks, "Racial Bias in Testing" in *The Black–White Test Score Gap*, edited by Christopher Jencks and Meredith Phillips (Washington, DC: The Brookings Institution, 1998), pp. 55–85.

23 Walt Gardner, "The SAT-ACT Duel," *Education Week*, June 25, 2010.

24 Neil J. Dorans, C. F. Lyu, M. Pommerich, and W. M. Houston, "Concordance Between ACT Assessment and Recentered SAT I Sum Scores," *College and University,* 73 (1997), pp. 24–34.

25 Roy Freedle, "Correcting the SAT's Ethnic and Social-Class Bias: A Method for Reestimating SAT Scores," *Harvard Educational Review,* 73 (2003), pp. 1–43.

26 Jay Mathews, "The Bias Questions," *The Atlantic* 292 (2003), pp. 130–140; Maria Veronica Santelices and Mark Wilson, "Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning," *Harvard Educational Review,* 80 (2010), pp. 106–134.

27 Michael Kirst and Henry Rowan, "Why We Should Replace Aptitude Tests with Achievement Tests," *Education Week*, September 8, 1993, citing, John Bishop, "Incentives for Learning: Why American High School Students Compare So Poorly to their Counterparts Overseas" *Research in Labor Economics, Vol. 11* (1990), pp. 17–52.

28 http://nces.ed.gov/nationsreportcard/about/ltt_main_diff.asp

29 Pamela M. Jakwerth, Frances B. Stancavage, and Ellen D. Reed, *NAEP Validity Studies: An Investigation of Why Students Do Not Respond to Questions* (Washington, DC: American Institutes for Research, 1999).

30 For example, see *The Black–White Test Score Gap*, edited by Christopher Jencks and Meredith Phillips (Washington, DC: The Brookings Institution, 1998); *Bridging the Achievement Gap*, edited by John E. Chubb and Tom Loveless (Washington, DC: The Brookings Institution, 2002); Abigail Thenstrom and Stephan Thenstrom, *No Excuses: Closing the Racial Gap in Learning* (New York, NY: Simon & Schuster, 2003); and *Unfinished Business: Closing the Racial Achievement Gap in Our Schools*, edited by Pedro A. Noquera and Jean Yonemura Wing (San Francisco, CA: Jossey–Bass, 2006).

31 Alan Vanneman, Linda Hamilton, Janet Baldwin Anderson, and Taslima Rahman, *Achievement Gaps: How Black and White Students Perform in Mathematics and Reading on the National Assessment of Educational Progress* (Washington, DC: NCES, 2009).

32 See Jaekyung Lee "Racial and Ethnic Achievement Gap Trends: Reversing the Progress Toward Equity?" *Educational Researcher* 31, pp. 3–12; Larry V. Hedges and Amy Nowell "Test Score Convergence Since 1965," *The Black–White Test Score Gap*, edited by Christopher Jencks and Meredith Phillips (Washington, DC: The Brookings Institution, 1998), pp. 149–181.

33 See Anna Habash Rowan, Daria Hall, and Kati Haycock, *Gauging the Gaps: A Deeper Look at Student Achievement* (Washington, DC: The Education Trust, 2010) and state profiles prepared by the Center on Education Policy (www.cep-dc.org).

34 The LTT does have five performance levels, anchored at 50 point intervals on the LTT scale, but they are rarely referred to by analysts. The Main NAEP achievement levels are set at points describing what students know and are able to do, differ by each grade and subject, and employ categories (e.g., proficient) that are also used by state tests.

35 David T. Burkam and Valerie E. Lee, *Inequality at the Starting Gate* (Washington, DC: Economic Policy Institute, 2002).

36 http://nces.ed.gov/ecls/kinderassessments.asp

37 David Baker and Gerald K. LeTendre, *National Differences, Global Similarities: World Culture and the Future of Schooling* (Stanford, CA: Stanford University Press, 2005).

38 Application, analysis, synthesis, and evaluation are the four highest levels of Bloom's Taxonomy of Educational Objectives. Recall and comprehension are the two lowest levels and usually considered to be apt descriptors of basic skills. Benjamin S. Bloom, *Taxonomy of Educational Objectives* (Boston, MA: Allyn and Bacon, 1956).

39 This section is based on an address given to the 52nd IEA General Assembly in Dublin, Ireland (October 10–13, 2010). Appreciation is extended to Jan-Eric Gustafsson, co-presenter and helpful collaborator on the project. For a history of international testing by IEA, See Ina V.S. Mullis and Michael O. Martin "TIMSS in Perspective," *Lessons Learned*, edited by Tom Loveless (Washington, DC: The Brookings Institution, 2007), pp. 9–36.

40 Tom Loveless, *The Tracking Wars* (Washington, DC: The Brookings Institution, 1999).

41 Executive Summary, *PISA 2006: Science Competencies for Tomorrow's World* (OECD, 2007), p. 39.

42 Executive Summary, *PISA 2006: Science Competencies for Tomorrow's World* (OECD, 2007), p. 39.

43 The World Bank, "Successful Education Reform: Lessons from Poland," 2010, *Europe and Central Asia Knowledge Brief*, 34 (2010), p. 3.

44 Eric A. Hanushek and Ludger Woessmann, "Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries," *Economic Journal, Royal Economic Society*, 116 (2005), pp. C63–76, 03. The study's coding the U.S. as a late tracking nation (after age 15) does not make sense for two reasons. Based on the European-style tracking that the study uses to define tracking (separating students by ability into separate schools) the U.S. basically *never* tracks, except for a few elite exam schools in a small number of big cities. Moreover, coding the U.S. as a non-tracking nation before age 15 obscures the huge amount of tracking that goes on in middle school mathematics courses and the common practice of within-class ability grouping beginning with first grade reading instruction. Bottom line: the U.S. tracks and ability groups students a lot before age 15 but almost never uses the between-school model of separation and differentiation.

45 OECD, "Improving Performance: Leading from the Bottom," *PISA In Focus*, 2 (2011), p. 1.

46 Miroslaw Sielatycki, Under-Secretary of State, Ministry of National Education, "Poland: Successes and Challenges Educational Reform," 14th OECD Japan Seminar, Tokyo, Japan, June 28–29, 2011.

47 Miroslaw Sielatycki, Under-Secretary of State, Ministry of National Education, "Poland: Successes and Challenges Educational Reform," 14th OECD Japan Seminar, Tokyo, Japan, June 28–29, 2011.

48 Christine Armario, "'Wake-Up Call': U.S. Students Trail Global Leaders," *Associated Press Wire*, December 7, 2010.

49 "The U.S. Falls In World Education Rankings, Rated 'Average,'" *The Huffington Post*, December 7, 2010.

50 Means reported from tests given to census populations—in other words, to all students—do not have sampling error. They are simply the average of all test scores. State tests given to all students in a particular grade or for high school graduation are examples of census populations. This does not automatically make them better tests because tests may possess other types of error. The proper term for the "wiggle room" of an estimate is "confidence interval."

51 Samuel E. Abrams, "The Children Must Play: What the United States Could Learn from Finland About Education Reform," *The New Republic*, January 28, 2011.

52 Hechinger Report, "What Can We Learn from Finland?" *Hechinger Report*, December 9, 2010, available at hechingerreport.org.

53 *PISA 2006 Science Competencies for Tomorrow's World* (OECD, 2006), pp. 240–244, 276–277. *PISA 209 Results: What Students Know and Can Do* (OECD, 2009), pp. 46–47.

THE BROOKINGS INSTITUTION

STROBE TALBOTT
President

DARRELL WEST
Vice President and Director
Governance Studies Program

BROWN CENTER STAFF

GROVER "RUSS" WHITEHURST
Senior Fellow and Director

TOM LOVELESS
Senior Fellow

MATTHEW CHINGOS
Fellow

MICHAEL GALLAHER
Research Analyst

SARAH WHITFIELD
Staff Assistant

PAUL T. HILL
Non-resident Senior Fellow

DIANE RAVITCH
Non-resident Senior Fellow

Views expressed in this report are solely those of the author.

BROWN CENTER on
Education Policy
at BROOKINGS

BROOKINGS
1775 Massachusetts Avenue, NW • Washington, D.C. 20036
Tel: 202–797–6000 • Fax: 202–797–6004
www.brookings.edu

The Brown Center on Education Policy
Tel: 202–797–6090 • Fax: 202–797–2480
www.brookings.edu/brown