Note: Please also use this writer as reference for the DQ as well.

Williamson, S. C. (2009). Assessment and evaluation 10/11. Boston: McGraw Hill.

Article 1 Assessments and Accountability (Condensed version)

Robert L. Linn

Assessment and accountability have played prominent roles in many of the education reform efforts during the past 50 years. In the 1950s, under the influence of James B. Conant's work on comprehensive high schools, testing was used to select students for higher education and to identify students for gifted programs. By the mid-1960s test results were used as one measure to evaluate the effectiveness of Title I and other federal programs. In the 1970s and early 1980s, the minimum competency testing movement spread rapidly; 34 states instituted some sort of testing of basic skills as a graduation requirement. Overlapping the minimum competency testing movement and continuing into the late 1980s and early 1990s was the expansion of the use of standardized test results for accountability purposes.

Assessment is appealing to policymakers for several reasons: it is relatively inexpensive compared to making program changes, it can be externally mandated, it can be implemented rapidly, and it offers visible results. This Digest discusses significant features of present-day assessment programs and offers recommendations to increase positive effects and minimize negative ones.

What Are the Characteristics of Current Reform Efforts?

Although a number of other important features might be considered in any discussion of assessment and education reform (e.g., the emphasis on performance-based approaches to assessment, the concept of tests worth teaching to, and the politically controversial and technically challenging issue of opportunity to learn), I focus on the following three:

• An emphasis on the development and use of ambitious content standards as the basis of assessment and accountability.

• The dual emphasis on setting demanding performance standards and on the inclusion of all students.

• The attachment of high-stakes accountability mechanisms for schools, teachers, and sometimes, students.

Content standards. The federal government has encouraged states to develop content and performance standards that are demanding. Standards-based reform is also a central part of many of the state reform efforts, including ones such as Kentucky and Maryland that have been using standards-based assessments for several years and ones such as Colorado and Missouri that have more recently introduced standards-based assessment systems. A great deal has been written about the strengths and weaknesses of content standards (e.g., Education Week, 1997; Lerner, 1998; Olson, 1998; Raimi & Braden, 1998).

It is worth acknowledging that content standards vary a good deal in specificity and in emphasis. Content standards can, and should, if they are to be more than window dressing, influence both the choice of constructs to be measured and the ways in which they are eventually measured.

Performance standards. Performance standards are supposed to specify how good is good enough. There are at least four critical characteristics of performance standards. First, they are intended to be absolute rather than normative. Second, they are expected to be set at high, world-class levels. Third, a relatively small number of levels (e.g., advanced, proficient) are typically identified. Finally, they are expected to apply to all, or essentially all, students, rather than a selected subset such as college-bound students seeking advanced placement.

Should the intent be to aspire not just to high standards for all students, but to the same high standards for all students and on the same time schedule for all students (e.g., meet reading standards in English at the end of Grade 4)? Coffman (1993) sums up the problems of holding common high standards for all students as follows: “Holding common standards for all pupils can only encourage a narrowing of educational experiences for most pupils, doom many to failure, and limit the development of many worthy talents” (p. 8). Although this statement runs counter to the current zeitgeist and may not even be considered politically correct, it seems to me a sensible conclusion that is consistent with both evidence and common sense. Having high standards is not the same as having common standards for all, especially when they are tied to a lock step of age or grade level.

High-stakes accountability. The use of student performance on tests in accountability systems is not new. Examples of payment for results such as the flurry of performance contracting in the 1960s can be found cropping up and fading away over many decades. What is somewhat different about the current emphasis on performance-based accountability is its pervasiveness. As Elmore, Abelmann, and Fuhrman note, “What is new is an increasing emphasis on student performance as the touchstone for state governance” (1996, p. 65). Student achievement is being used not only to single out schools that require special assistance, but also to provide cash incentives for imrovements in performance. Yet several fundamental questions remain about the student assessments, the accountability model, and the validity, impact, and credibility of the system.

As noted earlier, for example, the choice of constructs matters. Content areas (and subareas within those content areas) that are assessed for a high-stakes accountability receive emphasis while those that are left out languish. Meyer (1996) has argued that “in a high-stakes accountability system, teachers and administrators are likely to exploit all avenues to improve measured performance. For example, teachers may ‘teach narrowly to the test.’ For tests that are relatively immune to this type of corruption, teaching to the test could induce teachers and administrators to adopt new curriculums and teaching techniques much more rapidly than they otherwise would” (p. 140).

It is unclear, however, that there is either the know-how or the will to develop assessments that are sufficiently “immune to this type of corruption.” It is expensive to introduce a new, albeit well-equated, form of a test on each new administration. And if ambitious performance-based tasks are added to the mix, still greater increases in costs will result.

A second area of concern regarding high-stakes assessments relates to what data the basic model should employ. Some possibilities include current status, comparisons of cross-sectional cohorts of students at different grades in the same year, comparisons of cross-sectional cohorts in a fixed grade from one year to the next, longitudinal comparisons of school aggregate scores without requiring matched individual data, and longitudinal comparisons based only on matched student records. Should simple change scores be used or some form of regression-based adjustment? And, if regression-based adjustments are used, what variables should be included as predictors? In particular, should measures of socioeconomic status be used in the adjustments?

Elmore, Abelmann, and Furhman (1996) present both sides of this issue, noting that on the one hand, schools can fairly be held accountable only for those factors they can control, but on the other, controlling for student background or prior achievement institutionalizes low expectations for poor, minority, low-achieving students (pp. 93–94). Kentucky's interesting approach to this dilemma has been to set a common goal for all schools by the end of 20 years, thus establishing faster biennial growth targets for initially low-scoring schools than initially high-scoring schools (Guskey, 1994).

The biggest question of all is whether the assessment-based accountability models that are now being used or being considered by states and districts have been shown to improve education. Unfortunately, it is difficult to get a clear-cut answer to this simple question. Certainly, there is evidence that performance on the measures used in accountability systems increases over time, but that can also be linked to the use of old norms, the repeated use of test forms year after year, the exclusion of students from participating in accountability testing programs, and the narrow focusing of instruction on the skills and question types used on the test (see Koretz, 1988; Linn et al., 1990; Shepard, 1990). Comparative data are needed to evaluate the apparent gains. The National Assessment of Educational Progress provides one source of such data. Comparisons of state NAEP and state assessment results sometimes suggest similar trends; for example, increases in numbers of students scoring at or above basic or proficient levels on NAEP may track with improved state test scores over time. In other cases, the trends for a state's own assessment and NAEP will suggest contradictory conclusions about the changes in student achievement. Divergence of trends does not prove that NAEP is right and the state assessment is misleading, but it does raise important questions about the generalizability of gains reported on a state's own assessment, and hence, about the validity of claims regarding student achievement.

How Can Assessments Be Used More Wisely?

Assessment systems that are useful monitors lose much of their dependability and credibility for that purpose when high stakes are attached to them. The unintended negative effects of the high-stakes accountability uses often outweigh the intended positive effects. It is worth arguing for more modest claims about uses that can validly be made of our best assessments and warning against the over-reliance on them that is so prevalent and popular. To enhance the validity, credibility, and positive impact of assessment and accountability systems while minimizing their negative effects, policymakers should:

1. Provide safeguards against selective exclusion of students from assessments.

2. Make the case that high-stakes accountability requires new high-quality assessments each year that are equated to those of previous years.

3. Don't put all of the weight on a single test. Instead, seek multiple indicators. The choice of construct matters and the use of multiple indicators increases the validity of inferences based upon observed gains in achievement.

4. Place more emphasis on comparisons of performance from year to year than from school to school. This allows for differences in starting points while maintaining an expectation of improvement for all.

5. Consider both value added and status in the system. Value added provides schools that start out far from the mark a reasonable chance to show improvement while status guards against institutionalizing low expectations for those same students and schools.

6. Recognize, evaluate, and report the degree of uncertainty in the reported results.

7. Put in place a system for evaluating both the intended positive effects and the more likely unintended negative effects of the system.

References

Coffman, W. E. (1993). A king over Egypt, which knew not Joseph. Educational Measurement: Issues and Practice 12(2), 5–8.

Education Week (1997, January 22). Quality counts: A report card on the condition of public education in the 50 states. A Supplement to Education Week, vol. 16.

Elmore, R. F., Abelmann, C. H., & Fuhrman, S. H. (1996). The new accountability in state education reform: From process to performance. In H. F. Ladd (Ed.), Holding schools accountable: Performance-based reform in education (pp. 65–98). Washington, DC: The Brookings Institution.

Guskey, T. R. (Ed.) (1994). High stakes performance assessment: Perspectives on Kentucky's reform. Thousand Oaks, CA: Corwin Press.

Koretz, D. (1988). Arriving at Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? American Educator 12(2), 8–15, 46–52.

Lerner, L. S. (1998). State science standards: An appraisal of science standards in 36 states. Washington, DC: Thomas B. Fordham Foundation.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that “everyone is above average.” Educational Measurement: Issues and Practice 9(3), 5–14.

Meyer, R. H. (1996). Comments on chapters two, three, and four. In H. F. Ladd (Ed.), Holding schools accountable: Performance based reform in education (pp. 137–145). Washington, DC: The Brookings Institution.

Olson, L. (1998, April 15). An “A” or a “D”: State rankings differ widely. Education Week 17, 1, 18.

Raimi, R. A., & Braden, L. S. (1998). State mathematics standards: An appraisal of science standards in 46 states, the District of Columbia, and Japan. Washington, DC: The Thomas B. Fordham Foundation.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? Educational Measurement: Issues and Practice 9(3), 15–22.

From Practical Assessment, Research & Evaluation, January 2001, pp. 1–3. Copyright © 2001 by Robert L. Linn. Reprinted by permission of the author.

Article 2 Why Has High-Stakes Testing So Easily Slipped into Contemporary American Life?

Sharon L. Nichols

David C. Berliner

Ms. Nichols and Mr. Berliner suggest five reasons that high-stakes testing has become such a popular approach to “fixing” schools. What is interesting is that none of the reasons has to do with evidence that this approach will actually work in its intended ways.

High-stakes testing is the practice of attaching important consequences to standardized test scores, and it is the engine that drives the No Child Left Behind (NCLB) Act. The rationale for high-stakes testing is that the promise of rewards and the threat of punishments will cause teachers to work more effectively, students to be more motivated, and schools to run more smoothly—all of which will result in greater academic achievement for all students, but especially those from poverty and minority backgrounds. Although it is certainly arguable, we believe that, to date, there is no convincing evidence that high-stakes testing has the intended effect of increasing learning.1 By contrast, there is a growing literature suggesting that the unintended consequences are damaging to the education of students.2

Collateral Damage of High-Stakes Testing

In our recent book, we use Donald Campbell's law to illustrate how the high-stakes testing provision of NCLB has wreaked havoc with our education system, causing irreversible harm to many of our nation's youths and educators. Campbell's law states: “The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures, and the more apt it will be to distort and corrupt the social processes it is intended to monitor.”3 Under the current system of high-stakes testing, this is exactly what is happening. The pressure to score well on a single test is so intense that it leads to nefarious practices (cheating on the test, data manipulation), distorts education (narrowing the curriculum, teaching to the test), and ends up demoralizing our educators.

Perhaps the most visible and noticeable of the areas in which Campbell's law operates is the business world, where economists have long recognized the possibility for corruption when stakes are high. Despite the research, some businesses are structured such that incentives are especially weighty and salient. Such incentives as big bonuses for increased sales or for spending less time with patients increase the likelihood of corruption. Salespeople or physicians in such situations often take short cuts to obtain the incentives available. Of course, it is not surprising that many examples exist of how incentives in business can corrupt individuals. The pursuit of money, prestige, and power, as we all know, often leads to behavior that is unseemly, if not immoral or even illegal. Enron, sadly, is not an anomaly.

But it isn't only the business world where Campbell's law plays out so predictably. Corruption, cheating, gaming the system, taking short cuts, and so forth—all exist in the fields of medicine, athletics, academe, politics, government agencies, and the military. Given the widespread applicability of this social science law regarding corruption in the presence of a single highly valued indicator, we asked ourselves, Why has high-stakes testing so easily become a part of contemporary American life? We offer five reasons—and our thoughts on each—for why high-stakes testing has been so easily embraced by a culture looking for a way to judge and monitor the progress of the public schools.

The ‘Business’ of Education

First, and the most popular explanation, is one that notes the co-evolution of the prominence of business and accountability in our daily lives. In recent decades business has come to dominate a great deal of American cultural life through its influence on the media and on a broad range of policy at all levels of government. Tax policy, government spending, health care, employment training, and education policy have all been strongly influenced by business through the efforts of lobbyists and highly visible CEOs. As the influence of business on government has risen over the last few decades, so have business' interest in the skill set possessed by graduates of our schools and its concern for how tax dollars are used to support education. So basic business 101 models were applied to our schools: namely, ways were found first to monitor productivity, then to increase it, and finally to do so without spending any more money.

Tests were chosen as the means of measuring productivity. It was believed by the business community that productivity could be increased without spending more money simply by holding schools and educators accountable through the practice of high-stakes testing. Lazy teachers and students would be discovered and made to work harder. The models of accountability used in business could be applied to the inefficient school systems of America and, voilà, the schools would improve. Or they could be closed down or turned over to private entrepreneurs. For many Americans, these policies seemed sensible and worth pursuing, so it was easy to buy into the high-stakes accountability movement.

But the analogy doesn't really fit, because it is easier to judge the number and quality of widgets coming off an assembly line than to determine the knowledge and skill possessed by students. A widget is a widget, but a well-educated student is both a good citizen and a caring person, as well as someone with aesthetic sensibilities, good habits of health, and so forth. These are outcomes our citizens demand that we produce through our schools, but they are never assessed by tests.

Thus productivity for our teachers and our schools has a vastly different meaning than does productivity in a manufacturing plant or in the delivery of routine services. Furthermore, when inputs cannot be controlled, it is hard to assess a process by its outputs. Measuring the production of widgets assumes control over the quality of the raw materials needed to produce widgets. But in education we have little control over the input side. A class that contains two emotionally disturbed children or two English-language learners or many more boys than girls will inevitably affect its teacher's productivity, as measured by test scores. In addition, mobility rates of 40% or 50% at the school level, and much higher rates in particular classrooms, mean we are holding schools and teachers accountable for students they never had much chance to teach. So while ordinary ways of measuring productivity appear to be sensible, they do not work as well in educational settings. The high-stakes tests, with their threats and incentives to boost productivity, are not well matched to the ways our schools operate. Thus scores on tests will mislead us about genuine productivity. But it all sounds quite sensible and so appeals to many citizens who end up supporting the use of high-stakes testing programs for our schools.

The World Is Flat ... Isn't It?

A second and related reason high-stakes testing has slipped into the routines of our culture is the emerging belief on the part of both business and government that the future economy depends on a highly educated work force. This belief took on new urgency after Thomas Friedman's book The World Is Flat became a best seller.4 Large numbers of Americans now believe that we need to push all our children to the highest levels of education, moving most students to high school graduation with a degree that guarantees mastery of a rigorous curriculum. After that, the story goes, most graduates need to move into degree-granting two- and four-year colleges. Obviously, the demand for a rigorous curriculum and college-level preparation means a seriousness about testing in our public schools never before required. High-stakes testing is compatible with these national ambitions. High-stakes testing fits neatly into the American mindset that to be competitive in the global economy we need high rates of college graduation, especially in the STEM (science, technology, engineering, and math) fields.

In fact, this whole theory may be wrongheaded. As Dennis Redovich has reported in article after article, the employment profile of the future does not support the need for a big increase in the mathematical and scientific knowledge of our youths.5 We may well be demanding more than we will need in these areas and already producing enough scientists and enough college graduates for the needs of the economy. Certainly a scientifically sophisticated citizenry is in our national interest, but making advanced mathematics and science a major goal of U.S. education may be counterproductive. Creating a rigorous, high-quality science and math curriculum for those who will not be majoring in one of the STEM fields may be a better goal than putting all high school students through courses designed for the future college majors in these fields. The current system contributes to both student anomie and the already too high dropout rate.

But earnings are also an issue. Decades ago, those who failed to graduate from high school experienced a drop in real wages. That drop was followed only a short time later by a drop in real wages for those who had only a high school education. Now, even those with college degrees are suffering the same fate. Earnings for workers with four-year degrees fell 5.2% from 2000 to 2004, when adjusted for inflation, according to White House economists.6 Apparently, large percentages of recent college graduates are taking jobs for which no college degrees are necessary, and the trend may be accelerating. Nevertheless, we continue to demand that the education system produce ever-increasing numbers of high school and college graduates, though we may actually now be near record levels of high school graduation rates.7

Despite the fact that our national productivity is much more dependent on our tax structures, relative lack of corruption, and remarkable entrepreneurship, the citizenry believes that we need better schools to be competitive in the world economy. And although the goal of better schools should be a national priority, bringing them into conformity with what our colleges and universities demand should not. Yet high-stakes testing virtually ensures that schools will force students into submitting to these uniform goals. Unknowingly, high-stakes testing has easily slipped into our everyday life as the solution for the misguided goals of advanced achievement for all students in a narrowed curriculum.

Old, White, and Self-Serving

The third reason for the ease with which high-stakes tests have become commonplace in our culture is changing demographics. We can now see clearly the shape of an emerging gerontocracy. An older citizenry, much whiter than the youths of the nation and relatively well off financially, is now likely to outlive its resources and is beginning to act politically in its own best interests.8 As a powerful political and economic force, these folks will want income and services. They will demand medical, pharmaceutical, and social services; full payment of social security; and some form of housing support as their income stays relatively fixed. They will not want to spend much on youths—especially youths of color—whom they perceive as lazy and unappreciative. For many people in this category, high-stakes testing separates the deserving poor from the undeserving poor. It becomes, in effect, a policy mechanism to preserve social status more than to improve our schools. High-stakes testing subtly fits the mindset of this growing demographic group and thus makes it easier for this policy to gain purchase in our contemporary society.

Power Elite

A fourth reason is related to a new and larger power elite among the citizenry, along with the vast middle and upper-middle class whose children now attend good public schools and who see high-stakes testing working to their own children's advantage. While they bristle that their own children must suffer through these tests (e.g., the Scarsdale, Westchester County, mothers' rebellion9), the schools their children attend are not much bothered by the tests, and the pass rates for their children are very high. Thus on a day-to-day basis, many of these citizens are largely unconcerned about the impact of high-stakes tests.

But we think that the unnoticed slipping of high-stakes testing into our culture has taken place partly because it hits our poorest and most racially diverse student body hardest and thereby forces the kind of education on the children of the poor that ensures that they cannot compete successfully with the children of the wealthy. The drill-and-test-prep education we see in schools for the poor does not prepare them for knowledge workers' jobs or for the best universities in the nation. This makes room for the children of the more privileged. Since the status of children from advantaged families is preserved though high-stakes testing, it is easy for these folks to defend their use.

Middle- and upper-class Americans largely saw no reason to oppose high-stakes testing for accountability when it was first proposed because they knew that their children would do well. But even if their children were in danger of not succeeding on such tests, middle-class families always had the intellectual and financial resources needed to ensure their children's success. Thus high-stakes tests slipped easily into the culture because, by and large, the power elite didn't fore see a problem for themselves.

Middle- and upper-class Americans saw no reason to oppose high-stakes testing for accountability when it was first proposed because they knew that their children would do well.

Five years later, many middle-class parents and students are speaking out against high-stakes tests. Some do it because of how it affects them, but, thankfully, others argue that the system is unfair and unjust for others. For example, high school student John Wood refused to take his high school exit exam on grounds that the test is biased and unjust. Even though he would certainly pass, the decision cost him his diploma. In spite of these impassioned voices, relatively wealthy, higher-social-status politicians on both sides of the aisle continue to defend high-stakes tests as the solution to all our educational problems.

Sports Enthusiasts

Fifth—and least noted by any commentators we have read on the subject—is the fit between high-stakes testing and other spectacles that the public enjoys, such as baseball, football, basketball, or hockey. We are a game-playing, competition-seeking nation, and high-stakes testing fits easily into such a culture.

As is true of many sports, high-stakes testing has a tournament-like quality to it, bringing seasonal excitement to fans who now can follow the heavily publicized “winning” and “losing” streaks of their local schools, as they have often followed their local teams. Every summer when spring test results are released, there is a flood of publicity and great fanfare about how well (or poorly) our nation's teachers and students performed in the previous year. And like rabid fans who delight in watching rivals have a losing season, the American media feed on whatever bad news exists.10 Those who follow the news ponder endlessly about why certain schools/teams fail. How many times has this school/team failed in the past? What is its track record? What schools/teams might need to be reconstituted or even closed down or moved? What will we do to get rid of the bad teachers/players, and precisely which ones are they? Is it the science teacher or the first baseman, the English teachers or the defensive line, the coach or the principal? Exactly whom can we pin this failure on?

Numerous similarities between sports and testing explain the country's fascination with testing. After all, a match in the sport of cricket is called a test. Professional athletes in cricket and in most other sports practice hours and hours, repeating the same activities endlessly so that their responses at “test” time will be automatic. In the high-stakes-testing game, teachers also engage their students in endless repetitive activities to better ensure that students' responses are accurate and automatic come test time. In professional sports, teams with the highest-paid athletes are more likely to have winning seasons. Similarly, schools with more resources and those that serve the most affluent students tend to perform better academically.11 In professional sports, fans are immersed in statistics that highlight the successes and failures of their favorite teams and players; in the testing game, parents, politicians, and other community members are immersed in media coverage of academic data showing who is winning and who is losing.

Of course, we know stats say little about a player's many other contributions to the team, such as level of dedication, commitment, morale, and leadership. Similarly, when teachers and administrators are judged by their students' scores, we don't take any account of teachers' many other contributions, such as their nurturance of a love for learning, individual counseling of students in times of need, extra time spent meeting with students' families, provision of money from their own pockets for classroom items, and so forth.

High-Stakes Testing: We Are against It

High-stakes testing is now a part of our culture, and we are against it. It has come to prominence, we think, because it fits easily into contemporary ways of thinking about our nation and ourselves. We are a political and an economic system dominated by the interests of big business, and so business models of accountability for our schools naturally follow. High-stakes testing seems to be a hard-headed business practice brought to bear on the schools, despite the fact that no one uses such a system in knowledge-oriented businesses. And unless we are greatly mistaken, schools still fall into that category.

High-stakes testing also seems to help with preparing us for the vicissitudes of a competitive world economy, and so it is easily embraced. The argument that the new American economy may be vastly more service oriented than previously believed and that it may not require nearly as many college graduates as is now thought necessary is a point of view that is ignored.

The needs of the emerging gerontocracy and those who already have some status in society are also served by high-stakes testing. And high-stakes testing fits neatly into the gaming and spectacle seeking that so permeate the U.S. cultural scene.

For all these reasons high-stakes testing has grown to be an acceptable part of the culture. Those who oppose the spread of high-stakes testing are seen as status-quo oriented, against quality in education, against school improvement, obstructionist, anti-efficiency, anti-George W. Bush, and so forth.

But we are actually against high-stakes testing for none of these reasons. We oppose it for the same reason we are against forcing everyone to participate in extreme sports. If any person voluntarily chooses to jump the Grand Canyon on a motorcycle, scale Everest, or BASE jump, we wish them luck. We just don't think everyone should be required to engage in the same high-stakes sports because, if everyone did, lots of people would be hurt. We are against high-stakes testing for the same reason. If a person volunteers to take exams for the medical boards, the bar, or a pilot's license, that individual should be encouraged to follow a dream. But not all of us should be forced to take and fail such exams. In the current high-stakes environment, teachers, students, parents, and American education are being hurt by required high-stakes testing. This policy is corrupting our education system and needs to be stopped.

References

1. See, for example, Sharon L. Nichols, Gene V. Glass, and David C. Berliner, “High-Stakes Testing and Student Achievement: Does Accountability Pressure Increase Student Learning?,” Education Policy Analysis Archives, vol. 14, 2006, http://epaa.asu.edu./epaa/v14n1; and Sharon L. Nichols, “High-Stakes Testing: Does It Increase Achievement?,” Journal of Applied School Psychology (in press). By contrast, a more recent study suggests that high-stakes testing under NCLB is associated with increases in student achievement. See Answering the Question That Matters Most: Has Student Achievement Increased Since No Child Left Behind? (Washington, D.C.: Center on Education Policy, June 2007). Available at www.cep.org, or readers may google the title.

2. Monty Neill et al., Failing Our Children: How No Child Left Behind Undermines Quality and Equity in Education and an Accountability Model That Supports School Improvement (Cambridge, Mass.: FairTest, 2004); Gary Orfield et al., Losing Our Future: How Minority Youth Are Being Left Behind by the Graduation Rate Crisis (Cambridge, Mass.: Civil Rights Project at Harvard University, 2004); Jaekyung Lee, Tracking Achievement Gaps and Assessing the Impact of NCLB on the Gaps: An In-Depth Look into National and State Reading and Math Outcome Trends (Cambridge, Mass.: Civil Rights Project at Harvard University, 2006); and M. Gail Jones, Brett Jones, and Tracy Hargrove, The Unintended Consequences of High-Stakes Testing (Lanham, Md.: Rowman & Littlefield, 2003).

3. Donald Campbell, “Assessing the Impact of Planned Social Change,” in Gene Lyons, ed., Social Research and Public Policies: The Dartmouth/OECD Conference (Hanover, N.H.: Public Affairs Center, Dartmouth College, 1975).

4. Thomas L. Friedman, The World Is Flat: A Brief History of the Twenty-First Century (New York: Farrar, Straus & Giroux, 2006).

5. Dennis W. Redovich, The Big Con in Education: Why Must “All” High School Graduates Be Prepared for College? (New York: Universe, 2005).

6. Molly Hennessy-Fiske, “That Raise Might Take 4 Years to Earn as Well: Those with Bachelor's Degrees Are Finding Their Incomes Stagnate Despite a Growing Economy,” Los Angeles Times, 24 July 2006.

7. Lawrence Mishel and Joydeep Roy, Rethinking Graduation Rates and Trends (Washington, D.C.: Economic Policy Institute, 2006); and idem, “Accurately Assessing High School Graduation Rates,” Phi Delta Kappan, December 2006, pp. 287–92.

8. Gene V. Glass, “Fertilizers, Pills, and Robots: The Fate of Public Education in America,” lecture presented at the annual meeting of the American Educational Research Association, San Francisco, 2006.

9. “Scarsdale Parents Call Test Boycott a Success,” New York Times, 4 May 2001, p. B-1; and Michael Powell, “In NY, Putting Down Their Pencils: Parent Rebellion Against Standardized Testing Strikes at Heart of Bush Plan,” Washington Post, 18 May 2001, A-1.

10. See Sharon Nichols and Tom Good, American Teenagers—Myths and Realities: Media Images, Schooling, and the Social Costs of Careless Indifference (Mahwah, N.J.: Erlbaum, 2004), for a discussion on how the media exaggerate, distort, and overrepresent the bad news about American youths. See also Michael Males, Kids and Guns: How Politicians, Experts, and the Press Fabricate Fear of Youth (Philadelphia: Common Courage Press, 2000); and idem, The Scapegoat Generation: America's War on Adolescents (Philadelphia: Common Courage Press, 1996). See also David C. Berliner and Bruce J. Biddle, “The Lamentable Alliance Between the Media and School Critics,” in Gene Maeroff, ed., Imaging Education: The Media and Schools in America (New York: Teachers College Press, 1998).

11. David C. Berliner, “Our Impoverished View of Educational Research,” Teachers College Record, vol. 108, 2006, pp. 949–95; and Bruce J. Biddle, ed., Social Class, Poverty, and Education (New York: Routledge Falmer, 2001).

Sharon L. Nichols is an assistant professor in the Department of Counseling, Educational Psychology, and Adult and Higher Education at the University of Texas, San Antonio. David C. Berliner is a Regents' Professor of Education in the Department of Educational Leadership and Policy Studies at Arizona State University, Tempe. They are the authors of Collateral Damage: How High-Stakes Testing Corrupts America's Schools (Harvard Education Press, 2007). © 2008, Sharon L. Nichols.

From Phi Delta Kappan, May 2008, pp. 672–676. Copyright © 2008 by Sharon Nichols. Reprinted by permission of the author.

Article 3 Assessment around the World

Iris C. Rotberg

How does NCLB fit in an international context? Here's what's happening in the rest of the world.

Standardized testing is controversial everywhere, regardless of its purpose. Most countries use testing for tracking and for selecting students for admission into academic secondary schools or universities, but generally not for holding educators accountable. Many countries don't even administer standardized tests until the later grades. In fact, most Canadian universities don't require the Scholastic Aptitude Test (SAT) or other standardized admissions tests—except for students applying with a U.S. high school diploma! (Ghosh, 2004)

In a recent collection of studies of education systems worldwide, which I edited,1 numerous experts discussed current education policies in their countries, including the role that standardized testing plays in their public schools (Rotberg, 2004). I draw on these overviews here to set No Child Left Behind (NCLB) in the context of testing across the globe.

The current preoccupation with test-based accountability in the United States is founded on several misperceptions about other countries' practices as well as about international test score comparisons and the extent to which test scores are valid indicators of the quality of education or the state of the economy. These assumptions have dominated U.S. public policy dialogue for decades.

Assumption: The Rest of the Developed World Is One (High-Achieving) Country

Much of the rhetoric about international test score comparisons treats the rest of the developed world as though it were one mythical country that does a better job of educating students than the United States does. However, the rhetoric does not recognize the significant differences in student academic achievement among developed countries; the level and distribution of education funding; the extent to which schools track students by academic ability; secondary school and university enrollment rates; and perhaps most important, the quality of education that each country offers low-income students, minority students, students with disabilities, language-minority students, and recent immigrants.

Assumption: Other Countries Have Found the “Right” Way to Improve Student Achievement

Many people in the United States assume that other countries have centralized education systems and that the resulting standardization is the magic bullet for improving student achievement. This assumption ignores the fact that many countries question that policy. France, for example, is reassessing its highly centralized education system because it doesn't meet the needs of an increasingly diverse immigrant population. Many other countries, such as China, Israel, and Sweden, are moving from a centralized to a decentralized system of governance. Australia, Canada, and Germany—countries with long-standing decentralized systems—envision little change. In addition, no evidence supports the contention that organizational structure, whether centralized or decentralized, bears any relationship to academic achievement or the ability to compete in the global economy.

Assumption: International Test Score Rankings Are Valid Measures of the Quality of Education

Data do not support the causal relationships that many people establish on the basis of international rankings. If a country ranks high on a given international comparison, people assume that its schools must be “good”; if the country ranks low, its schools must be “bad.” The problem is, international test score comparisons are virtually impossible to interpret, not only because of enormous differences among nations in poverty rates and in societal values and objectives, but also because of major sampling problems, which make it difficult to ensure that comparable samples of students, schools, and regions are being tested across countries.

Assumption: A Country's Ranking on Test Score Comparisons Predicts Its Ability to Compete in the Global Economy

This assumption has been repeated in various guises for the past 40 years, with little evidence to support it. The fact is, many countries typically perceived as high-scoring on international test score comparisons—such as Austria, France, Sweden, Switzerland, and the United Kingdom—are not the United States' main economic competitors. The United States outsources technical jobs because U.S. consumers are unwilling to accept higher prices and U.S. workers are unwilling to accept lower wages to compete with computer programmers in India or with computer manufacturers in China. These two countries enroll only about one-third of a given age group by the final years of secondary school. Moreover, the education systems of Western European countries and Japan have not immunized their economies against competition from less developed countries with significantly lower rates of literacy and lower enrollment in secondary and higher education.

Assumption: Testing Can Help the United States Address the Problems That Poverty Has Created

The rhetoric about NCLB ignores the overwhelming impact of poverty, the primary correlate of low academic achievement in every country (Grissmer, Kirby, Berends, & Williamson, 1994). Although the size of the achievement gap among students from different socioeconomic backgrounds may vary among countries, the existence of this gap is universal.

The link between poverty and achievement is higher in the United States than in many other industrialized countries. This is not surprising, given the fact that the United States has both larger income gaps between rich and poor (Blackburn, 1997) and fewer social support systems than most industrialized countries do. But high-stakes testing, accountability requirements, and centralization cannot cure the problems associated with poverty. As one educator put it,

We believe that schools solve the problem of poverty, and now this program [NCLB] assumes that tests solve the problem of schools. By implication, that means tests are supposed to solve the problem of poverty. (Rotberg, Bernstein, & Ritter, 2001, p. 14)

Assumption: Countries That Score High on International Test Score Comparisons Hold Their Educators Accountable for Students' Scores on Standardized Tests

In reality, few countries hold educators accountable for students' test scores. Many of the countries that the United States most admires for their rankings on international comparisons—for example, Canada, Finland, France, Japan, and Sweden—do not use tests to hold educators accountable. Some do not even administer standardized tests until secondary school.

Many of the countries that the United States most admires for their rankings on international comparisons do not use tests to hold educators accountable.

It is ironic that many countries throughout the world are attempting to reduce their emphasis on rote learning, whereas current testing pressures in the United States promote just that kind of learning. NCLB supporters believe that because the legislation makes schools' “failures” public, it encourages educators to try harder to focus on important academic subject matter and pay more attention to marginalized students. Those opposed to NCLB are concerned that the pressure to raise test scores will encourage educators to narrow the curriculum and make questionable decisions about student assignments and grade retention. For example, schools may be reluctant to recommend their highest-achieving students to gifted programs in other schools because they would lose the advantage of these students' test scores. Schools may also focus on students who are close to meeting proficiency goals rather than on the lowest-achieving students. Moreover, NCLB may further increase attrition rates of the most qualified teachers and principals, especially in high-poverty schools, because these educators may not wish to be publicly associated with schools designated as “needing improvement.”

Testing Practices in Other Countries

Do other countries use tests to evaluate educators' performance? What role does testing play internationally in tracking students and providing access to universities? How strong is the link in other countries between testing and classroom practice? The following examples from England, Turkey, Germany, Singapore, Japan, and China illustrate how these countries manage these issues.

England

Like the United States, England holds educators accountable for students' scores on standardized tests, although major differences exist between the two countries' accountability systems. England has a national curriculum, which serves as the basis for its tests and avoids the problem so prevalent in many U.S. school districts where, in the absence of a clear curriculum, the tests become the curriculum. England's national curriculum is one of Europe's most prescriptive. Tests are administered at several points throughout the students' schooling, beginning in early elementary school, with the scores used to rank primary and secondary schools.

The initial versions of the tests were designed to be “authentic,” to give a fuller picture of a student's learning and avoid the problems inherent in paper-and-pencil standardized tests. But these tests took up so much time and left so many students unsupervised as the teacher tested students individually that paper-and-pencil tests eventually replaced them.

The test-based accountability policy remains highly controversial and raises issues similar to those currently discussed in the United States. A major question is the validity of using test scores, which are strongly influenced by students' socioeconomic status, to evaluate the quality of education. This problem is endemic in national and international test score comparisons.

England has continued its tradition of administering examinations at age 16 to determine which students will move on to A-level (advanced level) upper secondary schools. Examination results at the end of upper secondary school then determine the universities that a student can attend and the students area of specialty. Students used to be tested at age 11 to determine admission to highly selective “grammar schools,” which served as a pipeline to selective universities. In an attempt to make the education system more egalitarian, England replaced the grammar schools with comprehensive schools. However, this move may have had the opposite effect by encouraging affluent families, particularly in center cities, to move out of the state system into private schools.

Turkey

Turkey's heavily bureaucratic and centralized education system is modeled after the French system. It has been called “more French than the French system” (Simsek & Yildirim, 2004, p. 155) because French schools have undergone changes in the past 20 years that have not taken place in Turkish schools. However, Turkey's attempts to reduce the emphasis on rote learning have had limited success.

Turkey is a developing country with limited resources, high poverty rates, and relatively low access to secondary and higher education. It also has one of the highest birthrates in the world, which stretches the country's scarce education resources thin. These factors affect how national examinations play out in the country.

Examinations in Turkey are first administered at the end of basic education, although they influence what schools teach long before that. These exams determine admission into the prestigious Anatolian and science high schools, which accept approximately one-quarter of the students who take the exam. Students who wish to enter a university must take another nationwide exam at the end of high school; but because demand outweighs available spaces, acceptance rates are low (around 20 percent). Because of these conditions, Turkish students experience “some of the world's worst exam anxiety” (Simsek & Yildirim, 2004, p. 165).

Germany

Germany has a highly stratified education system that tracks students, generally beginning in grade 5, into three types of schools: the Gymnasium, which provides an academic, university-track education; the Realschule, which provides a general and vocational/technical education and occasionally permits transfer to a Gymnasium; and the Hauptschule, which provides a lower-level general and vocational education that often leads to unemployment. Teachers and parents—not an examination—determine a child's placement.

Because socioeconomic status highly correlates with academic achievement, affluent students are disproportionately represented in the Gymnasium, whereas the children of migrant workers are often tracked into the Hauptschule. The 2003 Program for International Assessment (PISA) study showed that the performance of German students correlates more highly with socioeconomic status than does the performance of students from almost any other country, suggesting that Germany's trucking system magnifies the effects of socioeconomic status (Organisation for economic Cooperation and Development, 2004).

Students attending the Gymnasium through grade 13 receive a school-leaving certificate called the Abitur, which fewer than one-quarter of German students receive. The Abitur provides access to universities after students pass a final examination.

School rankings in Singapore include a measure of how students in each school performed on a physical fitness test, combined with the percentage of overweight students in the school.

Singapore

In Singapore, educators are only held accountable for their students' test scores in the sense that secondary schools and junior colleges are ranked in publicly reported “league tables”; the 40 highest-ranked secondary schools receive cash awards. But this “accountability” system bears little resemblance to NCLB. In addition to test scores and a “value-added” measure, the rankings include a measure of how students in each school performed on a physical fitness test, combined with the percentage of overweight students in the school.

The main purpose of testing in Singapore is to determine student placement in the education system and access to elite academic programs—not to evaluate teachers. The system is heavily tracked; in a 10-year span, students are “streamed” three times. The goal is to make the system as efficient as possible in training students to contribute to the national economy.

The Singaporean system places enormous pressure on students to score well in the national examinations, which play a major role in determining students' Futures. At the same time, Singapore is attempting to reduce its emphasis on rote learning and pay greater attention to critical thinking, problem solving, and creativity. Singapore's traditional classroom practices, however, have been difficult to change because many believe that a flexible learning environment is inconsistent with the demands of an examination system that requires students to memorize large amounts of material.

Japan

Japan has a highly competitive examination system, but it doesn't hold educators accountable for students' scores on standardized tests. Indeed, Japan specifically excludes student achievement on these tests as a criterion for the self-evaluations that Japanese schools conduct.

In Japanese public schools, elementary and lower secondary students do not take high-stakes tests nor are they assigned to schools by achievement. The examination pressures begin between lower and upper secondary school, when examination results determine the upper secondary school that students will enter. The pressures that students applying to universities face have been well publicized, as have the supplementary schools (juku) that many Japanese students attend to study for the examinations. In recent years, because of a dramatically declining population, Japanese students have not had a problem gaining admission into higher education institutions. However, competition for admission to the most prestigious universities remains severe because graduates of these universities usually fill the top jobs in government and industry.

Japan, like Singapore, is attempting to increase the flexibility of the learning environment to cultivate “Japanese people with ‘rich humanity’ and ‘rich creativity’ by letting individual abilities grow” (Watanabe, 2004, p. 237). One component of this reform has been to reduce the school week from six to five days to give students more time to explore nature and participate in community-based activities. However, many families appear to be using this “free” time to increase their children's participation in juku.

Although the response to Japan's reforms has generally been positive, conservative politicians and some parents are concerned about changing an education system that they believe played a major role in the country's rapid economic growth after World War II, about encouraging individualism at the expense of Japans traditional values of cooperation and consensus, about weakening nationalism, and—perhaps most important to parents—about making any changes that might decrease their children's test scores and chances of gaining admission into prestigious universities.

China

For many centuries, the Chinese have viewed their country's examination system, which dates back to the Shui dynasty in 603 CE, as the main route out of poverty for a child from a low-income family. However, like Singapore and Japan, China is attempting to reduce its reliance on rote learning. Realizing that examinations inevitably drive classroom practice, China has revised its highly competitive university entrance exams by requiring students to integrate knowledge from a wide range of fields. For example, a recent exam question on the increased number of private cars in China required students to draw on the diverse fields of statistics, comparative analysis, supply and production, urban traffic, pollution, and social studies.

China's reforms in classroom and examination practices have occurred in an exceptionally short period of time. China's practice of building on traditional culture—or “holding new wine with the old bottle” (Cheng, 2004, p. 16)—appears to have contributed to its unusual success in implementing change. China's reforms, however, are not without controversy. The new teaching approaches have not reached the majority of schools in China's decentralized education system, with its increasing gaps in school quality between the country's rich and poor areas. Some Chinese are concerned that if examinations reduce their emphasis on memorization, children from poor families will be at an even greater disadvantage than before because they will be tested on skills that their schools have not taught them.

Chinese students face a highly competitive and stressful examination system. Yet China, like many other countries, has concluded that national exams are the best way to ensure objectivity and avoid the favoritism that might occur if the system permitted greater subjectivity in university admissions decisions.

Another Assumption—Dispelled

People generally assume that the education policies most visible today will continue unabated into the future. However, evidence shows that many policies are cyclical or gradually weaken until they have little influence, Well before NCLB, some states' test-based accountability programs began with great fanfare only to become largely irrelevant to school policies. It was simply considered educationally unwise—and politically incorrect—to acknowledge large failure rates, with the burden falling disproportionately on students in high-poverty schools.

Despite its current dominance in U.S. education policy, NCLB may well suffer the same fate. Opinions differ, of course, on whether weakening test-based accountability would be a positive or negative outcome. If test-based accountability does lose its centrality to U.S. policy, however, we would move closer to education practices in other countries, few of which assess educators' performance on the basis of their students' standardized test scores.

Note

1. This article draws on Balancing Change and Tradition in Global Education Reform. Alison Wolf of King's College, University of London contributed the chapter on England; Hasan Simsek and All Yildirim from Middle East Technical University, Ankara, contributed the chapter on Turkey; Barbara Kehm from the University of Kassel contributed the chapter on Germany; Batia P. Horsky from Nanyang Technological University in Singapore and Phyllis Ghim-Lian Chew from the National Institute of Education contributed the chapter on Singapore; Ryo Watanabe from Japan's National Institute for Educational Policy Research contributed the chapter on Japan; and Kai-ming Cheng from the University of Hong Kong contributed the chapter on China.

References

Blackburn, M. L. (1997). Comparing poverty: The United States and other industrial nations. Washington, DC: AEI Studies on Understanding Economic Inequality, American Enterprise Institute.

Cheng, K. (2004). China: Turning the bad master into a good servant. In I. C. Rotberg (Ed.), Balancing change and tradition in global education reform (pp. 3–19). Lanham, MD: ScarecrowEducation.

Ghosh, R. (2004). Canada: A multicultural policy. In I. C. Rotberg (Ed.), Balancing change and tradition in global education reform (pp. 261–282). Lanham, MD: ScarecrowEducation.

Grissmer, D. W., Kirby, S. N., Berends, M., & Williamson, S., (1994), Student achievement and the changing American family. Washington, DC; RAND Institute on Education and Training.

Organisation for Economic Co-operation and Development. (2004). Learning for tomorrow's, world: First results from PISA 2003. Paris: Author.

Rotberg, I. C. (Ed.). (2004). Balancing change and tradition in global education reform. Lanham, MD: ScarecrowEducation.

Rotberg, I. C., Bernstein, K. J., & Ritter, S. B. (2001, July). No Child Left Behind: Views about the potential impact of the Bush Administration's education proposals. Washington, DC: Graduate School of Education and Human Development, The George Washington University.

Simsek, H., & Yildirim, A. (2004). Turkey: Innovation and tradition. In I. C. Rotberg (Ed.), Balancing change and tradition in global education reform (pp. 153–185). Lanham, MD: ScarecrowEducation.

Watanabe, R. (2004). Japan: Encouraging individualism, maintaining community values. In I. C. Rotberg (Ed.), Balancing change and tradition in global education reform (pp. 229–242). Lanham, MD; ScarecrowEducation.

Iris C. Rotberg is Research Professor of Education Policy in the Department of Educational Leadership, Graduate School of Education and Human Development, The George Washington University, Washington, DC.

From Educational Leadership, November 2006. Copyright © 2006 by ASCD. Reprinted by permission. The Association for Supervision and Curriculum Development is a worldwide community of educators advocating sound policies and sharing best practices to achieve the success of each learner. To learn more, visit ASCD at www.ascd.org