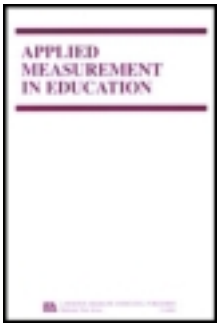

Applied Measurement in Education

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hame20

Controversies of Standardized Assessment in School Accountability Reform: A Critical Synthesis of Multidisciplinary Research Evidence

Lihshing Wang , Gulbahar H. Beckett & Lionel Brown
Published online: 07 Jun 2010.

To cite this article: Lihshing Wang , Gulbahar H. Beckett & Lionel Brown (2006) Controversies of Standardized Assessment in School Accountability Reform: A Critical Synthesis of Multidisciplinary Research Evidence, Applied Measurement in Education, 19:4, 305-328, DOI: 10.1207/s15324818ame1904_5

To link to this article: http://dx.doi.org/10.1207/s15324818ame1904_5

APPLIED MEASUREMENT IN EDUCATION, *19*(4), 305–328

Controversies of Standardized Assessment in School Accountability Reform: A Critical Synthesis of Multidisciplinary Research Evidence

Lihshing Wang
Division of Educational Studies and Leadership
University of Cincinnati

Gulbahar H. Beckett
Division of Teacher Education
University of Cincinnati

Lionel Brown
Division of Educational Studies and Leadership
University of Cincinnati

Standardized assessment in school systems has been the center of debate for decades. Although the voices of opponents of standardized tests have dominated the public forum, only a handful of scholars and practitioners have argued in defense of standardized tests. This article provides a critical synthesis of the controversial issues on state-mandated assessment of student achievement in the era of school accountability reform. By presenting an in-depth and balanced synthesis of published documents to date, this article critically reviews the 4 interlocking cornerstones of the reform's conceptual framework: assessment-driven reform, standards-based assessment, assessment-centered accountability, and high-stakes consequences. For each controversy, this article first presents the pro arguments that advocate the reform movement, followed by the con arguments that challenge the very core of the reform. The article then summarizes the debates by developing a critical synthesis of the available multidisciplinary evidence, including educational, psychometric, sociocultural, and neurocognitive research. The article concludes by proposing an action research

Correspondence should be addressed to Lihshing Wang, Division of Educational Studies and Leadership, Mail Location 0002, University of Cincinnati, Cincinnati, OH 45221. E-mail: Lihshing.Wang@uc.edu.

agenda for stakeholders of different arenas and outlining future directions for standardized assessment in schools.

Standardized assessment, defined as a large-scale, externally developed and mandated, uniformly administered and scored evaluation of student learning, has been a conspicuous part of the education reform landscape throughout American history. As early as the Committee of Ten in the late 1800s, to the National Defense Education Act in the 1950s, the Elementary and Secondary Education Act of 1965, *A Nation at Risk* in 1983, the *Goals 2000* in 1994, up to the most recent No Child Left Behind (NCLB) Act of 2001, each successive movement placed an increasing emphasis on standardized assessment as a reform catalyst and quality control mechanism (Horn, 2002; Linn, 2000).

The NCLB Act of 2001 sets unprecedented forceful provisions on using state-mandated assessments to hold schools accountable for their students' attainment of prescribed performance standards. The act ambitiously aims to close the achievement gap among *all* children regardless of their race, class, or disability status, and attaches high-stakes consequences to the assessment outcomes. Such unique features of the act garnered overwhelming political and public support, and mandatory state assessment programs have come to dominate the public school landscape (Crocker, 2003; Hamilton, Stecher, & Klein, 2002).

Although many stakeholders applaud the goals and missions of the NCLB Act of 2001 (Center on Education Policy, 2004), others have been skeptical of standardized assessment and outspoken in their criticisms (e.g., Jones, Jones, & Hargrove, 2003; Kohn, 2000; Lemann, 1999; McNeil, 2000; Sacks, 1999). This "war on standardized testing," as Phelps (2003) put it, has painted a distorted picture of a "totalitarian impulse toward absolute conformity" to the antitesting sentiment in both professional journals and popular press (p. 5). On this battlefield of overwhelming antitesting victory, only a handful of scholars have publicly defended standardized testing (e.g., Cizek, 2001a; Phelps, 2005; Thernstrom & Thernstrom, 2003). Ironically, as Cizek (2001a) pointed out, measurement professionals who know and make standardized tests have largely remained silent on this war front: "The laws of physics do not seem to apply—for every action in opposition to tests, there has been an equal and opposite silence" (p. 20).

Despite the controversies accumulated over 1 century of standardized testing (Linn, 2001), research on the consequences of standardized assessment and school reform has been described as yielding "scarce and equivocal" evidence (Mehrens, 2002, p. 151). Furthermore, most of the documented debates have presented polarized views without a balanced treatment of both sides. What seems to be missing from the existing literature, therefore, is a systematic review of the major controversies regarding standardized assessment in school accountability systems and a critical synthesis of the available research evidence from a multidisciplinary perspective. This article contributes to the literature by bridging that obvious gap and takes one step further to present policy recommendations for future action and research.

By presenting an in-depth and balanced synthesis of published documents to date, this article critically reviews the four interlocking cornerstones of the reform's conceptual framework: assessment-driven reform, standards-based assessment, assessment-centered accountability, and high-stakes consequences. For each controversy, we first present the pro arguments that advocate the reform movement, followed by the con arguments that challenge the very core of the reform. We then summarize the debates by developing a synthesized view based on the best available research evidence from educational, psychometric, sociocultural, and neurocognitive research. We conclude this article by proposing an action research agenda for stakeholders of different arenas and outlining future directions of standardized assessment in our schools.

CONTROVERSIES OF STANDARDIZED ASSESSMENT

Because of the massive amount of publications on standardized assessment and accountability reform, an exhaustive search and comprehensive treatment is clearly beyond the scope of this study. Instead, we focus our attention on the most influential works representing a wide spectrum of heated debates in the areas under investigation. Because of the far-reaching impacts of education reform, many of the publications that have exerted monumental influences on our society include both professional journals as well as popular press. As a result of the aforementioned considerations, our treatment on each controversy is necessarily sketchy and selective. However, the evidential basis of these arguments was scrutinized from a multidisciplinary perspective, and our synthesized views are always based on the best available research evidence to date.

After the selected publications have been carefully reviewed, we categorized the points of view into pro arguments and con arguments into four areas of controversy, yielding eight categories for content analysis. We then present both sides of the arguments, making sure that each side receives equal length and weight. Based on a careful review of the pro and con arguments, we then develop a critical synthesis of research evidence from educational, psychometric, sociocultural, and neurocognitive perspectives.

Before we proceed further, a conceptual clarification of the definition of standardized assessment may be in order. In this article, standardized assessment extends beyond norm-referenced tests to include standards-based tests typically used for high-stakes purposes. This means that such assessment (a) is externally imposed by the state government; (b) assesses state-prescribed content standards; (c) follows a uniform procedure in administering, scoring, and interpreting the test; and (d) the results are often used to determine rewards and sanctions for students, teachers, schools, or districts. Note that under this definition, standardized assessment is not necessarily limited to multiple-choice questions, as some antistandardized testing advocates have assumed. Authentic performance assessments can

and should be standardized to minimize abuses and inequities and to give us a common language for describing different children in different classrooms (Meisels, Dorfman, & Steele, 1995), although such work has been shown to be more challenging and less successful than multiple-choice tests (Reckase & Welch, 1999).

Assessment-Driven Reform

The various waves of education reform were launched, in part, on the basis of the claims that the U.S. leadership position in the world was being undermined by the "tragic legacy of educational mediocrity" (Sykes, 1995, p. 20). However, such claims have been challenged by some scholars and professionals who believe them to be a "manufactured crisis" predicated on false data (e.g., Berliner & Biddle, 1995). Even if such achievement lags did signal the need for reform, many wonder whether assessment should play such a central role in effecting school reform to close the achievement gaps. We now turn to these pro and con arguments on the controversy of assessment-driven reform.

Pro arguments. Ever since the widely circulated document, *A Nation at Risk,* was published in 1983 (National Commission on Excellence in Education, 1983), the entire nation has had its eyes fixated on the failures of the American school systems in domestic performance and international competition. In the following decades, more reports of failing and deteriorating U.S. performance were documented on both national and international examinations and cited by policymakers as evidence of a national crisis in need of reform.

In the national arena, the Scholastic Aptitude Test (SAT) average scores have shown a declining trend in the 1960s and 1970s, and stayed mostly stable with slight increase in the 1980s (College Entrance Examination Board, 1993). This overall downward trend was not reversed until the 1990s, and by early 2000s, the average scores have returned to their previous highs in the 60s for math, but only to the highs in the 80s for verbal (College Entrance Examination Board, 2003). Data from another national test, the American College Testing (ACT), paint a similar but even more disturbing picture. Between the 1960s and 1980s, the national average score fluctuated constantly, with decreases outnumbering increases. Despite its steady increase in the 1990s and continuing record high in the 2000s, a recent report found that only 22% of the 1.2 million high school graduates who took the ACT in 2004 were ready for college coursework (ACT, 2004). A surprisingly similar pattern can be seen in the National Assessment of Educational Progress (NAEP) long-term trend data. Generally, mathematics and science are characterized by declines in the 1970s, followed by increases during the 1980s and early 1990s, and mostly stable performance since then. Some gains are also evident in reading, but they are modest (Campbell, Hombo, & Mazzeo, 2000).

In the international arena, the Trends in International Mathematics and Science Study (TIMSS) ranked the United States at 12th out of 25 countries in 4th-grade mathematics and tied at 6th with 5 other countries in 4th-grade science, with no measurable change between 1995 and 2003. For eighth graders, the United States tied at 15th with 10 others out of 45 countries in mathematics and tied at 9th with 5 other countries in science, with modest improvement from 1995 to 2003 (Gonzales et al., 2004). Although the TIMSS data portrayed a not-so-dismal profile of the American students, another international assessment system comparing mostly industrialized countries tells us a quite different story. The Program for International Student Assessment reported U.S. 15-year-olds performance in mathematics literacy and problem solving to be lower than the average performance for most industrialized countries, and the changes from 2000 to 2003 were insignificant. The U.S. science literacy was also below the average and stayed so from 2000 to 2003. The only less discouraging finding is that the U.S. reading literacy was about the average of the other industrialized countries and stayed unchanged from 2000 to 2003 (Lemke et al., 2004).

These less-than-encouraging assessment outcomes have prompted many to view standardized testing as a potentially major force for school improvement. The rationale behind this belief is the recognition of the powerful influences standardized testing has on school curriculum and instruction. This view was called *measurement-driven instruction* (MDI) by Popham and his colleagues in 1985 (Popham, Cruse, Rankin, Sandifer, & Williams, 1985) and, despite some critics of MDI (e.g., Airasian, 1988; Cizek, 1993), continues to enjoy popularity today (Rothman 1995; Hamilton, Stecher, & Klein, 2002).

However, even defenders of MDI also acknowledge that for assessment-driven reform to have positive impacts on curriculum and instruction, tests must be carefully designed so as to be consistent with the kinds of learning desired in the classroom. Schafer (2002) asserted that standardized assessment has so far failed to deliver on the promise of effective school reform because we do not now establish a tight connection between the cognitive learning theory, the curriculum, the classroom activities, and the assessment items. Authentic assessment that bridges this link between classroom learning and standardized assessment is held by many to be the key to assessment-driven reform (e.g., Wiggins, 1990).

Con arguments. Some scholars who argue against standardized testing dispute the popular perception that America schools are failing. Berliner and Biddle (1995) called the test-score evidence a "hysterical fraud" (p. 14) and the idea of school failures a "big lie" told by politicians and a "manufactured crisis" of the media (p. 127). Bracey (2003) even went so far as to accuse the NCLB Act of 2001 assessment agenda of "educational terrorism" that brings death to childhood and destruction to schools and the nation (p. 16). Strickland and Strickland (1998) argued that the American education system is not a dismal failure by citing U.S. Depart-

ment of Education statistics that show student improvement in areas such as a decrease in dropout rates, increase in SAT scores, and an increase in high school students taking advanced courses and Advanced Placement examinations. More comprehensive data largely corroborate the aforementioned claims, with the dropout rates declining during the 1970s and 1980s and remaining unchanged during the 1990s; advanced coursework almost doubled from 1982 to 2000, and an international comparison on the Progress in International Reading Literacy Study found the average U.S. 4th-grade reading literacy score to be above the international average of the 35 countries (U.S. Department of Education, 2004). Looking at the same SAT, ACT, and NAEP data, opponents of the "crisis" claim attribute the pre-1990 score decline to non-performance factors (e.g., more high school students aspired to go to college) and stress the upward trend in the 1990s and into the 2000s (Berliner & Biddle, 1995).

Even if reform is justified, some opponents of assessment-driven reform see the logic of using assessment as the lever to effect change as fundamentally deficient (Haertel, 1999). They condemn standardized tests as unhumanistic and accuse them of undervaluing the daily, humanly sensitive interaction between teachers and their students in the complex social system of the classroom—the key to solving the nation's educational problems (Bauer-Sanders, 2000). They argue that an "overemphasis on assessment can actually undermine the pursuit of excellence" (Maehr & Midgley, 1996, p. 7). Others have articulated the technical complexity of using assessments, particularly performance-based assessments, as a basis for both educational reform and improvement (Frederiksen & Collins, 1989; Linn, Baker, & Dunbar, 1991; Moss 1992). For them, testing reform does not go to the heart of the problem: the fundamental misdesign of schools, lack of qualified teachers, and the instability of the families and communities from which students come.

A synthesized view. Despite some encouraging statistics on domestic educational performance in recent years, there seems to be credible evidence to support the widespread public perception that American school children have not been doing well internationally. Whether this lag was ever concerning enough to justify a crisis alert and a call to reform is, of course, a judgment call, but it seems clear that in the world of increasing globalization, the U.S. educational system can and should do better. However, Roeber (1995) reminded us that tests alone cannot create improvement:

> A program of systemic change begins with the content standards. Rather than simply assessing these standards and reporting the information, the systemic change process also seeks to develop appropriate alternative instructional materials, alternative instructional and learning models appropriate to the desired outcomes, and the staff development programs or courses needed by teachers to make the shifts in the desired teaching and learning. (p. 284)

Citing the limits of assessment as a tool of reform, Koretz (1995) also acknowledged that, "Assessment clearly has a great deal to offer educational reform if used prudently" (p. 156). The question, of course, is what is considered prudent use of standardized tests in the assessment-driven reform topic to which we now turn.

Standards-Based Assessment

Standards-based assessment is concerned with how well a student performance is relative to a prescribed set of content standards rather than relative to a norm group of peer students. Although this movement from norm-referenced to criterion-referenced, domain-referenced (for definitions of these terms, see Anastasi & Urbina, 1997), and more recently, *standards-based* assessment, is generally embraced by educators and measurement experts, the policy of externally imposing one set of standards for all students across a state has been rigorously debated as one of the most controversial policies in the NCLB Act of 2001. Questions often raised during the debate include the following: Should there be state or even national standards? Who has the authority to determine content and performance standards? Do externally imposed standards undermine teacher autonomy and student creativity? Should all students be held against the same set of rigorous standards? In this section, we discuss some of the debates surrounding these questions.

Pro arguments. A common theme of the current reform is the adoption of more rigorous and measurable standards and higher expectations for student performance (NCBL Act of 2001). Proponents of standards-based assessment contend that it is not only possible, but also desirable, to reach a common core of valued knowledge that teachers should teach and students should learn (Ravitch, 1995; Rothman, 1995). They argue that specified standards remove the secrecy and ambiguity surrounding traditional testing and provide a focal point on which teachers and students can collaborate in reaching common goals (Schiller, 2000). According to Phelps (2003), "In setting common standards, the public is expressing its opinion that there is nothing sacrosanct about what happens in the classroom independent of external quality control" (p. 39). Without such standards, it is impossible to compare grades across teachers and schools because they are locally normed to idiosyncratic standards (Shanker, 1995). Studies on the national trends of grade inflation in schools (e.g., Camara, Kimmel, Scheuneman, & Sawtell, 2003) further justify the need for measuring student learning on a common standardized metric.

The most controversial aspect of standards-based assessment reform seems to center on whether the same set of standards should be expected of all students, regardless of their socioeconomic status, race, or disability. The NCLB Act of 2001 is predicated on the ideology that all students should be expected and challenged to meet common standards that are more than just minimum requirements. Those standards must be anchored in the challenging content and skills that students need

to succeed, and this expectation should be "non-negotiable" (Gandal & McGiffert, 2003, p. 40). This "all-children-can-learn" philosophy is embraced by many educators who have strongly advocated for closing the achievement gaps among racial groups that have crippled the American education system (Thernstrom & Thernstrom, 2003).

Perhaps the most convincing body of evidence in support of the all-children-can-learn ideology comes from neuroplasticity research in the past decade. Contrary to the previously held belief that the human brain is hardwired and genetically fixed, this emerging line of research argues for its neurobiological malleability to the inner (e.g., mental force) and outer (e.g., enriched experience) environment (Restak, 2003; Schwartz & Begley, 2002). In contrast to the century-old nature versus nurture debate, which attempts to disentangle the two competing forces of heredity and environment, the latest nature via nurture research focuses on the interplay of the two complementary forces and suggests that human brain remains malleable well after birth, absorbing formative experiences and reacting to environmental cues (Ridley, 2003). What was previously considered irreversible, the critical period for learning is now considered regulatable through environmental enrichment and mental force throughout the adult life. As Hensch (2004) contended, "the critical period is not a simple, age-dependent maturational process but is rather a series of events itself controlled in a use-dependent manner" (p. 556). These groundbreaking research findings lend strong support to the very core value of the NCLB Act of 2001—that *no child should be left behind.*

It is interesting to note that the idea of having in place a set of common-core standards is not only consistent with the latest neurocognitive research findings but also embraced by many teachers and educators, despite its potential threat to teacher autonomy and student diversity. For example, a nationwide survey conducted by the National Board on Educational Testing and Public Policy showed that a majority of teachers actually supported their state content standards, and more than one half reported that their state-mandated test is based on a curriculum that all teachers should follow (Abrams, Pedulla, & Madaus, 2003). Another more recent but localized survey conducted by Harvard University largely corroborated the aforementioned findings, with the majority of teachers agreeing that the standards are "challenging, attainable and measurable," and their school curriculum is of high quality and aligned with state tests (Sunderman, Tracey, Kim, & Orfield, 2004, pp. 20–22). Turning to the public polls, we see even more passionate support from the parents, business leaders, and politicians in using standardized assessment for measuring students achievement (Phelps, 2005). A majority of the American public has been reported to endorse standardized testing and higher standards because it uncovers problems that need to be solved (Business Roundtable, 2001; Public Agenda, 2001).

Con arguments. The counterarguments against standards-based assessment center around issues related to intellectual freedom, student diversity, local

autonomy, and teacher empowerment. Opponents of standards-based assessment believe that by imposing standards on youngsters' minds and hearts, we are in effect depriving them of their "fundamental intellectual freedom" by gauging pluralistic knowledge against one standard set of knowledge (Sizer, 1995, p. 34). They further argue that standardized tests oversimplify knowledge and do not test higher order thinking skills (Hillocks, 2002). For example, through an analysis of over a dozen state tests designed to support standards-based instruction, Gandal and McGiffert (2003) reported that many tests are unbalanced in that they oversampled lower level standards and undersampled higher level ones.

The argument against standards-based assessment in relation to local autonomy and teacher empowerment is that state assessment standards are an externally imposed mechanism on local schools (Kohn, 2000). Unless teachers understand and accept the philosophical underpinnings out of which such externally imposed assessment systems grow, mandatory state assessment cannot work (Strickland & Strickland, 1998). Kubow and DeBard (2000) made a case for this argument by citing a survey study of teacher attitudes toward state testing that indicated, "the most pronounced concern for more than 96 percent of teachers is that proficiency testing has been imposed upon, rather than seeking input from, the school district" (p. 3).

One of the most serious attacks on the NCLB Act of 2001 is its "one-size-fits-all" standards imposed on all students (except those who meet some stringent exclusion criteria)—thus, no child should be left behind. Koretz (1995), for example, raised his objection based on the argument that because students vary markedly in their capabilities, presenting them all with the same tasks under the same conditions would entail "either dumbing instruction down to the lowest common denominator or condemning low-ability students to frequent failure" (p. 159). It is argued that the devastating effect of such uniformly high standards is that many "nonstandard kids" simply do not fit into that standard mold. "In the name of quality, the Standardistos offer a curriculum of death to children not already on the advanced placement track to elite universities" (Ohanian, 1999, p. x). Once these kids are labeled as "below-standard" and washed away in the "standardistos" stream, we are in danger of losing them forever (Cohen & Rogers, 2000; Sacks, 1999).

A synthesized view. Few would argue against the noble goal of bringing all children to the same set of high standards. The emerging evidence of neurocognitive research also seems to speak unequivocally to the adaptability of the human brain well into adulthood. Put succinctly by Schwartz and Begley (2002),

> The basic principle is this: genetic signals play a large role in the initial structuring of the brain. The ultimate shape of the brain, however, is the outcome of an ongoing active process that occurs where lived experience meets both the inner and the outer environment. (p. 117)

What this means is that environmental factors such as socioeconomic inequity and teacher quality only partially explain the achievement gaps prevailing in the American education system; it is "learned helplessness" (see Schwartz & Begley, 2002, p. 147) institutionalized by setting different standards that has left our children behind. It is interesting to note that although these neuroscientists probably never set their eyes on education reform issues, their research actually provides compelling support to the NCLB stipulation of uniform high standards.

However, even the most outspoken advocates of brain plasticity admit that, "It seems unlikely that cells could continue to enlarge and add synapses indefinitely" (Kolb, Gibb, & Robinson, 2003, p. 4) and that the genetic switches may stop functioning after reaching a certain critical period, some sooner than others, suggesting that the human mind may lose its plasticity in learning a certain task after reaching a certain age (Ptashne & Gann, 2002). It is no wonder that critical period regulation, especially in the reactivity to sensory input (e.g., musical and language learning), has been shown to be impossible without extensive intervention (Hensch, 2004).

The implication of the aforementioned body of research is that although we can exhaust our resources to provide the most nurturing possible learning environment for our students, there is a learning cap compounded by genetic as well as social–economic factors that determines how far and how fast an individual student can climb on the achievement ladder during their school years. The NCLB's stipulation that all children must reach the same set of high standards at the same time fails to acknowledge the diversity and pluralism embodied in our genes and embraced in our society.

Assessment-Centered Accountability

Assessment-centered accountability refers to using the results of some standardized assessment to hold students, teachers, schools, or all three accountable for their success or failure to reach the prespecified standards. Although there is little dispute that teachers and school administrators have the responsibility to help their students learn, there is much debate surrounding (a) how much of that responsibility should rest on the shoulders of the educators given the immutable factors such as socioeconomic status and family structure and (b) whether a standardized test is a proper instrument used as an indicator of educational effectiveness and thus an adequate lever for accountability. We now turn to the pro and con arguments on these issues.

Pro arguments. The idea that we can hold a school accountable for the results of their activities by linking testing to reform efforts began to emerge in the educational literature since at least the end of the 19th century (Resnick, 1982). After the publication of the *Coleman Report* (Coleman et al., 1966), the use of stan-

dardized achievement tests for accountability purposes became a common theme in the subsequent waves of educational reform. These accountability movements promote the idea that those involved in teaching and learning must answer for children outcomes to the legislative bodies that allocate tax revenues to education and to the government agencies that provide funding (Ravitch, 2002).

Among accountability proponents, it is believed by some that to subject schools to public audit of their performance, progress toward measurable outcomes must be monitored and reported. In an international study that looked at the effects of dropping and reintroducing standardized tests in 29 industrialized countries, it was concluded that after the standardized tests were dropped, academic standards declined, students studied less, curricula became incoherent, and selection and promotion became subjective and arbitrary (Phelps, 2000). In another study that compared the frequencies of using standardized tests to monitor educational quality across 30 countries, a positive linear correlation was found between the frequency of testing and the TIMSS performance, and the linear trend was even replaced by a concaved-up curvilinear trend after controlling for country wealth, suggesting an accelerated effect of testing on performance (Phelps, 2001, cited in Phelps, 2003, pp. 221–223). Although it is difficult to derive unequivocal causal inference from nonexperimental research data, such comparative studies provide encouraging preliminary evidence of positive impacts of assessment-centered accountability.

To serve the purpose of comparing student performance across different education systems, therefore, standardized assessment is believed to be the best alternative for providing such information because it reduces the role of judgment in human decision making that has been shown to be error prone. According to Cizek (2001a), "Decades of evidence have been amassed to support the contention that the quality of teacher-made tests pales compared with more rigorously developed, large-scale counterparts" (p. 25). If assessment results are to be used for accountability purposes, standardized tests can provide much more objective and less ambiguous evidence of student performance than classroom tests (Frary, Cross, & Weber, 1993).

Con arguments. Although holding schools accountable for student achievement is a popular notion that even antitesting advocates cannot resist, many criticize the use of state assessment as a lever for accountability. Kohn (2000), for example, reminded us that "endorsing the idea of accountability is quite different from holding students and teachers accountable specifically for raising test scores … . [G]enuine accountability and authentic standards are undermined by a myopic emphasis on testing" (p. 46). For those accountability proponents who object to standardized assessment, important learning outcomes that do not render themselves easily to an external mechanism for ensuring performance must also be valued and documented. They argue that only self-generated professional responsibility for continuous school and student improvement can sustain fundamental

improvement (Smith, 1995). According to these authors, classroom teachers should constantly look for evidence from a variety of sources, piece it together to make sense of what is happening, and use this information to support their decisions for the sake of student learning rather than using the standardized tests as an incentive or yardstick for learning (Strickland & Strickland, 1998).

Another line of antitesting argument challenges the claim that test score variation is caused by the quality of instruction, observing that this is a "confounded causation" (Popham, 1999, p. 12). Standardized tests, according to this view, measure little more than socioeconomic status, and teachers and school administrators should not be held responsible for the composition of their students (Popham, 2000). Even to the extent that the scores do reflect school experience, that experience is hardly limited to the current year. Therefore, "it seems difficult to justify holding a fourth-grade teacher accountable for her students' test scores when those scores reflect all that has happened to the children before they even arrived at her class" (Kohn, 2000, p. 20).

Another rationale for the arguments against standardized tests is that a single instrument cannot be expected to serve a multitude of purposes such as fostering good teaching and learning, making high-stakes decisions about individuals, holding schools and districts accountable, and monitoring national progress toward educational goals (Madaus, 1995). Shepard (1989), among others, advocated strongly against mixing accountability purposes with instructional assessment because a test used for accountability purposes is likely to be distorted by the incentives to teach to the test. Popham (1999) also pointed out that, "although educators need to produce valid evidence regarding their effectiveness, standardized achievement tests are the wrong tools for the task" (p. 15).

Lastly, Koretz (1995) argued that standardized tests fail to recognize the problem of "finding reasonable methods of differentiating instruction for different kinds of kids without condemning relatively low-achieving students to boring and unproductive schooling" (p. 163). Under such circumstances, instead of achieving equal excellence, the rich get richer, and the poor simply get ignored.

A synthesized view. Despite our reservations about the lofty goal of leaving no child behind by year 2014, we believe that educators have the sacred duty to break all the hereditary and environmental barriers to reach for every child and help them learn. In a political context of tight economy and global competition, educational accountability holds great appeal to taxpayers and funding agencies. However, a well-established accountability system must make sure that the process to accountability is both legal and substantial (Parkes & Stevens, 2003). Without adequate funding for test development, personnel training, and opportunity to learn, the accountability mandate is likely to be and has been challenged on legal grounds (National Education Association, 2005). To accomplish the accountability mission, we also need an evaluation mechanism that can capture the unique

contribution of individual teachers in the learning process of a child while acknowledging the preexisting differences in individual students. Without such an evaluation mechanism in place, the accountability mandate is likely to be challenged on psychometric grounds.

When embracing the accountability ideology, let us also not lose sight of the unfortunate reality that schooling only plays a small, albeit statistically significant, part of the nurturing process among all competing forces of familial, peer, and other sociocultural influences (Nye, Konstantopoulos, & Hedges, 2004). Despite our best intention, the current NCLB goal of bringing *all* children to the *proficient* level or above by year 2014 has been empirically projected to be an unattainable goal (Kifer, 2001; Linn, Baker, & Betebenner, 2002). As a result, this stipulation has placed an unprecedented burden on school constituencies. In fact, in a pre-NCLB context, Linn (1994) already foresaw the consequences of imposing uniform high expectations on all students: "The dual goals of setting performance standards for student certification that are both 'world class' and apply to 'all' students are laudable, but it cannot simply be assumed that only positive effects will result from this press" (p. 8). Holding students, teachers, and administrators accountable for reaching an unattainable goal will inevitably lead to unintended negative consequences—a topic we turn to in the next section.

High-Stakes Consequences

What makes the NCLB Act of 2001 an unprecedented—and most controversial—reform endeavor is its forceful provisions on attaching high-stakes consequences to the assessment. The stakes can be either sanctions or awards; they can be leveled on students, teachers, or schools; and their severity can be low, moderate, or high (Heubert & Hauser, 1999). According to a survey study cited by Goertz and Duffy (2003), all 50 states currently produce or require local school districts to produce and disseminate district or school report cards. In another national survey, 19 states were reported to attach high-stakes sanctions and rewards to their assessment results, including accreditation and financial incentives (Abrams et al., 2003). Many of the negative unintended consequences of standardized assessment can be directly attributed to the stakes attached to assessment. If assessment-based accountability is to be endorsed, we must find answers to the following questions: Should high-stakes consequences be attached to the results of accountability assessment? What kinds of stakes are most likely to serve the accountability function? What consequential evidence of the impact of high-stakes testing has been documented to date? These questions are explored in this section of the article.

Pro arguments. Proponents of high-stakes testing contend that assessment-based accountability is possible only when high stakes are associated with the assessment results. "Because many tests now have stakes associated with them,

it has become de rigeur for educators to inform themselves about their content, construction, and consequences" (Cizek, 2001a, p. 24). This has had a "trickle-down effect" making teachers "more reflective, deliberate, and critical in terms of their own classroom instruction and assessment" (Cizek, 2001a, p. 24). Using a growth-modeling approach, Stone and Lane (2003) empirically demonstrated that longitudinal changes in a high-stakes state assessment were found to be positively correlated with instructionally sound school and classroom practices.

One of the strongest advocates of high-stakes accountability systems is Shanker (1995), who argued that the American educational system has failed because "the United States has an education system in which very little counts" (p. 147). According to him, teachers who are used to that kind of freedom are not willing to follow a prescribed curriculum unless stakes are attached. He further stated:

> Stakes for kids go right to the heart of what motivates them to work and learn … . I, too, would prefer a world in which youngsters would open up a play by Shakespeare because they were eager to get into it instead of being forced to read the play because it is on the final examination. However, the last great experiment with a system that dismissed incentives—and relied instead on the goodness of people's instincts and motives—went down in flames recently, and the survivors are now trying to build a system based on incentives. (p. 149)

Despite its overwhelming public support, high-stakes accountability systems have met with strong resistance and vocal opposition from educators, on whom most of the stakes are leveled. Amidst the "testing backlash" sentiment, Cizek (2001a) reminded us that high-stakes testing actually has unintended positive consequences, such as improvement of professional development, in particular, knowledge about classroom assessment and sound testing practices. In a 50-state analysis, Carnoy and Leob (2002) found a strong positive relation between the level of stakes attached to standardized assessment and the NAEP improvement. Similarly, in an international study comparing countries with high-stakes exit examination systems and countries with no such systems in place, Bishop (1998) found the high-stakes countries to outperform the no-stakes countries on the TIMMS and International Assessment of Student Progress test scores, after controlling for level of economic development and national culture. Even more encouraging is the finding reported in Bishop's study that students in high-stakes systems reported more time on pleasure reading, more time on watching educational television programs, and more time talking with parents about schoolwork.

Con arguments. Many outspoken opponents of high-stakes testing have expressed concerns over attaching rewards and sanctions to the assessment results. One line of counterargument focuses on the motivation theory. Kohn (1993) cautioned that the behaviorist theory that underlies high-stakes accountability systems

represents seductive simplicity regarding how human behaviors are conditioned by rewards and punishments. Relying on a comprehensive review of decades of research, Kohn (1993) concluded that the use of extrinsic sources of motivation such as stars, stickers, trophies, and grades actually undermines students' natural curiosity and causes them to fail to enjoy learning in its own right. Kohn (2000) ridiculed high-stakes accountability by stating that "It doesn't make sense to 'put some teeth' into the standards unless you think the way to improve education is by biting people" (p. 20). He believed that punitive consequences achieve temporary compliance at a substantial cost of demoralizing the teachers and students. "Teaching and learning alike," he claimed, "come to be seen as less appealing when someone has a gun to your head" (Kohn, 2000, p. 21). Along this line of argument, the fundamental flaw of high-stakes accountability systems, therefore, lies in its overreliance on extrinsic motivation at the expense of intrinsic motivation. The higher the stakes are, the less intrinsically motivating the teaching and learning become.

Another line of argument against high-stakes testing is that it is counterproductive to turn the yardstick into a beating stick by punishing teachers or students for something beyond their control. Popham (2001) believed that the pressure that comes with a high-stakes test turns accountability into a "score-boosting game that educators have no chance to win—at least, not without harming children" (p. 17). Numerous studies reporting such negative consequences have been documented to date, citing evidence of higher dropout and holdback rates, lower motivation and higher pressure for both students and teachers, unethical test preparation, teaching to the test, and dumbing down the curriculum (e.g., Darling-Hammond, 1995; Elford, 2002; Hamilton et al., 2002; Haney, 2000; Kohn, 2000; Meisels et al., 1995; Orfield & Kornhaber, 2001; Rothman, 1995; Stecher & Barron, 2001).

On the issue of test fairness and social equity, Leonardo (2003) questioned whether the reform process is "democratic in nature" by holding schools accountable without hearing voices from disadvantaged groups or addressing larger structural issues (p. 40). For example, the graduation rate gap between different racial groups has been found to correlate with segregation in schools, independent of poverty (Sunderman, & Kim, 2004). Other studies highlight the disproportionate negative effect on students of color when districts prevent those students they believe "may fail" from taking the test to prevent lower district scores (Valenzuela, 2004).

Perhaps the most massive attack launched by high-stakes testing critics is on the lack of positive intended consequences of raising test scores. Reports of performance gains on high-stakes tests have been discredited as test-polluting practices such as improved test-taking skills and higher dropout and exclusion rates (e.g., McGill-Franzen & Allington, 1993). Support for such claims mostly came from comparing the performance trends between state assessments and other national standardized tests such as NAEP, SAT, and ACT, which found that the score increase

on the state assessment was not duplicated on the other national assessments (e.g., Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998).

A synthesized view. The simplistic view of behavioral conditioning has indeed long been superseded by complex constructivism and intrinsic motivation theories. Basing rewards and sanctions decisions on outcome assessment without due regard for moderating or intervening factors is sure to bring unintended consequences to the high-stakes accountability movement. Assessing the net effect of positive and negative consequences is difficult because many consequences are difficult to measure and they are not measured on a common metric, compounded by the fact that researchers have not documented the desirable consequences of testing as clearly as the undesirable ones (Hamilton et al., 2002).

In a reanalysis of the gains comparison between state assessment and NAEP, Rosenshine (2003) added a control group with no stakes attached to the state assessment. The average NAEP increases in the high-stakes schools were much higher than those in the no-stakes schools, with effect sizes ranging from moderate to strong. In response to the attack on state assessment data that have shown remarkable progress only because of its higher dropout and exclusion rates, Phelps (2003, pp. 129–137) reanalyzed the data and actually found its dropout rate to be well below the national average and the exclusion rate to be the same as the rest of the nation.

Based on the available research evidence to date, it appears that there is emerging evidence that the high-stakes state assessment has been a relatively potent policy in terms of bringing about real positive changes in student learning. Whether such extrinsically motivated score improvement can sustain life-long learning and whether such positive effects offset the negative consequences, however, remain to be seen.

AN ACTION RESEARCH AGENDA

In his recounting of the hundred years' war between the partisans of nature and nurture, Ridley (2003) made the case that "the genome has indeed changed everything, not by closing the argument or winning the battle for one side or the other, but by enriching the argument from both ends until they meet in the middle" (p. 3). In this article, we have tried to present a balanced treatment of the two sides of the arguments so that proponents and opponents of standardized assessment can meet in the middle. The controversies over standardized assessment in accountability reform do not seem to be near a point of resolution, but it is probably safe to predict that the pendulum will continue to swing, and there will continue to be controversies of testing (Linn, 2001).

No standardized assessment—or any assessment, for that matter—is perfect. It is in this imperfection that controversies lie. Without attempting to win the battle for one side or the other, we call for all constituencies—government officials, school administrators, community leaders, teachers, parents, students, and researchers—to come together and engage in meaningful dialogue and genuine collaboration. The following is an action research proposal for stakeholders of different arenas and an outline for future directions of standardized assessment in our schools.

Complete Assessment Systems

Developing high-quality standardized tests requires funding. Whether the benefits that might be gained justify its costs is a highly subjective judgment, depending in part on how the costs are estimated (e.g., according to Phelps [2003], pp. 52–53, the per-student cost has been estimated to range from $15–$1,792). If used prudently, standardized tests can complement teacher-made tests to provide a more comprehensive description and valid assessment of student achievement. From this perspective, funding should not be channeled exclusively to standardized assessment. A complete assessment system should include classroom-level diagnostic tests for formative evaluation that are aligned with and complementary to state-level standardized tests for summative evaluation.

Setting Developmental Standards

Although much research on standard-setting procedures has accumulated over the past 2 decades (Cizek, 2001b), how to set "challenging and attainable" performance standards for all without slipping back to the minimum competency requirements that characterized the reform in the 1970s and 1980s remains to be resolved. How to align the standards with the brain plasticity theory at different stages of cognitive development and how to celebrate diversity and pluralism while maintaining uniform performance standards across all groups undoubtedly present an unprecedented challenge to standard-setting researchers. A formidable task for state officials, therefore, is to set reasonably attainable standards that can challenge *all* students to their maximum potential *and* can be measured reliably and validly. Findings of such research will shed light on how educational policies can be made to promote learning for all.

Democratic Evaluation

To better align state assessment with content standards and school curriculum, the representation of the state assessment panel must be broadened to include classroom teachers and cognitive–developmental and social psychologists, among others. Such a participatory process (Haertel, 2002) not only improves teacher em-

powerment by reducing threats to teacher autonomy, it also helps ensure that the content standards and passing standards are challenging and yet attainable by all, if indeed such an assessment goal is supported by theory and feasible in practice. Consensus-seeking procedures that build on the strength of diversity in the standards development process have been proposed by Moss and Schutz (2001). Democratic evaluation that examines power relations in the evaluation context by including different stakeholders' perspectives might be the answer to addressing the concerns over a top-down hierarchical approach to accountability (Ryan, 2004). This implies that in the attempt to achieve meaningful alignment of content standards and classroom curriculum, not only must teacher empowerment be affirmed at the local community level (Stewart, 1995), broadened representation of the standards-setting panel must also be extended to include not only parents and community leaders, but also cognitive psychologists, behavioral geneticists, and social scientists.

Computerized Adaptive Testing

Although the content and performance standards may be one-size-fits-all, the tests that individual students take at different levels of attainment do not have to be identical. Adaptive testing methodology can challenge students of diverse ability levels to meet individualized learning goals that are tailored to their current ability level (Wainer, 2000). By shortening the test length, it also allows for multiple testing without causing excessive test fatigue or taking up valuable instructional time. Technological advancement has also made great strides in computerizing complex assessment beyond multiple-choice responses (Zenisky & Sireci, 2002). Proposals for such adaptive systems linked to the state standards have been advanced (e.g., McNabb, Cradler, Freeman, & Cradler, 2002; Northwest Evaluation Association, 2003), but the federal government is hesitant to respond to such innovative propositions. With federal funding to support such research and development, computerized adaptive methodology should hold promising potential for large-scale, K–12 assessment.

Modified Value-Added Assessment

Perhaps one of the most exciting lines of research that has recently gained considerable momentum in accountability research is gauging education effects with the value-added methodology (VAM; Barton, 2004). It is now gaining popularity in many states as a more equitable approach to assessing schools' contributions to student achievement because it measures the residual gain (or loss) between a student's achievement score and his or her projected score (Sanders & Horn, 1998). The system isolates the teacher and school effects by applying a hierarchical linear model, a mixed model, or both. VAM is considered a promising tool for accountability assessment not only because it is directly compatible with the "Adequate

Yearly Progress" policy of the NCLB Act of 2001, but it also has the potential benefit of controlling for demographic variables (e.g., socioeconomic status) and contextual variables (e.g., school funding) that are beyond schools and teachers' control (Hibpshman, 2004). Current VAM approaches do not include contextual variables other than preexisting achievement. With modification, VAM can shift the focus from outcome assessment to process assessment by incorporating the input and contextual variables in the evaluation of school performance. In particular, what contextual moderators should be allowed to enter the value-added model requires considerable research as well as political debate. However, how to use the VAM scores for accountability purposes without defeating the very mission of the NCLB to leave no child behind remains to be explored.

Standardized assessment may be the focal point of the current accountability reform, but let us not forget that assessment alone cannot be expected to do the magic, much like taking the temperature of a sick child is not going to heal the pain. It is only through concerted efforts from all constituencies of the education system that this current wave of standards-based accountability reform can reap the harvest that so many reformers before us have planted.

ACKNOWLEDGMENTS

We thank Dr. Mary Brydon-Miller of the University of Cincinnati for providing both substantive and editorial advice on earlier drafts of this article.

REFERENCES

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice, 42*(1), 18–29.

Airasian, P. W. (1988). Measurement-driven instruction: A closer look. *Educational Measurement: Issues and Practice, 7*(4), 6–11.

American College Testing. (2004). *Crisis at the core: Preparing all students for college and work.* Iowa City, IA: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Barton, P. E. (2004). *Unfinished business: More measured approaches to standards-based reform.* Princeton, NJ: Educational Testing Service.

Bauer-Sanders, K. (2000). Voices from a native American classroom in Nebraska. In A. A. Glatthorn & J. Fontana (Eds.), *Coping with standards, tests, and accountability: Voices from the classroom* (pp. 37–50). Annapolis Junction, MD: National Education Association Teaching and Learning Division.

Berliner, D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America public schools.* Reading, MA: Addison-Wesley.

Bishop, J. H. (1998). The effect of curriculum-based external exit exam systems on student achievement. *Journal of Economic Education, 29,* 171–182.

Bracey, G. W. (2003). *On the death of childhood and the destruction of public schools.* Portsmouth, NH: Heinemann.

Business Roundtable. (2001). *Assessing and addressing the "testing backlash."* Washington, DC: Author.

Camara, W., Kimmel, E., Scheuneman, J., & Sawtell, E. A. (2003). *Whose grades are inflated?* (Research Rep. No. 2003–4). New York: College Entrance Examination Board.

Campbell, J. C., Hombo, C. M., & Mazzeo, J. (2000). *NAEP 1999 trends in academic progress: Three decades of student performance.* Washington, DC: National Center for Education Statistics.

Carnoy, M., & Leob, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24,* 305–331.

Center on Education Policy. (2004). *From the capital to the classroom: Year 2 of the No Child Left Behind Act.* Washington, DC: Author. Retrieved February 1, 2005, from http://www.ctredpol.org/pubs/nclby2/cep_nclb_y2_full.pdf

Cizek, G. J. (1993). Rethinking psychometricians' beliefs about learning. *Educational Researcher, 22*(4), 4–9.

Cizek, G. J. (2001a). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19–27.

Cizek, G. J. (2001b). *Setting performance standards.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J., & Rogers, J. (Eds.). (2000). *Will standards save public education?* Boston: Beacon.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., et al. (1966). *Equality of educational opportunity.* Washington DC: U.S. Government Printing Office.

College Entrance Examination Board. (1993). *News from the College Board.* New York: Author.

College Entrance Examination Board. (2003). *SAT verbal and math scores up significantly as a record-number students take the test.* New York: Author. Retrieved January 25, 2005, from http://www.collegeboard.com/press/article/0,,26858,00.html

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice, 22*(3), 5–11.

Darling-Hammond, L. (1995). Equity issues in performance-based assessment. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 89–114). Boston: Kluwer.

Elford, G. W. (2002). *Beyond standardized testing: Better information for school accountability and management.* Lanham, MD: Scarecrow.

Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary school teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice, 12*(3), 23–30.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Gandal, M., & McGiffert, L. (2003). The power of testing. *Educational Leadership, 60*(5), 39–42.

Goertz, M., & Duffy, D. (2003). Mapping the landscape of high-stakes testing and accountability programs. *Theory Into Practice, 41*(1), 4–11.

Gonzales, P., Guzman, J. C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., et al. (2004). *Highlights from the Trends in International Mathematics and Science Study (TIMSS) 2003.* Washington, DC: National Center for Education Statistics.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice, 21*(1), 16–22.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (Eds.). (2002). *Making sense of test-based accountability in education.* Santa Monica, CA: RAND.

Haney, W. M. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved January 30, 2005, from http://epaa.asu.edu/epaa/v8n41/

Hensch, T. K. (2004). Critical period regulation. *Annual Review of Neuroscience, 27,* 549–579.

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion and graduation.* Washington, DC: National Academies Press.

Hibpshman, T. (2004). *A review of value-added models.* Frankfort: Kentucky Education Professional Standards Board.

Hillocks, G., Jr. (2002). *The testing trap: How state writing assessments control learning.* New York: Teachers College, Columbia University.

Horn, R. A., Jr. (2002). *Understanding educational reform.* Santa Barbara, CA: ABC–CLIO.

Jones, M. G., Jones, B. D., & Hargrove, T. Y. (2003). *The unintended consequences of high-stakes testing.* Lanham, MD: Rowman & Littlefield.

Kifer, E. (2001). *Large-scale assessment: Dimensions, dilemmas, and policy.* Thousand Oaks, CA: Corwin.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* (IP–202). Santa Monica, CA: RAND.

Kohn, A. (1993). *Punished by rewards: The trouble with gold stars, incentive plans, A's, raises, and other bribes.* Boston, MA: Houghton Mifflin.

Kohn, A. (2000). *The case against standardized testing.* Portsmouth, NH: Heinemann.

Kolb, B., Gibb, R., & Robinson, T. E. (2003). Brain plasticity and behavior. *Current Directions in Psychological Science, 12*(1), 1–5.

Koretz, D. M. (1995). Sometimes a cigar is only a cigar, and often a test is only a test. In D. Ravitch (Ed.), *Debating the future of American education: Do we need national standards and assessment?* (pp. 154–166). Washington, DC: Brookings Institute.

Koretz, D. M., & Barron, S. I. (1998). The validity of gains on the Kentucky Instructional Results Information System (KIRIS) (MR–1014–EDU). Santa Monica, CA: RAND.

Kubow, P. K., & DeBard, R. (2000). Teacher perceptions of proficiency testing: A winning Ohio suburban school district expresses itself. *American Secondary Education, 29*(2), 16–25

Lemann, N. (1999). *The big test: The secret history of the American meritocracy.* New York: Farrar, Straus & Giroux.

Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., et al. (2004). *Outcomes of learning in mathematics literacy and problem solving: PISA 2003 results from the U.S. perspective.* Washington, DC: National Center for Education Statistics.

Leonardo, Z. (2003). The agony of school reform: Race, class, and the elusive search for social justice. *Educational Researcher, 32*(3), 37–43.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4–14.

Linn, R. L. (2000). Assessment and accountability. *Educational Researcher, 29*(2), 4–14.

Linn, R. L. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational Assessment, 7,* 29–38.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability system: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3–16.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Madaus, G. (1995). Technological and historic consideration of equity issues associated with proposals to change our nation testing policy. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 23–68). Boston: Kluwer.

Maehr, M. L., & Midgley, C. (1996). *Transforming school cultures.* Bolder, CO: Westview.

McGill-Franzen, A., & Allington, R. L. (1993). Flunk 'em or get them classified: The contamination of primary grade accountability data. *Educational Researcher, 22*(1), 19–22.

McNabb, M., Cradler, J., Freeman, M., & Cradler, R. (2002). On the horizon: Electronic student performance assessments for higher-order thinking [Electronic version]. *Learning and Leading with Technology, 30*(3), 50–59.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing.* New York: Routledge.

Mehrens, W. A. (2002). Consequences of assessment: What is the evidence? In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 149–177). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Meisels, S. J., Dorfman, A., & Steele, D. (1995). Equity and excellence in group-administered and performance-based assessment. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 243–261). Boston: Kluwer.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62,* 229–258.

Moss, P. A., & Schutz, A. (2001). Educational standards, assessment, and the search for consensus. *American Educational Research Journal, 38*(1), 37–70.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperatives for education reform.* Washington, DC: U.S. Department of Education.

National Education Association. (2005). *Pontiac vs. Spellings lawsuit.* Retrieved July 18, 2005, from http://www.nea.org/lawsuit/images/nclbcomplaint.pdf

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

Northwest Evaluation Association. (2003). Computerized testing and NCLB. *The Assessment Standard, 2*(1), 1. Retrieved July 16, 2005, from http://www.nwea.org/assets/newsletter/newsletter_sp03.pdf

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26,* 237–257.

Ohanian, S. (1999). *One size fits few: The folly of educational standards.* Portsmouth, NH: Heinemann.

Orfield, G., & Kornhaber, M. L. (Eds.). (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education.* Washington, DC: The Century Foundation Press.

Parkes, J., & Stevens, J. J. (2003). Legal issues in school accountability systems. *Applied Measurement in Education, 16,* 141–158.

Phelps, R. P. (2000). Trends in large-scale, external testing outside the United States. *Educational Measurement: Issues and Practice, 19*(1), 11–21.

Phelps, R. P. (2001). *That "backlash" that testing opponents so desperately crave.* Retrieved August 8, 2005, from EducationNews.org

Phelps, R. P. (2003). *Kill the messenger: The war on standardized testing.* New Brunswick, NJ: Transaction.

Phelps, R. P. (Ed.). (2005). *Defending standardized testing.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Popham, W. J. (1999). Why standardized test scores don't measure educational quality. *Educational Leadership, 56*(6), 8–15.

Popham, W. J. (2000). *Testing! Testing! What every parent should know about school tests.* Boston: Allyn & Bacon.

Popham, W. J. (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, R. L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66,* 628–634.

Ptashne, M., & Gann, A. (2002). *Genes and signals.* New York: Cold Spring Harbor.

Public Agenda. (2001). *Reality check 2001.* New York: Author. Retrieved January 31, 2005, from http://www.publicagenda.org/specials/rc2001/reality.htm

Ravitch, D. (Ed.). (1995). *Debating the future of American education: Do we need national standards and assessments?* Washington, DC: Brookings Institute.

Ravitch, D. (2002). Testing and accountability, historically considered. In W. M. Evers & H. J. Walberg (Eds.), *School accountability: An assessment by the Koret task force on K–12 education* (pp. 9–22). Palo Alto, CA: Hoover Institution, Stanford University.

Reckase, M. D., & Welch, C. (1999). Advances in portfolio assessment with applications to urban school populations. In A. L. Nettles & M. T. Nettles (Eds.), *Measuring up: Challenges minorities face in educational assessment* (pp. 71–96). Boston: Kluwer.

Resnick, D. P. (1982). History of educational testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies. Part II: Documentation section* (pp. 173–194). Washington, DC: National Academies Press.

Restak, R. (2003). *The new brain: How the modern age is rewiring your mind.* New York: Rodale.

Ridley, M. (2003). *Nature via nurture: Genes, experience, and what makes us human.* New York: HarperCollins.

Roeber, E. D. (1995). Using new forms of assessment to assist in achieving student equity: Experiences of the CCSSO State Collaborative on Assessment and Student Standards. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 265–288). Boston: Kluwer.

Rosenshine, B. (2003). *High stakes testing: Another analysis.* Retrieved January 30, 2005, from http://faculty.ed.uiuc.edu/rosenshi/

Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform.* San Francisco: Jossey-Bass.

Ryan, K. E. (2004). Serving public interests in educational accountability: Alternative approaches to democratic evaluation. *American Journal of Evaluation, 25,* 443–460

Sacks, P. (1999). *Standardized minds: The high price of America's testing culture and what we can do to change it.* Cambridge, MA: Perseus.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12,* 247–256.

Schafer, W. D. (2002). How can assessment contribute to an educational utopia? In R. W. Lissitz & W. D. Schafer (Eds.), *Assessments in educational reform: Both means and ends* (pp. 80–91). Boston: Allyn & Bacon.

Schiller, L. (2000). Politics, pedagogy, and professional development in Michigan. In A. A. Glatthorn & J. Fontana (Eds.), *Coping with standards, tests, and accountability: Voices from the classroom* (pp. 95–107). Washington, DC: National Education Association.

Schwartz, J. M., & Begley, S. (2002). *The mind and the brain: Neuroplasticity and the power of mental force.* New York: Regan.

Shanker, A. (1995). The case for high stakes and real consequences. In D. Ravitch (Ed.), *Debating the future of American education: Do we need national standards and assessment?* (pp. 145–153). Washington, DC: Brookings Institute.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, 46*(7), 4–9.

Sizer, T. R. (1995). Will national standards and assessments make a difference? In D. Ravitch (Ed.), *Debating the future of American education: Do we need national standards and assessment?* (pp. 33–39). Washington, DC: Brookings Institute.

Smith, M. S. (1995). Education reform in America public schools: The Clinton agenda. In D. Ravitch (Ed.), *Debating the future of American education: Do we need national standards and assessment?* (pp. 9–32). Washington, DC: Brookings Institute.

Stecher, B. M., & Barron, S. (2001). Unintended consequences of test-based accountability when testing in "milepost" grades. *Educational Assessment, 7,* 259–281.

Stewart, D. M. (1995). Holding onto norms in a sea of criteria. In D. Ravitch (Ed.), *Debating the future of American education: Do we need national standards and assessment?* (pp. 83–93). Washington, DC: Brookings Institute.

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16,* 1–26.

Strickland, K., & Strickland, J. (1998). *Reflections on assessment: Its purposes, methods, and effects on learning.* Portsmouth, NH: Boynton/Cook.

Sunderman, G. L., & Kim, J. (2004). *Inspiring vision, disappointing results: Four studies on implementing the No Child Left Behind Act.* Cambridge, MA: Harvard University Press, The Civil Rights Project.

Sunderman, G. L., Tracey, C. A., Kim, J., & Orfield, G. (2004). *Listening to teachers: Classroom realities and No Child Left Behind.* Cambridge, MA: Harvard University Press, The Civil Rights Project.

Sykes, C. J. (1995). *Dumbing down our kids: Why American children feel good about themselves but can't read, write, or add.* New York: St. Martin Griffin.

Thernstrom, A., & Thernstrom, S. (2003). *No excuse: Closing the racial gap in learning.* New York: Simon & Schuster.

U.S. Department of Education. (2004). *The condition of education 2004.* Washington, DC: Author.

Valenzuela, A. (2004, April). *The struggle for valid assessment in Texas.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Wainer, H. (Ed.).(2000). *Computer adaptive testing: A prime*r (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation, 2*(2). Retrieved January 26, 2005, from http://PAREonline.net/getvn.asp?v=2&n=2

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15,* 337–362.