Assessment for the Masses: A Historical Critique of High-Stakes Testing in Reading

Andrew P. Huddleston
Abilene Christian University

Elizabeth C. Rockwell
Abilene Christian University

Abstract
This historical critique of high-stakes testing in reading focuses on selected events from three historical movements: 1) the history of standardized testing, 2) the history of standardized reading tests, and 3) the history of high-stakes testing. These three interrelated histories have produced the high-stakes, standardized reading tests used in U.S. public schools today. Describing each historical movement from a critical perspective, the authors argue that concerns for objectivity, efficiency, and accountability, rather than assessment, have historically been the driving forces behind testing in U.S. schools. The authors conclude by encouraging educators to move beyond objectivity, efficiency, and accountability and instead advocate for the development of alternative indicators that focus on students' needs.

Raúl (pseudonym) was a student in my (first author's) reading and language arts class in Lubbock, Texas. As a fifth grader he knew that he would have to pass, what at that time, was the Texas Assessment of Knowledge and Skills (TAKS) to be promoted to sixth grade. Texas had passed the Student Success Initiative, initially requiring that students in grades 3, 5, and 8 pass the TAKS in reading and math for promotion. Reading was difficult for Raúl, and he had been receiving extra tutoring throughout the year. Nonetheless, his greatest challenge to passing TAKS was his asthma. Raúl had anxiety-induced asthma, and the stress involved in the first administration of the Reading TAKS had triggered his wheezing, thus hindering his concentration. Luckily for Raúl, he had two more opportunities to take the Reading TAKS before his retention was finalized. As Raúl and I prepared for the second exam, we worked with the nurse and his mother to develop a medical plan for the day of the test, certainly not a strategy I had learned about in education courses in college, but necessary for him nonetheless. Prior to the test I was to send Raúl to the nurse for a check-up. She would listen to his lungs, check his heart rate, and give him his medication. When he completed the test I would send him back for a post check-up.

All went as planned until Raúl began his test. As I monitored the students I noticed that Raúl had a white-knuckle grip on a rosary in his left hand. The TAKS administration manual clearly stated that students were to have nothing on their desks except for their test materials and pencil. However, as his teacher, knowing about his struggles with test-anxiety, I knew much better than to ask him to put it away.

Raúl managed to pass the test that day. Nevertheless, through that experience, and many others like it, we (first and second authors) have found ourselves questioning the necessity and value of applying such high stakes to testing. As teachers, we believe that standardized testing, used with additional assessments, can provide useful information, but we worry about the pressure the high stakes places on students and their families. Wanting to know more about these tests that are used to make such important decisions concerning our students, we pursued the following questions:

1. Where and when did high-stakes achievement tests originate?
2. How did they come to yield such power in public school classrooms?

Literature Review

What follows is not intended to be a comprehensive history of high-stakes testing. Rather, we focus on selected events from three historical movements: the history of standardized testing, the history of standardized reading tests, and the history of high-stakes testing. We describe how these three interrelated histories have produced the high-stakes, standardized reading tests used in U.S. public schools today.

Our purpose is intentionally critical, and thus, as we describe each historical movement, we argue that concerns for objectivity, efficiency, and accountability, rather than assessment, have historically been the driving forces behind testing in U.S. schools. We conclude by encouraging educators and policy makers to move beyond objectivity, efficiency, and accountability and instead advocate for the development of alternative indicators that focus on students' needs.

The History of Standardized Testing

It is believed that the roots of standardized testing originated in Chinese civil service positions as early as 2200 B.C. Candidates were required to demonstrate proficiency in music, archery, horsemanship, writing, arithmetic, and rites and ceremonies of public and social life. In archery, candidates had to shoot three arrows at a human–shaped object from horseback. If all three arrows hit the target, a perfect score was given; if two hit the target it was graded "good," and if only one hit, it was a "pass" (Miyazaki, 1976, p. 103).

Standardized IQ and Achievement Tests. In the U.S., however, standardized testing did not begin until the mid-1800s. Prior to the twentieth century, the primary form of assessment in U.S. schools was oral recitation. Instruction was highly individualized, and teachers took turns listening to students read aloud and recite information from memory to determine their progress (Giordano, 2005). Oral recitation remained the primary tool for assessment until 1845 when Horace Mann pushed for standardized written essay exams (Rothman, 1995). Supporters of written tests believed they had developed a tool that could assess the swelling enrollments produced by compulsory education more objectively and efficiently than oral recitation.

In the early twentieth century this trend continued with the development of standardized IQ and achievement tests. In 1905, French psychologists Alfred Binet and Theodore Simon created the first IQ test, a thirty-item test to identify students unable to succeed in school in France (Wolf, 1973). Then in 1914, psychologists Henry Goddard, Edmund Huey, and Lewis Terman published an American version of the Binet test, the Stanford-Binet, to use with American students (Resnick, 1982). Around the same time that Binet was developing his

intelligence test in France, psychologist Edward L. Thorndike was working in the United States on achievement testing. Between 1908 and 1916, Thorndike and his students at Columbia University developed standardized achievement tests in arithmetic, handwriting, spelling, drawing, reading, and language (Wigdor & Gardner, 1982). Thorndike's handwriting test, considered to be among the first norm-referenced tests, had a great impact on the development of future standardized achievement tests (Pearson & Stallman, 1994).

New Technologies and the Proliferation of Testing. Although intelligence and achievement tests were rapidly developing, the written answers and individual administrations required were tedious to administer and to graders. When the United States entered World War I, the military needed what they believed would be an objective and efficient method for assessing the intelligence of almost two million recruits to sort them by their abilities. In 1917, Arthur Otis, a graduate student at Stanford working with Lewis Terman, met the military's need by designing a pencil-and-paper version of the Stanford-Binet using a new technology he borrowed from Frederick J. Kelly's (1916) *The Kansas Silent Reading Test*: multiple-choice questions (Samelson, 1987). The hallmark of creating an intelligence test with multiple-choice questions was that it allowed an entire room full of participants to take an intelligence test at the same time, rather than individually. The army then used the IQ scores to determine where to place its recruits.

Despite the novelty and innovation of this new type of test, the creation of a written, multiple-choice test was the first of many steps away from the individualized, face-to-face assessments of the 1800s and toward the objective, impersonal testing we see in today's schools. Following World War I, standardized testing flourished both in schools and industries. Between 200-300 psychologists who had worked for the army's testing program found jobs in public schools and universities after the war and began applying their skills (Chapman, 1988). People were amazed by what they believed were objective and reliable instruments that could quickly assess individuals in a bias-free manner.

Public schools in particular were attracted by the efficiency and reliability these tests appeared to offer. In the early 1900s, compulsory education laws drastically increased enrollment numbers. Intelligence tests were used not just for identifying extremely high- and low-performing children; they were also used to arrange students into ability groups. In high schools, intelligence tests were used for tracking students into college- or vocational-bound courses. By helping schools sort students quickly and efficiently, educating the masses became a more manageable task (Wigdor & Gardner, 1982). Additionally, numerous studies (e.g., Starch & Elliott, 1912, 1913) began to question the "reliability" of teachers' grading practices (Starch & Elliott, 1912, p. 442). As the number of studies critiquing the subjectivity of teachers' grades increased, so did the demand for the use of standardized achievement tests. Such tests, it was argued, would provide an objective assessment of student learning and more accurately determine important decisions regarding promotion and retention (Giordano, 2005).

Although the multiple-choice question enabled standardized assessments to be given to large numbers of people at a time, the tests themselves still had to be graded by hand using stencils. This was a time-consuming process and vulnerable to human error. However, in 1931 a new technology appeared that would again revolutionize the testing industry. Reynold B. Johnson, a high-school science teacher in Ironwood, Michigan, created the Markograph, a test scoring machine (Lemann, 1999). The Markograph was able to electronically sense if pencil

marks were used to select the correct answers on tests. Johnson sold the Markograph to IBM who then introduced several test scoring machines throughout the 1930s-1940s, enabling tests to be graded much more rapidly and cutting the price of scoring to one tenth of the previous cost (Giordano, 2005; Resnick, 1982).

The Development of Standardized Reading Tests. Reading was among the last subject areas to be assessed with a standardized test. In fact, the first standardized reading test was not developed until 1915 (Smith, 1934/2002). Although the demand for a standardized assessment of reading was great, researchers struggled in designing a test they felt would adequately measure reading achievement. There were two significant factors that made reading more difficult to assess than other subjects. First, reading itself was recognized as a highly complex process, and researchers disagreed not only on what reading was but also on what aspects of reading should be assessed. Another difficulty in standardizing an assessment for reading was the fact that historically reading instruction had always involved reading aloud.

The Transition from Oral to Silent Reading. Researchers quickly realized that oral reading proved to be uneconomical to assess using standardized instruments. By its very nature, oral reading demanded individual assessments. However, in the early 1900s, a shift began to occur from teaching oral reading to teaching silent reading in schools (Smith, 1934/2002). This transition was related to various factors.

For one, an increasing number of reading researchers argued that silent reading was superior to oral reading in terms of speed and comprehension. Also, when printed books were rarer, oral reading was necessary for sharing books with large groups of people. As printed materials increased, silent reading became more necessary for quickly consuming vast amounts of text. However, another factor for the shift to silent reading was the development of standardized tests that specifically assessed silent reading. Researchers found that silent reading was much more suited to their testing purposes and began designing silent reading tests even though it conflicted with the oral reading model of the time. As soon as silent reading became what was tested, it did not take long for teachers to focus on it in classroom instruction (Smith, 1934/2002). Consequently, teaching to the standardized reading tests helped initiate the transition from oral to silent reading in schools.

Early Forms of Standardized Reading Tests. Standardized reading tests, as we recognize them today, evolved from of a variety of attempts to assess comprehension. Readence and Moore (1983) identified three types of standardized test questions that were used to assess comprehension on standardized reading tests in the early 1900s: "reproducing a passage tests," "solving written puzzles tests," and "answering questions tests" (p. 307).

Starch (1915), for example, created a standardized test that relied on reproducing a passage to assess comprehension. There were multiple passages of increasing difficulty. Students had a limited amount of time to read the passages silently and then write everything they could remember about them from memory. The teacher then assigned a comprehension score based on the number of words that correctly retold the passage. Teachers also had the option of having students reread the passage a second time orally to assess reading speed and pronunciation.

Kelly (1916) designed the Kansas Silent Reading Test. Unlike Starch (1915), Kelly believed that reproducing a passage did not adequately measure comprehension and thus assessed comprehension through the use of written puzzles. The Kansas Silent Reading Test had

multiple puzzles, each with increasing difficulty. Students were timed as they read them silently and answered them. Below are three puzzles for grades 3 through 5:

> 2. Think of the thickness of the peelings of apples and oranges. Put a line around the name of the fruit having the thinner peeling.
>
> apples oranges
>
> 9. Here are some names of things. Put a line around the name of the one which is most nearly round in every way like a ball.
>
> saucer teacup orange pear arm
>
> 15. Fred has eight marbles. Mary said to him: "If you will give me four of your marbles, I will have three times as many as you will then have." How many marbles do they both have together? (Kelly, 1916, pp. 66-67)

Arthur Otis drew on these types of questions used by Kelly in The Kansas Silent Reading Test to develop multiple-choice questions for the Stanford-Binet (Samelson, 1987).

Thorndike (1914) created a standardized reading test he called Scale Alpha using a third approach for assessing comprehension: answering questions. Scale Alpha consisted of short paragraphs with questions of increasing complexity. Most of the questions were literal-level and were limited to one single, correct answer. Students had as much time as needed to complete the test, and answers were graded on a 0-4 scale to assess the degree of correctness. Below is an example of one of the Scale Alpha paragraphs and questions:

> Long after the sun had set, Tom was still waiting for Jim and Dick to come. "If they do not come before nine o'clock," he said to himself, "I will go on to Boston alone." At half past eight they came bringing two other boys with them. Tom was very glad to see them and gave each of them one of the apples he had kept. They ate these and he ate one too. Then all went on down the road.
>
> 1. How many boys are told about?
> 2. For whom was Tom waiting?
> 3. When did they come?
> 4. Where do you think they were going?
> 5. How many apples did they eat?
> 6. What did they do after eating the apples?
> 7. How did Tom feel when he saw Jim and Dick?
> 8. Who else came besides Jim and Dick?
> 9. How long did Tom say he would wait for them?
> 10. What happened after the boys ate the apples? (Thorndike, 1914, pp. 241-242)

Readence and Moore (1983) noted that Thorndike's answering questions format became the standard model for assessing reading comprehension. Passage reproductions, such as those used by Starch (1915), required significant time to complete, and the retellings were thought to be difficult to assess objectively. Also, students' writing skills affected their scores. Although written puzzles, such as those used by Kelly (1916), were considered more objective and efficient to administer, they only provided limited information concerning reading comprehension. The third approach, answering questions, was easily adaptable to the multiple-choice format that greatly expanded in the 1930s and to the computerized scoring of the 1960s.

Testing Becomes a High-Stakes Political Tool. As we have shown, a concern for efficiency and objectivity fueled the growth of standardized testing as a means for sorting students in the early twentieth century. But how did standardized testing come to have such high-stakes? Tests are often considered "high-stakes" when their results are perceived by "students, teachers, administrators, parents, or the general public, as being used to make important decisions that immediately and directly affect them" (Madaus, 1988, p. 87). As it turns out, the origins of high-stakes testing are as old as public education itself, dating as far back as 1845 and Horace Mann's role in the transition from oral to written examinations.

Horace Mann and Written Examinations. In the mid-1800s, Horace Mann was secretary of the Massachusetts State Board of Education and was well aware of the increasing enrollments in schools. As more immigrants came to the United States, Mann believed that a common, public school system would provide them with the tools they needed to succeed. Mann argued that standardizing the curriculum and instruction among common schools would help address the challenges faced by swelling enrollments and a diverse student body (Smith, 1934/2002). However, he faced significant opposition. Many teachers and school administrators were opposed to a practical, mass education system and preferred a classical education only for the elite. According to Rothman (1995), as a clever politician, Mann chose a strategy "education policy makers would turn to again and again over the next century and a half: he created a test" (p. 33).

In 1845, Mann required that the Boston School Committee give written essay examinations in lieu of the oral exams to which students were accustomed. Many of the students scored poorly, giving Mann the "objective information" he needed to push for change (Rothman, 1995, p. 33). In addition to bolstering his support for public education, Mann believed that regular written tests could be valuable instruments in comparing the quality of teaching among schools (Caldwell & Courtis, 1925).

Written exams (essay and short answer) were soon used for judging the effectiveness of teachers and programs (Resnick, 1982). By the 1870s, exam results were being printed in newspapers and had replaced teacher recommendations for determining promotions. However, such high-stakes use of written essays did not go uncriticized. Emerson White, a leader of the National Education Association (NEA) in the late 1800s, passionately argued that written exams should not be used for comparing students and teachers, nor should they be used alone in promotion and retention decisions. Test-focused instruction, he argued, was detrimental to education:

> They [written tests] have perverted the best efforts of teachers, and narrowed and grooved their instruction; they have occasioned and made well-nigh imperative the use of mechanical and rote methods of teaching; they have occasioned cramming and the most vicious habits of study; they have caused much of the overpressure charged upon schools, some of which is real; they have tempted both teachers and pupils to dishonesty; and last but not least, they have permitted a mechanical method of school supervision . . . They [the teachers] shut their eyes to the needs of the pupil and put their strength into what will 'count' in the examination. (White, 1886, pp. 199-201)

Two years later White (1888) again expressed his disapproval of tying promotion to written exams and described his schools' plans to return to teacher recommendations based on daily work.

The Cold War, Test Score Decline, and the Minimum-Competency Movement. It was not until the 1950s-1960s that politicians once again took an interest in using testing for political purposes and not just assessment (Koretz, 2008). The Cold War escalated, and there was an increased concern nationwide in issues of national security and education. After the Soviet Union launched Sputnik in 1957, Congress passed the National Defense Education Act of 1958 (Haney, 1984) and then in 1965 the Elementary and Secondary Education Act (ESEA), both of which required an expanded use of standardized testing.

Two factors greatly influenced the increase of standardized testing during the 1960s and 1970s. The first was the public's growing concern over declining SAT (originally Scholastic Aptitude Test) scores. A special panel was assembled by the test's sponsors, the College Board and Educational Testing Service (ETS), to study the fourteen-year decline (Wirtz, 1977). From 1963 to 1977 the average score on the verbal section of the SAT dropped 49 points while the average score on the mathematics section decreased by 31 points. The panel listed several potential factors it believed contributed to the SAT decline after 1970 such as soft educational standards, the deterioration of the family, and excessive television watching (Wirtz, 1977).

The second factor that led to the growth of standardized testing was the development of the minimum-competency movement. Prior to World War II, 25% of high school students participated in a college track, 25% participated in a vocational track, and 50% were expected to drop out (Berliner & Biddle, 1995). After World War II, several curricular adjustments were made to keep a wider range of students in school. Classes such as civics, health, personal development, and recreation were added to provide more engaging classes, and academic standards were relaxed to encourage students to stay in school. Additionally, grading procedures were changed to make failing less likely. Although these changes did lower the dropout rate and likely increased the educational rigor for a majority of students, parents of elite students interpreted these changes as lowering educational standards.

Decreasing test scores, coupled with a growing belief that public schools were not making the grade, ushered in the minimum-competency movement. This movement called for tests designed to guarantee that high school graduates could read, write, and complete basic arithmetic problems. Between 1963 and 1974, 73 laws were passed in states such as Florida, Colorado, and Texas, designed to hold schools accountable and raise student achievement (Wise, 1979). In the mid-1970s, 36 states had passed legislation that used minimum-competency tests to measure basic skills (Rothman, 1995).

A Nation at Risk. In April of 1983, the Commission on Excellence in Education released its report on the status of education in the United States, *A Nation at Risk: The Imperative for Educational Reform* (National Commission on Excellence in Education, 1983). The commission had been created by Terrell H. Bell, then the Secretary of Education under President Ronald Reagan. Bell charged the commission to report on the condition of American schools and universities and address a growing concern that the American education system was inherently flawed. With an indicting tone, the commission portrayed a bleak picture of American schools, warning that "if an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war" (National Commission on Excellence in Education, 1983, p. 5).

The commission cited numerous reasons for alarm such as declining SAT scores, poor performance on international comparison tests, and low high school standardized test scores. It

identified grade inflation, social promotion, and lax standards as further evidence of educational mediocrity.

The commission urged educators to return to teaching the basics and incorporate standardized tests at various grade levels. *A Nation at Risk* had a tremendous influence on education policies and the growth of standardized testing. By 1987-1988, 45 states and the District of Columbia used statewide tests to assess student achievement (Marks, 1989).

No Child Left Behind and the Common Core Standards. After its initial passage in 1965, the Elementary and Secondary Education Act (ESEA) was amended and reauthorized numerous times, with each reauthorization of the law including tougher accountability (Duffy, Giordano, Farrell, Paneque, & Crump, 2008). On January 8, 2002, President George W. Bush re-authorized ESEA as No Child Left Behind (NCLB). One of the most significant differences between NCLB and previous versions of ESEA was that previous ESEA laws only called for accountability of achievement for students in schools receiving Title I funding. NCLB, however, mandated that all students in schools receiving federal funding be held accountable for educational achievement. Although many states were already providing standardized tests, only half had tests for grades 3 through 8 in math and reading. NCLB required schools to have these tests in place by 2005-2006 (Fritzberg, 2004).

Most recently, in 2009, a group of governors and state commissioners of education through the National Governors Association and the Council of Chief State School Officers launched the development of a national set of state standards called the Common Core. Currently, 43 states, two territories, and the District of Columbia have adopted the Common Core. Most states who have adopted the Common Core have already given one of two tests, which was to begin in 2014-2015: the Partnership for Assessment of Readiness of College and Careers (PARCC) or the SMARTER Balanced Assessment Consortium (SBAC) (Long, 2011). Although these tests will be used for high-stakes accountability decisions, they promise to differ from more traditional standardized tests in a few ways. Each will be computer-based and will contain aspects that are performance-based, asking students to complete a specific task rather than only answering multiple choice questions. Another unique component of PARCC and SMARTER Balanced will be the formative components. Both will provide assessment tools that generate data throughout the year to inform instruction, rather than simply relying on a single end-of-year test.

A Manufactured Educational Crisis. It is important to note that numerous education researchers have argued that public concern over the so-called educational crisis that has fueled the high-stakes accountability movement has largely been misguided. Berliner and Biddle (1995), for example, argued that despite the significant criticism public schools have received, the supposed crisis is largely manufactured. They examined test scores over the last thirty years on major tests such as the SAT, ACT (originally American College Testing), GRE (Graduate Record Examination), NAEP (National Assessment of Educational Progress), and IQ tests and found that scores have remained steady and even improved in some areas.

Similar evidence of steady and increasing scores was found in the Sandia Report (Sandia National Laboratories, 1993), a study commissioned by Secretary of Energy James Watkins in 1990. The Sandia Report analyzed the SAT-score decline between the late 1970s and 1990 by subgroups. What they found was that although the overall SAT scores declined during this time, the scores of each subgroup actually held steady or increased. The overall average SAT scores were decreasing because more impoverished and ethnic-minority students were taking the SAT

than ever before. Only when the scores were viewed by subgroup was it apparent that these students were scoring higher than they ever had. Although the Sandia Report conflicted much of the crisis rhetoric of the 1980s, it was buried in review until President Bush Sr. left office (Ansary, 2007; Berliner & Biddle, 1995). By the time the report was released, the crisis rhetoric had largely been accepted, and high-stakes testing had become the political mechanism for ensuring accountability.

Moving Beyond Objectivity, Efficiency, and Accountability

In light of this history, what might the future of high-stakes, standardized reading tests look like? Madaus and O'Dwyer (1999) argued that a desire for objectivity and efficiency in testing has fueled the changes in assessment technology over the last two centuries. As the number of examinees increased, the changes "from oral to written, from qualitative to quantitative, from short answer to multiple choice—were all geared toward increasing efficiency and making the assessment system more manageable, standardized, easily administered, objective, reliable, comparable, and inexpensive" (Madaus & O'Dwyer, 1999, p. 689). High-stakes, standardized testing in reading has followed a similar path. As we look ahead to what the future of testing in reading might look like, it seems probable that concerns with objectivity and efficiency will persist. Americans have an infatuation with numbers (Hechinger, 1977) and want to invest their money in assessments they believe are trustworthy and affordable.

However, as the use of high-stakes, standardized testing has increased, so have the critiques of its use. Along with a growing number of educators, we believe it is time for testing to move beyond traditional notions of objectivity and efficiency. Dissatisfied with the isolated, lower-level skills assessed on many standardized tests, we believe that performance-based assessments can serve as tools for assessing learning through multiple, holistic, real-life tasks that require students to construct responses and create products (Koretz, 2008). Both Graves (2002) and Nichols and Berliner (2007) have proposed alternatives to traditional high-stakes testing that still include a standardized test but include multiple performance-based indicators as well. The PARCC and SMARTER Balanced assessments of the Common Core, promise to make strides in this area. However, the test designers have admitted that despite the performance-based components, these tests will not assess the more sophisticated Common Core standards such as speaking, listening, and collaboration (Merrow, 2013, August 14). As the act of reading itself evolves to include additional digital literacies (Leu, Kinzer, Coiro, & Cammack, 2004), and reading research continues to be influenced by sociocultural perspectives, we as educators should push for assessments that better reflect the kinds of literacy tasks that students are expected to have in the 21st century workforce.

Politicians, however, will most likely continue to advocate for the use of traditional high-stakes, standardized tests as long as they have the political support of their constituents. Nonetheless, we believe that teachers can play an important role in bringing about change. The current testing climate has created numerous windows of opportunity for us as educators to inform families who are frustrated with high-stakes assessments about reasonable assessment alternatives. By using formative assessments such as portfolio assessments, anecdotal records, and other task- and performance-based assessments, we not only learn more about our students, but we also model to our students and their parents that learning cannot be represented by a single test score.

For teachers wanting to bring about change, organizations such as Texans Advocating for Meaningful Student Assessment (TAMSA; www.tamsatx.org/) offer updates on current state legislation as well as recommendations for voicing concerns to the Texas Education Agency, State Board of Education representatives, and state legislators. TAMSA is a statewide organization of concerned parents and community members and was instrumental in reducing the number of high school end of course exams from 15 to 5 in the 83rd legislative session.

It is for students like Raúl, the student mentioned in the vignette at the beginning, that we remain hopeful. Although Raúl managed to pass the second administration of the TAKS, others did not. We remain hopeful that as educators we will continue to emphasize the importance of using multiple, alternative indicators in assessment-based decisions and that future policymakers will heed the warnings of reading researchers as early as Gray (1917):

A great variety of methods may and should be used in a thoroughgoing test of one's ability to get meaning. Neither a reproduction test, nor answers to questions, nor Kelly's test, nor all combined, serves as a complete test of comprehension (p. 16). Much like those early test-makers in 1915, it is time again to ask: What is reading? What aspects of reading should be assessed? Are we relying on multiple indicators? What have we recently learned about reading, and how are our assessments reflecting those new understandings?

References

Ansary, T. (2007). Education at risk: Fallout from a flawed report. *Edutopia, 3*. Retrieved from http://www.edutopia.org/landmark-education-report-nation-risk

Berliner, D. & Biddle, B.(1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. New York City, NY: Basic Books.

Caldwell, O. & Courtis, S. (1925). *Then and now in education 1845-1923: A message of encouragement from the past to the present*. Yonkers-on-Hudson, NY: World Book.

Chapman, P. (1988). *Schools as sorters: Lewis M. Terman, applied psychology, and the intelligence testing movement, 1890-1930*. New YorkCity, NY: New York University Press.

Duffy, M., Giordano, V., Farrell, J., Paneque, O., & Crump, G. (2008). No Child Left Behind: Values and research issues in high-stakes assessments. *Counseling and Values, 53*(1), 53-66.

Fritzberg, G.(2004). No Child Left Behind?: Assessing President Bush's assessment law. *Educational Foundations, 18*(3), 7-24.

Giordano, G. (2005). *How testing came to dominate American schools: The history of educational assessment*. New York City, NY: Peter Lang.

Graves, D. (2002). *Testing is not teaching: What should count in education*. Portsmouth, NH: Heinemann.

Gray, W. (1917). Studies of elementary-school reading through standardized tests (Vol. Supplementary Educational Monograph, No. 1). Chicago, IL: University of Chicago Press.

Haney, W. (1984). Testing reasoning and reasoning about testing. *Review of Educational Research, 54*(4), 597-654.

Hechinger, F. (1977, May 1). Why schools use standardized tests. *New York Times*, p. 16.

Kelly, F. (1916). The Kansas Silent Reading Tests. *Journal of Educational Psychology, 7*(3), 63-80.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

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

Leu, D., Kinzer, C., Coiro, J., & Cammack, D. (2004). Toward a theory of new literacies emerging from the Internet and other information and communication technologies. In R. B. Ruddell & N. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 1570-1613). Newark, DE: International Reading Association.

Long, R. (2011). Common core state standards: Approaching the assessment issue. *Reading Today, 29*(1), 23-25.

Madaus, G. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh year-book of the National Society for the Study of Education* (pp. 83-121). Chicago, IL: University of Chicago Press.

Madaus, G & O'Dwyer, L.(1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80*(9), 688-695.

Marks, D. (1989). Statewide achievement testing: A brief history. *Educational Research Quarterly, 13*(3), 36-43.

Merrow, J. (2013, August 14). Can a computerized test measure complex 'Common Core' skills? from Retrieved from PBS Newshour website: http://www.pbs.org/newshour/bb/education-july-dec13-commoncore_08-14/

Miyazaki, I. (1976). *China's examination hell: The civil service examinations of Imperial China* (C. Schirokauer, Trans.). New York City, NY: Weatherhill.

National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: Author.

Nichols, S. & Berliner, D.(2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Educational Press.

Pearson, P. & Stallman, A.(1994). Resistance, complacency, and reform in reading assessment. In F. Lehr & J. Osborn (Eds.), *Reading, language, and literacy: Instruction for the twenty-first century* (pp. 239-251). Hillsdale, NJ: Lawrence Erlbaum.

Readence, J. & Moore, D.(1983). Why questions? A historical perspective on standardized reading comprehension tests. *Journal of Reading, 26*(4), 306-313.

Resnick, D. (1982). History of educational testing. In A. K. Wigdor & W. R. Gardner (Eds.), *Ability testing: Uses, consequences, and controversies* (Vol. 2, pp. 173-194). Washington, DC: National Academy Press.

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

Samelson, F. (1987). Was early mental testing (a) racist inspired, (b) objective science, (c) a technology for democracy, (d) the origin of multiple-choice exams, (e) none of the above? In M. M. Sokal (Ed.), *Psychological testing and American society 1890-1930* (pp. 113-127). London, England: Rutgers University Press.

Sandia National Laboratories. (1993). Perspectives of education in America [known as the Sandia Report]. *Journal of Educational Research, 86*(5), 259-310.

Smith, N. (2002). *American reading instruction* (Special ed.). Newark, DE: International Reading Association. (Original work published 1934)

Starch, D. (1915). The measurement of efficiency in reading. *Journal of Educational Psychology, 6*(1), 1-24.
Starch, D. & Elliott, E.(1912). Reliability of the grading of high-school work in English. *School Review, 20*(7), 442-457.
Starch, D. & Elliott, E. (1913). Reliability of grading work in mathematics. *School Review, 21*(4), 254-259.
Thorndike, E. (1914). The measurement of ability in reading. *Teachers College Record, 15*,(4) 207-277.
White, E. (1886). *The elements of pedagogy*. New York City, NY: American Book.
White, E. (1888). Examinations and promotions. *Education, 8*(1), 517-522.
Wigdor, A. & Gardner, W. (Eds.) (1982). *Ability testing: Uses, consequences, and controversies* (Vol. 1). Washington, DC: National Academy Press.
Wirtz, W. (1977). *On further examination: Report of the advisory panel on the Scholastic Aptitude Test score decline*. New York City, NY: College Entrance Examination Board.
Wise, A. (1979). *Legislated learning: The bureaucratization of the American classroom*. Berkeley, CA: University of California Press.
Wolf, T. (1973). *Alfred Binet*. Chicago, IL: University of Chicago Press.