Paradoxes of High-Stakes Testing

GEORGE MADAUS AND MICHAEL RUSSELL, BOSTON COLLEGE

Over the centuries and across nations, tests have been employed as bureaucratic tools for a variety of purposes. As far back as 200 BC, the Chinese used tests to help eliminate patronage and open access to the civil service. The Dead Sea scrolls describe the use of tests by the Qumran community to determine when a man was ready to become a formal member of the community. England, France, and Italy, among other nations, have used tests to ensure that students acquire certain skills and establish standards of performance. In fifteenth-century Italy, tests were used to hold teachers accountable for student learning. Since then, policy-makers have used tests to hold students and schools accountable and allocate scarce resources.

These various testing policies were not meant to be punitive. Instead, these policies were, and continue to be, sincere attempts to address perceived problems in education. Two facts help explain why policy-makers are attracted to testing as a solution to problems in society and education. First, policy-makers realize they cannot directly regulate instruction in classrooms, but they can indirectly influence instruction by attaching rewards or sanctions to the results of mandated tests. Policy-makers have always been aware that stakes tied to a test force teachers to adjust instruction to prepare students for the test.

Beyond controlling teaching and learning, tests have served as an accountability tool that ensures value for expenditures of taxpayers' money. The logic behind using tests to hold teachers, students, and schools accountable was expressed by Eamon DeValera, the Prime Minister of Ireland. Proposing a system of certification examinations at the end of primary school, he argued successfully before Parliament in 1941:

> . . . if we want to see that a certain standard is reached and we are paying the money, we have the right to see that something is secured for that money. The ordinary way to test it and try to help the whole educational system is by *arranging our tests in such a way that they will work in a direction we want*. [emphasis added] (Ireland. Dail Eireann, 1941, col. 119)

Like many of today's policymakers, DeValera believed that tests provide the evidence that determines whether taxpayers' money is well spent. This reasoning is reflected clearly in President Bush's and President Obama's reliance on tests to evaluate the success of educational programs. This use of tests to measure the outcomes of education reflects a larger belief in the use of metrics to determine the success of any policy.

For many politicians, and many in the business and testing communities, testing has morphed from a means of obtaining information about the education system to a key strategy for improving educational quality. Testing is viewed as both a system of *monitoring* student performance and a *vehicle of change* driving what is taught and how it is taught, what is learned and how it is learned. The rise of testing in the United States is rooted in the idea that the correct system of rewards and punishments will motivate obstinate, dispirited, lazy, or recalcitrant students, as well as their teachers, to try harder (Webb, Covington, & Guthrie, 1993). Today, testing is seen as essential to developing a world-class educational system, motivating the unmotivated, lifting all students to world-class standards, increasing the nation's productivity, and restoring global competitiveness. A reform tool that yields these outcomes would truly be manna from above—manna to improve our schools, and feed our teachers and students who are often depicted as wandering in a desert of mediocrity.

Manna is generally defined as valuable—bread from heaven—but manna has another meaning. Exodus recounts that when the Israelites awoke to find "a fine, flake-like thing," they naturally asked, *"manna?"* "What is it?" It is time again to ask, "Manna?" What are these tests our children must take? Is testing really the bread of reform or are there weevils in the flour?

To answer these questions, there are at least four characteristics of testing that must be understood. First, we must recognize that testing is a technology. Second, human and cultural factors interact with this technology and affect the accuracy with which a test measures student achievement. Third, the technology of testing has evolved and continues to evolve. Finally, like any technology, the effects of testing are paradoxical.

TESTING IS TECHNOLOGY

Testing is deeply ingrained in the American psyche. Our familiarity with testing is one reason most Americans do not think of it as a technology. But this is precisely what testing is—a technology with deep roots in our educational system (Hughes, 1989).

A common definition of technology is the application of science to satisfy a pressing and immediate need to solve a problem or serve as a means to an end (Basalla, 1988; Staudenmaier, 1988). Tests have long been used to solve a variety of social and educational problems: eliminating patronage; opening access to public service; ensuring that students acquire certain skills; establishing and maintaining standards of performance; holding teachers, students, and schools accountable for learning; and allocating scarce resources.

Technology is also defined as a set of *special* knowledge, skills, and procedures that create standardized techniques (Lowrance,

1986; Winner, 1977). For every important technology, a community of experts forms and develops its own specialized vocabulary and value system (Staudenmaier, 1985). Testing has its own techniques, special knowledge, and a community of practitioners—referred to as psychometricians. This community relies on specialized techniques and arcane algorithms, the most common of which are based on Item Response Theory that is used to determine which questions are included on a test and to calculate scores used to classify students.

The term "technology" often conjures up images of inventions, artifacts, machines, or devices such as cars, trains, power transformers, or computers. Although testing is not thought of in this way, it also has its own "hardware." For years, test booklets, answer sheets, and optical scoring machines were used to make the testing of large numbers of people efficient and economical. Today tests are also administered on computers to increase the speed of reporting results, improve convenience, and decrease administrative and scoring costs.

Like other technologies, testing has hidden values (Borgmann, 1984). Testing values measurement and quantification, and objectivity over subjectivity. Testing experts believe that a single trait can be measured by a test, and that the degree to which a person possesses the trait explains the performance on a test. The testing community places little value on the many social, cultural, and individual factors that also influence how a student performs on a test.

Technologies are developed to solve problems and make human activities easier to perform. With every technology, though, there is a dark side that has serious costs (Postman, 1992). The emergence of a technical elite, specialized languages, arcane algorithms, and hidden values make testing a seductively attractive friend, yet a potentially harmful enemy for some students, teachers, and schools.

Technologies affect the nature of our lives and society in significant ways. Consider how technologies like automobiles, telephones, and television have reshaped our lives and society. Not only have they affected the way we travel, communicate, and entertain, they have created massive industries, influenced politics, and reshaped the structure of society. The many ways in which technology shapes people's lives apply to testing, as well.

When used as a policy tool, high-stakes testing shapes society by promoting a variety of values that include utilitarianism, economic competitiveness, technological optimism, objectivity, bureaucratic control, accountability, administrative convenience, numerical precision, efficiency, standardization, and conformity. Testing also shapes important educational values.

Using test scores to classify students and schools reshapes our conceptions of student attainment and school quality. Attainment no longer focuses primarily on skills and knowledge. School quality ceases to focus first on teaching, resources, and opportunities for learning. Instead, student attainment and school quality become defined by individual and group test scores.

Testing subjects like mathematics, science, and language arts, but not testing other subjects like history, music, and art, defines the relative value of different curricular areas. Highstakes tests also reshape student–teacher relationships and define what an educated person should know, understand, and be able to do, and therefore, what should be taught and learned.

The predominant, often tacit, values in testing have been those of policy-makers, test sponsors, testing advocates, and the technical community. Their values have not been critically examined, nor weighed against the competing pluralistic values of teachers, test-takers, parents, critics, and other concerned citizens.

The introduction of a technology can require "workers to take on more responsibility, use more judgment, and have a broader understanding of the total work process" (Applebaum, 1992, p. 537), but technology can also deskill, routinize, and place workers under closer supervision. Unfortunately, the technology of high-stakes testing diminishes teachers' judgment and decreases their responsibility, and instead, routinizes instruction, deskills many teachers, and places them under closer supervision.

For some teachers, a high-stakes test simplifies their job. The test gives teachers a clear target and allows them to simply teach to the content of the test. At the same time, high-stakes testing can degrade teaching skills by reducing teaching to narrow test preparation. Rather than developing each child as an individual, focus is placed on improving test scores. When these test scores are then compared among teachers and schools, relationships and trust among teachers are endangered, leading some teachers to blame their colleagues at lower grade levels for poorly preparing students; others are humiliated when class or school averages are displayed publicly (Booher-Jennings, 2005).

Focusing solely on test scores also devalues teachers' judgments about the achievement of their students and their readiness for specific instructional interventions. Mistrust of teachers' qualitative judgments has long been a hidden value that underpins policies dictating the use of a single standardized test score to make decisions about students. Despite contradictory teacher insight about what a student knows or is able to do, decisions based solely on test scores are often non-negotiable. While there is much talk about the need to make teaching more attractive, high-stakes testing programs treat teachers and students as passive beneficiaries, which comes at the cost of other core values.

High-stakes testing is a minimalist technological strategy used to reform education. Although these tests can provide useful information about student attainment, they do not address the deeper underlying problems that are barriers to student learning, such as student health, nutrition and living conditions, class size, and teachers' pre- and in-service education (Kellaghan, 2000). Ironically this list, however long, is one of the reasons that the technological solution of testing is so attractive. Mandating high-stakes testing allows policy-makers to sidestep difficult ideological, economic, and political issues that complicate addressing these underlying problems. Making high-stakes testing the heart of reform paves over the surface and obscures the root causes of poor attainment.

HUMAN AND CULTURAL FACTORS

Testing is an integral part of American education. For most readers, a school where children never take a test is unimaginable, but it is important to recognize that the way testing is viewed and approached in our society is determined by class and culture (Henry, 1963; Hall, 1977).

For starters, family and cultural background influence the way students view and interact with tests. A recent analysis of National Assessment of Educational Performance (NAEP) reading test scores examined the relationship between home factors and test performance. This analysis found that single-parent families, parents reading to a child every day, hours a child spends watching television, and the frequency of school absences collectively explained two-thirds of the differences in reading scores (Barton & Coley, 2007). The relationship between test performance and home factors calls into the question the use of test scores to judge school quality without considering other factors that influence learning and test performance (Rothstein, 2004).

This relationship is also influenced by two culturally held values. The first is that achievement is an individual accomplishment. The second value is that individuals must display their accomplishment publicly. Middle-class children are socialized to accept these two values. Prior to entering school, they are "tested" by their parents. Babies are repeatedly asked to point to their nose, bottle, shoes, and so on. Toddlers are asked "Where is the truck?" or to point to the horse in a picture. Preschoolers are asked about stories in books, and people and events in their lives.

In contrast, young children from different backgrounds are not asked by adults to be "information givers." As a result, many of these children do not have a clear idea of what testing is about when they initially encounter it in school. This is true for many American Indian children. Rather than emphasizing the development of verbal skills, Indian cultures socialize children through nonverbal communication and emphasize spatial and motor skills and sequential visual memory. Tribal cultures also emphasize sharing and working together. The tests the child encounters in school, however, focus primarily on verbal skills and force the child to work alone (Locust, 1988).

American Indians are not the only people who have a culture clash with tests. Culture also influences the test performance of other minorities, recent immigrants, bilingual students, and females (National Commission on Testing and Public Policy, 1990). Cultural influences on test performance begin at a very young age and affect how an individual student and groups of students are perceived and treated throughout their time in school.

The name of a test conveys powerful cultural meanings. Words like "intelligence," "ability," "competence," "honesty," "aptitude," "readiness," and even "achievement" are used to name the construct a test purports to measure. While these words are familiar to everyone, a universally shared understanding of what they mean when used to name a test is by no means guaranteed. The interpretation of test performance is based on the affective, connotative, emotional, and metaphorical baggage associated with the word used to name the test.

Intelligence, for example, means different things to different people. For many, the name "intelligence" or "IQ" test conjures up the image of an innate, stable ability to reason. Many people believe that intelligence is largely genetic. For others, the word "intelligence" refers to street smarts. To still others, there are multiple intelligences that include such characteristics as verbal intelligence, musical intelligence, and visual/spatial intelligence (Gardner, 1997).

The experience of many minority children in school is a powerful example of the other side of the IQ coin. As early as 1918, Charles Judd, then director of the School of Education at the University of Chicago, argued that "unsatisfactory school results [can] be traced to the *native limitations in the ability* [emphasis added] of [the] child or to the home atmosphere in which the child grows up" (p. 152). When these children scored low on an English-language IQ test, they were often treated quite differently from high-scoring children, despite the fact that when they took the test they were at a disadvantage for cultural and linguistic reasons.

Simply changing the name of a test can alter perceptions of that test and *attitudes* about *test use*. For example, in the 1980s and '90s, many people recoiled at schools using "intelligence" test scores for admission to kindergarten or promotion to the first grade. Their attitudes, however, changed when a test with the same kinds of questions was named a "readiness test" (attributed to Lorrie Shepherd in Cunningham, 1989). It was acceptable to say that the child is not ready for school, but not acceptable to say the child is not intelligent enough to enter kindergarten or first grade.

A crucial factor in test development is the quality of the items or questions. Creating a test is an art, not a science. Test developers must maximize test reliability and validity by covering the different kinds of content and skills that form the domain, but they must also build a test that takes into account the limited attention span of children and that can be administered in a single class period. Further, adults write the items that children and adolescents at different developmental stages must answer. For a traditional paper-and-pencil test, each item must stand alone without the benefit of detailed context to further define the situation or problem. Test-takers scrutinize each item very carefully; each word, diagram, graph, or equation in an item must unambiguously establish the task. The presentation of an item or the directions can cause some examinees to get the item wrong even though they have the necessary knowledge, skill, or ability, while other students without the knowledge, skill, or ability get it right. As an example, an item designed to test a student's mathematics achievement, but which references the tilling of a vegetable garden, may create an unfamiliar context for students living in large urban areas.

Bruner (1966) insists that knowing what children *do* is not enough; we need to know what they *think* they are doing and their reasons for doing it. Consistent with this recommendation, a study by Walter Haney and Laurie Scott (1987) asked young children to

explain why they chose a particular answer to items on commonly used standardized achievement tests. The item displayed an image of a potted flower, a potted cactus, and a head of cabbage, and asked "Which plants need the least amount of water?" The option designated as correct by the item writer was the cactus. A number of students, however, selected the cabbage. When asked why they chose the cabbage, the students explained that the cabbage was not in a pot, had been picked, and therefore no longer needed water. Perfectly good reasoning on the children's part, but unanticipated by the item writers who designated the cactus as the correct answer. While it may seem trivial to focus on a single problematic item on a test, such an item can have an adverse impact on the classification and subsequent treatment of students who are within a point or two of a cut-score.

Another important design limitation of both selection and supply items is that little is known about how a standardized procedure performs in slightly different contexts, and how different contextual presentations affect answers. Very slight alterations in test design can lead to very different descriptions of student performance. An illustration of this problem is the seemingly dramatic drop in the National Assessment of Educational Progress (NAEP) reading scores from 1984 to 1986. This drop was so large that people thought it improbable, and it became known as the infamous "reading anomaly."

Researchers Albert Beaton and Rebecca Zwick (1990) showed that posing the same questions but making slight, innocuous changes to the test's appearance explained the drop in test scores. Minor changes like switching the order of the questions—e.g., item 3 becomes item 7—using stapled instead of saddle-stitched test booklets, having students fill in an oval rather than circle a letter to mark their answer, and using brown or black instead of blue print conspired to produce a very misleading description of progress in reading performance in our nation's schools. The changes in test scores, used to make inferences about changes in student achievement, were affected by subtle changes in the test format rather than changes in students' actual skills and knowledge. These findings are powerful given the tendency of many to accept a quantitative test score as an infallible pronouncement about students' attainment.

In short, a range of human and cultural factors influence testing. These factors explain in part why it is challenging to develop a single test that works across a very diverse body of students and schools.

THE TECHNOLOGY OF TESTING IS EVOLVING

Since its inception, the technology of testing has evolved dramatically. When first introduced, tests were delivered in an oral format. With the introduction of a paper format, tests took the form of open-ended essays that resulted in a qualitative judgment about the correctness of the response. The introduction of the quantitative mark replaced qualitative judgments with seemingly precise numbers. To simplify the quantification of performance, questions aimed at assessing rhetorical style were replaced by items with a single correct answer. In the mid-nineteenth century, Horace Mann capitalized on this change to quantification and introduced exams that posed an identical set of questions simultaneously and under similar conditions, to a large number of students to yield comparable scores. This approach effectively introduced the notion of a standardized test. Quantification of performance also led to the integration of statistics with test development and ultimately created the field of psychometrics. In the early twentieth century, the multiple-choice item was introduced and further solidified the focus on a single correct answer. In the early 1930s, the first scanner was developed, and during the 1950s high-speed scanners greatly increased the efficiency of multiple-choice testing. More recently, the introduction of computers, computer-adaptive tests, and automated essay scoring have further increased efficiencies.

In 2001, Randy Elliott Bennett predicted that computer-based testing would pass through three evolutionary stages before reaching its full potential. First, computers would be used to increase the efficiency of testing. Second, multimedia would be integrated into tests to increase the authenticity of items and tasks presented to students. Finally, computers would be used to deliver tests anywhere and at any time, so that testing becomes more integrated with instruction. To date, testing programs that have embraced computer-based testing have done so solely to increase efficiency. Their goals are quite simple—improve the efficiency with which tests are distributed, decrease the time required to score multiple-choice answers, and increase the speed with which results are reported. While achieving these goals saves time and money, it does not capitalize on the full benefits of computer-based testing.

There are at least two ways in which computer-based technologies could be applied to help address some of the cultural and human challenges to testing. First, principles of universal design can be applied to tailor the delivery of tests to improve student access to test items. Second, diagnostic tests can provide more detailed information about student thinking and misconceptions that may interfere with their demonstrations of achievement.

Universally Designed Test Delivery

According to test developers, an error associated with a student's test score is random. Feeling ill, being momentarily distracted by noise or movement in the room, mismarking an answer sheet, and guessing correctly on an item are all random events that may affect a student's test score. These random events produce random error.

Some error, however, is not random, but occurs systematically due to the specific attributes of an individual student. The design of a test often causes these attributes to limit the student's ability to demonstrate what he or she actually knows or can do. For example, a student with dyslexia struggles to read word problems on a mathematics test, and has trouble understanding what is being asked in the problem or does not have time to complete all problems on the test. A student with impaired vision has difficulty reading graphs or figures that contain fine lines or tightly packed information, and consistently makes errors or skips items entirely. Another student

with an information processing disorder becomes overstimulated when working on an item that contains a narrative, images, and multiple answer options and requires the use of a formula sheet and a calculator. All of these students may have a firm grasp of the concept being measured, but they have difficulty demonstrating their knowledge or ability due to the design of the test items.

In each of these examples, error occurs systematically due to an attribute unrelated to what the test is attempting to measure. Dyslexia negatively affects a student's ability to decode text, but the test is measuring mathematics ability. Vision and information processing are important, but mathematics tests are not designed to measure these attributes.

Applying principles of universal design can decrease the effect that these attributes—decoding, vision, information processing, etc.—have on test performance, When first introduced, the concept of universal design was a direct response to design flaws in buildings—staircases, narrow entrances, escalators, high sinks, etc.—that made it difficult for people with physical disabilities to access buildings or use the facilities within the buildings. Universal design overcomes these types of obstacles by purposefully designing buildings so that they provide a choice of convenient and appropriate options when accessing a building or using specific facilities.

Applying the principle of universal design to testing ensures that all students have access to a test in a manner that allows them to *demonstrate what they actually know and can do*. Just as it is no longer acceptable to design and build a structure that requires a person with a physical disability to use a doorway located in the back of the building, a testing program should not require students with disabilities or special needs to take a test that is separate and distinct from that which is taken by all other students. Just as it is unreasonable to require people with a physical disability to bring their own ramp to ascend a staircase, a testing program should not require the students or their school to purchase special software in order to take a test. Instead, a universally designed test should be usable by all students, regardless of their disability or special need.

Rather than creating and distributing separate versions of the test or requiring students to work in separate rooms with an individual test proctor, a universally designed computer-based system can deliver a test tailored to all students. Building in reading and signing ensures that all students receive a high-quality, tailored presentation of the text, with accurate pronunciation of all words, free from inadvertent (or intentional) clues. Students are then free to have the text read or signed as many times as needed without feeling embarrassed or believing that they are overtaxing the test proctor.

Clearly, combining principles of universal design with computer-based technologies holds tremendous potential to increase test validity for students with disabilities and special needs. Moreover, just as the option of displaying a television program with closed captioning has proven useful for many purposes beyond increasing access for the hearing impaired—watching television in a noisy restaurant or airport gate, or late at night while a spouse sleeps—it is likely that universally designed computer-based tests will come to benefit a wide spectrum of students in various unanticipated ways.

Diagnostic Tests

For most teachers, the current approach to high-stakes testing is like the first few minutes of a visit to the doctor. Imagine that you are feeling ill and believe you have a fever. You go to the doctor, she takes your temperature, gets a reading of 102, and says that you do indeed have a fever. She then tells you to go home. Like a doctor telling a person with a fever that the temperature is 102 and then ending the visit without a diagnosis, high-stakes tests provide information that confirms what most teachers already know. For students whose level of achievement is relatively low, however, current tests fail to provide diagnostic information about *why* they are struggling.

Concern about the lack of diagnostic information provided by achievement tests dates back several decades. In the 1930s Ralph Tyler developed a test that yielded four scores that provided teachers with information about the type of errors students made when answering items (Smith & Tyler, 1942). In the 1970s, researchers at MIT developed a computer-based tutoring system that presented students with mathematics items. The system, called BUGGY, then used a student's incorrect responses to identify procedural errors and offer additional instruction to help the student apply the correct procedure (Krige, 1998).

More recently, the National Research Council (2001) resurrected calls to improve the diagnostic information provided by tests. Specifically, the Council recommended that student assessment should provide timely and informative feedback about the strategies children use when solving problems and that students' thinking be made visible so that instruction can be tailored to support learning.

In 2004, the Technology and Assessment Study Collaborative launched the development of a diagnostic assessment system designed to increase the instructional value of testing. (O'Dwyer & Miranda, 2009). Known as the *Diagnostic Algebra Assessment System*, the initiative developed a comprehensive computer-based assessment and instructional system that has three key features. First, the system provides teachers access to a series of online tests, each of which focuses on a specific algebraic concept. For each test, items are designed to measure students' understanding of the concept (e.g., equality). For a student who performs poorly, each test also provides information about whether a known misconception is interfering with the student's understanding.

Second, the system provides immediate feedback to teachers. An initial report sorts students into three categories. In the first category are students who performed well and appear to have a solid understanding of the tested concept. In the second category are students who did not perform well and who appear to hold a specific misconception. In the third category are students who also did not perform well, but do not appear to hold the misconception (e.g., they make a variety of errors unrelated to the measured

misconception). These categories give teachers a better understanding of how well their students are performing and why some students are struggling with a given concept.

Third, the system links teachers to lessons and activities they can use to help students correct a given misconception. Students identified with a given misconception are also connected to learning activities that focus on that misconception.

While diagnostic assessments can take many forms and be used to measure learning in many ways, the *Diagnostic Algebra Assessment System* provides a look into the future of how testing programs can capitalize on computer-based technology to provide teachers with diagnostic information they can use to tailor instruction for individual students.

PARADOXES OF TESTING

Without a doubt, high-stakes testing policies are well intentioned. State and federal testing policies are intended to focus instruction and learning on the important content and skills that form state curriculum—and they do. In each state, the tests define standards and expectations for student achievement. While differences in the expectations and the difficulty of tests vary widely across states, the test scores nonetheless provide teachers and schools with information about student performance. They give communities information about the quality of their schools and help parents make informed decisions when choosing a school for their children. High-stakes tests also open doors of opportunity to those previously shut out by holding teachers and schools accountable for student achievement and helping them to focus attention on students who were poorly served in the past. These are all positive outcomes, and it is important to acknowledge the positive aspects of high-stakes testing.

However, it is the combination of these intended outcomes and unintended consequences that make high-stakes testing paradoxical. As seen throughout history, in recent research, and in literature of all genres, there are three predictable ways that a high-stakes test, even one that embraces the technological advances we just explored, adversely affects teaching and learning. First, teachers give greater attention to tested content and decrease emphasis on non-tested content. This narrows the content and skills taught and learned *within a discipline*. Second, a high-stakes test preempts time and coverage from disciplines not tested. This narrows the curriculum *across subject fields*. Third, there is a "trickle down" effect. The content and skills covered on the high-stakes tests at the upper grades displaces the content and skills of non-tested lower grades, altering the curriculum *across grades*.

Narrowing What Is Taught Within a Discipline

Over the past forty years, surveys of American teachers about testing have documented the effects of high-stakes tests on the practices and attitudes of teachers. Across a range of state testing programs, large percentages of teachers report considerable attention and time given over to material covered by a high-stakes test at the expense of non-tested content and skills. Teachers also report that they spend more time preparing students specifically for the test (Abrams, Pedulla, & Madaus, 2003).

For example, a 2001 national survey of more than 4,000 teachers found large differences between those teaching in high-stakes situations and those teaching where the stakes were not as high (Pedulla, Abrams, Madaus, Russell, Ramos, & Miao, 2003). In settings where a high stake was tied to test performance, 80% of teachers reported that there is so much pressure for high scores on the test that they had little time to teach anything that was not on the test; in contrast only 56% of teachers responded this way in settings that did not have high stakes linked to test performance. In high-stakes settings, 43% of teachers reported that they greatly increased time spent on instruction in tested areas. In contrast, only 17% of teachers in low-stakes settings reported greatly increasing time on tested areas. And, in high-stakes settings 63% of teachers reported using test preparation materials developed commercially or by the state, whereas only 19% of teachers in lower-stakes states reported doing so.

Beyond having an impact on what is taught, preparation for state tests also affects the methods of teaching and learning. As an example, instructional use of computers is adversely affected. To help students become accustomed to writing an essay for a paper-and-pencil state test, approximately 30% of teachers nationwide reported that they either decreased the amount of time students used word processors in the classroom or did not allow their use. The reduction or elimination of computer use occurred despite a large body of research showing that regular use of word processors improves the quality of student writing (Russell & Abrams, 2004).

The National Commission on Testing and Public Policy also describes how pressure to improve scores on reading and math tests can narrow teaching to test preparation. The Commission warned that the high stakes attached to test use are ". . . driving schools and teachers away from instructional practices that would help to produce critical thinkers and active learners." As one example, the report described how "[i]nstead of reading books, students in many classrooms read isolated paragraphs and practice answering multiple-choice questions about them" (1990, p. 19).

In response to high-stakes tests, teachers also narrow the focus of their instruction so that students respond to only the item types found on the test. This is particularly problematic when a test includes only multiple-choice items. Deborah Meier, a highly regarded principal in Manhattan, describes how students in reading classes were required to read dozens of short paragraphs about which they then answered multiple choice questions that resembled the tests given each spring. She also recounted that when synonyms and antonyms were dropped from the test, teachers promptly stopped using worksheets on synonyms and antonyms (National Institute of Education, 1981).

More recent studies have found that many teachers have decreased the use of time-intensive instructional strategies and more lengthy enrichment activities and increased the use of problems and

questions similar to those on the high-stakes test (Pedulla et al., 2003; Taylor, Shepard, Kinner, & Rosenthal, 2003).

Narrowing What Is Taught Across Subject Areas

Scientists report that cuckoo birds have developed an interesting strategy for survival. Mother cuckoo birds lay their eggs in the nests of other birds. When the cuckoo egg hatches, the nesting mother bird attends to the cuckoo chick. As the cuckoo chick grows, it throws the nesting mother's chicks and eggs from the nest and becomes the primary focus of the mother.

An analogous phenomenon occurs when a high-stakes test is introduced. Like the nesting mother, teachers gradually spend more time attending to tested subject areas. Given that there is a limited amount of time in the school day, this increased attention squeezes out time for other school subjects and activities (National Research Council, 2007).

Disregarding characteristics and abilities not tested was apparent in China as long ago as 1043. A critic of the civil service exams complained that because the examinations did not assess imagination and studies of practical utility these areas were neglected (Little, 1993). More recent evidence shows that teachers increase emphasis on tested subjects at the expense of non-tested subject areas.

As an example, a national survey of teachers revealed that approximately 80% of the teachers reported increasing time spent on subject areas that are tested, and nearly 50% reported decreasing time on subjects that are not tested, such as fine arts, physical education, foreign languages, and industrial/vocational education. In addition, teachers reported that testing decreased the amount of time spent on activities not directly related to specific subject areas such as field trips and other enrichment activities (Pedulla et al., 2003). Shortchanging time is not limited to non-tested subjects but also extends to recess. To provide more time for reading and mathematics, schools across the nation are cutting back on recess time. In 2006, encroachment on recess time prompted the National Parent Teachers Association (PTA) to launch a *Rescuing Recess* campaign (National PTA, 2006).

In general, the influence of state testing programs on teachers' instructional practices is stronger where the stakes are high for both schools and students than in settings where the stakes are lower. The impact of testing programs is also generally stronger in elementary and middle schools than in high schools (Pedulla et al., 2003).

A separate study by the Center on Educational Policy found that 71% of school districts reduced time in at least one subject to expand time for reading and math. More specifically, 33% of districts reduced time for social studies, 29% reduced time for science, and 22% reduced time for art and music (Center on Educational Policy, 2006b). The study also found that in some districts the amount of time struggling students spent on tested subjects doubled, at times causing them to miss other subjects altogether (Center on Educational Policy, 2006a).

Narrowing Teaching and Learning Across Grades

The demands of high-stakes tests not only affect what is and what is not taught within test grades, but also trickle down to non-tested lower grades. To better prepare students for high-stakes tests given in the upper grades, the curriculum in kindergarten and first grade is altered. More emphasis is placed on academic skills at the expense of social, emotional, and physical goals for children.

At the same time, some school districts have started to "red shirt" kindergarten students. In athletics, red shirting refers to the practice of holding back a scholarship athlete from playing for a year to let the student develop further in the given sport. In elementary schools, red shirting occurs when a student either is not allowed to enter kindergarten despite meeting the age requirement or is retained once in kindergarten. Red shirting attempts to capitalize on the cognitive growth that occurs as students get older. By delaying the start of kindergarten or first grade, the hope is that test scores will be higher down the road if for no other reason than redshirted students are a year older when they take a high-stakes test (Shepard & Smith, 1988).

Broader Effects on Educational Practices

In addition to the effects discussed above, some schools and teachers have adopted the practice of triaging students. A study conducted by researchers at Leeds University found that many teachers concentrate their efforts on those students who are most likely to succeed on the *Standardized Assessment Tasks*. These students are often referred to as "Bubble Kids" because they are on the bubble of passing the test or of moving up to the next performance level. Other researchers have found that teachers concentrate instruction on those most likely to succeed, to the detriment of those not expected to do well (Kellaghan & Greaney, 1992).

Jennifer Booher-Jennings (2005) investigated the treatment of bubble kids in the Texas high-stakes testing program. A teacher in her study describes the bubble kids this way, "The ones who miss by one or two points—they just need a little extra help to pass so we concentrate our attention on that group. The bubbles are the ones who can make it" (p. 241). Booher-Jennings calls this emphasis on bubble kids "educational triage." She reports that bubble kids receive a variety of benefits. These benefits include more teacher attention, more class time and extra help to prepare for the test, individual attention or small group instruction, help from literacy teachers, afterschool and Saturday tutoring, and test preparation from music, gym, and library teachers, instead of instruction in non-tested subjects. She also found that referring bubble kids for special education exempted them from the accountability requirements of the test, thereby improving the school's accountability rating.

Educational triage is not a new practice. There is evidence that "educational triage" occurred during the nineteenth-century payment-by-results era. At that time, many teachers concentrated on those pupils who were most likely to yield the full monetary reward for them (Rapple, 2004).

Recently, the practice of "educational triage" has caught the attention of the press. Joshua Benton of the*Dallas Morning News* describes the downside of the triage approach this way:

> But what if you're one of the 'remedial' kids—everyone below the bubble? . . . [Teachers] realize they're going to be judged on how many of their kids pass—not how much improvement they can squeeze out of their weakest kids. So they go after the low-hanging fruit: the bubble kids. (Benton, 2005)

Corrupting the Measure

In economics and sociology, experts acknowledge that the act of measurement distorts what is being measured. As a result, every measure that becomes a target ceases to be a good measure. When a quantitative indicator is used for social decision-making, it distorts and corrupts the indicator itself and the social process it was intended to monitor (Campbell, 1975). The same corrupting effect results when educational tests are used as social indicators and targets for accountability.

A legendary example of widespread test corruption was exposed in 1987 when it was found that most states and districts were reporting above-average scores. This study—nicknamed the *Lake Wobegon Report*, after Garrison Keillor's mythical town where "the women are strong, the men are good-looking, and all the children are above average"—concluded that these results were implausible and misleading. One of the explanations for the above-average results was that schools routinely taught directly to the test, and even to specific test questions (Cannell, 1987; 1989).

After examining high-stakes testing programs in 18 states, a more recent study concluded, "While a state's high-stakes test may show increased scores, there is little support . . . that such increases are anything but the result of test preparation and/or the exclusion of students from the testing programs" (Amerin & Berliner, 2002). Together, these findings demonstrate that emphasis on test preparation distorts the test's ability to validly portray the "true" achievement level of many students.

CONCLUSION

The term *iatrogenic* refers to physician-induced illness—that is, a negative, unanticipated effect on a patient of a well-intended treatment by a physician. The paradox of high-stakes testing might well be called *peiragenics,* that is, the negative, unanticipated effects on students, teachers, and schools of well-intended testing policies.

As has been described, the negative effects are many. They include narrowing the curriculum, decreasing attention on non-tested subjects, changing preschool and kindergarten curricula, narrow test preparation, corruption of test results, cheating, triaging "bubble" students, retaining students in grade, increased dropout rates, and increasing student stress and anxiety. All of these paradoxical negative consequences of high-stakes testing are chronic, predictable, and well documented over centuries and across continents.

It is important to recognize, however, that it is not the test per se that causes these disorders. Instead, it is the *stakes* associated with test scores that drive teachers, pupils, and other stakeholders into behavior that results in the many paradoxical unintended outcomes discussed above. The conundrum in this paradox is that the stakes attached to test results are the driving force of the reform policy. The stakes produce both the salutary effects and the unintended negative consequences.

The debate over the use of tests in the development of policy is really a debate over what we want from our schools. It is a debate over educational values and competing educational philosophies, and it is about means and ends. It is not a debate on technical matters related to testing. In fact, if testing is the answer, then we have done a poor job of stating the question. By merely focusing on the test results we sidestep the more crucial question of the proper role of testing.

Medicine offers another apt analogy, namely the systematic evaluation of the impact of new medical technologies or treatments. One such study offered the following advice:

> Good decisions cannot be made without an adequate assessment of the relevant facts. . . . This evaluation should assess the likelihood of a favorable outcome and the benefits and burdens to the patient of all possible outcomes. Further, there should be candor not only about what is known, but also about what is unknown. (LORAN Commission, 1988, p. 27)

This mentality is generally absent in discussions about high-stakes testing programs. Despite the fact that the advantages and disadvantages of such testing programs—for different kinds of students, at different grades and ages, in different kinds of educational settings—are predictable before implementation, too often policy-makers ignore them.

This does not mean that we must wait for the "perfect" test—there is no such thing. Nor does it mean that when we find harmful effects—and we will—the program must be scrapped. Medical technology is not perfect; there are potentially harmful side effects associated with treatments determined to be *generally* safe and efficacious. However, like physicians, educators should know the nature and extent of harmful side effects before adopting a high-stakes testing program.

Equally important, we need to know how the infrastructure of a high-stakes testing program will change our schools and other social systems. In short, we need a sociology of testing. We must satisfy ourselves that the benefits of the test and its accompanying infrastructure will clearly outweigh the harms before implementation. If we then decide to proceed, we must monitor the effects to ensure that the benefits continue to outweigh the harm, and to identify any unanticipated negative side effects.

References

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

Amerin, A. L., & Berliner, D. (2002). High-stakes *t*esting, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Retrieved from http://epaa.asu.edu/epaa/v10n18/.

Applebaum, H. (1992). *The concept of work: Ancient, medieval, and modern*. Albany, NY: State University of New York Press.

Barton, P. E., & Coley, R. J. (2007). *The family: America's smallest school*. Princeton, NJ:Educational Testing Service.

Basalla, G. (1988). *The evolution of technology*. New York, NY: Cambridge University Press.

Beaton, A. E., & Zwick, R. (1990). *The effects of changes in the national assessment:Disentangling the NAEP 1985–86 reading anomaly*. Princeton, NJ: Educational Testing Service.

Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives, 9*(5). Retrieved February 15, 2001, from http://epaa.asu.edu/epaa/v9n5.html.

Benton, J. (2005, September 19). TAKS push not so equal. *Dallas Morning News*.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal, 42*(2), 231–268.

Borgmann, A. (1984). *Technology and the character of contemporary life: A philosophical inquiry*. Chicago, IL: University of Chicago Press.

Bruner, J. S. (1966). *Toward a theory of instruction*. Cambridge, MA: Harvard University Press.

Campbell, D. T. (1975). Assessing the impact of planned social change. In *Social and Public Policies: The Dartmouth/OECD Conference*. Hanover, NH: Public Affairs Center, Dartmouth College.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average*. Daniels, WV: Friends for Education.

Cannell, J. J. (1989). *The "Lake Wobegon" report: How public educators cheat on standardized achievement tests.* Albuquerque, NM: Friends for Education.

Center on Educational Policy. (2006a). *From the capital to the classroom: Year 4 of the No Child Left Behind Act: Summary and recommendations*. Washington, DC: Center on Educational Policy.

Center on Educational Policy. (2006b). *From the capital to the classroom: Year 4 of No Child Left Behind*. Retrieved from www.cep-dc.org. Washington, DC: Center on Educational Policy.

Cunningham, A. (1989). *Enney, Meeny, Minie, Moe: Testing policy and practice in early childhood*, Paper commissioned by The National Commission on Testing and Public Policy. Chestnut Hill, MA: Boston College.

Gardner, H. (1997). *Frames of mind*. New York, NY: Basic Books.

Hall, E. T. (1977). *Beyond culture*. New York, NY: Anchor Books.

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test item ambiguity. In R. Freedle & R. Duran (Eds.), *Cognitive and linguistic analyses of testperformance*: Advances in Discourse Processes (Vol. 22, pp. 298–368). Norwood, NJ: Ablex Publishing Corporation.

Henry, J. (1963). *Culture against man*. New York, NY: Vintage Books.

Hughes, T. P. (1989). *American genesis: A century of invention and technological enthusiasm*. New York, NY: Penguin Books.

Ireland. Dail Eireann. (1941). Parliamentary Debates, Col 1119.

Judd, C. H. (1918). A look forward. In G. M. Whipple (Ed.), *The measurement of educational products* (pp. 152–160). Bloomington, IL: Public School Publishing Co.

Kellaghan, T. (2000). *Educational equity and inclusion*. Paper presented at the Mexico-Irish Conference, Ministry of Education, Mexico City, October 5–6, 2000.

Kellaghan, T., & Greaney, V. (1992). *Using examinations to improve education. A study infourteen African countries*. Washington, DC: World Bank.

Krige, A. (1998). *Intelligent tutoring systems: BUGGY*. Retrieved July 17, 2008 from http://tecfa.unige.ch/staf/staf-d/krige/staf11/buggy.html.

Little, A. (1993). *Toward an international framework of understanding assessment*. Paperpresented at the Conference on Learning, Selection and Monitoring: Resolving the Roles of Assessment, International Centre for Research on Assessment, Institute of Education, University of London.

Locust, C. (1988). Wounding the spirit: discrimination and traditional American Indian belief systems. *Harvard Educational Review, 58*(3), 315–330.

LORAN Commission. (1988). *Report of the LORAN Commission to the Harvard Community Health Plan*: Harvard Community Health Plan, Boston, MA.

Lowrance, W. W. (1986). *Modern science and human values*. New York, NY: Oxford University Press.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: transforming testing in America*. Chestnut Hill, MA: The National Commission on Testing and Public Policy, Boston College.

National Institute of Education. (1981, July 8, 9, 10). Transcript of the Minimum Competency Testing Clarification Hearings prepared by Alderson Reporting Co, Washington, DC.

National PTA. (2006). *Recess is at risk, New campaign comes to the rescue*. Retrieved April 4, 2006, from http://www.pta.org/ne_press_release_detail_1142028998890.html

National Research Council. (2001). *Knowing what students know*. Washington, DC: National Academy Press.

National Research Council. (2007). *Lessons learned about testing: Ten years of work at the National Research Council*. Washington, DC: National Research Council.

O'Dwyer, L., & Miranda, H. (2009). Diagnosing students' misconceptions in algebra: Results from an experimental pilot study. *Behavioral Research Methods 41*(2), 414–424.

Pedulla, J., Abrams, L., Madaus, G. F., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from anational survey of teachers*. Chestnut Hill, MA: National Board on Educational Testing and Public Policy, Boston College Retrieved from http://www.bc.edu/research/nbetpp/.

Postman, N. (1992). *Technoloply: the surrender of culture to technology*. New York, NY: Alfred A. Knopf.

Rapple, B. (2004). Standardized testing in America's schools: Lessons from Matthew Arnold's Britain. *Contemporary Review, 285*(1665), 193–198.

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform toclose the Black-White achievement gap*. New York, NY: Teachers College Press.

Russell, M., & Abrams, L. (2004). Instructional uses of computers for writing: How some teachers alter instructional practices in response to state testing. *Teachers College Record*,*106*(6), 1332–1357.

Shepard, L. A., & Smith, M. L. (1988). Flunking kindergarten: Escalating curriculum leaves many behind. *American Educator, 12*(2), 34–39.

Smith, E., & Tyler, R. (1942). *Appraising and recording student progress*. New York, NY:Harper & Row.

Staudenmaier, J. M. (1985). *Technology's storytellers: Reweaving the human fabric*. Cambridge,MA: MIT Press.

Staudenmaier, J. M. (1988). Technology and faith (audio cassette). Kansas City, MO: Credence Cassettes.

Taylor, G., Shepard, L. A., Kinner, F., & Rosenthal, J. (2003). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost.* Los Angeles, CA: University of California, Center for Research on Evaluation, Standards, and Student Testing (CSE Technical Report 588).

Webb, F. R., Covington, M. V., & Guthrie, J. W. (1993). Carrots and sticks: Can school policy influence student motivation? In T. M. Tomlinson (Ed.), *Motivating students to learn: Overcoming barriers to high achievement* (pp. 99–124). Berkeley, CA: McCutchan.

Winner, L. (1977). *Autonomous technology: Technic-out-of-control as a theme in political thought*. Cambridge, MA: MIT Press.

Authors' Note

Space does not allow us to provide evidence from all sources. For more details, see Madaus,G.,Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: Information Age Publishing.

George Madaus is Boisi Professor Emeritus at Boston College, Lynch School of Education. Professor Madaus can be reached at 323 Campion Hall, 140 Commonwealth Avenue, Chestnut Hill, MA 02467; madaus@bc.edu.

Michael Russell is Associate Professor at Boston College, Lynch School of Education. Professor Russell can be reached at 323 Campion Hall, 140 Commonwealth Avenue, Chestnut Hill, MA 02467; russelmh@bc.edu.