The Adverse Impact of High-Stakes Testing on Minority Students
2001

George Madaus and Marguerite Clark

Research Question

How do current high-stakes testing policies impact minority students and others who are traditionally underserved by the educational system?

Major Findings

- High-stakes, high-standards tests do not have a markedly positive effect on teaching and learning in the classroom.
- The higher the stakes, the more likely corruption will occur.
- Curriculum is narrowed as teachers teach to the test.
- The higher the number of minority students, the more test preparation takes over the curriculum.
- High-stakes do not motivate the unmotivated.
- High-stakes testing programs increase high school drop out rates, especially in minority student populations.

Policy Implications

Remove the high-stakes label and use multiple methods to assess students. Testing practices need to be closely monitored to ensure cheating and other negative practices are avoided.

Methods

This report is based on research done at Boston College over the past 30 years.

CHAPTER FIVE

The Adverse Impact of High-Stakes Testing on Minority Students

Evidence from One Hundred Years of Test Data

GEORGE MADAUS AND MARGUERITE CLARKE

It's a bull market for high-stakes testing programs in education, far surpassing the bull market days of minimum competency testing of the early 1970s. Now they are called assessments, not tests. However, while their look and feel may have changed, deep down the underlying technology has not, and the same issues about their impact and effects persist. The range of high-stakes testing programs is expansive: from readiness testing for entrance to kindergarten; to tests required for promotion and graduation; to accountability testing for teachers, schools, and districts; to testing teachers for certification. These high-stakes testing programs will not go away; if anything, they will become more important as policy tools and societal signaling devices. For example, policymakers in several states are setting high standards on state exams to deliver a "wake-up call" to students, teachers, parents, and the public to what they portray as a generalized crisis, deterioration, and failure of public education.[1]

Advocates of the standards movement that is now sweeping the country take for granted that high standards on high-stakes assessments—in current parlance, assessments "worth teaching to"—can reduce adverse impacts in test performance by influencing what is taught and learned. Paradoxically, they are both right and wrong. They are right because high-stakes tests do influence which and how things are taught and learned. The history of high-stakes testing over the last five centuries is testimony to that truism: test scores will go up. However, as

teaching turns into test preparation, test results cease to reflect what examinees really know or can do. Therefore, it is wrong to believe that we can test our way out of our educational problems. Quite the opposite is true: our fixation on test results deflects attention from fundamental educational problems and thus hinders reform.[2]

In this chapter, we examine four aspects of current high-stakes testing that impact minority students and others who are traditionally poorly served by the educational system. Our analysis is based primarily on research done at Boston College over the past thirty years. After reviewing the available evidence, we come to the following conclusions:

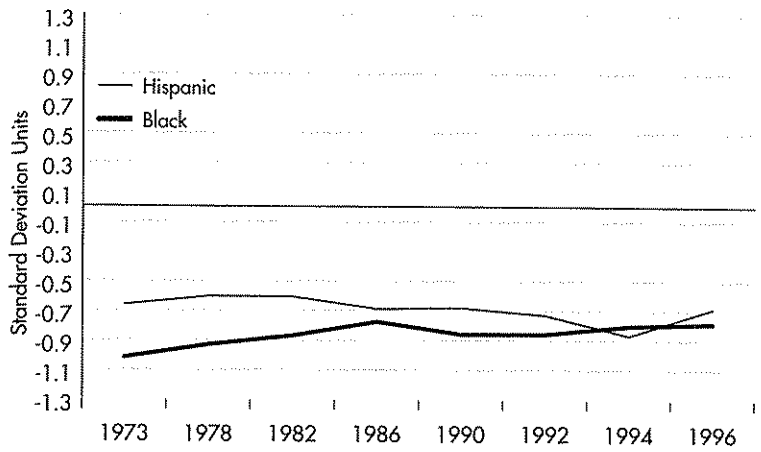
- high-stakes, high-standards tests do not have a markedly positive effect on teaching and learning in the classroom;
- high-stakes tests do not motivate the unmotivated;
- contrary to popular belief, "authentic" forms of high-stakes assessments are not a more equitable way to assess the progress of students who differ by race, culture, native language, or gender;[3]
- high-stakes testing programs have been shown to increase high school dropout rates, particularly among minority student populations.

Exploring the Achievement Gap

Before developing these points, it is worth looking at some of the trend data available on the test performance of minority students since the 1970s. Results from several studies, including the National Assessment of Educational Progress (NAEP), the National Education Longitudinal Study (NELS), and the SAT-I (formerly known as the Scholastic Assessment Test), indicate that while the achievement gap between black and Hispanic students and their white counterparts has been generally narrowing over the years, it is still "a substantial" one.[4]

For example, results from the 1996 trend NAEP provide evidence that averages for nine-, thirteen-, and seventeen-year-old black and Hispanic students in mathematics were generally higher than those for their counterparts in 1973 (see Figures 5.1–5.3, pages 87–89). In addition,

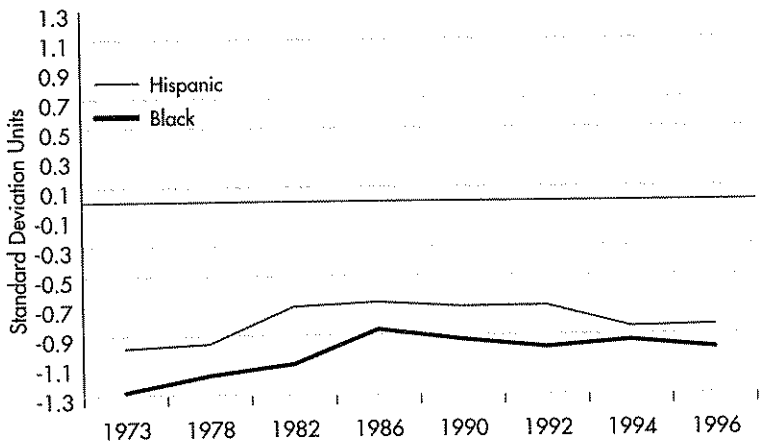
Figure 5.1. Trends in Racial/Ethnic Group Differences on the NAEP Mathematics Test: Nine-year-olds



Note: The performance of white students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse.
Sources: "Weighted Means, Standard Deviations, and Percentiles of Mathematics Distributions with Jackknifed Standard Errors," Educational Testing Service, Princeton, N.J., undated and unpublished tabulations; John Mazzeo, Educational Testing Service, personal communication, September 1997; Jay R. Campbell, Kristin E. Voelkl, and Patricia L. Donahue, *NAEP 1996 Trends in Academic Progress* (Washington, D.C.: National Center for Education Statistics, 1997).

gains made by these groups were generally larger than those made by white students.[5] However, tempering this positive trend is the realization that the average proficiency for white thirteen-year-olds as measured on the 1996 NAEP scale was around the level achieved by black seventeen-year-olds.[6] In addition, data from the main NAEP for 1996 indicate that at grades 4, 8, and 12, achievement gaps between white students and their black and Hispanic counterparts ranged from 0.8 to 1.1 standard deviation units. To put this in context, a difference of 1 standard deviation implies that just 16 percent of the low-achieving group perform at levels exceeded by 50 percent of the higher achieving group.

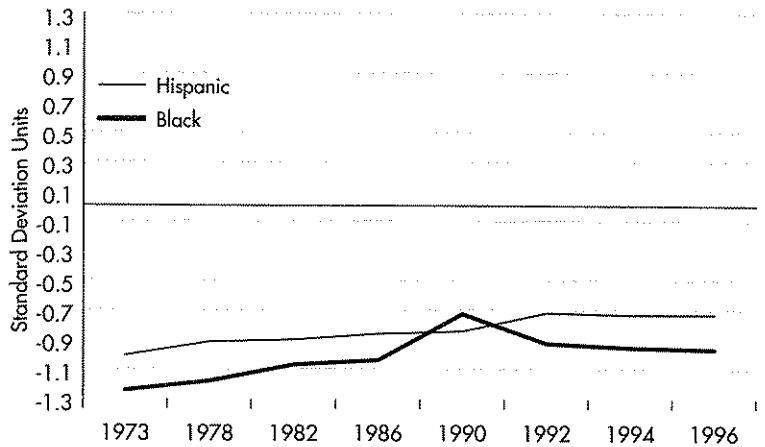
FIGURE 5.2. TRENDS IN RACIAL/ETHNIC GROUP DIFFERENCES ON THE NAEP MATHEMATICS TEST: THIRTEEN-YEAR-OLDS



Note: The performance of white students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse.
Sources: "Weighted Means, Standard Deviations, and Percentiles of Mathematics Distributions with Jackknifed Standard Errors," Educational Testing Service, Princeton, N.J., undated and unpublished tabulations; John Mazzeo, Educational Testing Service, personal communication, September 1997; Jay R. Campbell, Kristin E. Voelkl, and Patricia L. Donahue, *NAEP 1996 Trends in Academic Progress* (Washington, D.C.: National Center for Education Statistics, 1997).

Similar outcomes for racial differences in mathematics scores are found on longitudinal studies such as the NELS and High School and Beyond (see Figure 5.4, page 90) and on college admissions tests such as the ACT (formerly known as the American College Test) and the SAT-I. As the trend lines in Figures 5.5 and 5.6 indicate (pages 91–92), Asian American students were the highest performers on both the ACT assessment and the SAT-I from 1977 to 1997. White students, the second-highest performers, achieved at average levels that were about 0.3 of a standard deviation unit lower than Asian students. The average performance of black students was consistently lower than those of all other groups and stood well below those of Asian and white students in 1997. Nevertheless, it is important to emphasize that

FIGURE 5.3. TRENDS IN RACIAL/ETHNIC GROUP DIFFERENCES ON THE NAEP MATHEMATICS TEST: SEVENTEEN-YEAR-OLDS

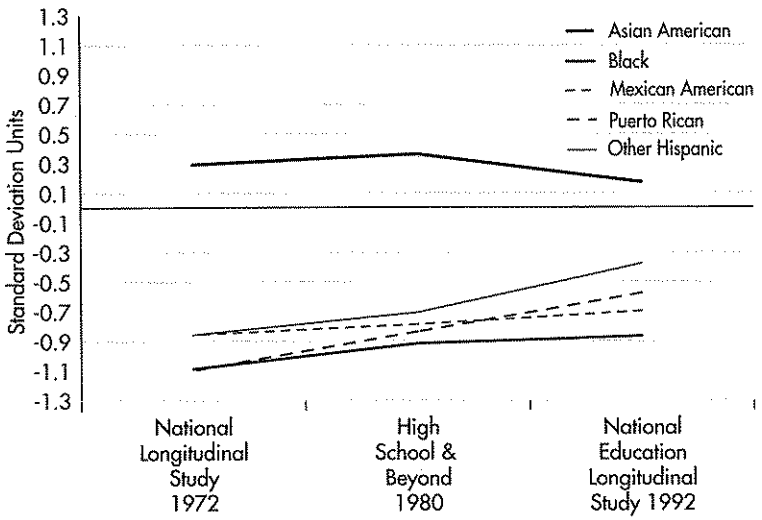


Note: The performance of white students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse.
Sources: "Weighted Means, Standard Deviations, and Percentiles of Mathematics Distributions with Jackknifed Standard Errors," Educational Testing Service, Princeton, N.J., undated and unpublished tabulations; John Mazzeo, Educational Testing Service, personal communication, September 1997; Jay R. Campbell, Kristin E. Voelkl, and Patricia L. Donahue, *NAEP 1996 Trends in Academic Progress* (Washington, D.C.: National Center for Education Statistics, 1997).

while all groups improved their performances from the mid-1970s on, some of the largest improvements on average were made by blacks. The improved performance of Mexican American students, especially on the ACT assessment, is also notable. An important finding in a study by Gail T. McLure, Anji Sun, and Michael J. Valiga of ACT scores between 1987 and 1996 was that the improved performance of minority groups in general on the ACT mathematics test coincided with a pattern of increases in the number of courses taken by students from these groups.[7]

Explanations for these achievement gaps have run the full gamut. A lot of attention has been given to the tests themselves and possible flaws therein. However, in today's high-stakes testing context, it is becoming

FIGURE 5.4. TRENDS IN RACIAL/ETHNIC GROUP DIFFERENCES ON THE NATIONAL LONGITUNDINAL STUDY, HIGH SCHOOL AND BEYOND, AND NATIONAL EDUCATION LONGITUDINAL STUDY MATHEMATICS TESTS: HIGH SCHOOL SENIORS

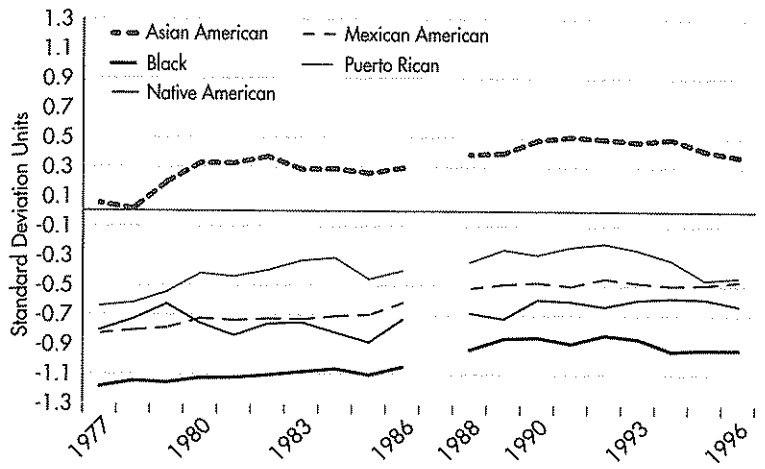


Note: The performance of white students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse.
Source: U.S. Department of Education, National Center for Education Statistics, "Trends among High School Seniors, 1972–1992," NCES 95-380 (Washington, D.C.: U.S. Government Printing Office, 1995).

harder and harder to blame the tests for large performance differences among groups—what the courts call "adverse impact." Today, there is a realization that differential results can reflect genuine group differences in whatever trait is being assessed.[8] These tests—even when faulty—are signaling that something is wrong. Sturm and Guinier's analogy of test results being the miner's canary is apt.[9] Further, there will be no breakthrough in the technology of testing to ameliorate the adverse impact we see on cognitive measures. Technically we can only tinker at the margins.

One caveat about the malleability of testing technology: a major weakness of the high-stakes, high-standards assessment movement is that the standard-setting process increasingly results in cut scores—the scores

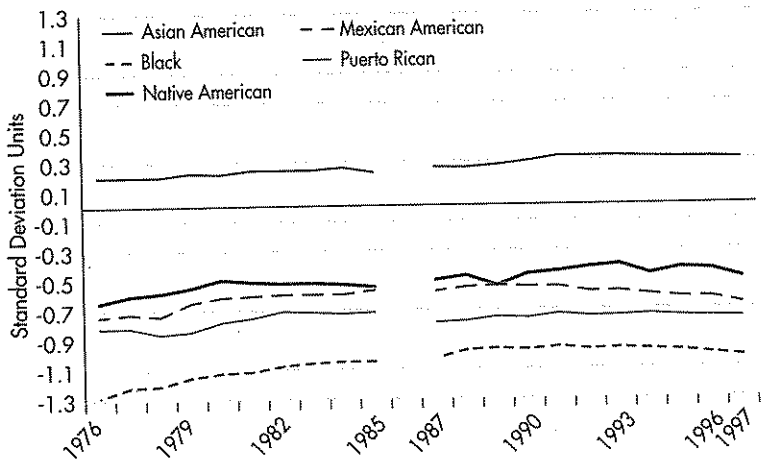
FIGURE 5.5. TRENDS IN RACIAL/ETHNIC GROUP DIFFERENCES ON THE ACT MATHEMATICS TEST: HIGH SCHOOL SENIORS



Note: The performance of white students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse. Data were not available for 1987.
Sources: "ACT Math Scores and ACT Score Means and SDs for Successive Years of ACT-Tested College-Bound Seniors 10 percent National Sample," American College Testing, Iowa City, Iowa, undated and unpublished tabulations; James Maxey, American College Testing, personal communications, August–September 1997.

that serve to demarcate passing from failing, or from one achievement level to the next—that are very high relative to current distributions of scores on standardized achievement tests. For example, according to the four cut scores used by the NAEP, the percentages of students at the advanced mathematics level never exceed 4 percent. Yet, independent achievement data from the SAT-I, ACT, and advanced placement tests in mathematics call into question the NAEP percentages.[10] Thus, the NAEP's description of national math attainment is a classic ipse dixit. The bleak picture is valid only so long as one accepts the cut scores as definitive. Another example is the Massachusetts Comprehensive Assessment System (MCAS), which results in large numbers of students falling into the "failure" and "needs improvement" categories

FIGURE 5.6. TRENDS IN RACIAL/ETHNIC GROUP DIFFERENCES ON THE SAT-I MATHEMATICS TEST: HIGH SCHOOL SENIORS



Note: The performance of white students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse. Data were not available for 1986.
Source: College Bound Seniors: National Report (New York: College Entrance and Examination Board, 1972–1997).

despite contradictory, standardized, norm-referenced data. In short, many policymakers fail to recognize that the arbitrary nature of cut-score methodologies cries out for the external validation of the categories produced and used to label students.

There are three additional problems with very high cut scores.[11] First, they result in schools and parents getting essentially no information other than "failure," "inadequate," or "needs improvement" for large numbers of children. Second, they provide a misleading view of change. Large changes that stay within one of the levels go unnoticed, and trivial changes that cross a cut-score boundary are treated as important. Finally, the school systems often set goals for overall or annual improvement that are simply not possible to meet by legitimate means. This gives teachers an additional incentive to teach to the test in inappropriate ways.

THE IMPACT OF HIGH-STAKES TESTING ON TEACHING AND LEARNING

Contemporary policymakers who advocate the use of tests as levers of educational reform certainly recognize the historical role of testing in controlling what is taught and learned. There seems to be little argument that tests affect the curriculum. Indeed, the power of an examination to shape what is taught and learned was noted at least as far back as the sixteenth century, when Philip Melanchthon, a Protestant, German teacher, wrote in *De Studiis Adolescentum*, "no academical exercise can be more useful than that of examination. It whets the desire for learning, it enhances the solicitude of study while it animates the attention to whatever is taught."[12]

The concept of the power of an important test is beautifully captured by Chief Inspector of Schools Edmond Holmes, writing about nineteenth-century school examinations in Great Britain. Victorian style apart, Holmes' observation remains true today for the United States:

> Whenever the outward standard of reality (examination results) has established itself at the expense of the inward, the ease with which worth (or what passes for such) can be measured is ever tending to become in itself the chief, if not sole, measure of worth. And in proportion, as we tend to value the results of education for their measurableness, so we tend to undervalue and at last ignore those results which are too intrinsically valuable to be measured.[13]

How do important tests exert such an influence? What are the mechanisms that vest high-stakes tests with their power to change instruction and learning? Four principles explain the importance of such tests on what is taught and learned; the latter three are special cases of the first as played out in education.[14]

1. The more any quantitative social indicator is used for social decisionmaking, the more likely it will be to distort and corrupt the social process it is intended to monitor. This very general principle comes from the work of Don Campbell and is a social version of Heisenberg's uncertainty principle.[15]

2. One of the necessary conditions for measurement-driven instruction to work is that valued rewards or serious sanctions are perceived to

be triggered by test performance. If teachers perceive that important decisions are related to the test results, they will teach to the test.[16]

3. When test stakes are high, past exams come to define the curriculum. Once a high-stakes testing program has been in place for several years, teachers see the kind of intellectual activity required by the previous test questions and prepare students to meet these demands.[17]

4. When teaching to the test, teachers pay attention to the form of the test as well as the content. When this occurs, the form of the questions can narrow the focus of instruction, study, and learning to the detriment of other skills.[18]

Today, advocates of "authentic" assessment (that is, assessment that requires active engagement, such as producing extended responses, generating material for portfolios, or carrying out experiments) presume that such assessments are somehow outside the purview of the latter three principles. Not so. A powerful illustration that these supply-type exams—exams that require students to produce their own answers to a question rather than choose from a set of possible answers—are also likely to distort instruction comes from samples of student essays from an Irish examination of the mid-1940s, the Primary Leaving Certificate.[19]

Box 5.1 illustrates how students learned to memorize stock responses that would be adaptable to any writing prompt. Note the similarity between the answers to three different writing prompts from three different students across three different years. Scores from such tests said more about students' memories and test-taking strategies than they did about students' ability to write.[20]

How do the four principles about the impact of tests on curriculum specifically affect minority students? A 1992 national study, sponsored by the National Science Foundation, of the effects of high-stakes tests—primarily standardized, multiple-choice tests—on math and science instruction found that students in classrooms with a high percentage of minority students are affected significantly more by such tests than are their peers in classrooms with few minority students.[21] Teachers with more than 60 percent minority students in their class, compared to teachers with less than 10 percent minority students, reported more reliance on mandated standardized tests for various uses, more test pressure, and more test preparation and influence on instruction. They more often reported that test scores were "very important" or "extremely important" either to themselves or

Box 5.1.

A Bicycle Ride (1946)

I awakened early, jumped out of bed and had a quick breakfast. My friend, Mary Quant, was coming to our house at nine o'clock as we were going for a long bicycle ride together.

It was a lovely morning. White fleecy clouds floated in the clear blue sky and the sun was shining. As we cycled over Castlemore bridge we could hear the babble of the clear stream beneath us. Away to our right we could see the brilliant flowers in Mrs. Casey's garden. Early summer roses grew all over the pergola which stood in the middle of the garden.

A Day in the Bog (1947)

I awakened early and jumped out of bed. I wanted to be ready at nine o'clock when my friend, Sadie, was coming to our house. Daddy said he would take us with him to the bog if the day was good.

It was a lovely morning. White fleecy clouds floated in the clear blue sky. As we were going over Castlemore bridge in the horse and cart, we could hear the babble of the clear stream beneath us. Away to our right we could see the brilliant flowers in Mrs. Casey's garden. Early summer roses grew all over the pergola which stood in the middle of the garden.

A Bus Tour (1948)

I awakened early and sprang out of bed. I wanted to be ready in good time for our bus tour from the school. My friend, Nora Greene, was going to call for me at half-past eight as the tour was starting at nine.

It was a lovely morning. White fleecy clouds floated in the clear blue sky and the sun was shining. As we drive over Castlemore bridge we could hear the babble of the clear stream beneath us. From the bus window we could see Mrs. Casey's garden. Early summer roses grew all over the pergola which stood in the middle of the garden.

Source: Primary Leaving Certificate, Irish examination answer papers, Department of Education, Dublin, Ireland.

to administrators for placing students in special services, determining graduation, recommending textbooks, planning curriculum and instruction, evaluating student progress, and giving feedback to students.

About 75 percent of both math and science teachers with classes that had high percentages of minority students reported pressure from their district to improve standardized test scores, in comparison to about 60 percent of teachers with classes that had low percentages. Teachers in classrooms with high percentages of minority students significantly more often reported teaching test-taking skills, teaching topics known to be on the test, increasing emphasis on tested topics, beginning preparation more than a month before the test, and including topics not otherwise taught. Case studies conducted in six urban districts with large minority enrollments confirmed the finding of the national survey that instruction by teachers facing high-stakes testing pressure is often heavily oriented toward test preparation. There is little question that high-stakes tests have greater consequences for minority and poor children than they do for majority and more affluent students, although the latter also feel the impact of these tests.

The Impact of High-Stakes Testing on Student Motivation

The idea that mandating a high-stakes assessment will improve students' motivation to learn is one that goes back at least to the eighteenth century.[22] A recent monograph sponsored by the American Educational Research Association examined the claim that high-stakes assessments will move obstinate, dispirited, lazy, or recalcitrant students to try harder in school.[23] Three corollaries associated with contemporary assertions about the motivational force of high-stakes testing emerged.[24] First, the greater the reward offered (or the more noxious the consequences of not complying), the harder students will try. Second, the meaning of rewards and punishments is essentially the same for all students (including poor, middle class, and minority). Third, student arousal is maximized when rewards are distributed on a competitive basis.[25]

Unfortunately, these assumptions about motivation are flawed. The concept of motivation used in claims about the power of examinations lacks definition and clarity.[26] The complexity of the construct, its variety of meanings, and the different mechanisms that elicit it are ignored in blanket claims about the examinations' motivational power. Investigators in cognitive psychology, while identifying goals as central to the motivational

process, point to the complexity of the relationships and processes involved, the idiosyncrasy of individuals' choices of goals and subgoals, and differences between students in their assessment of their ability and self-efficacy.[27] There seems to be little appreciation among reformers that motivation may be even more complex in the context of external (that is, imposed from outside the school), high-stakes, national examinations, embedded as they are in complex school, cultural, and social networks. In other words, advocates of the motivational potential of examinations have not paid enough attention to who will be motivated and who will not. This point is particularly relevant when the examinations are referenced to "world-class" standards that all students, regardless of grade level, circumstances, context, and individual differences, are expected to attain.

Furthermore, to become motivated students must first see that striving for the rewards attached to examination performance is not only important in their lives but also realistic.[28] In practice, however, some students immediately dismiss the examination because they feel they lack the ability to do what is necessary to pass. Others, while believing they may have the ability to pass, are not motivated to work toward examination success because they do not see the test credential necessarily resulting in jobs or college because of scarcity, competition, or lack of relevance in their social setting. Still others perceive not only that they lack the ability, but that the rewards in reality are illusory even for many who might pass.

In addition, there is no reason to believe that the motivational power of examinations will be the same at all grade levels. Since present reform proposals call for testing at elementary and middle grades, we need to consider how examinations over the course of many years will influence the behavior of students at these age levels.

There can be no doubt that some students in other countries do indeed work hard to pass high-stakes external examinations. Further, many pay considerable sums of money to attend test-preparation schools, and some, no doubt, internalize the competitive values embedded in the examinations. It may even be true that some students develop genuine, intrinsically motivated behavior or, at any rate, relatively autonomous forms of extrinsic motivation. However, experience with external examinations indicates that all too many students focus their efforts on mastering strategies to help them over the examination hurdle rather than on developing mastery of subject matter and honing lasting competencies.

This consequence does not seem to have been anticipated by some testing proponents in the United States. For example, a laudable objective

of the Learning Research and Development Center and the National Center on Education and the Economy is that their proposed examination system would lead students to see that school is a place to learn and not just a place to be labeled as smarter or slower than other students.[29] While implicitly recognizing the distinction between competence and performance, there does not seem to be an appreciation that high-stakes external examinations are likely to thwart rather than support the attainment of this objective. In his description of a cramming (or test preparation) school, Wilfred Sheed offers an amusing insight into the effects of focusing on performance rather than on learning in the context of external examinations. The crammer, Jenkins Tutorial Establishment in London, offered

> successful examination results as it might a forged passport [bypassing] education altogether. Their only texts were examination papers—all the relevant ones set in the last fifty years, with odds of repetition calculated and noted as in the Racing Form. Within six months, I was able to pass London matriculation without knowing any of the subjects involved; and by applying Jenkins' method later, to pass every exam that ever came my way afterwards. Hence I remain a profoundly uneducated man.[30]

There are further consequences of external examinations.[31] For example, many teachers, under pressure to help students secure good examination results, will be more controlling in their teaching. In fact, the "rhetoric from Washington continues to advocate greater accountability, greater discipline, and increased use of standardized testing, all of which are means of exerting greater pressure and control on the educational process."[32] The fact that examinations are bureaucratically controlled from the top down, rather than under the control of practitioners, parents, and the local community moving from the bottom up, has implications for adopting and pursuing goals. Further, insofar as it involves increased control of teacher and student behavior through the imposition of rewards and sanctions associated with test performance, other undesirable consequences are possible. When controlling events are perceived to determine behavior, students' need for competence, self-determination, conceptual learning, and creativity will not be met, but rather diminished.[33]

Examinations also can increase competition and cheating on the part of some students. We argued above that situations that emphasize test performance rather than learning per se narrow the curriculum to what

is embodied in the tradition of past examinations. We therefore expect the pursuit of high-stakes examination performance in time to corrupt the examination, resulting in invalid inferences about achievement based on performance.

Another important consideration is that students who may not be motivated to pursue examination success for sociocultural reasons or on the basis of estimates of their present levels of achievement or ability are likely to become alienated from both the examinations and the whole educational process.[34] The fact that around 10 percent of students, mostly from disadvantaged backgrounds, avoid taking any public examination before leaving school is regarded as a serious problem in several European countries. In America, research on dropouts also shows that many, while fully appreciating the importance of educational credentials, do not believe that such credentials are of much help in their social milieu. The motivational argument of proponents of high-stakes national examinations does not address these realities. We will examine the dropout issue later in this chapter. Suffice it to say here that motivation and dropout issues are intertwined.

Many of the reform arguments relating to motivation in education have their origins in the world of industry. However, the use of external motivation techniques in industry has little relevance to the high-stakes assessment situation in education. In industry, skills have already been acquired, and rewards are tangible and immediate and serve to reinforce and direct behavior. The feedback mechanisms in industry are also well developed and more immediate. None of these conditions hold for high-standards, high-stakes assessments. Other high-stakes exam situations—no-pass no-play or no-pass no-drive programs, for example—do not seem to motivate very well, despite the fact that they are more immediate and real for many students than are high-stakes school assessments.[35]

These considerations lead to the conclusion that it is incumbent on reformers to weigh more carefully the costs and benefits that are likely to be associated with the variety of outcomes of examination-induced motivation.[36] While there are possible positive aspects, such as the specification of clear goals and standards for teachers and students, greater consideration needs to be given to the nature of those goals and standards and how they might need to vary for different students. Other important issues also should be addressed, including the probability that many students will not be motivated at all. Moreover, for many who are motivated, the high stakes associated with the assessment may work toward focusing their efforts on improving their test performance rather than on

the more demanding job of developing general competence, higher-order thinking skills, problem-solving ability, and creativity. It is these latter traits that reformers claim will be the eventual outcome of standards-based reform. Yet, until such issues are addressed, we can have little confidence in high-stakes assessments as a panacea for the ills of American education.

The Impact of Authentic Assessments on the Performance of Students Who Differ in Race, Culture, Native Language, or Gender

An outgrowth of the high-stakes, high-standards movement is the belief that authentic assessments are more equitable for assessing the progress of students who differ in race, culture, native language, or gender.[37] This assumption, however, is questionable in light of data from a high-stakes authentic assessment implemented by the United Kingdom in 1992.[38]

These data permitted an examination of the relative performance of gender, linguistic, low-income, and special needs groups on the authentic assessments administered to seven-year-olds as part of the national curriculum.[39] No matter what the method of authentic assessment, once all other factors had been taken into account there were substantial differences between particular groups of students. Low-income students, students whose first language was not English, and special needs students performed at significantly lower levels than other students on English, mathematics, and science assessments. Gender was the only factor that varied in impact across the three subjects. In English and mathematics, girls performed at a significantly higher level than boys, although the difference in mathematics was small. However, in science there was no significant difference between boys and girls.

As would be expected, older students performed better than younger students on the United Kingdom exams. The United Kingdom data appear to validate the trend recognized by the United States in the 1980s—that improved high-stakes test performance was related to age. However, this fact, along with concerns about improving upper-level grade performance, created controversial incentives and practices at the school level in the United States during this period (see Chapter 8 by Robert Hauser in this volume). For example, one tactic used by schools to improve reading scores in upper primary grades was to put pressure on teachers to emphasize reading skills more strongly. This emphasis, however, came at the

expense of other, more traditional goals for children in kindergarten through third grade. A related tactic to improve upper-level performance was that of "red shirting" kindergarten students, or holding them back a grade. Educators realized that retaining children in kindergarten, or not letting them enter in the first place if they did not have the necessary "readiness," would ultimately contribute to higher test scores. If for no other reason, students would perform better because they would be a year older when they took the tests.[40]

The United Kingdom results are from the early stages of a new program, from a different country and educational system, and from seven-year-olds only. Nonetheless, at the very least, they point to the need to monitor carefully group differences in light of positive equity claims made about authentic assessment technology. These results are cautionary to those making such claims.

The Impact of High-Stakes Assessments on High School Dropout Rates

One aspect of high-standards, high-stakes assessments that needs considerably more attention is the relationship between such assessments and dropping out of high school. Five suggestive lines of evidence argue for a careful examination of this relationship.

The first intriguing data come from the minimum competency testing (MCT) era. An overall, albeit quite crude, view of the relationship between MCT and dropout rates comes from examining the ten states with the highest dropout rates in 1986 and the ten states with the lowest dropout rates that year.[41] Correlational data indicate a strong relationship between attrition, or dropout, rates and the existence of MCT programs in these states.[42] Five of the ten states with the lowest dropout rates had no minimum competency testing programs. The other five states with low dropout rates had MCT programs that could be characterized as involving relatively low stakes: four used the tests for decisions about remediation, and only one used them for accountability; none required the tests for critical decisions about graduation or grade promotion. Furthermore, in three of these five states, local, not state, education agencies set the standards.

States with the highest dropout rates, on the other hand, had MCT programs for which standards were set, at least in part, at the state level. Nine of the ten used the tests in decisions about high school graduation; four used them in decisions about promotion. In sum, the ten states with

the highest dropout rates employed minimum competency tests with higher stakes and less flexible standards than the states with the lowest dropout rates.

These data are not evidence of a causal relationship between high-stakes MCT programs and dropout rates. The states with the highest dropout rates differed in other important ways from the states with the lowest dropout rates. The latter were largely in the West and Midwest, and they had a relatively low proportion of minority and poor students among their school-age populations. Perhaps high dropout rates are symptoms of the educational system's failure, which spurred legislators to mandate MCT programs in the first place, or perhaps MCT does contribute in some way to the dropout problem.[43] In any case, crude as these data may have been, they underscore the need to explore further competing hypotheses about the connection, or lack of connection, between high-stakes testing and dropping out.

A second line of evidence comes from a correlational study that examined the relationship between MCT in eighth grade and early high school dropout patterns.[44] The results suggest that in schools with high proportions of students from families with low socioeconomic status, MCTs are linked to higher dropout rates. The dropout rates from these schools are two to six percentage points higher, on average, than from similar schools with no such requirement. The overall conclusion is, "It is the concentrated poverty of these schools and their communities, and their concomitant lack of resources, that link MCT policies to higher dropout rates, rather than other risk factors, such as student grades, age, attendance, and minority group membership."[45]

A third, more recent line of evidence on the connection between dropping out and high-stakes assessments, especially for minorities, comes from an analysis of data from the Texas Assessment of Academic Skills (TAAS).[46] One study suggested that the TAAS requirement was related to over 25,344 black and Hispanic students of Texas's 1993 sophomore cohort dropping out of school. For white students, the dropout figure was 14,809. Black and Hispanic students dropped out at a significantly higher rate than did white students even when accounting for socioeconomic status,[47] academic track, language program participation, and school quality.

While these data apply to one particular cohort, they fit well with patterns Walt Haney observed for Texas enrollment rates over a twenty-year period. After breaking down enrollment rates by ethnicity, Haney found that the grade 12/grade 9 enrollment ratio has dipped considerably for both black and Hispanic students since about 1990 but has remained quite stable for white students. For example, in 1989 (the year prior to implementation of the TAAS), the grade 12/grade 9 enrollment ratio for whites was around 0.76. Even after the TAAS was implemented in the 1990–91 school year, this ratio never dipped below 0.73. In 1989, the grade 12/grade 9 enrollment ratio for blacks was quite close to that for whites, at around 0.74. However, since the implementation of the TAAS this ratio has dipped dramatically, falling to a low of about 0.55 in 1994 and hovering around 0.57 in 1997. Hispanic students also experienced a similar drop in their enrollment ratio during this period. Again, we cannot argue causation from these correlational data, but these findings call out for more detailed interviews with a carefully chosen, random selection of dropouts to gauge the actual impact of the TAAS on their decisions to leave school.

A fourth line of evidence on the association between high-stakes tests and dropping out comes from work on the relationship between grade retention, being overage for one's grade, and dropout rates. Research on the effects of grade retention—whether or not it is coupled with high-stakes testing—has generally concluded that its detrimental effects outweigh any purported benefits.[48] In particular, research has found that being overage for their grade as a result of having been retained in an earlier grade eats away at students' sense of efficacy, with the impact especially severe for black students. Compared to on-grade students, these overage students are twice as likely to be retained again.[49] Many of these students ultimately become disengaged and drop out; and many, under the pressure of high-stakes testing, drop out earlier in their school careers than otherwise would be expected. In fact, being overage for one's grade is a better predictor of dropping out than below-average test scores.[50]

These findings were played out fifty years ago in Ireland in response to a mandatory primary school leaving certificate examination that was administered to all sixth-grade pupils between 1943 and 1967.[51] Teachers were employing a policy of not promoting weaker pupils in order to control the potential failure rate on the examination. This policy tended to take place at two points in the system—moving from grade 3 to grade 4, and from grade 5 to grade 6. A pupil that was held back at one or both points would be eligible to leave school before he or she reached the sixth grade—reducing the number of these overage students who ever sat for the primary examination.

A final piece of related—but little known—evidence about dropouts in the United States comes from data from the Third International Mathematics and Science Study (TIMSS). The data indicate that about 35 percent of the senior cohort was no longer enrolled in school when the

TIMSS tests were administered in the spring of 1995.[52] Further, a post hoc check revealed that the NAEP data on dropouts are quite consistent with the TIMSS data.[53] The "no longer enrolled" figure for TIMSS is quite a bit higher than the National Center for Education Statistics (NCES) dropout rate of 12 percent for the United States for 1995, but one possible explanation is that many students who are enrolled in October, the month used in calculating the NCES rate, were no longer in school by April/May when the TIMSS tests were administered. The higher rate of missing seniors revealed by the TIMSS and NAEP data comes at a time when all but one state has a state testing program, and when the state tests are increasingly tied to graduation. The TAAS results discussed above are a case in point. Once again, this conjecture is correlational, not causal, but it points to the need for a more carefully designed study to gauge the impact of high-standards, high-stakes assessments on decisions to drop out of school.

Conclusion

Testing is a powerful, but often blunt, tool. Like a medication, it may fail or have diverse, unintended negative consequences. A testing program, for example, may unfairly or unreasonably deny opportunities to certain classes of people. It may reward or punish the wrong individuals or institutions, or it may undermine the performance of institutions it is intended to strengthen. Further, the four issues described above—the impact of testing on teaching and learning, the motivational power of high-stakes assessments, claims that authentic assessments are fairer to minority students than traditional standardized multiple-choice testing, and the role of high-stakes assessments in decisions to drop out of school—all need to be carefully monitored. Policymakers and test users have been able to turn to extensive commercial, not-for-profit, and governmental infrastructures that have evolved over the past ninety years to assist them in test development, administration, scoring, and reporting. However, there has been no analogous infrastructure for independently evaluating a testing program before or after implementation or for monitoring test use and impact.

In its 1990 report *From Gatekeeper to Gateway: Transforming Testing in America*, the National Commission on Testing and Public Policy recognized "the need for sound, fair, and reasonably efficient mechanisms to help make difficult decisions about individuals and institutions."[54] It also noted that "although tests have become important instruments of

public accountability, there are few mechanisms to *audit* or *appraise* the quality of publicly sponsored tests, to monitor their use as instruments of social policy, and to assess their impact on individuals, groups, and institutions."[55] To remedy this, the commission called for "the development of *additional institutional means* to examine the quality of tests and assessment instruments and to provide oversight of test use."[56]

Since 1990, a variety of factors, including changes in statute and case law, national and state policy initiatives, and plebiscites, have both transformed educational testing and increased the need for oversight. All have altered the uses of tests, in some cases fundamentally, and all have the potential to affect profoundly both educational systems and individuals, particularly students currently ill-served by the educational system. These changes also raise difficult equity and technical questions, placing policymakers, educators, and even researchers in uncharted territory. Perhaps at no time in the past half century have questions and consequences surrounding educational testing been as widespread and as serious.

In many other areas where technology and policy intersect, the public insists on oversight to protect individuals from unintended negative effects. For example, faced with the policy decision to introduce a major new, untried medical technology to millions of children, the public would ask about the safety, efficacy, quality, and social and economic effects of the new technology or treatment. Public agencies have been established to address such concerns systematically. The effects of testing are now so diverse, widespread, and serious that it is necessary to establish similar mechanisms for catalyzing inquiry about, and systematic independent scrutiny of, them.

In response to the need for monitoring high-stakes educational testing, in September 1998 the Ford Foundation provided start-up funding for an independent, institutional oversight agency called the National Board on Educational Testing and Public Policy. Housed in the Lynch School of Education at Boston College, the board will review testing programs and encourage close consideration of the diverse uses of testing in education. It will be a permanent institutional entity, as opposed to a transitory commission that studies a situation and then issues a report and recommendations that are easily ignored and then soon forgotten.[57]

Few would disagree with the objectives of raising educational standards and improving the general quality of American education. It seems clear, however, that efforts to foster academic achievement should involve more than simply setting demanding standards and mandating examinations. The task remains of identifying strategies to achieve efficiently and effectively the desirable reform objectives—without having

a negative impact on any subpopulation of students. Those strategies will, among other things, need to address the issue of restructuring the academic experiences of students in ways that will help them appreciate the value of academic achievement, increase their expectations and aspirations, and enhance their sense of academic efficacy.[58] This is a much more difficult and, we dare say, expensive task than mandating tougher standards and a high-stakes examination referenced to them.

CHAPTER SIX

Do High-Stakes Graduation Tests Improve Learning Outcomes?

Using State-level NAEP Data to Evaluate the Effects of Mandatory Graduation Tests

MONTY NEILL, WITH KEITH GAYLER

The test-driven approach to school reform is sweeping the nation. More and more states are adopting high-stakes testing for graduation and grade promotion, and they are using test results to impose sanctions on schools, districts, teachers, and students.[1] Texas and North Carolina have been widely cited recently as examples for other states to follow as they reform elementary and secondary education.[2] Both states are praised for their rising scores on the National Assessment of Educational Progress (NAEP). Supporters of test-driven reform maintain that NAEP scores in Texas and North Carolina have risen because those states have implemented extensive testing coupled with strong accountability sanctions.[3]

Should the testing and accountability practices of these two states be held out as an example to other states? We conclude they should not. Our nationwide comparison of states with and without high-stakes graduation tests shows there is no clear relationship between high-stakes tests and gains on the NAEP. Indeed, at grade 8 the relationship is the inverse: states *without* high-stakes graduation tests have been more likely to show progress. Rather than rely on test-driven reform, states should support alternative means for improving education that have been developed and proved in practice at exemplary schools. These schools use alternate methods of assessment and model better approaches to accountability.