

---

**CONSEQUENCES OF LARGE-SCALE, HIGH-STAKES  
TESTING ON SCHOOL AND CLASSROOM PRACTICE**

*Brian M. Stecher*

---

- Why should we care about the effects of testing?
- What research has been done about the effects of high-stakes testing?
- What are the positive and negative effects of testing on classrooms? On schools?

This chapter examines the consequences of high-stakes testing on the educational system. We focus on the effects of high-stakes tests on students, teachers, and principals because the evidence of these effects is comparatively strong. High-stakes testing may also affect parents (e.g., their attitudes toward education, their engagement with schools, and their direct participation in their child's learning) as well as policymakers (their beliefs about system performance, their judgments about program effectiveness, and their allocation of resources). However, these issues remain largely unexamined in the literature. As a result, this chapter concentrates on the impact of large-scale, high-stakes testing on schools and classrooms and the adults and students who teach and learn in these environments.

**THE IMPORTANCE OF THE EFFECTS OF TESTING**

There are a number of reasons to be concerned about the consequences of testing on schools and classrooms, but two are particularly compelling:

- First, the goal of changing educational practice is one of the major justifications for implementing high-stakes tests. Advocates hope test scores will prompt schools to reform policy, encourage teachers to adopt more effective practices, and motivate students to work harder. Under these circumstances, the *Standards for Educational and Psychological Testing* make it clear that broad evidence about the consequences of testing is necessary for a thorough investigation of test validity (American Educational Research Association et al., 1999). Information about the broad impact of tests on the educational system is necessary when the purported benefits of tests include behavioral changes such as improved instruction or increased student motivation. Therefore, we need to examine whether these changes are occurring to ascertain whether high-stakes testing is meeting policymakers' goals for reform.
- Second, changes in behavior may, in turn, affect the validity of various interpretations of test scores. For example, some reactions to high-stakes tests, such as changes in the conditions under which tests are administered, will affect the relationship between test scores and achievement. These behaviors can lead to increases in scores without concomitant increases in knowledge—i.e., score inflation—which was discussed in Chapter Three. Without monitoring such changes in behavior, we will not know the extent to which gains in scores are due to real improvement in achievement rather than differences in testing conditions or other factors.

### **The Effects of Testing and Test Validity**

It is worth making a brief detour at this point to explain why information about changes in school and classroom practices is important in judging the validity of test scores. As we have discussed, large-scale tests measure an extremely limited sample of behaviors—only a few questions are asked and they are limited to those that can fit into a few formats. People who use test scores—from policymakers to parents—do so in the hope that performance on the test questions is indicative of performance in a broader domain, such as third-grade language arts or first-year algebra. Under appropriate conditions,

well-developed tests will support inferences from test scores to broader domains such as these.

However, certain practices can reduce the meaningfulness of test scores as indicators that students have mastered large subject-matter domains. For example, if the same test form is used repeatedly, teachers may become familiar with the specific items that appear on that form. If the test content becomes well known, teachers may shift their instruction accordingly. Such targeted teaching to those skills that are represented on a test can raise scores without increasing mastery of the broader domain. Although it is quite likely that students who learn the full curriculum will do well on a test that is sampled from that curriculum, the converse is not necessarily true. Students who do well on specific test questions that have been emphasized in their class may not have mastered the full curriculum. If this situation occurs, then the use one makes of the test score information—judging program quality, retaining students in grade, rewarding schools, and other such decisions—will be suspect.

### **Broadened Use of High-Stakes Testing to Promote Changes in School Practice**

As we described in Chapter Two, there was little concern about the effects of testing on teaching prior to the 1970s. The federal government and the states used large-scale tests to monitor the status of the educational system and provide information that might be helpful to teachers and students. However, specific rewards or sanctions were seldom associated with performance. For example, the National Assessment of Educational Progress (NAEP), which is the only large-scale federally commissioned achievement test, was designed solely with a monitoring role in mind. William Bennett, former U.S. Secretary of Education, described this role as “supplying the American people with prompt, reliable, and comprehensive data on the educational performance of our children” (National Assessment of Educational Progress, 1987, 3). He likened NAEP to a measuring tool: “It is the closest thing we have to a barometer of our educational performance . . . as a nation . . .” When tests are conceived in this manner by policymakers, there is little concern about their direct impact on practice.

Beginning with the minimum competency testing movement in the 1970s, policymakers began to use test results in new ways—specifically, as the basis for decisions about individual performance. Tests grew more common in the 1980s, and the rationale for large-scale testing expanded from judging performance to influencing practice (Popham, 1987). With the advent of formal accountability systems in the 1990s, policymakers embraced a new, more potent vision for the role of assessment. They envisioned tests (often in combination with standards) as a mechanism to influence changes in practice, something that could be used “to exert a strong positive effect on schooling . . .” (Achieve, Inc., 2000, 2). Testing programs built in this mold provide incentives and/or sanctions for individual students (e.g., graduation, retention-in-grade) and/or for schools (e.g., cash rewards, administrative review) on the basis of test scores. The incentives indicate that performance has become an important issue to policymakers.

Test-based accountability systems, such as those that provide incentives and sanctions for both students and schools, are designed to affect schooling in multiple ways. For example, the California Department of Education articulated five ways that high-stakes, standards-based reform would lead to positive school changes (California Department of Education, 1998, p. 4):

1. Signal important content to teachers so that they can improve instruction
2. Identify learning that is below what is expected of students, thus motivating students and parents to put more effort into school work
3. Raise public awareness and prompt citizens to bring pressure to bear on ineffective schools
4. Encourage greater parental involvement
5. Facilitate the targeting of resources to schools that are in trouble.

This list of positive outcomes is typical of the rationale that other states provide to justify their high-stakes testing programs to policymakers and the public.

## **GATHERING EVIDENCE ABOUT THE EFFECTS OF HIGH-STAKES TESTING**

In light of the changes that occurred in the uses of large-scale testing in the 1980s and 1990s, researchers began to investigate teachers' reactions to external assessment. The initial research on the impact of large-scale testing was conducted in the 1980s and early 1990s. In the mid-1990s, states began to implement statewide, test-based accountability systems, prompting renewed interest in the effects of testing on the practice of teaching. Large-scale studies of test validity and the effects of testing were conducted in a number of states that implemented such accountability systems. Research on these issues continues to the present day.

The bulk of the research on the effects of testing has been conducted using surveys and case studies. Typical of the survey research was a study conducted in 1991 by Shepard and Dougherty as part of a larger effort to examine the validity of test scores as well as the effects of testing on practice. The study was conducted in two large school districts with student populations in excess of 50,000. The researchers surveyed a sample of approximately 850 third-, fifth-, and sixth-grade teachers in approximately 100 schools. Surveys were administered in the spring, near the end of the school year. The surveys included questions about pressure to improve test scores, the effects on instruction, test preparation activities, controversial testing practices, use of the test results, and the effects—both positive and negative—of standardized testing.

More recently, the scope of survey research has been expanded to include statewide samples of teachers and principals, and the methods have expanded to include both written and telephone surveys. In the 1990s, researchers at RAND conducted a number of studies of state accountability systems, including those in Kentucky, Maryland, Vermont, and Washington. These studies usually included surveys of principals and teachers as well as quantitative analyses of test scores. For example, Koretz et al. (1996a) studied the effects of Kentucky's educational reform effort, which included a test-based accountability system that used portfolios of students' written work as well as more-traditional tests. A stratified random sample of 80 elementary schools and 98 middle schools was selected for the study. Both computer-assisted telephone interviews and written surveys were used to

collect data. Representative samples of 186 fourth-grade teachers and 175 eighth-grade mathematics teachers were interviewed and surveyed. In addition, principals in the sampled schools were asked to participate in a telephone interview. The interviews with the principals focused on general support for the reform effort, the principals' own responses to the reform, the effects of the reform on their schools, how the test scores are used, and the burdens imposed by the testing program. Teachers were questioned on some of the same issues and also on test preparation, instructional practices, and their understanding of the testing program, particularly the portfolio component. Similar methods were used in the other states studied by the RAND researchers.

Case studies have also been used to examine the effects of high-stakes testing on practice. In a study published in 1991, Smith et al. conducted detailed observations of teachers in two Arizona elementary schools whose students took tests that had significant consequences. During the fall 1987 semester, the authors conducted daylong observations in 29 classrooms. Lessons were also audiotaped. The researchers also observed and recorded staff meetings. In January 1988, they selected a subset of 20 teachers for detailed open-ended interviews covering the validity of the tests, the effects of the tests on teachers, test preparation methods, and the effects of the tests on pupils. Subsequently, six teachers were selected for more-extensive observations occurring one, two, or three days a week during the spring of that year. In total, the six classes were observed for 81 days. The purpose of the observations was to understand "ordinary instruction"; therefore, the observers focused on what was taught, the methods used, the allocation of time, language and interaction among teachers and pupils, teaching materials, and classroom interruptions. The researchers used a variety of techniques to review and summarize the data and compare the situation in these classrooms to the literature on testing and its effects.

Other researchers have used case study techniques to study teaching practices within the context of high-stakes testing. McNeil and Valenzuela (2000) accumulated information from hundreds of Texas teachers and administrators over a period of a decade while the state implemented a test-based accountability system. Their research included in-depth longitudinal studies in three high schools as well as many years of professional development work with hundreds of

teachers and dozens of principals and administrators. Borko and Elliott (1999) and Wolf and McIver (1999) focused their case studies of testing effects in a slightly different direction. They identified six “exemplary” elementary and middle schools in Kentucky (and later in Washington State) and conducted observations and interviews to see how the most respected administrators and teachers were reacting to testing mandates.

### **THE POSITIVE AND NEGATIVE EFFECTS OF HIGH-STAKES TESTING**

The earlier discussion suggests that researchers need to cast a wide net when examining responses to large-scale, high-stakes testing programs because the developers of those programs envision them operating through many different mechanisms. On the positive side, one might expect to find changes in school policies that are designed to make schools more effective, changes in teaching practice that will enhance student achievement, and changes that result in increased motivation on the part of students. However, one might also find changes that most would consider negative, such as narrowing of the curriculum to tested topics to the exclusion of other domains of learning, inappropriate test preparation, or even cheating. Table 4.1 provides a partial list of the potential effects of high-stakes tests on students, teachers, administrators, and policymakers, differentiating between those effects that would be considered positive and those that would be considered negative.

Two issues complicate the problem of judging the net effect of large-scale, high-stakes testing:

- Many of the effects suggested by Table 4.1 are difficult to measure. For example, it is difficult to assess psychological variables such as motivation or competitiveness with any accuracy. Similarly, it is difficult to track the influence of diverse factors on policymaking. Furthermore, while it is possible to measure the quantity of many of the potential effects of high-stakes testing (e.g., how many hours were spent teaching decimal place values?), measuring the quality of those effects can be vexingly difficult (e.g., how well were place-value lessons taught?). As a result,

**Table 4.1**  
**Potential Effects of High-Stakes Testing**

Positive Effects	Negative Effects
Effects on Students	
Provide students with better information about their own knowledge and skills	Frustrate students and discourage them from trying
Motivate students to work harder in school	Make students more competitive
Send clearer signals to students about what to study	Cause students to devalue grades and school assessments
Help students associate personal effort with rewards	
Effects on Teachers	
Support better diagnosis of individual student needs	Encourage teachers to focus more on specific test content than on curriculum standards
Help teachers identify areas of strength and weakness in their curriculum	Lead teachers to engage in inappropriate test preparation
Help teachers identify content not mastered by students and redirect instruction	Devalue teachers' sense of professional worth
Motivate teachers to work harder and smarter	Entice teachers to cheat when preparing or administering tests
Lead teachers to align instruction with standards	
Encourage teachers to participate in professional development to improve instruction	
Effects on Administrators	
Cause administrators to examine school policies related to curriculum and instruction	Lead administrators to enact policies to increase test scores but not necessarily increase learning
Help administrators judge the quality of their programs	Cause administrators to reallocate resources to tested subjects at the expense of other subjects
Lead administrators to change school policies to improve curriculum or instruction	Lead administrators to waste resources on test preparation
Help administrators make better resource allocation decisions, e.g., provide professional development	Distract administrators from other school needs and problems

**Table 4.1—Continued**

Positive Effects	Negative Effects
Effects on Policymakers	
Help policymakers to judge the effectiveness of educational policies	Provide misleading information that leads policymakers to suboptimum decisions
Improve policymakers' ability to monitor school system performance	Foster a "blame the victims" spirit among policymakers
Foster better allocation of state educational resources	Encourage a simplistic view of education and its goals

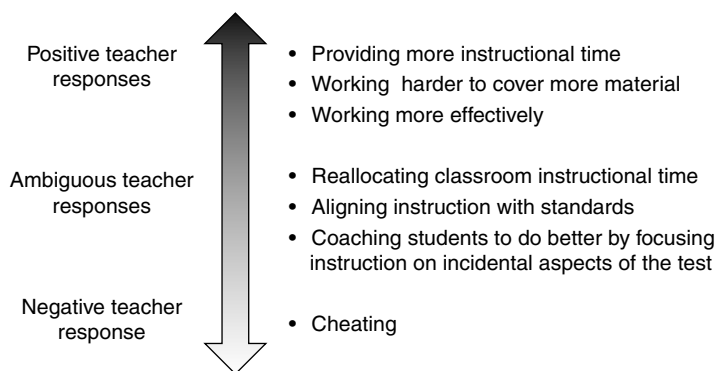
few of these potential consequences have been studied systematically. Research primarily focuses on school and classroom practices using teacher and principal surveys to gather information.

- The effects that are measurable are not measured in a common metric. For example, the amount of additional professional development teachers receive might be measured in hours, but relaxation of test administration rules to benefit students would have to be measured in some other way. As a result, there is no way to combine positive and negative effects to produce a "net" judgment about impact.

To make sense of the research on the effects of high-stakes testing on practice, it is helpful to differentiate among responses to high-stakes testing at different levels of the educational system. Most of the research has been conducted at the classroom level, and it has focused on changes in curriculum and instruction under the control of teachers. There is also some evidence about changes at the school level, including decisions about curriculum emphasis, teacher support, and programmatic changes. Less is known about the use of high-stakes test results by state policymakers.

### **Teacher Response to High-Stakes Testing**

It is also helpful to differentiate among types of responses to high-stakes testing. Koretz, McCaffrey, and Hamilton (2001) identify seven categories of teacher responses to high-stakes tests and their likely effects on test scores and student learning (see Figure 4.1). They differentiate between three types of teacher responses: those that are positive (i.e., they have beneficial effects on learning and lead to valid



**Figure 4.1—Seven Categories of Teacher Responses to High-Stakes Testing**

increases in scores), those that are negative (i.e., they lead to distortions of learning or inflated scores), and those whose impact is ambiguous (i.e., they can be positive or negative depending on the specific circumstances).

Some forms of teacher response, if handled effectively, are clearly positive: providing more instructional time, working harder to cover more material in a given amount of instructional time, and working more effectively by adopting a better curriculum or more-effective teaching methods. These are precisely the effects that proponents of high-stakes testing hope will occur. All of these effects have been documented to some extent by researchers, and all should generate real increases in student achievement.

Other forms of teacher response to high-stakes testing have ambiguous effects—that is, they can lead to real gains in student understanding and achievement or to inflation of scores (i.e., gains that do not generalize to other measures) or both, depending on the specific ways those gains are realized. Ambiguous responses include reallocating classroom instruction among topics or subjects to emphasize tested content instead of content that receives little or no emphasis on the test; aligning instruction with standards, which is a special case of curriculum reallocation motivated by attention to curriculum standards; and coaching students to do better on a test by focusing

instruction on aspects of the test that are partly or entirely incidental to the definition of the domain the test is intended to represent.

Reallocation of instruction, alignment of instruction with test content, and coaching can be classified as positive effects when they focus on important aspects of the domain the test is designed to measure or specific skills that help students demonstrate their actual achievement. Those effects will be negative when they focus on specific features of test content or format that are not broadly reflective of the domain. For example, reallocation of classroom time to emphasize topics covered by the test can be beneficial if the coverage that was reduced or eliminated is on topics that are clearly less important than those given added emphasis. Conversely, reallocation can be negative if classroom time is taken away from important aspects of the domain that do not happen to be represented in the test (for example, because they are difficult to assess in a multiple-choice format).

Similarly, efforts to improve alignment can lead to a focusing of instruction that may either be beneficial or lead to inflation of test scores. If teachers focus more intently on desired outcomes at the expense of relatively unimportant material and do so effectively, the result should be higher achievement in terms of the desired outcomes and higher scores. On the other hand, if the material being deemphasized as a result of this refocusing is important, scores may become inflated. The extent to which greater alignment—that is, sharper focus—or any other reallocation produces real gains in total achievement rather than score inflation depends in part on what goes *out of* focus as well as what *comes into* focus. The issue is further complicated because reasonable people may differ in their judgment about the relative merits of the topics that are emphasized or deemphasized. For example, some may think that greater emphasis on spelling, grammar, and punctuation is appropriate while others may think that the time should be spent on other topics related to good writing, such as studying literary genres or learning to write for different audiences and different purposes.

A similar principle applies to coaching. Reasonable efforts to familiarize students with the format and other aspects of a test can increase the validity of scores. If students do not understand the test instructions or the question formats, or how they should record their

answers, their scores will underestimate their actual learning. Removing these obstacles to performance by familiarizing students with the testing procedure makes their test results more valid. However, coaching can also inflate scores when it improves test performance by focusing on features of the test that are incidental to the domain the test is supposed to measure. Because these features are incidental, learning about them does not produce real improvements in students' knowledge of the domain.

A teacher can also respond to high-stakes testing by cheating, a response that is clearly negative and can only lead to inflation of scores.

### **Positive Classroom Effects**

On the positive side, there is evidence that high-stakes tests have led teachers to work more effectively. Case studies revealed how high-stakes testing that includes innovative forms of assessment can encourage teachers to change their instructional practices in positive ways (Wolf and McIver, 1999; Borko and Elliott, 1999). They also reveal how some schools can seize on assessment requirements to rededicate themselves to quality (Wolf et al., 1999) and how testing programs can influence schools to refocus professional development and support services (Borko, Elliott, and Uchiyama, 1999). Bishop (1986) cites evidence from Ireland to support the contention that curriculum-based external examinations promote "the development of mentoring relationships between teachers and students." He also found that teachers in "all Regents" high schools in New York (schools that require all students to take demanding Regents courses in five core subjects) were inspired to work harder by their school's commitment to student success on the high-stakes Regents examination (Bishop and Mane, 1999).

States have also had some success using high-stakes tests as "instructional magnets" (Popham, 1987) to persuade teachers to reallocate instructional time to include new elements of the state curriculum. For example, both Vermont and Kentucky used test-based accountability systems as pivotal elements in large-scale curriculum reform efforts. Statewide surveys revealed that teachers in Vermont increased the amount of time they spent teaching problem-solving and mathematical representations to prepare

students for the state's portfolio-based high-stakes assessment (Koretz et al., 1994). Similar survey results in Kentucky showed that the high-stakes, performance-based assessments in writing and mathematics strongly influenced teachers to make their instruction more consistent with the state curriculum in these areas (Stecher et al., 1998; Koretz et al., 1996a).

In addition, testing can provide useful information for curriculum and instructional decisionmaking. For instance, the majority of teachers in two high-stakes testing districts surveyed by Shepard and Dougherty (1991) said test results were helpful in identifying student strengths and weaknesses and in attracting additional resources for students with the greatest needs.

### **Neutral and Negative Classroom Effects**

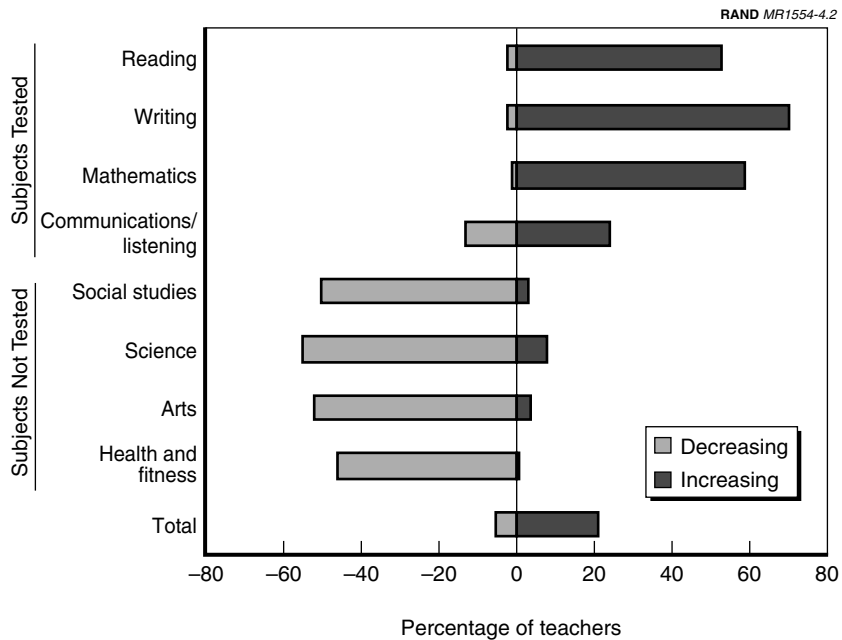
Despite these positive findings, a large share of the published research on the impact of high-stakes testing on educational practice describes neutral or deleterious effects.

Firestone, Mayrowetz, and Fairman (1998) found that high-stakes testing in Maryland and Maine had little effect on instructional practices one way or the other. Similarly, Jones et al. (1999) reported mixed effects of tests on teaching methods in North Carolina. For example, roughly equal percentages of teachers said they had either increased their use of inquiry projects (thought to have educational benefits but not necessarily useful in preparing students for the tests) or decreased their use. The same was true for the percentage of teachers who increased or decreased their amount of lecturing, use of textbooks, and use of worksheets.

**Negative Curriculum Reallocation.** In contrast, the evidence on negative reallocation of classroom instruction among certain topics or subjects is widespread. Researchers first began to notice that high-stakes tests led to negative reallocation in the late 1980s; the effect was described at the time as "narrowing" of the curriculum (Shepard and Dougherty, 1991). Moreover, the greater the stakes, the more likely that such narrowing would occur (Corbett and Wilson, 1991). For example, one of the first studies of the effects of testing (conducted in two Arizona schools in the late 1980s) showed reallocation among subjects that reduced the emphasis on important

material. The study revealed that teachers neglected subjects such as science, social studies, and writing that were not part of the mandated testing program (Smith et al., 1991). Similar declines in instructional time for nontested subjects have been observed in statewide studies in other states including Maryland, North Carolina, and Washington (Koretz et al., 1996b; Jones et al., 1999; Stecher et al., 2000a). Figure 4.2 shows the shifts in instructional emphasis reported by fourth-grade teachers in Washington State, which has high-stakes testing in four of the eight subjects covered by state standards.

Research in Kentucky shows that the size of subject-to-subject shifts in emphasis can be substantial. Table 4.2 shows the average number of hours per week that fourth- and fifth-grade Kentucky teachers spent on seven different subjects. What makes the table interesting is that Kentucky tested some subjects in fourth grade and others in fifth



SOURCE: Stecher et al., 2000a, 21.

**Figure 4.2—Percentage of Teachers Increasing or Decreasing Instructional Time in Tested and Nontested Subjects**

**Table 4.2**  
**Mean Weekly Classroom Hours per Subject, Self-Contained Kentucky**  
**Fourth-Grade and Fifth-Grade Classrooms**

	Fourth Grade	Fifth Grade
Subjects tested in fourth grade		
Reading	5.2	4.7
Writing**	5.8	4.0
Science**	5.2	3.5
Subjects tested in fifth grade		
Mathematics**	4.9	6.4
Social studies**	3.5	5.6
Arts and humanities**	1.5	2.4
Practical living/vocational education**	1.4	2.4

Note: \*\*Significant at  $p < 0.01$ .

SOURCE: Stecher and Barron, 1999.

grade. Teachers responded accordingly, leading to between-grade differences of an hour and a half per week or more in student exposure to subject content. Reallocating instructional time across grades to better align the available time with the subjects that are tested in each grade runs the risk of inflating scores on a grade-by-grade basis.

Negative reallocation can also occur within a subject area when teachers change their emphasis on specific topics in response to a test. Early research found that teachers tend to change course objectives and the sequence of the curriculum to correspond to the content and timing of new tests (Corbett and Wilson, 1988; Herman and Golan, [n.d.]; Darling-Hammond and Wise, 1985). Teachers also place more emphasis on topics that appear on the test and less emphasis on topics that are not tested.

For example, Romberg, Zarinia, and Williams (1989) surveyed a national representative sample of eighth-grade mathematics teachers and found that they increased coverage of basic skills, paper-and-pencil computation, and topics emphasized on their local tests while decreasing coverage of extended project work, work with calculators, and topics not emphasized on these tests. Shepard and Dougherty (1991) found that two-thirds to three-quarters of all teachers in two districts gave more emphasis to basic-skills instruction, vocabulary lists, word recognition skills and paper-and-pencil computation as a result of mandated tests that emphasized these topics.

Middle school teachers in Maryland and Maine also shifted their topic emphasis to correspond to the primary topic areas covered on the state test, although researchers reported that the extent of the change was not as dramatic as had been reported in other studies (Firestone, Mayrowetz, and Fairman, 1998). Opposite shifts were observed in Kentucky, where the testing program was designed to emphasize problem-solving in mathematics and extended writing in language arts. Teachers reduced their emphasis on computation and algorithms in mathematics and on the mechanics of writing (Koretz et al., 1996a). Researchers also found that pressure to improve test scores caused some Arizona teachers to neglect important curriculum elements that were not tested, including “reading real books, writing in authentic context, solving higher-order problems, creative and divergent thinking projects, longer-term integrative unit projects, [and] computer education” (Smith et al., 1991).

**Adapting Teaching Styles to Test Formats.** A more subtle type of negative reallocation—one that can shade into negative coaching—occurs when teachers adapt their teaching styles to make classroom presentations more like the format of the test or adopt instructional approaches that resemble testing methods.

For example, Shepard and Dougherty (1991) found that many teachers in two high-stakes testing districts were asking students to practice finding mistakes in written work rather than producing writing of their own. Smith and Rottenberg (1991) reported that teachers they studied in two Arizona schools had students solve only the type of math story problems that are found on the Iowa Test of Basic Skills (which was mandated in Arizona at the time). Stodolsky (1988) studied math and social studies instruction in 11 school districts in the Chicago area and found that high-stakes testing discouraged teachers from using joint- or team-teaching approaches and from changing their methods to facilitate serious student learning.

In Vermont, where the portfolio testing program encouraged teachers to include mathematical problem-solving in their curriculum, researchers found that many teachers focused narrowly on the aspects of problem-solving that would result in higher scores with the specific rubrics used in the tests rather than on problem-solving in the broadest sense (Stecher and Mitchell, 1995). This approach, which the authors labeled “rubric driven instruction,” is an instance in

which the distinction between substantively important and incidental aspects of the test is vague and the distinction between reallocation and coaching is blurred. If students' performance appears to improve when scored with one set of rubrics but does not appear to improve as much using another reasonable rubric, then scores may be inflated.

**Negative Coaching.** The literature contains other examples of negative coaching, i.e., activities that focus excessive amounts of time on incidental aspects of a test. For example, several studies have shown that "test preparation" activities (such as becoming familiar with the format of the test questions and learning how to record answers) can consume substantial amounts of limited instructional time. Herman and Golan (n.d.) surveyed upper-elementary schoolteachers in nine states and found that between one and four weeks of class time were diverted away from other learning activities and given to test preparation. Similar amounts of test preparation time (up to 100 hours per class) were reported in Arizona (Smith, 1994).

More recently, Jones et al. (1999) reported that 80 percent of teachers in North Carolina said their students spent more than 20 percent of their total instructional time practicing for end-of-grade tests. In these instances, the phrase "test preparation" was not clearly defined and exactly what activities occurred in preparing for the end-of-grade test is uncertain. However, this amount of coaching would certainly entail the loss of considerable learning time. In general, it is very difficult to quantify the extent of coaching without monitoring instruction for extended periods of time. An activity that uses a test-like format or representation may be quite appropriate in the short run, but the continuing use of such approaches to the exclusion of others constitutes coaching. In part because it is so difficult to detect, there is little research evidence about the extent of negative coaching. However, research on score inflation suggests that coaching is widespread in high-stakes testing situations.

**Cheating.** Cheating is the most extreme negative reaction to high-stakes testing. Cheating can take many forms: providing the actual test items in advance, providing hints during test administration, suggesting revisions, making changes to answer sheets before scoring, leaving pertinent materials in view during the testing session, and so on. Cheating scandals surface frequently. For example, in a

recent case in New York City, investigators charged that dozens of teachers had cheated over a period of five years by giving students answers to the mathematics and reading tests that are used both as promotional gates and to rank schools. Educators told students which answers to change, had them put their initial answers on scrap paper and then corrected the students' answers before transferring them to the answer sheet, and gave them practice tests containing questions from the operational test (Goodnough, 1999).

Data on the incidence of cheating are scarce, but high-stakes testing can be expected to increase cheating. In a study of Kentucky educators' responses to the high-stakes Kentucky Instructional Results Information System (KIRIS) assessment, Koretz et al. (1996a) found that 36 percent of teachers reported seeing test questions rephrased during testing time either occasionally or frequently. Twenty-one percent reported seeing questions about content answered during testing time, and the same percentage reported seeing revisions recommended either during or after testing. Seventeen percent reported observing hints provided on correct answers. The corresponding percentages were somewhat lower in a parallel study of the lower-stakes Maryland School Performance Assessment Program (Koretz et al., 1996b).

### **School-Level Effects**

Less is known about changes in policies at the district and school levels in response to high-stakes testing, but mixed evidence of some impact has appeared.

**Positive Effects.** Positive changes include revising district curriculum and testing programs to be consistent with state curricula and providing professional development opportunities for teachers (Stecher et al., 2000a). Bishop (1986) argues that external examination can also lead districts and schools to use their resources more effectively—for example, by hiring more qualified teachers and by providing essential instructional materials.

Testing programs can also be credited with helping focus resources on students or schools most in need. For example, about one-half of the principals in Washington State indicated that their school had added summer sessions in response to low test scores (Stecher and

Chun, 2001). Schools also reported adding after-school sessions and Saturday school to address the needs of low-performing students (Stecher et al., 2000a). State accountability systems can formalize this reallocation of resources based on test results. For example, California's accountability system provides additional financial resources as well as professional assistance to schools with low test scores to help them improve their effectiveness. Similarly, the new Elementary and Secondary Education Act (ESEA) legislation requires poor-performing schools to provide funds for tutoring for students whose test scores do not show adequate progress toward proficiency. Testing programs can also influence attitudes among staff; for example, they can promote greater cohesion, openness to new ideas, and esprit de corps, although these effects have only been documented anecdotally.

**Negative Effects.** On the other hand, researchers have also documented changes that appear to be designed to improve scores with little regard to their larger educational implications. For example, Koretz et al. (1996b) found that about one-third of principals in Maryland reassigned teachers among grades to improve the relative quality of teaching in the assessed grades. Because such shifts do not improve the quality of teaching across the grades, it is likely to inflate scores. Many Washington principals offered incentives to students in the form of parties and field trips for good test performance (Stecher et al., 2000a).

Other potential negative effects of high-stakes testing have recently come to the public's attention. Scores can be increased artificially, for example, by excluding low-scoring groups of students (e.g., students with disabilities, limited proficiency in English, or just low performance); by retaining low-scoring students in grades below those in which the test is administered; by allowing an increase in absences on test days; by granting waivers (exemptions from testing) demanded by parents; and by increasing dropout rates. Other potential effects would not inflate scores but could be important nevertheless. For example, Hauser, Pager, and Simmons (2000) argue that while current racial/ethnic differences in rates of retention in grade can be "almost entirely explained by social and economic deprivation among minority youth," group differences in test scores are larger than generally expected as a result of social and economic fac-

tors. Therefore, they suggest that tying promotion to test scores could increase racial/ethnic disparities in retention rates.

The extent to which these negative effects have occurred and the factors that may influence their occurrence remain uncertain, but there is a clear need for further monitoring of these effects and research on them. Although numerous news articles have addressed the negative effects of high-stakes testing, systematic research on the subject is limited. For example, a recent article in the Texas media argued that “schools that climbed in the state’s accountability ratings in 1999 had substantially larger increases in TAAS [Texas Assessment of Academic Skills] exemptions for special education students than did other schools.” However, the article also suggested that more investigation is needed to clarify the relationship between exclusions and accountability ratings (Dallas Morning News, 2000). Parent hostility to testing has led to increased requests for waivers for students to be exempted from testing in California and increased absences during the testing period (Neufeld, 2000).

Certainly, testing that creates “gates” through which students must pass in order to be promoted will lead to an increase in retention in grade, but it is unclear to what extent other forms of high-stakes testing will do the same. Grade retention has increased in Texas in recent years, particularly for African-American students (see, for example, Haney, 2000). The timing of this increase is not consistent with the implementation of the most recent high-stakes testing program in Texas (Carnoy, Loeb, and Smith, 2000), but it may be related to earlier waves of high-stakes testing in Texas. A study of the first years of Kentucky’s high-stakes KIRIS assessment program found no evidence of increased retention in grade (Koretz and Barron, 1998).

### **Effects on Equity**

It is important to note that while one of the rationales for test-based accountability is to improve educational equity, it is not clear that these accountability policies lead to more-equal educational opportunities or outcomes for students from different backgrounds. Indeed, some observers have argued that the negative effects of high-stakes testing on curriculum and instruction appear to be greater for low-performing students and low-scoring schools than they are for

high-performing students or high-performing schools (Shepard, 1991).

Research regarding the effects of test-based accountability on equity is very limited. For example, McNeil (2000) reports that high-stakes testing in Texas is widening the gap between what is taught in historically low-performing schools (with predominantly minority and poor students) and what is taught in high-performing schools. Other observers have argued that the Texas system of statewide testing has improved the quality of instruction for students from all backgrounds. They point to the decreasing gap between racial/ethnic groups in their mean scores on the Texas state test (TAAS). However, recent evidence shows that the relative gains of Texas minority students on the TAAS were not echoed in the NAEP (an external, low-stakes test); the racial/ethnic gap in Texas did not narrow at all on the NAEP. This might reflect greater inflation of the scores of minority students (Klein et al., 2000) or artifacts of the testing program, such as “ceiling effects” that occur when a test is relatively easy for the students who are taking it, leading to an excessive number of students answering most questions correctly.

Another study showed that in Arizona, the amount of class time spent on test preparation and test administration was greater in urban, low-income, high-minority districts (Smith, 1997). Conversely, a high-stakes accountability system in Kentucky led some teachers to increase their academic expectations for students; however, more teachers reported increased expectations for high-achieving students than reported increased expectations for low-achieving or special education students as a result of the high-stakes tests (Koretz et al., 1996a).

## SUMMARY

The net effect of high-stakes testing on policy and practice is uncertain. Researchers have not documented the desirable consequences of testing—providing more instruction, working harder, and working more effectively—as clearly as the undesirable ones—such as negative reallocation, negative alignment of classroom time to emphasize topics covered by a test, excessive coaching, and cheating. More important, researchers have not generally measured the extent or

magnitude of the shifts in practice that they identified as a result of high-stakes testing.

Overall, the evidence suggests that large-scale high-stakes testing has been a relatively potent policy in terms of bringing about changes within schools and classrooms. Many of these changes appear to diminish students' exposure to curriculum, which undermines the meaning of the test scores. It will take more time and more research to determine on balance whether the positive impact on teaching practice and student learning outweigh the negative ones. Similarly, a new type of research will be needed to try to determine how to design accountability systems to maximize benefits and minimize negative consequences. In Chapter Six, we offer some recommendations for changing accountability systems to maximize the positive effects and minimize the negative consequences identified in this chapter.