CHAPTER 2

REVIEW OF RELATED LITERATURE

"...and justice for all"

> Natalie J. Martinez had one foot in the college door. She had completed her senior year at a San Antonio public high school, and her talent as a soprano, honed in the school choir, had won her a full scholarship to study music at the University of the Incarnate Word. But she could not overcome one last hurdle: the Texas Assessment of Academic Skills, a standardized achievement test that the state requires public high-school students to pass before they can graduate. Unable, despite repeated attempts, to earn a passing score on the mathematics portion of the test–mainly due to her difficulties in working with fractions–Ms. Martinez was denied a high-school diploma in 1997. As a result, Incarnate Word, a local Roman Catholic institution, would not admit her. Her plans for college were stifled on a sour note. "I was very ashamed of what happened to me," says Ms. Martinez, who had been "more or less an A or B student, except in math." (Schmidt, 2000)

The accountability movement evident in Virginia's Standards of Learning (SOL) and Standards of Accreditation (SOA) rests on a foundation that relies on the empirical validation of student learning. The format is simple. A curriculum is developed. Teachers are told to teach it. Students are tested on their attainment of the curriculum. And, lastly, rewards are distributed to students, teachers, and schools who meet the state's assessment targets while threats and punishments are meted out to those who do not. However, those advocating accountability seem to assume that Virginia's proposed program of reforms will have only positive outcomes. I compiled this review of literature to give a more complete account of the potential consequences of high-stakes testing and measurement-driven instruction.

Outcomes for Students

Ample evidence exists to document the position that some students will suffer negative outcomes in a high-stakes testing environment. Pedroza (1997) reported that in 1993, 90 percent of the students in 333 schools identified as Texas' lowest performing schools were largely Hispanic or African-American. Some 80 percent of these students were identified as low income as well.

Additional evidence from Sacks (1999) illustrates the profound connection between race, class, and student achievement. His analysis of 1997 data from the Texas Assessment of Academic Skills (TAAS) shown in Table 2 illustrates that while 81 percent of white students passed the 10^{th} grade exit test (allowing the granting of a high school diploma) only half of poor and minority students scored at this level. Similarly, white students were about three times more likely to pass the Algebra I end-of-course test and were approximately thirty percent more likely to pass the Biology I end-of-course test.

Table 2

TAAS Pass Rates by Race and Class in Texas

	African American %	Mexican American %	Economically disadvantaged %	White %
TAAS Exit (10th Grade, Spring 1997	48	52	50	81
Algebra I End of Course (Spring 1996)	11	14	14	40
Biology I End of Course (Spring 1996)	59	61	59	90

Note. From Standardized Minds (p. 112), P. Sacks, 1999, Cambridge, MA: Perseus Books. Reprinted with permission.

Natriello and Pallas (1999) found that in 1998-99 while 52 percent of white students passed the TAAS Algebra I end-of-course test, only 20 percent of African-American students and 26 percent of Hispanic students passed. Similarly, Strang (1981) analyzed data from 13 states that were in the process of implementing test-based graduation requirements. Findings indicated that failure rates were higher for minority students than for non-minorities on both initial and subsequent test administrations.

Shepard (1991) explained that high-stakes testing is harmful to disadvantaged children because they are assigned greater amounts of "deadly boring drill for longer periods" (p. 236). She described an increasing negative attitude toward schoolwork by low achievers.

Darling-Hammond (1994a) determined that high-stakes testing can result in disproportionately large percentages of low income and minority students being placed in low instructional tracks or in remedial programs. Natriello and Pallas (1999) found that New York's dual track diploma program leading to either a local (minimum competency) diploma or a college preparatory (Regents) diploma consigned large proportions of poor and minority students to lower-level tracks, and, therefore, to lower-level, non-college preparatory diplomas.

Smith, Edelsky, Draper, Rottenberg, and Cherland (1987) described the effects of standardized tests on elementary school students as seen by their teachers. The authors reported that such tests were harmful to students' psychological well-being. During the test week, teachers reported an increase in absenteeism, physical illnesses, worry, wetting, and refusal to participate.

High-stakes testing can also affect *who* receives instruction. The International Reading Association (1999) opined that educators sometimes respond to high-stakes testing by focusing their efforts only on particular students. This may mean that only low readers receive instruction

or assistance while average learners get relatively little teacher time. The association also stated that students at certain "critical" achievement levels are sometimes targeted for intervention at the expense of other students. For example, "sometimes there is an attempt to raise test scores by focusing instructional initiatives on those students scoring just below cut-off points, and ignoring those above or far below cutoff points" (p. 259).

McDonnell, McLaughlin, and Morison (1997) described inequities for students with disabilities. First, since most high-stakes tests were not standardized with the needs and characteristics of the disabled in mind, these tests can be quite unfair for this population. Second, high-stakes testing is based on the premise that all children can achieve to high academic standards. This argument, however, is not based on fact or research. "The expectations of those advocating standards-based reforms currently exceed the limits of existing professional practice and expert knowledge" (p. 64).

Reigeluth (1997) stated the problem succinctly:

> Because students differ greatly in ability (ranging from severely learning disabled to highly gifted), as well as in mastery of learning skills, prior knowledge, home environment, and so forth, it seems likely that standards that are challenging for some students will be easy for others. Therefore, uniform standards cannot be uniformly challenging (rigorous) for all students. (p. 203)

In a 1997 study of nine states and 25 school districts considering or implementing standards-based reforms, Massell, Kirst, and Hoppe reported that equity issues had not been carefully considered. Special educators in most of the states reported that they had been only marginally involved in the development of content standards and even then their participation was relegated to the latter stages of the process. The authors argued that students who are homeless, poor, hungry, disabled, and those who live in violent neighborhoods will need policies to address their needs if they are to have equal opportunities to learn.

Pedroza (1997) corroborated the lack of attention to equity issues when considering the allocation of resources to educational programs and when comparing student outcomes. Simply put, accountability testing programs focus almost exclusively on the comparison, reward, or condemnation of schools. They rarely concern themselves with the need for equitable levels of staffing, with innovative programs, with the improvement of inadequate facilities, or with staff development to improve the effectiveness of teaching personnel.

Natriello and Pallas (1999) reported consistently negative consequences of high-stakes testing for secondary school students. Disproportionately large percentages of poor and minority students in Texas, New York, and Minnesota were reported to have failed their state's tests leading to graduation, leaving the potential to drop out as a realistic response. McDonnell et al. (1997) reported that students with disabilities dropped out of school at an alarming rate–some 38 % by grade 12 compared to about 12% for students in the general population.

Archer and Dresden (1987) described a new population of dropouts as students who have not done well academically but--

> ...who have the willingness to stay in school and graduate but who do not meet the minimum standards. These students may become frustrated and quit school or they may remain in school only to receive a certificate of attendance. (p. 278)

While minimum competency tests are designed to ensure that students meet the minimum requirements for entry into the job market, the real requirements of complex work environments are unlikely to be captured in multiple-choice tests. “In fact...neither employability nor earnings are significantly affected by students’ scores on basic skills tests, while chances of employment and welfare dependency are tightly linked to graduation from high school” (Darling-Hammond, 1994a, p. 14). While high-stakes graduation tests do not necessarily reflect the real requirements of the world of work, they often have the effect of requiring all graduates to meet college entrance requirements (Jorgensen, 1999).

Positive outcomes for students have been described as well. New (1995) reported that teachers believed the Arkansas Minimum Performance Test (MPT) was an effective tool for measuring student performance in basic skills. Davison, Schleisman, Koeppen, Wu, and Kwak, (2001) reported that summer school programs for students who failed Minnesota’s Basic Skills Tests were effective in helping students to achieve their achievement levels on the MPT. Schleisman (2000) found that students who did not pass Minnesota’s Basic Skills Tests were receiving more instruction in basic skills.

Outcomes for Instructional Practices

High-stakes testing is frequently reported to have negative effects on instructional programs themselves. Herman and Abedi (1994) and Shepard (1991) concluded that teachers of low SES students feel more pressure to focus on test content in their teaching than do teachers of high SES students, thus limiting the breadth of curriculum taught to students from families with low incomes.

Madaus (1988) wrote that high-stakes tests drive teaching. Teaching to the test corrupts the test’s ability to accurately assess the skill domains it was intended to measure and we are, thus, no longer able to make meaningful inferences from the tests to the domains that concern us. While rising scores on high-stakes tests may make policy framers feel better, they do not necessarily signify increased learning on the part of students. Madaus summarized the effects of high-stakes testing on teachers’ practices and the curriculum: “When the teacher’s professional worth is estimated in terms of test success, teachers will corrupt the measured skills by reducing them to the level of strategies in which the examinee is drilled” (p. 40).

The International Reading Association (1999) echoed this sentiment: “Another problem is that high-stakes tests have a tendency to narrow the curriculum and inflate the importance of the test. Schools should address a broad range of student learning needs, not just the subjects or parts of subject areas covered on a particular test” (p. 259). Smith and Rottenberg (1991) also reported

curriculum narrowing as a response to high-stakes testing in Arizona. Corbett and Wilson reported the same phenomenon in their 1991 study of Pennsylvania and Maryland schools as did New in her 1995 study of Arkansas schools.

Cameron (1997) stated that the teachers in a (high-stakes) elementary school in New York had developed a bunker mentality in which professional survival was the driving force. A prevailing attitude among teachers was that if "low" students did not receive extensive drill and practice on basic skills, they would drag down test scores for the whole school.

Teachers who are pressured for test scores will make every adaptation possible to ensure higher scores. One adaptation is to pay particular attention to the format of the questions in high-stakes tests and then to adjust teaching and assessment accordingly (Madaus, 1988). Textbook companies have been quick to note the potential market for materials that correlate with the high-stakes tests of individual states and have produced mounds of material to satisfy the need. This wide availability of material, along with pressure to raise test scores, will inevitably lead to more student time being spent completing worksheets and workbooks designed to improve measured "skills" (Madaus, 1988). In fact, reading instruction in some schools and school districts is reduced almost exclusively to the reading of short passages followed by multiple-choice questions (Madaus, 1988). Smith and Rottenberg (1991) found that high-stakes testing encourages the use of instructional materials that resemble the tests themselves rather than materials teachers may have found valuable or appropriate.

Darling-Hammond (1991) reported that as the influence of testing has increased, the use of teaching methods that foster the development of higher-level thinking has decreased. The decline has been noted in the use of instructional methods such as student-centered discussions, research projects, laboratory work, and theme writing. Shepard (1992) offered the following opinion about the effect of high-stakes testing on the curriculum: "I argue that massive effort (in preparing for testing) has produced small gains in rote skills at the expense of teaching students to reason and apply what they have learned."

A 1990 study by Herman, Dreyfus, and Golan demonstrated the impact of high-stakes testing on teaching and learning. The researchers analyzed questionnaire responses from 85 teachers in a K-12 urban school district. Among their conclusions about standardized testing were the following: (1) Schools serving low SES students paid the most attention to improving test results. (2) Elementary teachers in the study spent several weeks each year in studying test-taking strategies and practicing for tests. Low SES schools spent more time in test preparation than did mid-to-high SES schools. (3) Teachers did not feel that the tests fairly reflected their skill in teaching. (4) Teachers felt there was a discrepancy between what they were being forced to teach because of tests and what they felt they should be teaching. (5) Teachers, particularly those working with low SES students, did not feel that tests helped their schools to improve.

The International Reading Association (1999) is clear in its condemnation of high-stakes testing: "Our central concern is that testing has become a means of controlling instruction as opposed to a way of gathering information to help students become better readers" (p. 257).

While supporters of high-stakes tests have pointed to the potential for increased student learning, studies have shown that teachers feel pressured to teach to the test (Darling-Hammond & Wise, 1985), thus leaving less time available for discretionary teaching and other learning experiences. In an 18-month observational study in two schools Smith, Edelsky, Draper, Rottenberg, and Cherland (1990) reported that high-stakes testing reduced the amount of time available for ordinary instruction. At least three to four weeks of school were observed to be devoted to test preparation, testing, and recovery from testing. Smith and Rottenberg (1991) also found that high-stakes testing reduced the time available for ordinary instruction. In one school, for example, the annual testing required about 18 hours. However, teachers in this school spent about three hours in test preparation for every hour actually spent testing.

High-stakes tests have been described as deficient in their ability to provide useful diagnostic information about students. School-wide assessments should have the capacity to continually inform teachers about the effectiveness of their instructional practices. Darling-Hammond (1994a) offered "Assessment should increase the overall amount of learning and good practice across all schools, rather than merely measuring how much of a nonexpanding pool of knowledge is claimed by different students and schools" (p. 18).

While the primary value of testing might be assumed to be the improvement in learning of individual students, testing in high-stakes environments is actually designed to evaluate and rank schools rather than to diagnose student strengths and weaknesses. In a 1995 study, New found that schools testing students' minimum competencies did not use obtained scores to implement intervention plans for low-achieving students, nor did they use this information to formulate future budgets to meet documented student needs.

Test use in a high-stakes environment differs from test use in a traditional environment in important ways, including the end use of test scores, district autonomy in scoring and interpreting tests, and the consequences associated with test scores (Airasian, 1987b, 1988). "Approaches to learning are not well supported by external testing programs that send secret, secured tests into the school and whisk them out again for machine scoring that produces numerical quotients many months later" (Darling-Hammond, 1994b, p. 496).

A variety of positive outcomes for instructional programs have been noted as well. New (1995) determined that teachers believed the Arkansas Minimum Performance tests gave teachers valuable information on low-achieving students. Some two-thirds of teachers said they used this information to modify curriculum and instructional priorities, and some 60 percent said the tests had caused districts to divert extra funds for student remediation. About 80 percent agreed that eighth graders who had failed the tests were being evaluated by assessment teams as required by the state.

Schieisman (2000) found that Minnesota's Basic Standards Tests (BST) had caused schools to identify students who might otherwise have "slipped through the cracks." The BST resulted in additional emphasis on the teaching of basic skills. She found that the tests had increased the availability of compensatory opportunities for students including after school programs, tutoring sessions, and summer school.

Outcomes for Schools

High-stakes testing and measurement-driven instruction can have subtle but powerful effects on the way schools make decisions and on the consequences they experience. These effects are often determined by the resources that are available to support education in individual communities. In a 1999 study of Virginia school achievement on SOL tests, Wilkins concluded that four demographic opportunity structures predicted between 34 and 54 percent of the variance in school success on SOL tests. His research indicated that students in communities whose members are largely financially secure, predominantly white, well-educated, and who live in close proximity to large metropolitan areas are likely to do well on SOL tests, and their schools are more likely to attain state accreditation. Conversely, students from communities whose members are poor, minority, and less educated are likely to do less well on SOL tests, thus making it more difficult for their schools to become accredited. Wilkins concluded his study by stating, "With the inequities that have been shown in this study to exist across demographic opportunity structures, it is unreasonable to accredit schools based solely on their success on the SOL tests" (p. 18).

When pressure to increase test scores is high enough, schools may look for ways to reduce the test participation of students who are most likely to produce low scores–those who have learning handicaps, who have limited English proficiency, and those from educationally disadvantaged environments (Allington & McGill-Franzen, 1992a, 1992b). The same authors described school practices such as delayed entry into kindergarten, retention, assigning students to "transitional" classes, and assigning more students to special education classes as strategies used to reduce the numbers of "marginal" students who will be tested. For example, when Allington and McGill-Franzen (1992b) studied "significantly improving" schools in New York, they found that many of the cohort group for testing grades were never tested at all, making the test scores of improving schools appear to be much higher than they actually were.

In a study of New York elementary schools, McGill-Franzen and Allington (1993) found that from 5 percent to 25 percent of students were classified as handicapped by the end of second grade and in one school district, "about 15 percent of the children had been identified as handicapped before entry into kindergarten" (p. 21). Fries noted in her 1998 study of the state-mandated testing program in Illinois that many students with disabilities were routinely excluded from the state testing program to increase school-level test scores. Because these students' actual achievement levels are never reported., their learning characteristics and needs could not become a part of fiscal or programmatic planning.

Pedroza (1997) and Allington and McGill-Franzen (1992a, 1992b) reported that schools often use the strategy of removing students from the test cohort to increase their scores. Strategies used to "hide" these students included retention, assigning students to special education classes, and exempting students who are limited English proficient (LEP).
Pedroza reported the effects of these strategies in Texas. In 1994, less than seventy-two percent of the approximately three hundred thousand fourth grade students in Texas actually took the reading portion of the Texas Assessment of Academic Skills (TAAS). Pedroza "concluded that as fewer students took the test, scores in Texas increased" (p. 41).

High-stakes testing may also result in deception or in outright cheating. Deception in testing occurs whenever schools or school divisions deliberately hide certain groups of students from the testing process by retaining them, assigning them to special education classes, enrolling them in transitional grades or classrooms, or any other actions designed to reduce the percentage of the actual cohort group that will be tested. For example, one middle school in Virginia cited for its improvement status in 1999 tested only about 82 percent of its students compared to a statewide average of over 95%. This phenomenon was also described by Pedroza (1997), Fries (1998), and by Allington and McGill-Franzen (1992a, 1992b).

High-stakes testing and measurement-driven instruction have been shown to reduce schools' abilities to find and solve school-level problems concerning teaching and learning. Newman, King, and Rigdon (1997) reported that strong external accountability systems have been associated with low organizational capacity in schools. That is, strong external accountability systems tended to produce schools with low capacity to evaluate needs and to invoke school-level interventions. Similarly, Richards (1988) argued that governmental entities ignore the capacity of low performing schools and districts to react to sanctions because leadership, money, and talent are not distributed equally among school divisions.

Outcomes for Public Confidence in Teachers and Schools

Though one objective of the Virginia Department of Education's Regulations Establishing Standards for Accrediting Public Schools in Virginia (July 28, 2000) is to foster public confidence, evidence of the effectiveness of the SOA in achieving this goal is difficult to find. In a recent *Washington Post* poll, Matthews and Benning (2000) reported that 51% of respondents to a statewide poll said that the SOL testing program is not working. When asked what should be done about the tests, 43 percent said they should be changed substantially, and 21% said they should be ended entirely.

National opinion poll information is also enlightening. Rose and Gallup's (2000) *32nd Annual Phi Delta Kappa/Gallup Poll of the Public's Attitudes Toward the Public Schools* indicated what all preceding PDK polls have suggested, that people generally like their local schools. In fact the authors offered that support for public education is approaching an all-time high. Rather than indicating public dissatisfaction with the quality of teachers, schools, or student achievement, respondents ranked the lack of financial support for public schools as their biggest concern followed by lack of discipline, overcrowded schools, fighting and violence, and use of drugs. Only five percent of respondents listed a concern for standards or quality as their highest concern. The same poll indicated that the vast majority of respondents (69%) thought that the primary purpose of public schools is to provide a balanced education rather than to teach basic subjects.

Similarly, another national poll in 2000 sponsored by Sylvan Learning Center and the Association for Supervision and Curriculum Development (Moore, 2000) was one of the first to ask parents' opinions about state-mandated testing. The growing state-level demand for high academic standards was not supported by most parents who responded to this poll. More than

half did not believe that mandated achievement tests accurately measure what their children have learned (52% of respondents in high-stakes states and 60% in states without high-stakes testing). Most parents (57%) believed that important skills are being neglected because of the emphasis on state-mandated curricula.

The first three years of SOL testing in Virginia have been accompanied by extensive media coverage of the test score performance of individual schools and school divisions. Newspapers regularly indicate which schools have passed, which have improved, which have experienced declines in student achievement on the tests, and which will be accredited schools. Superintendents and principals are quoted about why their schools scored as they did. Those representing high-scoring schools offer reasons for their success while those representing low-scoring schools offer plans for improvement the following year. Though such publicity undoubtedly affects public perceptions of schools and teachers, information is lacking to document how these perceptions have been interpreted by Virginia teachers.

Positive outcomes for public confidence have been noted as well. Noll (1999) interviewed Minnesota citizens (including educators) to determine their views about the impact of Minnesota's Graduation Standards (MGS). The author found that over 76 % felt that the MGS would prepare Minnesota's youth for the work force. Some 90 % of government leaders shared this conviction compared to about 60 % of educators. The state emphasis on both basic skills and high standards was viewed as a strength by about 80% of respondents. Young (1996) found that Tennessee superintendents and state officials thought that the Tennessee Value-Added Assessment System (TVAAS) would improve students' achievement scores.

Outcomes for Teachers

Research literature is clear in reporting that teachers in high-stakes environments suffer undue pressure, stress, and loss of satisfaction in their jobs. Herman and Abedi (1994) and Shepard (1991) concluded that teachers of low SES students feel more pressure to improve test scores than do teachers of high SES students. Smith (1991) also found that standardized tests took a toll on teachers. Teacher respondents in her study reported anxiety, pressure, shame and embarrassment resulting from the publishing of test results in the media. They also reported guilt that student test scores were not higher and expressed the concern that published test scores did not accurately represent the quality of their teaching. Corbett and Wilson (1991) found teachers in Maryland (a high-stakes testing state) were under greater stress than teachers in Pennsylvania where testing stakes were less significant.

Koretz, Baron, Mitchell, and Stecher (1996) reported that Kentucky teachers experienced high levels of stress and a diminishment of morale as a result of that state's accountability testing program. Similarly, Smith and Rottenberg (1991) concluded that external tests negatively affected teachers. Teachers in their study reported feeling shame and embarrassment if their students scored low on district tests or if they did not achieve improvement benchmarks set by the school district. Fish (1988) concluded that pressure to achieve higher test scores through questionable test preparation activities caused teachers to feel guilt, anger, and low professional esteem. Teachers reacted negatively to the pressure associated with the public display of school and

classroom scores on high-stakes tests. Similarly, Young (1996) found that the Tennessee Value-Added Assessment System contributed to teacher stress and lowered morale.

Measurement-driven instruction can affect teachers' abilities and willingness to make instructional decisions. Madaus (1985) and Darling-Hammond and Wise (1985) opined that high-stakes testing requirements remove important powers of decision-making from classroom teachers. Corbett and Wilson (1991) found that teachers in a high-stakes state experienced a diminished reliance on their own professional judgement and felt greater discontinuity between what they were required to teach and what they thought should be taught. Smith and Rottenberg (1991) reported that teachers in high-stakes schools were reluctant to use their own professional judgement in developing and delivering curriculum, emphasizing, rather, what they felt would be assessed. Pedroza (1997) found that high-stakes testing resulted in a reduction in teacher decision-making in the instructional process. Koretz et al. (1996) reported that about half of elementary school principals and a third of middle school principals had moved teachers to and from "test target grades" to increase the likelihood of higher scores on Kentucky's KIRIS tests.

Darling-Hammond (1991) expressed concern that publicity and sanctions associated with low test scores may drive principals and teachers from low-scoring schools to other environments where rewards are more easily attained. Daring-Hammond (1994a) stated:

> These policies further exacerbate existing incentives for talented staff to opt for school placements where students are easy to teach and school stability is high. Capable staff are less likely to risk losing rewards or incurring sanctions by volunteering to teach where many students have special needs and performance standards will be more difficult to attain. (p. 15)

While schools may engage in unethical practices to hide students from testing, perhaps no form of cheating is less palpable than when educators purposefully act to change test scores directly. Examples of overt forms of cheating are readily available in the media: The June 10, 1999 edition of the *Courier-Journal* in Louisville, KY reported that "The award-winning principal of Bell County High School has been stripped of his administrative and teaching certification for 18 months in the most severe penalty ever issued for violating state school-testing rules." The school board found the principal guilty of encouraging teachers to assist students on the state's proficiency test. The June 24, 1999 issue of the *Chicago Sun-Times* reported the firing of an elementary school principal for improperly recruiting high-scoring students for her school and for refusing to enroll wards of the state. High-scoring students from 12 other schools were sent letters stating, "Congratulations. You've been accepted at Clay School," when the students had not applied there. The July 6, 1999 edition of the *New York Daily News* reported the attempt by the school board to fire a Bronx principal for allegedly helping students cheat on tests. The principal reportedly gave teachers answers to test questions and urged them to make sure students scored well. On December 8, 1999, the *New York Times* reported that "dozens of teachers and two principals across New York City's public school system had given students the answers on standardized reading and mathematics tests that help determine how schools are ranked and whether students move on to the next grade." And, in a recent incident reported on June 3, 2000, in the *Roanoke Times & World News*, an elementary principal from Montgomery County,

Maryland resigned after protests from parents that students at her school had been given help in completing standardized tests. On July 30, 2000, the *Roanoke Times & World News* also reported that a Virginia teacher had resigned for allegedly coaching her students on the state's SOL tests by using questions from the previous year's tests.

The lesson from these reports seems clear. When stakes are high enough for educators– when jobs may be lost, when careers may be destroyed, when licenses may be revoked, when the risk of embarrassment is high enough, when publicity is sufficiently critical, when pressure from superiors for increased scores is pronounced and continuous, some educators will engage in deception or dishonesty to increase test scores.

Individual Research Studies

Corbett and Wilson

Corbett and Wilson (1991) conducted an extensive study of the intended and unintended consequences of high stakes testing in local school divisions. The authors noted Airasian's (1987b) claim that while standardized testing once served the general purposes of evaluating instructional programs and of determining where they might need to improve, a new purpose has emerged. The (new) use is most correctly termed *state-mandated certification testing* (Airasian, 1987b, p. 403). "In this approach, testing is not used to guide classroom instruction or to monitor educational policy. Rather, state-mandated certification testing has made testing and test results a crucial aspect of educational policy itself" (Airasian, 1987b, p. 403). Corbett and Wilson undertook their study to assess the degree to which state-mandated testing causes schools and school districts to change how they approach decision-making and policy development.

Corbett and Wilson selected Pennsylvania and Maryland for the study. Pennsylvania represented a relatively "low stakes" testing environment, while Maryland was judged as relatively "high stakes," requiring students to pass tests in reading, writing, math, and citizenship to graduate from high school.

The study had three phases. First, preliminary qualitative field work was conducted in which the researchers spent several days interviewing a wide variety of staff members. In the second phase, the results of these interviews were used to construct a questionnaire that would be used in the districts to be studied. Third, the interviews were used to structure a second round of qualitative interviews to check the validity of the survey. Only phases one and two were directly pertinent to my research.

In phase one of the study six sites were visited by experienced field workers in each of the two states. Over 250 educators (central office administrators, principals, teachers) and students participated in the interviews which included such topics as the local testing program, state testing program, standards, local consequences, comparisons, and curriculum revision patterns.

In phase two of the study, interview responses were used to develop three major questionnaire domains that were believed to impact heavily on the context and results of testing in

the states and school divisions. These were *organizational adjustments, technological adjustments, and cultural adjustments.*

Organizational adjustments included perceptions about the flow of information concerning the testing program and the degree to which organizational members saw the testing program as being useful for system accountability. Items constituting this domain described the degree to which districts communicated with constituents. These items were used to determine whether parents and teachers were told how their children performed on tests, whether districts had informed staff about test content, and whether the tests were a topic of discussion. Items assessing the use of tests for accountability were used to ask respondents whether test scores were used to compare the performance of individual classrooms, whether test results were used to compare schools, and whether test results were used to compare school districts.

A review of responses indicated that there was a considerable flow of information in both states. Most respondents felt the communication statements were probably true or definitely true. However, with respect to the accountability items, Maryland respondents reported more frequent use of tests as a measure of accountability than in the low-stakes state of Pennsylvania.

The testing programs in the two states are summarized in Table 3. Means were computed for each respondent by combining the six information flow items into a single scale and the four accountability items into another. An analysis of variance was conducted to make comparisons between the two states. An F-value was computed with the larger the F-value, the greater the variance in responses between the two states (Corbett & Wilson, p. 59). The authors combined responses from the three major respondent categories: central office administrators, building principals, and teachers. No significant differences were noted among responses from the three groups. Mean scores for information flow were almost identical in the two states. The mean scaled scores for the information items were noted to be very high, 4.5 on a 5.0 point scale. These data indicate that tests and test scores were talked about equally (and quite frequently) in the two states. However, information from the accountability scale showed significant differences between the Maryland, the high-stakes state, and Pennsylvania, the low-stakes state. Maryland educators reported more frequent use of tests as a means of making judgements about the effectiveness of local schools. One Maryland teacher summarized the emphasis in this way: “The tests have become the focus for judging the total system” (Corbett & Wilson, 1991, p. 60).

Table 3

Analysis of Variance Comparison of Organizational Adjustment Scores for Pennsylvania and Maryland (N=1038)

	Mean			
Adjustment	PA	MD	F	Scale
Information flow	4.52	4.51	0.02	1.00 to 5.00
Benchmark	2.00	2.89	104.7*	1.00 to 5.00

* Indicates significance beyond the .001 level. Note: Scale scores of 5 indicated high respondent agreement that there had been much information flow about testing and much discussion about how districts, teachers, and schools had fared in reaching "benchmark" scores. Scores of 1 indicated little discussion about testing and little emphasis on the comparison of districts, teachers, and schools.
Note. From Testing, Reform, and Rebellion (p. 59), by H. D. Corbett and B. L. Wilson, 1991, Norwood, N. J.: Ablex Publishing Corporation. Reprinted with permission.

Technological adjustments referred to "the set of practices and knowledge bases through which a school district is able to accomplish its work"(Corbett & Wilson, 1991, p. 66). The authors paid attention to three clusters of items representing this category: (a) strategies used by districts to improve test scores, (b) changes in the content of courses and in instructional activities (curriculum and instruction) to correspond to test objectives, and (c) the magnitude of changes due to high-stakes testing. Items used to assess this category asked whether students took a practice test or tests before taking the state test, whether test content was reviewed prior to taking the test, whether staff development had emphasized test score improvement, and whether special efforts had been made to get lower-scoring schools to improve their test scores.

Changes in curriculum and instruction included items related to the degree that adjustments had been made in course content and teaching practices. Items used for this purpose asked whether the test was used to determine when instructional objectives and content needed to be added, whether students should be placed in instructional groups, and whether students should be placed in homogeneously grouped courses or classes.

Items assessing the magnitude of change that had occurred in schools and school divisions due to high-stakes testing included those asking whether teachers had altered class content, whether testing had caused teachers to implement new teaching strategies, and whether basic skills instruction had been added to the curriculum. Two final items assessed the perceived magnitude of changes by asking if the curriculum had been narrowed and if the curriculum had been improved.

The following differences were found between Maryland (the high-stakes state) and Pennsylvania (the low-stakes state): (1) Educators in Maryland made extensive use of activities to

prepare students for tests while those in Pennsylvania made little such use. (2) Other district-wide strategies such as emphasizing staff development and work with under-achieving schools were more evident in Maryland than in Pennsylvania. (3) The educators in high-stakes Maryland reported more changes than did those in low-stakes Pennsylvania on items dealing with student placement, changes in class content, exposure to new ideas, new instructional methods, and emphasis on basic skills instruction. (4) With regard to the single item on narrowing the curriculum, there was a notable difference between the two states. In Pennsylvania, two-thirds of the respondents indicated that there was no change in emphasis to improve test scores, while in Maryland only one in every seven respondents indicated no change. The data indicated that Maryland educators were devoting much more time to improving test scores than those in Pennsylvania. The degree of curriculum change between the two states was of moderate magnitude.

Cultural adjustments are the ways individuals, organizations, and constituents believe about issues. The questionnaire items that dealt with cultural adjustments measured student worklife and teacher worklife. Student worklife was measured by statements that assessed how serious students were about their studies, whether teachers had more empathy for low-achieving students, and whether staff members were more informed about the needs of low-achieving students.

Teacher worklife was sampled with statements that assessed whether educators' professional judgement had been devalued, whether teachers' time demands had increased, whether teachers and administrators had been reassigned, whether staff members felt pressured to improve test scores, and whether staff members were worried about lawsuits. Mean scores for each respondent were determined by combining student worklife items in one scale and teacher worklife items in another. Educators in Maryland reported that state-mandated tests had resulted in a greater impact on teachers' and students' worklives than did educators in Pennsylvania (see Table 4).

Table 4

Analysis of Variance Comparison of Culture Adjustment Scores by State (N = 346)

	Mean		
Adjustment	PA	MD	F
Student Worklife	1.29	1.48	6.3**
Teacher Worklife	0.81	2.17	152.2***

**p ≤ .01 ***p ≤ .001 The scale was 5.0 = total change, 4.0 = major change, 3.0 = moderate change, 2.0 = minor change, 1.0 = no change.
Note. From Testing, Reform, and Rebellion (p. 85), by H. D. Corbett and B. L. Wilson, 1991, Norwood, N. J.: Ablex Publishing Corporation. Reprinted with permission.

Summary of findings:

In comparison to Pennsylvania respondents, those from Maryland reported (1) greater alteration of their curriculum, (2) greater use of test scores for comparison purposes both within and outside the school district, (3) greater proportions of students who were serious about school, (4) teachers who were under greater stress, (5) teachers who had more paperwork, (6) teachers who experienced a diminished reliance on their professional judgement, (7) a curriculum that had become more narrowed (and improved), (8) greater emphasis on test scores than on learning, and (9) greater discontinuity between what was being emphasized and tested and what educators thought should be taught.

Smith and Rottenberg

Smith and Rottenberg (1991) studied Arizona's high-stakes testing program which included the use of the reading, language, math, and materials subtests of the Iowa Tests of Basic Skills (ITBS). required by the state. Other district-required tests which were the Basic Skills Test (BST) and the Continuous Uniform Evaluation System (CUES) which was given periodically with results reported to the district office three times each year. While the ITBS is a normed instrument, both the BST and the CUES are criterion-referenced instruments. The stakes of these tests were high in that they contributed to the evaluation of principals and to curriculum decisions made at the district level.

The authors reported a case study of two elementary schools in Arizona. The schools were similar in that both served low SES students, both had achievement scores that ranked at the bottom of the school district, and both had four classes per grade in kindergarten through grade six. Each was different, however, in terms of instructional and curricular contexts. One school, Hamilton, had a program-centered, phonics-based language arts program while the other, Jackson, had a student centered, literature-based language arts program.

The researchers studied two classrooms at each grade level in each school with periodic observations from August to December. Observations were recorded by detailed notes which were later transcribed into scripts. The researchers conducted clinical interviews with 19 of the 49 teachers who had been observed. Open-ended questions such as the following were used: "What are some of the things that go through a teacher's mind when she/he sees the scores on the ITBS from her/his class?" (Smith & Rottenberg, 1991). Such questions are neutral and should not have led the respondents to provide any particular answer. At the same time, they provided important information about teachers' opinions that could not be obtained in other ways. In addition to observations and interviews, the researchers attended pupil placement meetings and staff meetings.

From the teachers observed and interviewed, four were selected for intensive observation over the spring semester–one second and one sixth grade teacher from each of the two participating schools. The criteria for the selection of teachers for this phase of the project was their willingness to be subjected to the required level of intense scrutiny. The researchers observed each class once, twice, or three times per week for a total of 81 days of observation. Staff meetings and schoolwide activities continued to be observed periodically. The researchers

interviewed district and building administrators, and perused many documents, including curricula, tests, and test results. Following the collection of data, themes or principles were deduced. Themes derived were checked against the views held by participants.

The following perceived consequences of externally generated high-stakes testing were listed and explained (Smith & Rottenberg, 1991, pp. 8-11):

(1) *External testing reduced the time available for ordinary instruction.* The required testing programs were served on top of an already crowded curriculum. In Hamilton, for example, testing itself required about 18 hours. However, teachers also spent about 3 hours of test preparation for every hour that students actually spent testing. In addition, teachers provided "recovery time" following the tests due to their perceived stress of students and because the teachers felt they needed to reward students with time off to ensure their best performance (effort) on the test. The time spent on these activities in total exceeded 100 hours, or approximately 3 to 4 weeks of instructional time.

(2) *Testing affected what elementary schools taught; in high-stakes environments, schools neglected material that external tests exclude*. Teachers in these schools tended to spend their instructional time on "testworthy" activities–that is, they taught what they knew would be tested on state and district-mandated tests. Arithmetic computation, word recognition, comprehension of short passages, and the recognition of spelling errors were emphasized over other potential learning activities such as reading trade books, participating in long-term, integrated projects, composing original writing, or engaging in problem-solving activities.

The authors cited the example of a sixth grade teacher whose teaching specialty was writing. From September through December he had devoted about 40 minutes per day to creative writing activities including the composition of poetry, essays, narratives, and personal journals. From January to May (the time before the testing program), time devoted to writing diminished to nothing while approximately 40 minutes a day were spent on grammar exercises and memorization of rules of usage. Science instruction, similarly, went from hands-on investigative laboratory experiences in the fall to memorization of science facts from the text book in the spring. Both changes in instruction were clearly undertaken to improve student performance on upcoming standardized tests.

(3) *External testing encouraged the use of instructional methods that resemble tests*. Especially at Hamilton, the perceived need to improve ITBS test scores resulted in the practice of creating many opportunities for students to take school-generated tests that were virtually indistinguishable from the ITBS itself. Unfortunately, this practice totally eliminated other possibilities for instructional reform which might have included more or different kinds of staff development for teachers, the provision of more time for students to read or write, the employment of teacher aides to assist students, and so forth. What resulted were test-driven instructional improvements that grew and prospered without any public discussion or debate as to their merit.

(4) *External testing affected school organization by placing general boundaries on placements and instructional opportunities.* Students of fourth grade age who were learning disabled or who scored "below grade level" were placed in a transitional fourth grade based on test scores alone. This decision not only delayed students' eventual graduation by a year, but also prevented them from having access to enriched curricula and the interaction of higher-achieving peers. It also eliminated them from consideration for gifted programs, and guaranteed that they would be placed in low-level tracks at the highly stratified junior high school.

(5) *By teachers' definitions, external testing negatively affected pupils.* Teachers reported that tests caused a variety of worrisome effects, including test anxiety, the development of "failure mentality," misbehavior and fighting, psychological distress, stress, and even physical illness.

(6) *External tests negatively affected teachers.* Teachers reported feeling shame and embarrassment if their students scored low on district tests or if they did not achieve improvement benchmarks set by the school district. They also reported feeling relieved rather than proud when their students posted high test scores because they felt that student achievement was mostly a matter of student effort and that their contributions to eventual test results were minimal. In this environment teachers and principals continually looked for other ways to demonstrate their competence as evidenced by alternative measures of student growth.

Smith and Rottenberg ended their report by stating: "Our research report shows that mandated testing programs [also] have consequences that are both problematic and contrary to the general goal of improving schools" (p. 11).

Allington and McGill-Franzen

Richard Allington and Anne McGill-Franzen, professors of reading at State University of New York, Albany, conducted a study in New York State (1992a) that focused largely on the issue of retention as an unintended consequence of high-stakes testing. The researchers cited Gottfredson (1986) who found that school divisions who seemed to have achieved significant gains in minimum reading competency for their students–from 70 percent to 90 percent by one large school division over a five-year period–had not improved students' reading prowess at all. The deceptive achievement claims had been masked by the district's retention policies that had only 60 percent of the original grade-level cohort group taking the reading competency test on time.

The authors cited Walker and Levine (1988) who studied the relationship between retention and student achievement in an urban elementary school. These researchers found that retention increased measured student achievement in both the retention grade and the promotion grade. However, as the gains had been artificially achieved when compared to the actual cohort, there was no indication of an actual increase in the reading proficiency of any students in the primary grades, nor was there any indication of increased school effectiveness.

The study was focused on the New York State public education system due to its high-stakes characteristics. Special emphasis was placed on Grade 3 reading achievement test scores as

students in this grade were targeted for remedial instruction when they scored poorly on the required state reading test. Since 1985 New York school divisions have been required to release test score information to the public in accountability profiles each fall. These releases have been accompanied by much media attention as school rankings are compared widely (Allington & McGill-Franzen, 1992a, p. 3). The combination of public test scores, school and district comparisons, and required remediation has moved New York's third grade testing program in reading into the realm of high-stakes testing.

Allington and McGill-Franzen sought to determine if school passing rates on the third-grade reading assessment were accurate in assessing the school's instructional program. In other words, did the level of test scores accurately portray the level of student achievement? To answer this question, the authors studied seven schools to determine how their students' achievement on state accountability tests had been shaped by forces *other than* the instructional quality of the schools–namely policies affecting retention and special education.

Since 1980 New York State had been experiencing a slow but steady increase in the percentage of students scoring at or above the minimum competency level set by the state. At the same time, the proportion of students identified with disabilities in the primary grades showed a significant increase. In fact, the proportion of students identified as disabled in New York State schools almost doubled in the 15 year period prior to the study, jumping from six percent to over ten percent of the population of students (Allington & McGill-Franzen, 1992a, p. 4). The level of increase was most dramatic in grades K-2, the period just before students would have to take the mandated grade 3 reading test. These circumstances led to the specific research question: "How does a school's policy on retaining students or identifying them as handicapped influence the reading achievement levels reported by that school for high-stakes testing" (Allington & McGill-Franzen, 1992a, p. 4)? The task of the authors was to determine the reading achievement of members of the original kindergarten cohort in each school and then to estimate the degree to which retentions and special education placements had affected the third-grade reading achievement scores actually reported on the required state assessment profile.

The researchers began by locating the original kindergarten cohort lists for each school and reviewing student lists for evidence of special education placements and retentions. They recorded the reading achievement scores for each student according to whatever standardized reading test the student took at the end of their fourth year in school. For most students this was the statewide Degrees of Reading Power Test administered in grade 3.

Comparability between the state reading test and the various other standardized tests was estimated by equating the third-grade minimum competency level with the 26th percentile ranking for a third grader and a 61st percentile ranking for a second grader. The authors observed that when examining four other standardized group reading tests, it was noted that any student tested at the end of the second grade year who achieved at the 61st percentile would earn a grade equivalent score (GE) of 3.0 or higher.

Analysis of data produced widely varying results. One school identified as the Orton School had a practice of promoting almost all students. Of the 53 children represented in the

kindergarten cohort, only two were removed from the tested cohort group; one child had been identified as learning disabled in third grade, the other had been retained in kindergarten. Parents in this school were solidly middle class in an established community with upscale homes. Fewer than five percent of Orton children qualified for free or reduced price lunches. The school had many ancillary programs to help students, including parent volunteers, teacher tutoring outside of school hours, and a reading teacher who helped children. Virtually all of Orton's children reached the required achievement level on the state-required reading test in grade 3. Because the retained student achieved an end-of-second-grade reading score of 3.3 (GE), the authors concluded that only one student would have likely not passed the state-required reading test in third grade. This would have indicated a passing rate of 98%.

At the other end of the spectrum was Riverton Elementary School. This school served a mixed community of middle class and working class families who worked in both skilled and semi-skilled jobs. Housing quality varied from a mix of older homes and apartments to trailer parks to middle class and even some expensive dwellings. The district served relatively few poor children; only about 10 percent are eligible for free or reduced price lunches. However, Riverton was markedly different in its use of retention practices. In one recent year, some 35 percent of students in first grade were retained. This practice reflected a long-established history of having some children spend an extra year in a large "prefirst" traditional program. The program was ostensibly for children who needed another year to grow and mature and for those whose literacy skills were delayed or low (Allington & McGill-Franzen, 1992-a, p. 7). Some children were also retained in second grade, and some were identified as handicapped when their achievement did not keep pace.

Student achievement at Riverton had been improving over recent years. Some 96% of its students were reported to have achieved the grade three reading standard. However, 34 of 68 kindergarten students were either assigned to pre-first grade classes, were retained elsewhere prior to grade 3, or were identified with disabilities before grade 3. So, a full 50 percent of the original kindergarten class of 68 were no longer with their cohort and were not tested on the mandatory grade 3 reading assessment. Given this information, the authors conservatively concluded that the actual passing rate was no greater than 78% and was likely to have been even lower as some of the 11 children who moved out of the area were not likely to have passed.

This pattern was repeated at the Beech Street School where only 18 of the original kindergarten cohort "survived" to the third grade to be tested on the mandated grade 3 reading test. Twenty-three of these students had been assigned to a pre-first grade class after kindergarten (so only 29 went to first grade). By the end of their second year in school, 20 percent of the cohort had been assigned to special education and three children failed after being promoted to first grade. Shockingly, the school district reported that 98 percent of the third grade students at Beech Street passed the mandated grade three reading test, but only a third of the original cohort group was there to be measured. Allington and McGill-Franzen estimated the actual percentage of the cohort group who would have passed the test at 63%.

Research findings are summarized as follows:

(1) The "snapshot" findings of student achievement levels as presented in New York's accountability profiles are subject to interpretation. Schools that made heavy use of retention and placement in special education classes reported significantly higher percentages of students who passed the state's grade 3 reading test than would have been the case had the retained or "labeled" students been tested with their cohort. (2) Instead of rewarding legitimate instructional quality, publicized school rankings of high-stakes tests often reward questionable pedagogical practices. (3) Responding to the demands of high-stakes tests via retentions and special education placements may actually meet the needs of schools rather than the needs of the children they are supposed to serve. (4) The responses of retention and special education placement to meet the needs of at-risk children cost more. When considering the actual educational cost per student of the seven schools profiled in this study, it is estimated that the additional cost of keeping a child in school another year was over $6,000. Additionally, serving students with learning handicaps typically costs from 1.5 to 2.5 times the cost of regular education programs (Allington & McGill-Franzen, 1992a, citing LINC Resources, 1989).

Finally, the authors proposed that if accountability practices such as student retention and special education placements are consistent responses to the demands of high-stakes testing, then testing of this type cannot be considered to be a viable method to improved student achievement.

New

New (1995) conducted a study that evaluated the effects of Arkansas' mandated testing program known as the Minimum Performance Test (MPT) on teaching and learning. New sought to determine the outcomes of the enabling legislation requiring high-stakes testing and graduation standards in Arkansas. Arkansas standards were considered high-stakes for three reasons: (1) Eighth grade promotion was predicated on successful completion of the MPT. (2) MPT scores were used as a proxy for school effectiveness and were reported widely in state newspapers, and (3) School districts with a passing rate of less than 85% were required to initiate and complete an improvement plan.

A survey instrument was developed to assess teacher perceptions of the MPT. The instrument included 23 statements that required response in a Likert-type, four choice format. Statements were designed to measure five domains: (1) the effect of the MPT on students and learning, (2) the effect of the MPT on teaching, (3) the effect of the MPT as a diagnostic tool, (4) the effect of the MPT on school districts, and (5) the effect of the MPT on remediation. Four open-ended questions were also asked to help further explain teacher perceptions about the five domains. New reported that most teachers responded to the 23 statements, while fewer responded to the four open-ended questions. Descriptive statistics were used to analyze responses. A Chi-Square statistical analysis was used to assess the relationships between the responses and region and grade level taught. "Qualitative thematic analysis" (New, p. 54) was used to describe major response patterns in the open-ended data. Both quantitative and qualitative data were presented to advisory groups to check the accuracy and validity of conclusions reached.

New defined the population for the study as 11,000 currently employed teachers as listed by the Arkansas State Department of Education. A random stratified sample was selected to be surveyed from this group. Approximately 2,500 were chosen from grades three through seven,

and 1,000 were chosen from eighth grade. Of the 3,500 surveys mailed, 2,044 (58%) usable surveys were returned.

The following are selected findings from New's study: (1) Regarding students and learning: "Most (86%) of the teachers reported that the MPT does not improve learning for gifted students and 66% stated that it does not benefit average students" (New, p. 60). When asked to respond if the MPT motivates students to learn more than basic skills, 86% disagreed. And, 75% reported that the MPT did not result in positive outcomes for minorities. (2) Regarding teaching and learning: Two-thirds (67%) of teachers did not agree that the MPT had improved the quality of teaching in their school districts. A similar percentage (65%) reported that they no longer had time to teach higher level thinking skills while 60% said they spent too much time in activities that they classified as teaching to the test. (3) Concerning the MPT as a diagnostic tool: About two-thirds (64%) of teachers said that the MPT was an effective tool for measuring student performance in basic skills. Similarly, 61% of respondents said that the MPT provided teachers with valuable information about the needs of low-performing students. (4) Concerning the impact of the MPT on school district policy: Two-thirds (67%) of teachers said that their districts use MPT results to direct curriculum modification and priorities, and 59% reported that their districts provided extra funds to address the needs of low-performing students based on MPT results. (5) Concerning the impact of the MPT on remediation: Most (81%) respondents agreed that eighth graders who failed the MPT were being evaluated by an assessment team as required by the enabling legislation in the state. And, the same percentage of respondents replied that low-performing students are receiving adequate remediation from their school districts.

Koretz, Barron, Mitchell, and Stecher

Koretz et al. conducted a 1996 study entitled The Perceived Effects of the Kentucky Instructional Results Information System (KIRIS). KIRIS was one component of the Kentucky Education Reform Act (KERA). It was a high-stakes program for educators (though not for students) in that all schools are held to the same performance standards and schools are held responsible for increasing the percentage of students who reach achievement benchmarks. There were financial rewards for schools where KIRIS scores improved and planned sanctions for schools that did not improve.

This report was based on surveys of randomly-selected fourth grade teachers, eighth grade math teachers, and fourth and eighth grade principals across Kentucky. All groups were surveyed in the spring of 1994-95, in advance of the administration of the KIRIS, by means of telephone interviews and a lengthy written survey of teachers. The surveys looked at perceived support for KIRIS, changes in school organization and instruction, time spent in preparation for KIRIS, and teachers' views about the implementation of a new portfolio assessment program. Interviews were conducted with 216 teachers and 115 principals. Mailed surveys were completed and collected from 209 teachers.

The following were among the findings resulting from the study: (1) While a majority (about 60%) of math teachers and principals were supportive of KIRIS , about half of fourth grade teachers were opposed to the program. (2) About three-fourths of principals said that dealing with KIRIS had imposed more than a minor burden on their schools–though about two-thirds of principals said that it had become easier to accommodate the program after it had been in

place for several years. (3) Teachers reported that KIRIS had created high levels of stress and most agreed strongly that teachers had been put under undue pressure. (4) Most teachers responded that KIRIS had resulted in a diminishment of morale in their schools. (5) Though this was a program without high-stakes consequences for students, a significant minority reported that KIRIS had resulted in decreased morale among their students. (6) Teacher opinions were evenly divided on a major tenet of the program–that all students can learn at high levels. However, the same teachers felt this was the right message to send to Kentucky's students even if it was not realistic or attainable. (7) Only about a fourth of teachers supported the accountability component of KIRIS–again, in a program with relatively minor consequences for educators and none for students. (8) About half of teachers believed that KIRIS scoring standards were inconsistent over time, and that the curriculum content for the assessments had not been defined well enough to allow them to prepare students for the tests. (9) Principals and teachers strongly agreed that schools with high percentages of transient populations were at considerable disadvantage on KIRIS. (10) While about two-thirds of teachers reported that KIRIS has led to a change in expectations for student achievement, they also indicated that expectations had risen substantially more for high-achieving students than for low-achieving and special education students. (11) A substantial majority of teachers reported that their schools had found ways to improve test scores without improving the quality of education for students. Such strategies included increased time on test-taking skills and increased emphasis on the narrowed curriculum assessed on KIRIS. (12) About half of elementary school principals and a third of middle school principals reported moving teachers to and from target grades to increase the likelihood of higher KIRIS scores.

Noll

Noll (1999) conducted a qualitative study in Minnesota to determine how various stakeholders viewed the effectiveness of Minnesota's graduation standards. In 1993 Minnesota law required the development of a set of K-12 requirements known as the Minnesota Graduation Standards (MGS). The MGS required that students demonstrate competency in basic skills. Students must pass basic literacy tests in math, reading, and writing to graduate. Schools were required to provide continuous remediation to students who did not pass the MGS. Additionally, the MGS required students to complete competencies in various application skills. "The Graduation Standards require learning and assessment in which a student will demonstrate the ability to apply complex ideas and concepts in real-world situations" (Noll, p. 12). Many of these competencies were assessed when students completed products, performances, or portfolios. Application areas to be assessed included:

1. Read, view, and listen
2. Write and speak
3. Literature and the arts
4. Math applications
5. Inquiry
6. Scientific applications
7. People and cultures
8. Decision-making, and
9. Resource management (Noll, p. 13)

Noll's sample for face-to-face interviews included "homogeneous sample groups of influential people from business, government, and education" (Noll, p. 58) to determine if Minnesota's graduation standards were adequately preparing students for the world of work. A total of 48 participants were selected to participate in the study. These individuals were asked six open-ended questions which focused on their knowledge of the MGS initiative, their beliefs about the strengths and weaknesses of the MGS initiative, and their recommendations to improve the roles of Minnesota stakeholders (including businesses) in Minnesota K-12 education. All interviews were completed between July and October of 1998. The QSR NUD*IST 4 computer software package was used to help the researcher code and search the data for response categories. Response categories were given hierarchical titles, and related themes were assigned to each category.

Noll's study produced several findings germane to my research: (1) The majority of respondents (57%) thought the legislature was confident that the MGS would improve the quality of Minnesota's K-12 system of education. (2) A sizeable minority (37%) thought that the MGS would result in a "healthy competition" among schools and school districts. (3) Some 30% of respondents thought that teacher commitment was a positive factor in the implementation of the MGS, but 43% thought it was a negative factor. Noll speculated that, though teachers were supportive of higher standards, many felt the program was too "top down" and that more funds should be made available for funding and training. (4) Funding, in fact, became an important issue for a sizeable minority of respondents. Some 33% of interviewees expressed concern with the extra work associated with implementing standards and assessments. These respondents requested that more money, time, and training be devoted to implementation of the MGS. (5) Over 76% of respondents felt that the MGS would adequately prepare Minnesota's young people for the work force. The strength of this conviction varied from 90% of government leaders to only about 60% of educators. (6) The state emphasis on both basic standards and high standards was viewed as a strength of the MGS by 80% of respondents. (7) When asked about weaknesses of the MGS, 67% of respondents replied that teachers needed more time and resources to convert from the traditional (Carnegie units) system to a system based on standards and performance.

Young

In a 1996 study, Young assessed the opinions of Tennessee teachers, principals, superintendents, legislators, and department of education officials concerning the consequences of the Tennessee Value-Added Assessment System (TVAAS). The TVAAS used score gains from an annually administered norm-referenced test to determine student gains and teacher effectiveness. The purpose of the study was to determine if these groups varied in their opinions about the impact of TVAAS. The author used three hypotheses to guide the research: (1)There will be significant differences in the views of school district employees (superintendents, principals, and teachers) concerning the consequences of TVAAS compared to legislators and department of education officials. (2) There will be significant differences in how Tennessee principals, teachers, and superintendents view the consequences of the TVAAS. (3) There will be significant differences in how demographic variables affect how principals, teachers, and superintendents view the outcomes of TVAAS. Each hypothesis was tested with analysis of

variance. Eleven research questions guided the study. Information was gathered with a Likert-type questionnaire with eleven questions and three demographic variables: race, educational experience, and size of school.

Study findings were as follows: (1) The views of principals, teachers, and superintendents varied significantly about the impact of TVAAS. (2) The views of educators varied significantly from the perceptions of state legislators and from members of the state department of education. (3) Educators' opinions did not vary significantly based on race or size of school. (4) Over 85% of teachers did not believe that TVAAS was effective in assessing the impact of teacher behavior on student learning. (5) All groups agreed that TVAAS contributes to teacher stress and lowered teacher morale. (6) State officials agreed and educators strongly disagreed that TVAAS has a positive effect on instruction, curriculum, and student achievement. (7) Educators disagreed and state officials agreed that high and low achievers have the same opportunity to achieve on the TVAAS. (8) Educators disagreed and state officials agreed that the TVAAS is fair to all teachers. (9) While principals and teachers felt that TVAAS would not improve test scores, superintendents and state officials thought that they would.

Cameron

Cameron's 1997 dissertation made important contributions to the topic of high-stakes testing in an elementary school setting. The author conducted a single case study of a large urban elementary school in New York state. Everett Elementary is a large school, housing some 850 students in pre-kindergarten through sixth grade. The student population is approximately 42% white, 33% African American, and 25% other. Approximately 90% of Everett's students live at or below the poverty level as evidenced by subsidized lunch eligibility. Everett has historically demonstrated poor test performance on the New York State Pupil Evaluation Program tests (PEPs). During the three year period of 1987-1989 the school's reading scores were below the state goal for each grade tested (3^{rd}, 5^{th}, 6^{th}). The Everett faculty had been the target of pointed criticism from the district superintendent for low test scores. The superintendent had threatened to consolidate Everett with a nearby school and divide the students into primary and elementary levels–effectively closing Everett as it then existed–if scores did not improve. The superintendent suggested the adoption of a continuous improvement model as an acceptable intervention. Given the choice of the continuous improvement model or school reorganization, Everett's teachers chose the continuous improvement model.

The notion of continuous improvement in this context meant that children would remain with their age-level peers whether or not they have met grade level expectations. A selling point for this approach was that students would remain with their same teacher for two years in a row, thus reducing the amount of time necessary for the teacher to understand student strengths and weaknesses. Continuous improvement, if implemented, would effectively end the practice of retention. However, it might also lead to relatively lower test scores than in schools without this model because nearly all students would be tested with their age cohort rather than only a portion of the cohort–i.e., those who have not been retained.

The author's purpose initially was to identify problems related to teaching and learning in an urban elementary school and to do so from the perspective of students, teachers, and administrators. Eighteen questions were generated to guide the study–approximately half before the study began and half after it was underway. Various ethnographic strategies were used to gather data, including participant observation, in-depth interviews, document review, and key informants. Cameron focused on the classrooms of four cooperative teachers who allowed the researcher to observe and to gather extensive amounts of data. Cameron acted essentially as a classroom volunteer during the study period to enhance her acceptance and "invisibility." As working hypotheses were developed, the author requested feedback from the participants to determine their potential validity. Fourteen of seventeen teachers in the related "units" were interviewed as were all three building administrators, the school psychologist, the school social worker, and the PTO president.

Data were gathered and hypotheses were generated according to accepted ethnographic design. Interview themes were developed from topics clearly important to the teachers themselves. Activity codes were used to discover patterns (i.e., variation of ability, behavioral issues, teacher behaviors). Data were then coded further according to frequency. The researcher continually looked for major themes and eventually divided these into internal and external categories. Four major themes were developed that became the basis of the remainder of the study: (1) The continuous progress model instituted at Everett Elementary was poorly conceived and implemented. Teachers needed time and training to have a chance to succeed with this model, but they were given too little of either. (2) Many teachers could only see their students as "at-risk" students–rather than as students with potential. The perceived lack of parental support and perceived lack of student ability were telling in teacher attitudes and in resultant instructional practices. (3) The continued pressure for standardized test results was an obstacle to the development of potentially effective instructional strategies and practices. The teachers in this school had developed a bunker mentality in which the name of the game was survival. If "low" students did not receive extensive drill and practice on skills, they would drag down test scores. (4) The absence of a clear vision for reform in a high-stakes environment caused ability grouping to continue as the dominant instructional design for the school. The powerful pressures exerted for improved test scores led teachers to retreat to the instructional strategies they knew best. No time or incentives were allowed or provided to enhance the attractiveness of heterogeneous grouping in a continuous progress setting.

Rather than the forced reform described in this study, the author concluded that meaningful reform will not occur unless substantive changes are made in individual schools and at the level of the school division and of the state. Schools must meet the following conditions to enhance the probability of success of meaningful reforms (Cameron): (1) All stakeholders must engage in meaningful collaboration that results in "a collective vision of reform" (p. v). (2) All members of the school staff must share a vision that students can learn–that though they lack experiences, they can learn. (3) Teachers and administrators must be competent in their understandings and use of alternate instructional methodologies and assessments. (4) To accomplish the above recommendations, however, teachers and administrators must have relief from the stress of high-stakes testing. Teachers and administrators in this school were so occupied with deflecting criticism and threats that they did not have any reasonable opportunity to study,

discuss, or effect meaningful change. Success was defined as higher test scores and only in that manner.

The author concluded with several recommendations to improve low SES, culturally diverse, large urban elementary schools: (1) Teacher education programs should be implemented to train teachers to see the promise in all students and to recognize and avoid negative "teacher culture." (2) Schools must find ways to bring (and welcome) parents and community members into the public schools. (3) Districts must demonstrate to teachers that *they* value students by committing the necessary personnel and resources to under-resourced schools. These schools do not need equitable treatment; they need greater access to resources than privileged schools to have a chance to reach parity. Schools serving populations like those in Everett need greater access to staff development and assistance to help them transition from traditional to innovative methods of instruction. (4) Effective instructional leadership cannot occur in a school with over 800 students and a staff of 100. A smaller campus size would allow the principal the opportunity to engage teachers in cooperative decision making and curriculum development. It would allow the principal to be visibly engaged in classrooms and in interaction about student learning, and it would allow the principal the time to confront "unengaged teachers" (p. 229).

Massell, Kirst, and Hoppe

Massell, Kirst, and Hoppe (1997) led an extensive review of standards-based reform for the Consortium for Policy Research in Education in nine states and 25 districts during 1994 and 1995. The researchers orchestrated in-depth case studies in these locations with the help of some 15 on-site interviewers and observers. Over the research period the authors studied policy reform initiatives in California, Connecticut, Florida, Georgia, Kentucky, Minnesota, New Jersey, South Carolina, and Texas. District case studies were developed through a routine of field visits, telephone interviews, and document review. Some 12 to 20 policymakers, interest-group representatives, teachers, and school administrators were contacted for each case study. Respondents included the following: superintendents or deputy superintendents, curriculum specialists, assessment specialists, special education directors, certification specialists, budget directors, state board of education president, governor's executive aide, legislators serving on education and appropriation committees, teachers' union representatives, business representatives, and education journalists.

The researchers found that standards-based reform remained a strong and viable component of policy agendas in 1994-95. And, this momentum seemed to survive despite key changes in leadership at the level of the school division, state department of education, and federal government. Despite a series of setbacks, policymakers had been able to continue a steady path toward major systemic reform.

Reformers faced several challenges to their work. The first of these was the need to provide higher quality support to teachers. While those supporting higher standards were asking teachers to move students to ever higher levels of competency on established academic standards, teachers seemed to have been the recipients of relatively little worthwhile training on how to accomplish this goal. This deficiency was readily illustrated by fiscal shortcomings. For example,

budget reductions caused Florida to eliminate financial support for summer institutes for teachers, and Georgia's teacher training institutes for academic standards were reduced from a budget of $3 million to $500,000. Kentucky was the only state in the sample to maintain significant increases in professional training dollars to be spent for district-level professional development.

The second major finding was that equity issues had not been carefully addressed. The authors argued that if a goal of standards-based reform is truly for all students to achieve to high standards, then considerable effort must be focused on finding ways to ensure that both curriculum and testing standards are fair to all students and that all students have an equal opportunity to learn. Special educators in most states reported that they had been involved only infrequently in the development of content standards or frameworks and that their actual participation was normally in the latter stages of standards review.

Texas respondents reported that special educators were rarely involved in standard-setting and assessment policy development at all. In addition to the traditional meaning of *opportunity to learn* initiatives (meaning that all students will have the opportunity to be exposed to the same, high quality/high level academic standards), the authors concluded that other students–those who are homeless, poor, hungry, or those who live in violent families or neighborhoods–will need policies to address their needs if they are to have equal opportunities to learn.

Third, the authors concluded that as policy makers engage in systemic reform, they will have to try to balance the interests of various groups. This will be particularly true with polarized groups, some of whom may believe that schools should teach a rigorous, skills-based curriculum while others believe instruction should emphasize problem-solving and creative thinking. The authors predicted a scenario in which curriculum and assessment practices designed to please such competing groups will make little sense in the classroom.

Fourth, various non-governmental change agents or groups have had a significant impact on the progress of the standards-based movement. A host of entities such as the National Council of Teachers of Mathematics and the American Association for the Advancement of Science as well as various conservative religious groups have played significant roles in forcing the standards debate in new directions. These groups have helped to energize a vast array of professional networks dealing with sub-components of the standards-reform initiatives, but they have also added complexity to the system. Rather like the proverbial squeaky wheel, well-funded and articulate interest groups can send state and local agencies in a variety of directions at the same time, thus preventing the likelihood of coherent, well-researched policy initiatives.

Last, the authors reported that state and local policymakers will need to do more than simply present preordained policies to their publics. Rather, they must learn to value the thinking of all citizens and to find ways to include the public in reform efforts from their inception. Specifically, the researchers found that while all states and many districts have developed strategies for involving the public in the process of developing standards, these attempts were typically of only brief duration and they tended to look more like public relations campaigns than honest efforts to engage citizens in meaningful dialogues about standards and assessment. For example, though Kentucky made strenuous efforts to *inform* the public about Kentucky's educational reforms, a 1994 poll indicated that nearly half of those responding hadn't even heard

of the reforms. The authors pointed to Florida's school improvement councils (which include non-parental public representatives) as one mechanism that seemed to have been successful in raising public awareness. Focus on the school rather than the state or district as the most viable level of communication seemed to have paid dividends (Massell, Kirst, & Hoppe, p. 14).

Newman, King, and Rigdon

Newman, King, and Rigdon (1997) authored a study that dealt with schools' capacity to respond to restructuring initiatives. The authors posited that when politicians and policymakers link accountability to school performance, they incorrectly draw inferences from the corporate world and assume that strong external accountability expectations will inevitably cause schools to improve student achievement.

The authors examined twenty elementary schools that were involved in restructuring efforts during a time of increased expectations for accountability regarding student achievement in an attempt to answer the key question: *Will increased accountability of schools to external agents* (i.e., state departments of education and boards of education) *improve school performance?* (Newman, King & Rigdon, p. 42).

The study was accomplished by two week-long visits by research teams to each school within a one year period. Researchers gathered data by conducting a series of observations and interviews designed to evaluate the accountability systems operating in the school districts. This process was followed by the administration of surveys designed to measure schools' *organizational capacity.*

For purposes of this study, *accountability* was defined as "the process by which school districts and states–or other constituents such as parents–attempt to ensure that schools and school systems meet their goals" (Rothman, 1995, cited in Newman, King, & Rigdon, p. 43). The authors proposed that a complete school accountability system should include at least these components:

(1) Data about test scores
(2) Comparison to other schools' test scores
(3) Rewards or punishments
(4) A constituency that gathers achievement information and judges schools' success in meeting standards. (Newman, King & Rigdon, p. 43)

The four parts of this accountability system were assumed to enhance school performance. The strength of each school's accountability program was quantified by having two researchers from each school research team answer these eleven specific questions:

(1) Does the school participate in a required district testing program?
(2) Does the school participate in a required state testing program?
(3) Are individual teachers judged by standards related to student academic performance, other student outcomes, or the teachers' instructional behavior?
(4) If yes, are the standards explicit or implicit?

(5) Are teachers subject to any consequences for meeting or failing to meet standards?
(6) If yes, indicate the most significant consequences teachers would face.
(7) Is the school as a whole, or the principal, judged by any standards related to student academic performance, other student outcomes, or teachers' instructional behavior?
(8) Is the school as a whole, or the principal, subject to any consequences for meeting or failing to meet standards?
(9) If yes, indicate the most significant consequences the school or principal would face.
(10) To what extent have the district and state influenced the accountability system of the school (no impact to major impact)?
(11) To what extent have the district and state made efforts to influence the accountability system of the school (no effort to significant effort)?
(Newmann, King & Rigdon, p. 70).

Schools having positive responses to questions 1-9 were considered to have strong accountability systems. Schools were judged to have medium-level accountability systems if they had some combination of information, standards, and consequences but not all three of these. Schools that collected only standardized test data and and had no standards or consequences were considered to have weak accountability systems.

The authors measured schools' *organizational capacity* according to the criteria of teacher knowledge and skill, school autonomy to act according to circumstances (termed power and authority), and shared collaboration toward clear purposes about student learning. Data were gathered by two methods; first, the coding of questions that could be answered by studying written research reports for each school and, second, the administration of a written questionnaire.

When interpreting the research reports, the knowledge and skills domain was measured by three questions associated with the quantity of staff development for schools and teachers and the focus of this development on achieving school goals. The power and authority domain was assessed with 12 questions that determined the amount of influence teachers and the principal had over curriculum, instruction, student assessment, budgeting, hiring, and staff development. The shared commitment domain was assessed with five questions on this topic.

The researchers then used a written survey to further evaluate these domains for each school. A Likert-type scale was used to assess opinions. The sample questions below reflect the construction of questions used for each domain.

Knowledge and skills: (two actual questions)
Teachers in this school are continually learning and seeking new ideas.
Power and Authority: (14 actual questions)
How much control do you feel you have in your target class over selecting textbooks and other instructional material?

Shared Commitment: (16 actual questions)

> Most of my colleagues share my beliefs and values about what the central mission of the school should be. (Newmann, King & Rigdon, pp. 72-74)

Internal accountability was identified as (1) The staff had identified clear standards for student performance. (2) They collected information to inform themselves about their levels of success, and (3) They exerted strong peer pressure within the faculty to meet these goals (Newman, King & Rigdon, p. 48).

The study resulted in significant findings concerning accountability and school restructuring. The first was that of the 20 schools studied, only seven had strong accountability systems. Thirteen of the twenty were judged deficient in that they did not have explicit standards for student performance, and they lacked significant incentives or consequences related to student success. Table 5, adapted from Newman, King, and Ridgon, illustrates the dispersion of strength of accountability among the schools.

Table 5

Strength of Accountability Systems

	Strong	Mid-range	Weak
Elementary	Careen (E) Humboldt (I)	Ashley Lamar Sumpter	Eldorado Falls River
Middle	Okanagon (I)	Copan Red Lake Selway	Morris Shining Rock
High	Cibola (I) Flinders (B) Island (E) South Glen (E)	Huron Wallingford	Fremont

Note. Strong accountability systems were externally required (E), internally generated (I), or both (B). All school names are pseudonyms. Adapted from Newmann, F., King, B., and Rigdon, M. Accountability and school performance: Implications from restructuring schools. Harvard Educational Review, 67 (1) (February, 1997), (41-74). Copyright 1997 by the President and Fellows of Harvard College. All rights reserved.

The following examples illustrate the power of this ranking system. Fremont High School gave the required state tests as well as district-level criterion-referenced tests. However, no one at the school, district, or state level seemed to make any significant use of the test results. No scores

were published, and there were no consequences for students, teachers, or the school as a result of student performance. Most staff members felt little or no pressure to affect student performance on the tests. South Glen High School, however, was required by the state to administer end-of-course tests in English, math, social studies, and science. The tests were viewed by many as a way to hold schools accountable for teaching the state's mandated curriculum. South Glen's school district also required that final exams be given for compulsory courses and required that exam scores count for 20 percent of a student's final grade. The state mandated that students pass competency tests in reading, writing, and grammar to graduate. To further heighten accountability, South Glen was required to develop an annual school improvement plan with ten performance indicators other than test results. Teachers in South Glen received up to a 3 percent salary supplement for meeting individual and school-wide goals.

The second finding was that schools varied significantly in the extent of their organizational capacity. Table 6 shows schools clustered by level and their mean organizational capacity scores using combined standardized scores from teacher surveys and researchers' codings. Both the codings of individual school reports and teacher surveys were used to measure the key components of organizational capacity: (1) teacher knowledge and skills, (2) power and authority at the school, and (3) shared commitment and collaborative activity to achieve a clear purpose (Newmann, King, & Rigdon, p. 68). The overall sample mean was determined to be -.01, with a standard deviation of .90.

As a group, elementary schools scored considerably higher on these measures of organizational capacity than did middle and high schools. The overall mean for elementary schools was .78. The overall mean for middle schools was -.09 and for high schools, -.21. Lamar Elementary was the school with the highest score on organizational capacity. Teachers there used a variety of innovative strategies, including team-teaching in multi-age clusters, team planning of units, and team feedback on teaching. Teams focused on implementing the inquiry-based philosophy of instruction that was embraced by the whole staff. The faculty had significant input into decisions involving the hiring of staff, curricular themes, and other issues.

Sumpter Elementary, by contrast, had almost the lowest score for elementary schools in the study. Though innovation was valued in this school, it was characterized by largely independent activities by teachers. The result was a plethora of fragmented programs that–while good in themselves–did not contribute to a common purpose or to instructional collaboration.

Table 6

Level of Organizational Capacity

Elementary Schools			Middle Schools			High Schools		
	Org. Cap.	Overall Rank 1-20		Org. Cap.	Overall Rank 1-20		Org. Cap.	Overall Rank 1-20
Lamar	1.64	1	Okanagon	1.17	4	Cibola	.66	6
Ashley	1.51	2	Red Lake	.73	5	Huron	.00	10
Humboldt	1.50	3	Copan	.11	9	Flinders	-.11	11
Careen	.58	7	Shining Rock	-.35	13	S. Glen	-.84	17
Eldorado	.48	8	Morris	-.39	14	Fremont	-.89	18
Sumpter	-.13	12	Selway	-.74	16	Walfd.	-1.24	19
Falls River	-.41	14				Island	-1.81	20

Note . From Newman, et al. Accountability and school performance: Implications from restructuring schools. Harvard Educational Review, 67 (1) (February, 1997), (p.53).

The third finding of the study was that school accountability and organizational capacity were not necessarily related. For example, of the seven schools rated strong in terms of their accountability systems, only two (Careen & Humboldt) were rated as high in organizational capacity. The data, therefore, did not support the notion that school accountability requirements will inevitably cause schools to engage in more organizational planning and practices to accomplish specified goals.

Finding four was that schools that had strong external accountability tended to have low organizational capacity. Three schools–Careen, Island, and South Glen–were characterized by strong external accountability expectations. Two of the three were ranked among the lowest group for organizational capacity. The Island score of -1.81 and the South Glen score of -.84 suggest that strong external accountability systems could actually hinder the development of organizational capacity in schools.

Finding five was that strong internal accountability enhanced organizational capacity in individual schools. Three schools had strong internal accountability systems: Humboldt Elementary, Okanagon Middle, and Cibola High. Of the seven schools ranked highest in terms of accountability systems, these three schools were also ranked highest for their organizational capacity.

Conclusions were as follows: (1) External accountability alone did not offer assurance that schools would have the required resources, technical knowledge, skill, or authority to achieve goals. (2) External accountability did not necessarily result in a shared vision to enhance student learning. (3) Strong external accountability tended to be associated with low organizational

capacity, and (4) Some schools became highly accountable for student achievement without externally imposed requirements.

Pedroza

Pedroza (1997) studied the consequences of high-stakes testing on historically disenfranchised students (those attending what she termed "border schools") in a dissertation completed at the University of Texas at Austin. The dissertation was a five-year, qualitative, study with four guiding research questions. The questions (following) were designed to discover the effects of high-stakes testing policy: (1) What patterns of change are evident in student achievement? (2) What patterns of change are evident in student retention? (3) What patterns of change are evident in student program placement? (4) What local administrative responses are evident from the state's mandated high-stakes testing policy?

The tests in question were referred to as the Texas Assessment of Academic Skills or TAAS. These assessments were used by the Texas Department of Education to evaluate school and district performance. The state examinations were given in grades 3 through 8 and in grade 10. The tenth grade test was referred to as the exit test. The stakes were indeed high for students in that if they did not pass the exit test by the end of their senior year, they would not graduate or receive a diploma, regardless of their grade point average. Following Madaus' 1988 definition, the testing program was high-stakes because test results were published and school and school divisions were publicly compared.

The Buena Vista School District was chosen for evaluation in this dissertation. The site for the study was selected for three reasons: (1) In 1995 the high school in the school division had been identified as low performing by the state while the elementary school had been identified as an exemplary campus, (2) an international bridge was located quite close to the central office and the elementary campus, and (3) the school district was characterized as growing.

Data were collected through in-depth interviews, document analysis, trend analysis, and participant observations. A total of 65 persons were interviewed, including teachers, board members, administrators, and community members. Document analysis included newspapers, school accreditation reports, census documents, and financial reports. Trend analysis included cohort graduation rates, retention rates, and special education placements as well as student performance on the state's high-stakes tests over time. Participant observation allowed data to be collected first-hand about the effects of high-stakes testing on students who attended a border school.

During data analysis, data were coded according to eight general categories:

> (1) background information, (2) opinions about the Texas accountability system, (3) perceptions regarding policy assumptions, (4) perceptions about students' academic improvement and testing, (5) challenges experienced by participants, (6) initiatives implemented in the campus and district, (7) programmatic or procedural changes, and (8) policy changes. (Pedroza, pp. 96-97)

Fifteen themes were identified from the data and these were reduced to six: language development, curriculum and instruction, resources and materials, programmatic changes, organizational structures, and support structures. Data were then analyzed across the participating school campuses by use of a cross-case matrix. This information was then shared with the campus principals, the division superintendent, and the director of research and evaluation for their perceptions. Qualitative data from interviews, document analysis, and trend analysis were used to triangulate conclusions drawn from the study.

Findings

Selected study findings are summarized below:

Student outcomes. (1) After five years of participation in Texas' high-stakes testing program, almost 70% of participating students had difficulty mastering all sections of the test. (2) Performance by Buena Vista students on college admissions tests showed no achievement gains despite the fact that more students actually took the SAT and ACT exams between 1990 and 1995 and despite the state's high-stakes testing environment. (3) Exemptions from testing played a major role in how schools were perceived by the wider public. Buena Vista Elementary had been recognized by the state as an *exemplary campus* for two years while Buena Vista High School was identified as a *low-performing campus*. The elementary school's rating, however, was based on the performance of only about 40% of the third graders. The remainder had either been retained or had been placed in special programs and were not tested.

Curriculum and instruction. (1) The curriculum for both regular and special education students had been narrowed and limited as a result of high-stakes testing. (2) There was a pervasive perception that special education students were exempt from the state's testing program.

Learning and teaching environment. (1) Teachers across all grade levels consistently reported that the state's high-stakes testing program had decreased the joy they found in teaching. (2) Teachers consistently reported that they would teach differently and spend more time teaching concepts rather than practicing test-taking skills were it not for the state testing program. (3) Teachers and administrators had become involved in a process termed "impression management" (citing Goffman, 1959) in which the perception of improvement and change was infinitely more important than actual improvement or change.

Program quality. (1) The school district had experienced considerable difficulty in employing qualified special education, remedial, and bilingual education teachers. The remote location and low teacher salaries contributed to this condition and was found to have a significant negative effect on student achievement. (2) Professional development for teachers had become minimal in recent years due to budget constraints. This limited both teacher skill as well as understanding and appreciation of the indigenous culture. (3) Lack of appropriate academic assistance, bilingual education issues, lack of understanding of cultural differences, and the requirements of Texas' high-stakes testing program resulted in fewer students remaining in school.

Student skill development. (1) The schools participated in "downward blaming" in which elementary schools blamed the home environment for the lack of student success on the state testing program while the intermediate schools blamed the elementary school and the high school blamed all the schools below them. (2) Teachers were ill-advised about allowable test modifications for students with disabilities. So, disabled students who did take the state tests found the experience harder than it needed to be. (3) Special education students were routinely exempted from the exams, and this diminished their exposure to the same content as their non-disabled peers. When disabled students did take the TAAS and did not do well, their performance was used as a reason to exempt them from the next round of testing.

Student retention. (1) Student retention data was difficult to quantify as grade-level retention information was available for only a three-year period. However, the author did note that the between 1992 and 1994 the greatest level of retention took place in grade 9–the year before the grade 10 exit testing would take place and when results would be published. (2) While the elementary school appeared not to have adopted the policy of retaining large numbers of students, the high school did have an increase in the number of students who were "not advanced" or students whose placement was pending the completion of summer school. (3) While the absolute number of students retained did not show an increase, the number of students *placed* in the next grade (rather than being promoted) had increased. Pedroza questioned, then, the hypothesis that high-stakes testing motivates students to achieve.

Fries (1998) studied the inclusion of students with disabilities in the state-mandated testing program in Illinois. The author's major thesis was that students with disabilities are excluded regularly from the state's testing program and that this exclusion is to their detriment. This exclusion, Fries argued, made this group of students invisible in terms of accountability and sent the message that they cannot achieve and that their education is not a concern to the state or locality.

Fries

Fries presented a simple quantitative assessment of the percentage of students with disabilities who actually participated in the grade three Illinois Goal Assessment Program (IGAP), a state-developed and state-mandated testing program. Percentages of students participating were summarized for a six year period in Table 7.

Table 7

Student Participation in the Illinois Goal Assessment Program, 1987-1993

	Percentage tested		
Year	Reading	Math	Writing
1987-1988	89.75		
1988-1989	88.59	89.31	
1989-1990	87.11	87.96	85.62
1990-1991	87.01	88.10	85.52
1991-1992	86.57	87.52	84.00
1992-1993	85.60	83.50	85.90

Note. Adapted from The Inclusion and Exclusion of Students with Disabilities in State and National Testing Programs: A Policy Study with a Focus on Illinois (p. 64), by R.K. Fries, 1998, Unpublished doctoral dissertation,, University of Illinois at Urbana-Champaign. Reprinted with permission.

In 1994 Illinois passed legislation that required all students–both those with and without special needs–to take the IGAP unless this was contraindicated on their IEP's. A full year later–the 1995-1996 testing year--the percentage of special needs students taking the IGAP was essentially unchanged. The author's conclusion was that the percentage of students taking the IGAP had not changed despite specific legislation requiring greater participation.

Fries reported two conflicting state laws in Illinois, the second of which was likely to result in a lower number of students with disabilities taking the state-mandated tests. The first was the 1994 legislation requiring greater participation by disabled students. A separate 1998 law, however, required that any school superintendent who entered into a new contract for more than one year to have student performance indicators directly linked to his or her contract. This type of pressure alone was unlikely to cause superintendents to take actions to include special needs students in the IGAP testing program as their participation would likely have caused a reduction in the overall achievement scores for the school division.

The author recommended that state testing programs should insist on the inclusion of virtually all students with disabilities as participants. If such decisions are left to educators, she predicted that they would find reasons to decrease significantly the number of disabled students who do participate to keep their school's achievement scores as high as possible. Fries recommended that school districts be given financial incentives for each student who participates in the state testing program. This incentive should not be linked to scores, but simply to student participation.

Natriello and Pallas

Natriello and Pallas (1999) provided information pertinent to a discussion of the results of high-stakes testing at the state level. Their report, *The Development and Impact of High-Stakes Testing,* analyzed student achievement data in three high-stakes states: Texas, New York, and Minnesota, with particular emphasis on disparities among the achievement levels of students who are white, minority, and poor.

Students in Texas had to pass the Texas Assessment of Academic Skills (TAAS) testing program to graduate from high school. The exit-level exam of the TAAS was administered in the spring of the 10th grade year and subsequently each fall and spring thereafter. Students could pass the TAAS at one of the benchmark test periods, or they could satisfy graduation requirements by passing end-of-course tests in Algebra I, English II, and either Biology or U.S. History. Special education students were offered exemptions and accommodations as stipulated in their Individualized Education Plans. TAAS pass rates for 1998-99 were telling (only data for non-special education students were reported). While 52% of White students passed the Algebra I end-of-course test, only 20 percent of African-American students passed. The test was passed by only 26 percent of Hispanic students. Further, cumulative pass rates for the classes of 1996 through 1998 showed that approximately 18 percent of African-American and Hispanic students had not met TAAS requirements to graduate, while only 7 percent of White students had been denied diplomas for this reason.

New York State was described as having a dual track graduation system in which students could meet the requirements for a local diploma by passing Regents Competency Tests, relatively low-level assessments designed to measure basic skills. Students could prepare for the more prestigious Regents diploma by studying for and passing Regents examinations. The result of this system was the creation of two very different curricular tracks in which students preparing for the local diploma studied a much less challenging curriculum than did the students preparing for the Regents diploma.

The authors reported the effects of this dual tracking system on poor and minority students as follows: First, disproportionally fewer African-American and Hispanic students qualified for Regents diplomas. In schools in which Black students made up more than 50 percent of the student body, fewer than 40 percent earned Regents diplomas. Similarly, when the proportion of Hispanic students in a school exceeded 40 percent, the Regents diploma pass rate was below 40 percent. Second, a comparison of students' socio-economic status with the Regents diploma pass rate showed consistently strong negative correlations. The correlation between school rates for awarding the Regents diploma and school participation rates in the free lunch program was -.62. In schools in which half or more of a school's students qualified for free or reduced price lunches, fewer than 35 percent of students qualified for Regents diplomas. Last, the correlation between the awarding of Regents diplomas and the percentage of students classified as Limited English Proficient (LEP) was -.45. Correlations indicated that schools were substantially less likely to award Regents diplomas as the minority, poor, or LEP populations increased.

Minnesota's graduation requirements were described as a combination of basic skills (minimum competency) assessments in reading, mathematics, and writing along with requirements for students to demonstrate proficiency in a broad range of competencies. Beginning in 2002 students will have to pass even higher requirements to graduate. The first administration of the Minnesota Basic Standards (MBS) tests occurred in 1996. Relative performance data for Caucasian and minority students showed substantial differences. Math scores for grade 8 students indicated a pass rate of 80 percent for Caucasian students, 65 percent for Hispanic students, and only 58 percent for African-American students. The pattern for reading at the eighth grade level was similar with 70 percent of Caucasian students receiving passing scores when compared to 54 percent of Hispanic students and 50 percent for African-American students.

The authors concluded their study by questioning a basic platform upon which high-stakes testing rests–that high-stakes tests have a positive motivational effect on students. They offered the opinion that if the motivational consequences of high-stakes tests are not equally distributed across the entire population of students, then educators and parents should be concerned about their potential to increase the disparity of educational opportunity and achievement among affected groups.

Wilkins

Wilkins (1999) conducted a study directly related to student achievement as measured on Virginia's Standards of Learning (SOL) tests. The author cited Guiton and Oakes (1995) in stating that performance assessments such as Virginia's SOL tests assume that students, teachers, and schools have all had equal opportunities and resources to foster achievement and that "before students and schools can be held accountable for their achievement levels and be compared to a standard, they must be given comparable starting points in terms of educational and demographic opportunities" (Wilkins, p. 13).

The author posited two major types of opportunity structures associated with student achievement–*educational opportunity structures* and *demographic opportunity structures*. Educational opportunity structures are exemplified by such factors as the quality of instruction, the amount of instruction, qualifications of teachers, and the courses available to students. Demographic opportunity structures (DOS) refer to the characteristics of the people in the community in which schools are located. Wilkins' study focused only on these demographic structures and their association with student achievement on SOL tests.

DOS variables measured included *financial capital*, *human capital*, *cultural capital*, and *geographic capital*. Financial capital referred to the fiscal resources available to parents to provide for the needs of their children, including needs that helped ensure that they were ready to participate effectively in school. Financial capital, then, referred not only to food, clothing, and shelter, but also to the provision of an appropriate place to study, computers, books, and materials for school projects. Cultural capital referred to "an embodiment of status and expectancy that is related to being a member of a dominant group or class" (Wilkins, p. 13). The dominant group or class (those in the "White" mainstream) were assumed to have more privileges and opportunities than those usually available to those not members of the dominant group (i.e.,

non-Whites). Here one may expect to see that a history of racial segregation and oppression has limited educational and employment opportunities for those not in the dominant group. Human capital was explained by the level of parents' education. Parents who have higher levels of education are better able to provide stimulating cultural and educational experiences for their children and to help them with homework. Geographic capital referred to the opportunities available to children by virtue of their geographic location. Students who lived near urban areas were more likely to have access to enriched experiences such as those available in museums and libraries.

In the design of this study, the author examined the relationship between the various components of DOS and school-level passing rates on Virginia's SOL exams. The sample was composed of 1,560 public schools, 1016 of which were elementary, 247 were middle schools, 251 were high schools, and 46 were combined schools. School success was measured by virtue of passing rates in the four content areas of math, English, science, and history. "A composite score for each grade level was created by standardizing each of the separate content scores ($\underline{M}$ = 50, $\underline{SD}$ = 10) and then averaging the four scores by grade level" (Wilkins, p. 14). This composite score mirrored the Virginia Department of Education policy of considering that all four areas must be passed for schools to be accredited.

Proxies were devised for each type of DOS. Financial capital of a community was associated with the median-household income of a community as measured in thousands of dollars. Human capital of a community was calculated by determining the percentage of community members 25 years of age or older who had achieved a college degree. Cultural capital was indicated by determining the level of urban influence affecting the community. Cities and counties were divided into five groups according to size as indicators of this variable. Groups ranged from large metropolitan areas such as Fairfax County to semi-rural areas not adjacent to metro areas.

Multiple regression was used to estimate the relationships between school passing rates on SOL tests and the DOS of each school. Standardized regression coefficients were calculated to allow a comparison of the four measures of opportunity. Study results included the following: (1) The four demographic opportunity structures were found to predict between 47 and 54 percent of the variance in school success on the SOL tests. (2) "For all grade levels, opportunities associated with financial, human, and cultural capital were consistently found to significantly predict school success" (Wilkins, p. 14).

In summary, Wilkins confirmed that an understanding of demographic opportunity structure can help predict school-level SOL success. Demographic opportunity structures predicted about half the variance in student achievement. Stated simply, schools whose students are largely financially secure, predominantly White, and well-educated are more likely to do well on SOL tests. The results indicated that school performance on SOL tests and, therefore, school accreditation can by predicted without reliance on educational opportunity structures–including such aspects as qualification of teachers, course offerings, and the quality of instruction. Wilkins concluded by stating, "With the inequities that have been shown in this study to exist across

demographic opportunity structures, it is unreasonable to accredit schools based solely on their success on the SOL tests" (Wilkins, p. 18).

Schleisman

Schleisman (2000) studied how schools and districts responded when students failed to meet Minnesota's basic standards for high school graduation. Schleisman explained that Minnesota had recently passed a results-oriented graduation rule. Beginning with the class of 2001, all Minnesota students must pass the Minnesota Basic Standards Tests (BST) which are administered beginning in the eighth grade as well as the Basic Standards Written Composition Test, administered to all students beginning in tenth grade, to graduate from high school. BST results are used to measure the success of schools and school districts in the state in teaching basic skills. Policymakers have required that school districts must offer a variety of compensatory services to students who do not pass the BST. Schleisman sought to learn if school districts did, in fact, offer the remediation services required in state legislation and in what manner this remediation was provided.

The author stated two primary research questions to guide the study: First, "What are the school-and district-level responses when a state institutes a high-stakes testing policy? In particular, how do schools and districts respond for students who do not initially meet the basic requirements?" (Schleisman, p. 5). Second, what specific instructional opportunities are provided to students who do not meet the BST standards in reading and math in eighth grade?

Schleisman chose to use qualitative methods to research her questions. Interview questions were developed to gather interviewees' perceptions of the following: (1) how BST had changed schools and school districts, (2) present or future changes that result from BST, and (3) identified needs at the school and district levels. The population for the study consisted of two urban school districts with PK-12 enrollments of approximately 50,000 students, two suburban districts with PK-12 enrollments of approximately 10,000, two "outstate" school districts with enrollments of approximately 7,000 and 11,000 respectively, and two smaller "outstate" school districts with enrollments of approximately 400 and 1300 each. The researcher identified proportionate numbers of educators to interview from each locality. Twenty-seven interviews were conducted in all. Eleven open-ended questions were asked of each interviewee with specific probes specified to elicit thorough responses. Interviews were all audiotaped and transcribed. Data were then summarized into tables representing the following themes:

1. General information about a program or school
2. Responses specific to the subject areas of reading, math, and writing
3. Summer school, after-school, or tutoring opportunities
4. Communication to parents
5. Staff development, and
6. Assessments or tools. (Schleisman, p. 100)

These tables were sent back to participants to allow verification of the data and perceptions about the themes which had been developed. Based on this information and a more

thorough analysis of coded interview transcripts from one suburban district of about 14,000 students, three more general themes emerged regarding the adoption and implementation of BSTs."

> 1. Helps schools identify students who may otherwise have 'slipped through the cracks'
> 2. Brings greater curricular coherence to the system, and
> 3. Prompts schools to offer a variety of services for students who need remedial work. (Schleisman, p. 127)

In a second phase of the study, Schleisman analyzed the data to determine the types of additional instructional or remedial offerings that were being made to students as a result of the BSTs and related legislation. Seven major themes were identified: (1) Schools and districts *were* placing a greater emphasis on teaching of basis skills, especially at the high school level. (2) Schools were offering a wide variety of compensatory learning opportunities including remediation during the regular school day, after school and weekend programs, and summer school. (3) Schools and districts were struggling with the issues of student attendance and participation in remedial programs. (4) Schools and districts were struggling to find strategies to help students who moved into their school or district near to or after the time when the BSTs were administered. (5) Schools and districts were struggling with strategies to address the need for both basic standards and high standards. (6) Schools and districts faced new challenges related to hiring, transportation, and scheduling related to remediation programs. (7) Schools and districts were struggling to determine effective methods of assessing the value of remediation programs and strategies.

Schleisman offered a caveat about the benefits from the standards movement that seem to be reflected in her study results. She cautioned that, in an effort to bring all children up to the same basic standard of proficiency, a new system of tracking may ensue in which poor and minority students may receive a different curriculum because they will spend all of their time in classes emphasizing basic skills or test-taking strategies.

<u>Davison, Schleisman, Koeppen, Wu, and Kwak</u>

A 2001study by Davison, Schleisman, Koeppen, Wu, and Kwak assessing the consequences of Minnesota's high school graduation test found that state-funded summer school experiences for weak students had some positive outcomes. The authors identified students who failed to meet the Basic Standards Test (BST) requirements in grade 8-10 reading or math in the February 1998 BST test administration and then tracked student achievement gains on the BSTs in a subsequent test administration in July of 1998. Both students who did and did not attend summer school were allowed to take the July administration of the BST.

Work on the study was complicated by the fact that Minnesota had no clear definition of a state-funded summer school program. Some divisions may have met this remediation requirement by providing only a one week summer school. However, summer schools in Minneapolis/St.Paul were a uniform seven weeks duration, and the curriculum focus was uniform as well. Therefore,

data from the Minneapolis/St.Paul district were analyzed by the researchers separately from those derived from the rest of the state. As these data are the only ones from the study in which "summer school" is defined in the same way (a seven week, state-funded experience), I am reporting only information from the Minneapolis/St.Paul summer school sites in this summary.

The study was organized around four research questions: (1) Do students who have not passed the reading or math BST sign up for summer school at a higher rate than those students who have already passed the BST? (2) Of the students who did not pass the reading or math sections of the BST in February of 1998, are there differences between the students who did not enroll in the subsequent state-funded summer school? (3) For students who did not pass the February 1998 BST, are there differences in score gains between students who did enroll in a state-funded summer school and students who did not enroll in a state-funded summer school? (4) Are students who failed the reading or math sections of the BSTs in February of 1998 and who then participated in a state-funded summer school passing future administrations of the BST at a higher rate than students who did not attend a state-funded summer school?

With respect to research question number one, the researchers found that Minneapolis/St. Paul students who had failed the BST attended summer school at a higher rate than students who had passed. In fact, attendance rates were strikingly higher for those not passing the BST compared to those who had passed. For example, while only 15.9 percent of eighth graders passing the reading test enrolled in summer school, 54.6 percent of eighth graders failing the reading test enrolled. Though this is not surprising in the context of high-stakes consequences for students, the fact is that many students needing remediation in basic skills did attend a rather lengthy summer school in Minneapolis and St Paul school district. The positive implication of this information is that, under the right conditions, schools may expect to attract many of the students needing remediation to summer school.

Research question two was designed to explain any differences in demographics among students who failed the February 1998 BST and who then did or did not enroll in state-funded summer school. Independent variables in this assessment included socio-economic status, special education status, gender, ethnicity, mobility (whether or not a student was new to the district), and limited English proficiency (LEP) status. Some notable differences among attendees and non-attendees were found, the most significant of which were higher percentages of non-enrolled special education and Black students. LEP students were also more likely to be in the enrolled group than the non-enrolled group.

Data from this part of the study indicate that, regardless of demographic classification, the mean achievement score from the February 1998 administration of the BST was about 50 %. Since the 50% score is likely to represent the *easiest* part of the content for students to master, the authors noted the importance of this information for future diagnostic and instructional planning. That is, summer school planning for most students should likely focus on the more complex material from a given domain rather than on basic skills.

In research question three the researchers asked if there were differences in score gains between students who did and who did not enroll in state-funded summer school. To answer this

question only scores for students who did not pass the February 1998 administration of the BST were analyzed. The mean scores of students who did not pass the February administration were compared to students who had completed the 1998 post-summer school administration. The data indicated that, on average, students who had enrolled in summer school gained about five percentage points in reading and about eight percentage points in math. However, students who did not enroll in summer school varied from losing about three percentage points (9^{th} grade reading) to a gain of about three percentage points (8^{th} grade math). The analysis showed statistically significant gains as a result of enrolling in summer school in all areas except for 10^{th} grade reading.

In research question four the authors tried to determine if there were differences in future BST passing rates for students who did and did not attend a state-funded summer school. That is, were students *as a group* more likely to pass the BST if they had attended summer school? The data indicated that, whether students had enrolled in the 1998 summer school or not, few passed the BST at the end of the summer school session. Passing percentages ranged from 9% of 9^{th} graders who did not attend summer school before taking the BST math test to 25% of 8^{th} graders who did attend summer school before taking the BST reading test. Students who attended summer school almost always passed the 1998 summer administration in higher percentages, albeit only slightly higher, than students who did not attend summer school. Paradoxically, however, when the time period to pass the BST test was extended to include the February 1999 administration, students who had failed the February 1998 administration of the BST and had *not* attended summer school passed at significantly higher percentages than students who had failed the February 1998 BST and then *had* attended summer school. One possible explanation for this phenomenon is that many students who initially failed the BST and did not attend summer school may have simply applied themselves more diligently at subsequent test administrations. It is also possible that some students who failed the February administration and did not attend a state-financed summer school were able to participate in a private or non-state supported summer school, thus preparing themselves for the test in an alternate fashion.

The researchers offered the following conclusions for the Minneapolis and St. Paul school districts: (1) There were few demographic differences between students who did and did not attend summer school. (2) Summer school was effective in helping students improve their skill level as evidenced by score gains between February 1998 and July of 1998. (3) Though helpful, summer school programs may not be able to do enough to help students to pass the BST. As students were, on average, 20-25 percentage points below a passing score, more than one summer session will be needed to help most students achieve passing scores, and some will require other differentiated programming throughout the school year.

Summary

Five categories of outcomes that are likely to result from high-stakes testing and measurement-driven instruction are described in Chapter 2: outcomes for students, outcomes for instructional practices, outcomes for schools, outcomes for public confidence in teachers and schools, and outcomes for teachers.

Among the outcomes for low-scoring students were assignment to low-level drill and practice activities, tracking toward lower level, non-college preparatory diplomas, and greater instances of retention in grade and assignment to special education classes. Students were also reported to suffer degraded psychological health as a result of participation in high-stakes testing programs. Other potentially harmful outcomes included less teacher time being devoted to students who were farthest away from passing state-mandated tests, unfair testing requirements for disabled students, and increased numbers of drop-outs.

Outcomes for instructional practices included a narrowing of the curriculum to help ensure higher test scores and an increase in the time devoted to drill and practice, thus leaving less time for creative or higher-level activities. Schools serving low SES students were reported to spend the most time on rote drill and practice. High-stakes testing was an important factor in reducing the time available for ordinary instruction. The significant time devoted to preparation for and administration of high-stakes tests limited the time available for discretionary learning experiences. Reading programs were sometimes determined to be little more than the reading of short paragraphs followed by comprehension questions. And, finally, test scores were reported to be used for the evaluation of schools rather than to assess the strengths and weaknesses of individual students.

Schools themselves were found to engage in a variety of unsound practices to hide students from the testing process, including retention, assignment to transitional classes, and referral to special education classes. Strong external accountability systems were associated with lower problem-solving capacity on the part of individual schools. Schools were seen to shift their strongest teachers to targeted test grades, and the entire focus of schools' missions often shifted from education to image management.

As has been the case for years, the public was found to express continued faith in the quality of their own local schools. A variety of health and safety issues were of prominent parental concern as expressed in polls, but the need for higher academic standards or school quality was a strong concern for very few citizens. Parents in one poll were skeptical that mandated state achievement tests accurately measured what their children knew, and they also expressed concern about the consequences associated with narrowed curricula. Another poll suggested that Virginia citizens have noteworthy levels of concern about their state's Standards of Learning tests.

A range of possible outcomes for teachers was reported, including undue job stress, loss of self-esteem, loss of job satisfaction, and guilt for low student scores. Measurement-driven instruction has also been associated with a diminishment of decision-making authority on the part of teachers. Concern was expressed that teachers with high percentages of poor or minority students may gravitate to schools where higher test results are easier to obtain. Numerous reports have indicated that some teachers (and principals) have resorted to cheating on high-stakes tests to ensure acceptable scores.

Positive outcomes for high-stakes testing and measurement-driven instruction were described in the literature as well. Teachers felt that tests gave them valuable information about low-achieving students. They felt that the tests helped them modify the curriculum and to alter

instructional priorities. Teachers noted that extra funds were allocated to their schools as a result of student assessments. Students were reported to be more serious about school. Some teachers reported a greater emphasis on the teaching of basic skills, while others indicated the availability of a wide variety of compensatory opportunities for students. Some members of the public indicated stronger beliefs in the effectiveness of instructional practices as a result of curricular expectations and testing practices. Teachers' opinions about testing practices were found to depend on the scope and comprehensiveness of districts' initiatives and directives related to assessment.