

This article has been retrieved **35041** times since March 28, 2002

[other vols.](#) | [abstracts](#) | [editors](#) | [board](#) | [submit](#) | [book reviews](#) | [subscribe](#) | [search](#)

Education Policy Analysis Archives

Volume 10 Number 18

March 28, 2002

ISSN 1068-2341

A peer-reviewed scholarly journal

Editor: Gene V Glass

College of Education

Arizona State University

Copyright 2002, the **EDUCATION POLICY ANALYSIS ARCHIVES**.
Permission is hereby granted to copy any article
if **EPAA** is credited and copies are not sold.

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the [ERIC Clearinghouse on Assessment and Evaluation](#) and are permanently archived in *Resources in Education*.

High-Stakes Testing, Uncertainty, and Student Learning

Audrey L. Amrein

Arizona State University

David C. Berliner

Arizona State University

Citation: Amrein, A.L. & Berliner, D.C. (2002, March 28). High-stakes testing, uncertainty, and student learning *Education Policy Analysis Archives*, 10(18). Retrieved [date] from <http://epaa.asu.edu/epaa/v10n18/>.

Related articles:

[Vol. 11 No. 24](#)

[Vol. 11 No. 25](#)

Abstract

A brief history of high-stakes testing is followed by an analysis of eighteen states with severe consequences attached to their testing programs. These 18 states were examined to see if their high-stakes testing programs were affecting student learning, the intended outcome of high-stakes testing policies promoted throughout the nation. Scores on the individual tests that states use were not analyzed for evidence of learning. Such scores are easily manipulated through test-preparation programs, narrow curricula focus, exclusion of certain students, and so forth. Student learning was measured by means of additional tests covering some of the same domain as each state's own high-stakes test. The question asked was whether transfer to these domains occurs as a function of a state's high-stakes testing program.

Four separate standardized and commonly used tests that overlap

the same domain as state tests were examined: the ACT, SAT, NAEP and AP tests. Archival time series were used to examine the effects of each state's high-stakes testing program on each of these different measures of transfer. If scores on the transfer measures went up as a function of a state's imposition of a high-stakes test we considered that evidence of student learning in the domain and support for the belief that the state's high-stakes testing policy was promoting transfer, as intended.

The uncertainty principle is used to interpret these data. That principle states "The more important that any quantitative social indicator becomes in social decision-making, the more likely it will be to distort and corrupt the social process it is intended to monitor." Analyses of these data reveal that if the intended goal of high-stakes testing policy is to increase student learning, then that policy is not working. While a state's high-stakes test may show increased scores, there is little support in these data that such increases are anything but the result of test preparation and/or the exclusion of students from the testing process. These distortions, we argue, are predicted by the uncertainty principle. The success of a high-stakes testing policy is whether it affects student learning, not whether it can increase student scores on a particular test. If student learning is not affected, the validity of a state's test is in question.

Evidence from this study of 18 states with high-stakes tests is that in all but one analysis, student learning is indeterminate, remains at the same level it was before the policy was implemented, or actually goes down when high-stakes testing policies are instituted. Because clear evidence for increased student learning is not found, and because there are numerous reports of unintended consequences associated with high-stakes testing policies (increased drop-out rates, teachers' and schools' cheating on exams, teachers' defection from the profession, all predicted by the uncertainty principle), it is concluded that there is need for debate and transformation of current high-stakes testing policies.

The authors wish to thank the Rockefeller Foundation for support of the research reported here. The views expressed are those of the authors and do not necessarily represent the opinions or policies of the Rockefeller Foundation.

This is an era of strong support for public policies that use high-stakes tests to change the behavior of teachers and students in desirable ways. But the use of high-stakes tests is not new, and their effects are not always desirable. "Stakes," or the consequences associated with test results, have long been a part of the American scene. For example, early in the 20th century, scores on the recently invented standardized tests could, for immigrants, result in entrance to or rejection from the United States of America. In the public schools test scores could uncover talent, providing entrance into programs for the gifted, or as easily, provide evidence of deficiencies, leading to placement in vocational tracks or even in homes for the mentally inferior. Test scores could also mean the difference between acceptance into, or rejection from, the military. And throughout early twentieth century society, standardized test scores were used to confirm the superiority or inferiority of various races, ethnic groups, and social classes. Used in this way, the consequences of standardized tests insured maintenance of the status quo along those racial, ethnic and class lines. So, for about a century, significant consequences have been attached to scores on standardized tests.

A Recent History of High-stakes Testing

In recent decades, test scores have come to dominate the discourse about schools and their accomplishments. Families now make important decisions, such as where to live, based on the scores from these tests. This occurs because real estate agents use school test scores to rate neighborhood quality and this affects property values. (Note 1) Test scores have been shown to affect housing prices, resulting in a difference of about \$9,000 between homes in grade "A" or grade "B" neighborhoods. (Note 2) At the national and state levels, test scores are now commonly used to evaluate programs and allocate educational resources. Millions of dollars now hinge on the tested performance of students in educational and social programs.

Our current state of faith in and reliance on tests has roots in the launch of Sputnik in 1957. Our (then) economic and political rival, the Soviet Union, beat the United States to space, causing our journalists and politicians to question American education with extra vigor. At that time, state and federal politicians became more actively engaged in the conduct of education, including advocacy for the increased use of tests to assess school learning. (Note 3)

The belief that the achievement of students in U.S. schools was falling behind other countries led politicians in the 1970s to instigate a minimum competency testing movement to reform our schools. (Note 4) States began to rely on tests of basic skills to ensure, in theory, that all students would learn at least the minimum needed to be a productive citizen.

One of these states was Florida. After some hasty policy decisions, Florida implemented a statewide minimum competency test that students were required to pass prior to being graduated. Florida's early gains were used as an example of how standards and accountability systems could improve education. However, when perceived gains hit a plateau and differential pass rates and increased dropout rates among ethnic minorities and students from low socioeconomic backgrounds were discovered, Florida's testing policy was postponed. (Note 5)

In the 1980s, the minimum competency test movement was almost entirely discarded. Beyond what was happening in Florida, suggestions that minimum competency tests promoted low standards also raised concerns. In many schools the content of these tests became the maximum in which students, particularly in urban schools, became competent. (Note 6) It was widely perceived that minimum competency tests were "dumbing down" the content learned in schools.

In 1983, the National Commission on Education released *A Nation at Risk*, (Note 7) the most influential report on education of the past few decades. *A Nation at Risk* called for an end to the minimum competency testing movement and the beginning of a high-stakes testing movement that would raise the nation's standards of achievement drastically. Although history has not found the report to be accurate, (Note 8) it argued persuasively that schools in the United States were performing poorly in comparison to other countries and that the United States was in jeopardy of losing its global standing. Citing losses in national and international student test scores, deterioration in school quality, a "diluted" and "diffused" curriculum, and setbacks on other indicators of U.S. superiority, the National Commission on Education triggered a nationwide panic regarding the weakening condition of the American education system.

Despite its lack of scholarly credibility, *A Nation at Risk* produced massive effects. The National Commission on Education called for more rigorous standards and accountability mechanisms to bring the United States out of its purported educational

recession. The Commission recommended that states institute high standards to homogenize and improve curricula and rigorous assessments be conducted to hold schools accountable for meeting those standards. The Commission and those it influenced intended to increase what students learn in schools. This report is an investigation of how well that explicitly intended outcome of high-stakes testing programs was achieved. We ask, below, whether increases in school learning are actually associated with increases in the use of high-stakes tests? Although it appears to be a simple question, it is very difficult to answer.

The Effects of *A Nation at Risk* on Testing in America

As a result of *A Nation at Risk*, state policymakers in every state but Iowa developed educational standards and every state but Nebraska implemented assessment policies to check those standards. (Note 9) In many states high-stakes, or serious consequences, were attached to tests in order to hold schools, administrators, teachers, and students accountable for meeting the newly imposed high standards.

In fixing high-stakes to assessments, policymakers borrowed principles from the business sector and attached incentives to learning and sanctions to poor performance on tests. High performing schools would be rewarded. Under performing schools would be penalized, and to avoid further penalties, would improve themselves. Accordingly, students would be motivated to learn, school personnel would be forced to do their jobs, and the condition of education would inevitably improve, without much effort and without too great a cost per state. What made sense, in theory, gained widespread attention and eventually increased in popularity as a method for school reform.

Arguments in Support of High-stakes Tests.

At various times over the past years different arguments have been used to promote high-stakes tests. A summary of these follows:

- students and teachers need high-stakes tests to know what is important to learn and to teach;
- teachers need to be held accountable through high-stakes tests to motivate them to teach better, particularly to push the laziest ones to work harder;
- students work harder and learn more when they have to take high-stakes tests;
- students will be motivated to do their best and score well on high-stakes tests; and that
- scoring well on the test will lead to feelings of success, while doing poorly on such tests will lead to increased effort to learn.

Supporters of high-stakes testing also assume that the tests:

- are good measures of the curricula that is taught to students in our schools;
- provide a kind of "level playing field," an equal opportunity for all students to demonstrate their knowledge; and that
- are good measures of an individual's performance, little affected by differences in students' motivation, emotionality, language, and social status.

Finally, the supporters believe that:

- teachers use test results to help provide better instruction for individual students;
- administrators use the test results to improve student learning and design better professional development for teachers; and that
- parents understand high-stakes tests and how to interpret their children's scores.

The validity of these statements in support of high-stakes tests have been examined through both quantitative and qualitative research, and by the commentary of teachers who work in high-stakes testing environments. A reasonable conclusion from this extensive corpus of work is that these statements are true only some of the time, or for only a modest percent of the individuals who were studied. The research suggests, therefore, that *all* of these statements are likely to be false a good deal of the time. And in fact, some research studies show exactly the opposite of the effects anticipated by supporters of high-stakes testing. (Note 10)

The Heisenberg Uncertainty Principle Applied to the Social Sciences

For many years the research and policy community has accepted a social science version of Heisenberg's Uncertainty Principle. That principle is *The more important that any quantitative social indicator becomes in social decision-making, the more likely it will be to distort and corrupt the social process it is intended to monitor.* (Note 11) When applied to a high-stakes testing environment, this principle warns us that attaching serious personal and educational consequences to performance on tests for schools, administrators, teachers, and students, may have distorting and corrupting effects. The distortions and corruptions that accompany high-stakes tests make inferences about the meanings of the scores on those tests uncertain. If there is uncertainty about the meaning of a test score, the test may not be valid. Unaware of this ominous warning, supporters of high-stakes testing, particularly politicians, have caused high-stakes testing to proliferate. The spread of high-stakes tests throughout the nation is described next.

Current High-stakes Testing Practices

Today, twenty-two states offer schools incentives for high or improved test scores. (Note 12) Twenty states distribute financial rewards to successful schools, and nineteen states distribute financial rewards to improved schools.

Punishments are attached to school scores twice as often as rewards, however. Forty-five states hold schools accountable for test scores by publishing school or district report cards. Twenty-seven of those states hold schools accountable through rating and ranking mechanisms; fourteen have the power to close, reconstitute, or take over low performing schools; sixteen have the authority to replace teachers or administrators; and eleven have the authority to revoke a school's accreditation. In low performing schools, low scores also bring about embarrassment and public ridicule.

For administrators, threats of termination and cuts in pay exist, as does the potential for personal bonuses. In Oakland, California, for example, city school administrators can receive a 9% increase in pay for good school performance with a potential for an additional 3% increase—1% per increase in reading, math and language arts. (Note 13)

For teachers, low average class scores may prevent teachers from receiving salary increases, may influence tenure decisions, and in sixteen states may be cause for dismissal. Only Texas has linked teacher evaluations to student or school test results, but more states have plans to do so in the future.

High average class scores may also bring about financial bonuses or raises in pay. Eleven states disperse money directly to administrators or teachers in the most improved schools. For example, California recently released each school's Academic Performance Index (API). This is based almost entirely on Stanford 9 test scores. Schools showing the biggest gains were to share \$677 million in rewards while low performing schools in which personnel did not raise student achievement scores were to face punishments. (Note 14) In addition, teachers and administrators in 1,346 California schools that

demonstrated the greatest improvements over the past 2 years were to share \$100 million in bonus rewards, called Certificated Staff Performance Incentive Bonuses, ranging from \$5,000 to \$25,000 per teacher. Although over \$550 million had already been disbursed to California schools, the distribution of the staff bonuses was deferred because some teachers who posted gains on the API scale, but felt they were denied their share of the reward money, filed a lawsuit against the state. (Note 15) The court found in favor of the state.

Schools and teachers were not the only targets of rewards and punishments for test performance. Policy makers also attached serious consequences to performance on tests for individual students.

Although test scores are often promoted as diagnostic tools useful for identifying a student's achievement deficits and assets, they are rarely used for such purposes when they emanate from large-scale testing programs. Two major problems are the cause of this. First, test scores are often reported in the summer after students exit each grade and second, there are usually too few items on any one topic or area to be used in a diagnostic way. (Note 16) As a result of these factors, scores on large-scale assessments are most often used simply to distribute rewards and sanctions. This contributes to the corruptions and distortions predicted by the social science version of Heisenberg's Uncertainty Principle.

The special case of scholarships

The distortions and corruptions predicted by the Uncertainty Principle find fertile ground for developing when high scores on a test result in special diplomas or scholarships. Attaching scholarships to high performance on state tests is a relatively new concept, yet six states have already begun granting college scholarships and dispersing funds to students with distinguished performance on tests. (Note 17) Michigan is a perfect example of the corruptions and distortions that occur when stakes are high for a quantitative social indicator.

The Michigan imbroglio. In spring 2000, Michigan implemented its Merit Award Scholarship program in which 42,700 students who performed well on the Michigan Educational Assessment Program high school tests were rewarded with scholarships of \$2,500 or \$1,000 to help pay for in-state or out-of-state college tuition, respectively. (Note 18)

There is quite a story behind these scholarships, however. (Note 19) In 1996, Michigan became the 13th state to sue the nation's leading cigarette manufacturers to recover health care costs encumbered by the state to treat smoking-related diseases developed by Michigan's poor and disadvantaged citizens. The care and treatment of these citizens placed a financial burden on the states, so they sued the tobacco companies for financial compensation. Michigan won approximately \$384 million to recover some of these health care costs and then decided to distribute approximately 75% of this money among high school seniors with high test scores as Merit Award Scholarships. The remainder of the money went to health related needs and research, more or less unrelated to smoking or disease treatment. Thus, the monies that were awarded to the state did not go to the victims at the center of the lawsuit—Michigan's poor and indigent suffering from tobacco related diseases—but went instead to those students who scored the highest on the Michigan Educational Assessment Program high school test. These were Michigan's relatively wealthier students who had the highest probability of enrolling in college even without these scholarships. (Note 20)

Approximately 80% of the test-takers in an affluent Michigan neighborhood earned

scholarships while only 6% of the test-takers in Detroit earned scholarships. (Note 21) One in three white, one in fourteen African American, one in five Hispanic, and one in five Native American test takers received scholarships. (Note 22) In addition, from 1982 to 1997, while education spending for needy students increased 193%, education spending for merit based programs such as the merit scholarships increased by 457% in Michigan. (Note 23) Tests have often been defended because they can distribute or redistribute resources based on notions of "merit." But too often the testing programs become thinly disguised methods to maintain the status quo and insure that funds stay in the hands of those who need them least.

Michigan is now being sued by a coalition that includes students, the American Civil Liberties Union of Michigan (ACLU), the Mexican American Legal Defense and Education Fund (MALDEF), and the National Association for the Advancement of Colored people (NAACP). They are arguing that Michigan is denying students scholarships based on test scores that are highly related to race, ethnicity, and educational advantages. Michigan appears to be a state where high-stakes testing has had a corrupting influence.

The satisfying effects of punishing the slackers. Connecting high-stakes tests with rewards for high performance, such as in the example above, is not nearly as prevalent as have been punishments attached to student scores that are judged to be too low. Punishments are used three times as often as rewards. Policy makers appear to derive satisfaction from the creation of public policies that punish those they perceive to be slackers.

Throughout the nation low scores are used to retain students in grade, using the slogan of ending "social promotion." Promotion or retention is already contingent on test performance in Louisiana, New Mexico, and North Carolina, while four more states have plans to link promotion to test scores in the next few years. (Note 24)

Low scores may also prevent high school students from graduating from high school. Whether a student passes or fails high school graduation exams – exams that purportedly test a high school student's level of knowledge in core high school subjects – is increasingly being used as the *only* determinant of whether some students graduate or whether students are entitled to a regular high school diploma or merely a certificate of attendance.

In fact, high school graduation exams are the assessments with the highest, most visible, and most controversial stakes yet. When *A Nation at Risk* was released, only three states (Note 25) had implemented high school graduation exams, then referred to as minimum competency tests on which students' *basic* skills were tested. But in *A Nation at Risk*, the commission called for more rigorous examinations on which high school students would be required to demonstrate mastery in order to receive high school diplomas. (Note 26) Since then, states have implemented high school graduation exam policies with greater frequency.

Now, almost two decades later, eighteen states (Note 27) have developed and employed high school graduation exams and nine more states (Note 28) have high-school graduation exams underway. The frequency with which high school graduation exams have become key components of states' high-stakes testing policies has escalated almost linearly over the past twenty-three years and will continue to do so for at least the next six years (see Figure 1).

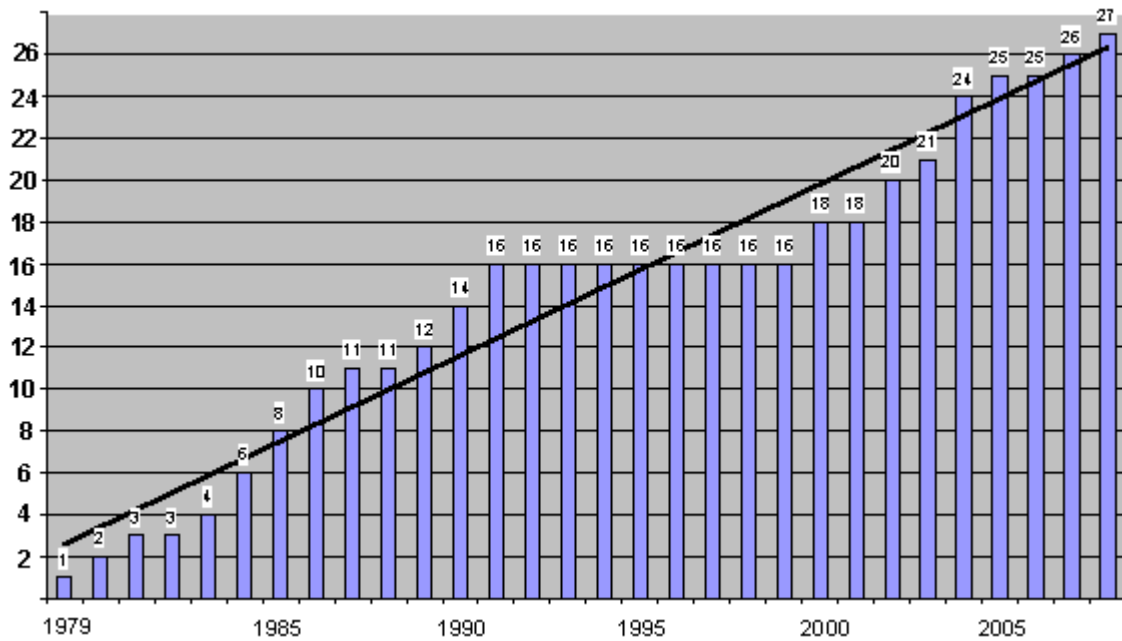


Figure 1. Number of states with high school graduation exams 1979–2008 (Note 29)

Who Uses high-stakes Tests?

Analyses of these data reveal that high school graduation exams are:

- more common in states that allocate less money than the national average per pupil for schooling as compared to the nation. High school graduation exams are found in around 60% of the states in which yearly per pupil expenditures are lower and in about 45% of the states in which yearly per pupil expenditures are higher than the national average. (Note 30)
- more likely to be found in states that have more centralized governments, rather than those with more powerful county or city governments. Of the states that have more centralized governments, 62% have or have plans to implement high school graduation exams. Of the states that have less centralized governments, only 37% have or have plans to implement high school graduation exams. (Note 31)
- more likely to be found in the highly populated states and states with the largest population growth as compared to the nation. (Note 32) For example, 76% of the country's most highly populated states and only 32% of the country's smallest states have or have plans to implement high school graduation exams. Looking at growth, not just population we find that 76% of the states with the greatest population growth and only 32% of the states with the lowest population growth from 1990–2000 have or have plans to implement high school graduation exams. (Note 33)
- most likely to be found in the Southwest and the South. High school graduation exams are currently in use in 50% of the southwestern states and 66% of the southern states. Analyses also suggest that high school graduation exams will become even more common in these regions in the future. By the year 2008, high school graduation exams will be found in 75% of the southwestern and southern states.

High school graduation exams will probably continue to be randomly dispersed throughout 50% of the states in the Northeast and least likely to be found in 33% of the

mid-western states. The western states, over the next decade, will have the greatest increase in the number of states with high school graduation exams by region. While 10% percent of the western states have already implemented high school graduation exam policies, 50% of these states will have implemented these exams by the year 2008. (Note 34)

More important for understanding high-stakes testing policy is that high school graduation exams are more likely found in states with higher percentages of African Americans and Hispanics and lower percentages of Caucasians as compared to the nation. Census Bureau population statistics helped to verify this. (Note 35) Seventy-five percent of the states with a higher percentage of African Americans than the nation have high school graduation exams. By 2008 81% of such states will have implemented high school graduation exams. Sixty-seven percent of the states with a higher percentage of Hispanics than the nation have high school graduation exams. By 2008 89% of such states will have implemented high school graduation exams. Conversely, 13% of the states with a higher percentage of Caucasians than the nation have implemented high school graduation exams. By 2008 29% of such states will have implemented high school graduation exams. *In other words, high school graduation exams affect students from racial minority backgrounds in greater proportions than they do white students.* If these high-stakes tests are discovered not to have their intended effects, that is, if they do not promote the kinds of transfer of learning and education the nation desires, the mistake will have greater consequences for America's children of color.

Similarly, high school graduation exams disproportionately affect students from lower socioeconomic backgrounds. High school graduation exams are more likely to be found in states with the greatest degrees of poverty as compared to the nation. Economically disadvantaged students are most often found in the South and the Southwest and least often found in the Northeast and Midwest. As noted, states in the South and the Southwest are most likely to have high-stakes testing policies. Further, 69% of the states with child poverty levels greater than the nation have or have plans to implement high school graduation exams. Seventy percent of the states with the greatest 1990–1998 increases in the number of children living in poverty have or have plans to implement such exams. (Note 36) That is, *high school graduation exams are more likely to be implemented in states that have lower levels of achievement, and the always present correlate of low achievement, poorer students.* Again, if these high-stakes tests are discovered not to have their intended effects, that is, if they fail to promote transfer of learning and education in its broadest sense, as the nation desires, the mistake will have greater consequences for America's poorest children.

Matters of national standards and implementation of high-stakes tests are less likely to be of concern for the reform of relatively elite schools, (Note 37) that are more often found in regions other than the South and Southwest. Perhaps this helps to explain the more extensive presence of high-stakes tests in the South and Southwest. This seems a reasonable hypothesis especially when one purpose of high-stakes testing is to raise student achievement levels in educational environments perceived to be failing.

It should be noted, however, that there is considerable variability in these data. All states with high rates of children in poverty have not adopted high-stakes testing policies while some states with lower rates of children in poverty have. In states with higher or lower levels of poverty, however, schools that exist within poor rural and urban environments are still more frequently targeted by these policies. Although legislators promote these policies, claiming high standards and accountability for all, schools that already perform well on tests are not the targets for these policies; poor, urban, under performing schools are. But, for different reasons, support for high-stakes testing receives support in both high and low achieving school districts. In successful schools and districts, high-stakes testing policies are acceptable because the scores on those tests merely confirm the

expectations of the community. Thus, in successful communities, the tests pose little threat and also have little incentive value. (Note 38) In poorer performing schools high-stakes testing policies often enjoy popular support because, it is thought, at the very least, that these tests will raise standards in a state's worst schools. (Note 39)

But if high-stakes testing policies do not promote learning, that is, if they do not appear to be leading to education in the most profound sense of that term, then the tests will not turn out to have any use in successful communities and schools, nor will they improve the schools attended by poor children and ethnic minorities. If, in addition, the tests have unintended consequences such as narrowing the curriculum taught, increasing drop out rates and contributing to higher rates of retention in grade, they would not be good for any community. But these unintended negative consequences would have a greater impact on the families and neighborhoods of poor and minority students.

Faith in testing. The effects of high-stakes tests on students is well worth pursuing since it is unquestionably a "bull market" for testing. (Note 40) The faith state legislators have put into tests, albeit blind, has increased dramatically over the past twenty years. (Note 41) The United States tests its children more than any other industrialized nation, has done so for well over thirty years, (Note 42) and will continue to depend on even more tests as it attempts to improve its schools. At the national level, President Bush has been unquestionably successful in passing his "No Child Left Behind" plan that calls for even more testing – annual high-stakes testing of every child in the United States in grades 3 through 8 in math and reading. Republicans and Democrats alike have endorsed high-stakes testing policies for the nation making this President Bush's only educational proposal that has claimed bipartisan support. (Note 43) According to the President and other proponents, annual testing of every child and the attachment of penalties and rewards to their performance on those tests, will unequivocally reform education. Despite the optimism, the jury is still out on this issue.

Many researchers, teachers and social critics contend that high-stakes testing policies have worsened the quality of our schools and have created negative effects that severely outweigh the few, if any, positive benefits associated with high-stakes testing policies. Because testing programs and their effects change all the time, reinterpretations of the research that bears on this issue will be needed every few years. But at this time, in contradiction to all the rhetoric, the research informs us that states that have implemented high-stakes testing policies have fared worse on independent measures of academic achievement than have states with no or low stakes testing programs. (Note 44) The research also informs us that high-stakes testing policies have had a disproportionate negative impact on students from racial minority and low socioeconomic backgrounds. (Note 45)

In Arizona, for example, officials reported that in 1999 students in poor and high-minority school districts scored lower than middle-class and wealthy students on Arizona's high-stakes high school graduation test, the AIMS (Arizona's Instrument to Measure Standards). Ninety-seven percent of African Americans, Hispanics, and Native Americans failed the math section of the AIMS, a significantly greater proportion of failures than occurred in the white community, whose students also failed the test in great numbers. (Note 46) Due to the high failure rates for different groups of students, as well as various psychometric problems, this test had to be postponed.

In Louisiana parents requested that the office for civil rights investigate why nearly half the children in school districts with the greatest numbers of poor and minority children had failed Louisiana's test, after taking it for a second time. (Note 47) In Texas, in 1997, only one out of every two African American, Mexican American, and economically disadvantaged sophomores passed each section of Texas' high-stakes test the TAAS –

Texas' Assessment of Academic Skills. In contrast, four out of every five white sophomores passed. (Note 48) In Georgia, two out of every three low-income students failed the math, English, and reading sections of Georgia's competency tests. *No* students from well-to-do counties failed any of the tests and more than half exceeded standards. (Note 49)

The pattern of failing scores in these states are quite similar to the failure rates in other states with high school graduation exams and are illustrative of the achievement gap between wealthy, mostly white school districts and poor, mostly minority school districts. (Note 50) It appears that a major cause of these gaps is that high-stakes standardized tests may be testing poor students on material they have not had a sufficient opportunity to learn.

Education, Learning, and Training: Three Goals of Schooling

In this report we look at just one of the distorting and corrupting possibilities suggested by Heisenberg's Uncertainty Principle applied to the testing movement, namely, that *training* rather than *learning* or general *education* is taking place in communities that rely on high-stakes tests to reform their schools. As will become clearer, if we have doubt about the meaning of a test score, we must be skeptical about the validity of the test.

Our interest in these distinctions between training, learning and education stems from the many anecdotes and research reports we read that document the narrowing of the curriculum and the inordinate amount of time spent in drill as a form of test preparation, wherever high-stakes tests are used. The former president of the American Association of School Administrators, speaking also as the Superintendent of one of the highest achieving school districts in America, notes that:

The issue of teaching to these tests has become a major concern to parents and educators. A real danger exists in that the test will become the curriculum and that instruction will be narrow and focused on facts.

... Teachers believe they spend an inordinate amount of time on drills leading to the memorization of facts rather than spending time on problem solving and the development of critical and analytical thinking skills. Teachers at the grade levels at which the test is given are particularly vulnerable to the pressure of teaching to the test.

Rather than a push for higher standards, [Virginia's high-stakes] tests may be driving the system toward mediocrity. The classroom adaptations of "Trivial Pursuit" and "Do You Want to be a Millionaire?" may well result in higher scores on these standardized tests, but will students have acquired the breadth and knowledge to do well on other quality benchmarks, such as the SAT and Advanced Placement exams? (Note 51)

This is our concern as well. Any narrowing of the curriculum, along with the confusion of training to pass a test with broader notions of learning and education are especially problematic side effects of high-stakes testing for low-income students. The poor, more than their advantaged peers, need not only the skills that training provides but need the more important benefits of learning and education that allow for full economic and social integration in our society.

To understand the design of this study and to defend the measures used for our inquiry requires a clarification of the distinctions between the related concepts of *education*,

learning (particularly *school learning* and the concept of *transfer of learning*), and *training*. For most citizens it is education (the broadest and most difficult to define of the concepts) that is the goal of schooling. Learning is the process through which education is achieved. But merely demonstrating acquisition of some factual or procedural knowledge is not the primary goal of school learning. That is merely a proximal goal.

The proper goal of school learning is both more distal and more difficult to assess. The proper goal of school learning is transfer of learning, that is, the application or use of what is learned in one domain or context to that of another domain or context. School learning in the service of education focuses deliberately on the goal of broad (or far) transfer. School instruction that can be characterized as training is ordinarily a narrow form of learning, where transfer of learning is measured on tasks that are highly similar to those used in the training. Broad or far measures of transfer, the appropriate goal of school learning, are different from the measures typically used to assess the outcomes of training.

More concretely, training in holding a pencil, or of doing two-column addition with regrouping, or memorizing the names of the presidents, is expected to yield just that. After training to do those things is completed students should be able to write in pencil, add columns of numbers, and name the presidents. The assessments used to measure their newly acquired knowledge are simple and direct. On the other hand, learning to write descriptive paragraphs, arguing about how numbers can be decomposed, and engaging in civic activities should result in better writing, mathematics and citizenship. To inquire whether that is indeed the case, much broader and more distal measures of transfer are required and these kinds of outcomes of education are much harder to measure.

Although enormously difficult to define, almost all citizens agree that school learning is designed to produce an "educated" person. Howard Gardner provides one voice for these aspirations by claiming that students become educated by probing, in sufficient depth, a relatively small set of examples from the disciplines. In Gardner's curriculum teachers lead students to think and act in the manner of scientists, mathematicians, artists, or historians. Gardner advocates deep and serious study of a limited set of subject matter to provide students with opportunities to deal seriously with the genuine and profound ideas of humankind.

I believe that three very important concerns should animate education; these concerns have names and histories that extend far back into the past. There is the realm of *truth*—and its underside, what is false or indeterminable. There is the realm of *beauty* — and its absence in experiences or objects that are ugly or kitschy. And there is the realm of *morality* — what we consider to be good, and what we consider to be evil. (Note 52)

Gardner's "educated" student thinks like those in the disciplines because the students learn the forms of argument and proof that are appropriate to a discipline. Thus tutored, students are able to analyze the fundamental ideas and problems that all humans struggle with. It is a discussion and project-oriented curriculum, with minimum concern for test preparation as a separate activity. Gardner's discipline-based curriculum is explicitly concerned with transfer to a wide array of human endeavors. Despite the difficulty in obtaining evidence of this kind of transfer of learning, there is ample support for this kind of curriculum. Earl Shorris recently demonstrated the effect of this kind of curriculum with desperately poor people who were given the chance to study the disciplines with excellent and caring teachers. (Note 53) The experience of studying art, music, moral philosophy, logic, and so forth, transformed the lives of these

impoverished young adults.

Minnesota Senator Paul Wellstone also understands that school learning is not an end in itself. For him, our educational system should be designed to produce an "educated" person, someone for whom transfer of what is learned in school is possible:

Education is, among other things, a process of shaping the moral imagination, character, skills and intellect of our children, of inviting them into the great conversation of our moral, cultural and intellectual life, and of giving them the resources to prepare to fully participate in the life of the nation and of the world." (Note 54)

Senator Wellstone, however, sees a problem with this goal:

Today in education there is a threat afoot,...: the threat of high-stakes testing being grossly abused in the name of greater accountability, and almost always to the serious detriment of our children." (Note 55)

The Senator, like many others, recognizes the possible distorting and corrupting effects of high-stakes testing. He worries about compromising the education of our students, because of "a growing set of classroom practices in which test-prep activities are usurping a substantive curriculum." (Note 56) The Senator is concerned that test preparation for the assessment of narrow curricular goals will turn out to be more like training than like the kind of learning that promotes transfer. And if that were to be the case, the test instruments themselves are likely to be narrow and near measures of transfer, as befits training programs. If this scenario were to occur, then broad and far measures of transfer, the indicators, we hope, of the educated person that we hold as our ideal, might not become part of the ways in which we assess what is being learned in our schools.

To reiterate: education (in some broad and hard-to-define way) is our goal. School learning is the means to accomplish that goal. But, as a recent National Academy of Science/National Research Council report on school learning makes clear, schooling that too closely resembles training, as in preparation for testing, *cannot* accomplish the task the nation has set for itself, namely, the development of adaptive and educated citizens for this new millennium. (Note 57) Of course, school learning that promotes transfer is only a necessary, and not a sufficient condition, to bring forth an educated person. The issue, however, is whether high-stakes tests, with their potential for distorting and corrupting classroom life, can overcome the difficulties inherent in such systems, and thereby bring about the transformation in student achievements sought by all concerned with public education. One of the nation's leading experts on measurement has thought about this issue:

As someone who has spent his entire career doing research, writing, and thinking about educational testing and assessment issues, I would like to conclude by summarizing a compelling case showing that the major uses of tests for student and school accountability during the past 50 years have improved education and student learning in dramatic ways.

Unfortunately, I cannot. Instead, I am led to conclude that in most cases the instruments and technology have not been up to the demands that have been placed on them by high-stakes accountability. Assessment systems that are useful monitors lose much of their dependability and credibility for that purpose when high-stakes are attached to them. The unintended negative effects of high-stakes accountability uses often outweigh the intended

positive effects." (Note 58)

Transfer of learning and test validity. This report looks at one of the effects claimed for high-stakes testing: that states with high-stakes tests will show evidence that some kind of broad learning, rather than just some kind of narrow training, has taken place. It is well known that test preparation, meticulous alignment of the curriculum with the test, as well as rewards and sanctions for students and other school personnel, will almost always result in gains on whatever instrument is used by the state to assess its schools. Scores on almost all assessment instruments are quite likely to go up as school administrators and teachers *train* students to do well on tests such as the all-purpose widely-used SAT-9s in California, or the customized Texas Assessment of Academic Skills (TAAS), the Arizona Instrument to Measure Standards (AIMS), or the Massachusetts Comprehensive Assessment System (MCAS). We ask a more important question than "Do scores rise on the high-stakes tests?" We ask whether there is evidence of student learning, *beyond the training that prepared them for the tests they take*, in those states that depend on high-stakes tests to improve student achievement? We seek to know whether we are getting closer to the ideal we all hold of a broadly educated student, or whether we are instead developing students that are much more narrowly trained to be good test takers. It is important to note that this is not just a question of how well the nation is reaching its intended outcomes, it is also an equally important psychometric question about the validity of the tests, as well.

The National Research Council cautions that "An assessment should provide representative coverage of the content and processes of the domain being tested, so that the score is a valid measure of the student's knowledge of the broader [domain], not just the particular sample of items on the test." (Note 59)

So the score a student obtains on a high-stakes test must be an indicator of transfer or generalizability or that test is not valid. The problem is that:

1. tests almost always are made up of fewer items than the number actually needed to thoroughly assess the entire domain that is of interest;
2. testing time, as interminable as it may seem to the students, is rarely enough to adequately sample all that is to be learned from a domain; and
3. teachers may narrow what is taught in the domain so that the scores on the test will be higher, though by doing this, the scores are then invalid since they no longer reflect what the student knows of the entire domain.

These three factors work against having high-stakes test scores accurately reflect students' domain scores in areas such as reading, writing, science, etc. Because of this constant threat of invalidity, attaching high-stakes to achievement tests of this type may be impossible to do sensibly. (Note 60)

How might this show up in practice? Unfortunately there is already research evidence that reading and writing scores in Texas may not reflect the domains that are really of interest to us. The Heisenberg Uncertainty Principle applied to assessment seems may be at work distorting and corrupting the Texas system. The ensuing uncertainty about the meaning of the test scores in Texas requires skepticism about whether that state obtained valid indicators of the domain scores that are really of interest. That is, we have no assurance that the performance on the test indicates what it is supposed to, namely, transfer or generalizability of the performance assessed to the domain that is of interest to us. For example,

... high school teachers report that although practice tests and classroom drills have raised the rate of passing for the reading section of the TAAS at

their school, many of their students are unable to use those same skills for actual reading. These students are passing the TAAS reading section by being able to select among answers given. But they are not able to read assignments, to make meaning of literature, to complete reading assignments outside of class, or to connect reading assignments to other parts of the course such as discussion and writing.

Middle school teachers report that the TAAS emphasis on reading short passages, then selecting answers to questions based on those short passages, has made it very difficult for students to handle a sustained reading assignment. After children spend several years in classes where "reading" assignments were increasingly TAAS practice materials, the middle school teachers in more than one district reported that [students] were unable to read a novel even two years below grade level. (Note 61)

A similar phenomenon exists in testing writing, where a single writing format is taught—the five paragraph persuasive essay. Each paragraph has exactly five sentences: a topic sentence, three supporting sentences, and a concluding sentence much like the introductory sentence. The teachers call this "TAAS writing," as opposed to "real writing."

Teachers of writing who work with their students on developing ideas, on finding their voice as writers, and on organizing papers in ways appropriate to both the ideas and the papers' intended audience find themselves in conflict with this prescriptive format. The format subordinates ideas to form, sets a single form out as "the essay," and produces predictably, rote writing. Writing as it relates to thinking, to language development and fluency, to understanding one's audience, to enriching one's vocabulary, and to developing ideas has been replaced by TAAS writing to this format. (Note 62)

California also has well documented instances of this. The curriculum was so narrowed to reflect the high-stakes SAT 9 exam, and the teachers under such pressure to teach just what is on the test, that they voluntarily felt obliged to add a half hour a day of unpaid teaching time to the school schedule. As one teacher said:

This year [we] ... extended our day a half hour more. And this is exclusively to do science and social studies. ... We think it's very important for our students to learn other subjects besides Open Court and math ... because in upper grades, their literature, all that is based on social studies, and science and things like that. And if they don't get that base from the beginning [in] 1st [and] 2nd grade, they're going to have a very hard time understanding the literature in upper grades There is no room for social studies, science. So that's when we decided to extend our day a half hour But this is a time for us. With that half hour, we can teach whatever we want, and especially in social studies and science and stuff, and not have to worry about, "OK, this is what we have to do." It's our own time, and we pick what we want to do. (Interview, 2/19/01) (Note 63)

In this school the stress to teach to the test is so great that some teachers violate their contract and take an hourly cut in pay in order to teach as their professional ethics demand of them. Such action by these teachers—in the face of serious opposition by some of their colleagues—is a potent indicator of how great the pressure in California is to narrow the curriculum and relentlessly prepare students for the high-stakes test. The paradox is, that by doing these things, the teachers actually invalidate the very tests on

which they work so hard to do well. It is not often pointed out that *the harder teachers work to directly prepare students for a high-stakes test, the less likely the test will be valid for the purposes it was intended.*

Test preparation associated with high-stakes testing becomes a source of invalidity if students had differential test preparation—as often happens in the case of rich and poor students who take the SAT for college entrance. But even if all the students had intensive test preparation the potential for invalidity exists because the scores on the test may then no longer represent the broader domain of knowledge for which the test score was supposed to be an indicator. Under either of these circumstances, where there is differential preparation for the tests by *different* groups of students, or intensive test preparation by *all* the students, there is still a way to make a distinction between training effects and the broader more desirable learning effects. That distinction can be made by using transfer measures, that is, other measures of the same domain as the high-stakes test but where no intensive test preparation occurred. The scores of students on tests of the same or similar domains as those measured by the high-stakes test can help to answer the question about whether learning in the broad domain of knowledge is taking place, as intended, or whether a narrow form of learning is all that occurs from the test preparation activities. If scores on these other tests rise along with the scores on the state tests then genuine learning would appear to be taking place. The claim that transfer within the domain is occurring can then be defended, and support will have been garnered for the high-stakes testing programs now sweeping the country. We will now examine data that help to answer these questions about whether broad-based learning or narrow forms of training are occurring.

Design of the Study

The purpose of this study is to inquire whether the high-stakes testing programs promote the transfer of learning that they are intended to foster. A second report in this series inquires if there have been negative side-effects of high-stakes testing for economically disadvantaged and ethnic minority students (see "The Unintended Consequences of High-Stakes Testing by A. L. Amrein & D. C. Berliner, forthcoming, at <http://www.edpolicyreports.org/>). The sample of states used to assess the intended and unintended effects of high-stakes testing are the eighteen states that have the most severe consequences, that is, the highest stakes associated with their K–12 testing policies: Alabama, Florida, Georgia, Indiana, Louisiana, Maryland, Minnesota, Mississippi, Nevada, New Jersey, New Mexico, New York, North Carolina, Ohio, South Carolina, Tennessee, Texas, and Virginia. Table 1 describes the stakes that exist in each of these states at this time.

Table 1
Consequences/"Stakes" in K–12 Testing Policies in States that
Have Developed Tests with the Highest Stakes (Note 64)

States	Total Stakes	Grad. exam ^a	Grade prom. exam ^b	Public report cards ^c	Id. low perform. ^d	\$ awards to schools ^e	\$ awards to staff ^f	State may close low perform. ^g	State may replace staff ^h	Students may enroll elsewhere ⁱ	\$ awards to students ^j
Alabama	6	X		X	X	X		X	X		
Florida	6	X		X	X	X	X			X	
Georgia	5	X	2004 (Note 65)	X	X	X	X	2004			
Indiana	6	X		X	X	X		X		X	

Louisiana	7	X	X (Note 66)	X	X			X	X	X	
Maryland	6	X		X	X	X		X	X		
Minnesota	2	X		X							
Mississippi	3	X		X	X	2003		2003			
Nevada	6	X		X	X			X	X		X
New Jersey	4	X		X	X	X					
New Mexico	7	X	X (Note 67)	X	X	X		X	X		
New York	5	X		X	X			X	X		
North Carolina	8	X	X (Note 68)	X	X	X	X	X	X (Note 69)		
Ohio	6	X	2002 (Note 70)	X	X	X	X				X
South Carolina	6	X	2002 (Note 71)	X	X	X		X	X		
Tennessee	6	X		X	X	X	X	X			
Texas	8	X	2003 (Note 72)	X	X	X	X	X	X (Note 73)	X	
Virginia	4	X		X	X			X			

- ^aGraduation contingent on high school grad. exam.
- ^bGrade promotion contingent on exam.
- ^cState publishes annual school or district report cards.
- ^dState rates or identifies low performing schools according to whether they meet state standards or improve each year.
- ^eMonetary awards given to high performing or improving schools.
- ^fMonetary awards can be used for "staff" bonuses.
- ^gState has the authority to close, reconstitute, revoke a school's accred. or takeover low performing schools.
- ^hState has the authority to replace school personnel due to low test scores.
- ⁱState permits students in failing schools to enroll elsewhere.
- ^jMonetary awards or scholarships for in- or out of state college tuition are given to high performing students.

These states have not only the most severe consequences written into their K–12 testing policies but lead the nation in incidences of school closures, school interventions, state takeovers, teacher/administrator dismissals, etc., and this has occurred, at least in part, because of low test scores. (Note 74) Further, these states have the most stringent K–8 promotion/retention policies and high school graduation exam policies. They are the only states in which students are being retained in grade because of failing state tests and in which high school students are being denied regular high school diplomas, or are simply not graduating, because they have not passed the state's high school graduation exam. These data on denial of high school diplomas are presented in Table 2.

Table 2
Rates at Which Students Did Not Graduate or Receive a High School

Diploma Due to Failing the State High School Graduation Exam (Note 75)

State (Note 76)	Grade in which students first take the exam	Percent of students who did not graduate or receive a regular high school diploma because they did not meet the graduation requirement (Note 77)	Year
Alabama*	10	5.5%	2001
Florida*	11	5.5%	1999
Georgia*	11	12%	2001
Indiana*	10	2%	2000
Louisiana	10 & 11	4%	2001
Maryland	6	4%	2000
Minnesota	8	2%	2001
Mississippi*	11	n/a (Note 78)	n/a
Nevada	11	3%	2001
New Jersey	11	6%	2001
New Mexico*	10	n/a	n/a
New York	n/a (Note 79)	10%	2000
North Carolina*	9 (Note 80)	7%	2000
Ohio	8	2%	2000
South Carolina	10	8%	1999
Tennessee	9	2.5%	2001
Texas	10	2%	2001
Virginia*	6	0.5%	2001

The effects of high-stakes tests on *learning* were measured by examining indicators of student learning, academic accomplishment and achievement *other* than the tests associated with high-stakes. These other indicators of student learning serve as the transfer measures that can answer our question about whether high-stakes tests show merely training effects, or show transfer of learning effects, as well. The four different measures we used to assess transfer in each of the states with the highest stakes were:

1. the ACT, administered by the American College Testing program;
2. the SAT, the Scholastic Achievement Test, administered by the College Board;
3. the NAEP, the National Assessment of Educational Progress, under the direction of the National Center for Education Statistics and the National Assessment Governing Board; and
4. the AP exams, the Advanced Placement examination scores, administered by the College Board.

In each state, for each test, participation rates in the testing programs were also examined since these vary from state-to-state and influence the interpretation of the

scores a state might attain.

Transfer measures to assess the effects of high-stakes tests. As noted above, psychometricians teach us that one facet of validity is that the scores on a test are indicators of performance in the domain from which the test items are drawn. Thus, the score a student gets on a ten-item test of algebra, or on their driving test, ought to provide information about how that student would score on any of the millions of problems we could have chosen from the domain of algebra, or on how that student might drive in innumerable traffic situations. The score on the short classroom assessment, or on the test of driving performance, is actually an indicator of the students' ability to transfer what they have demonstrated that they have learned to the other items and traffic situations that are similar to those on the assessment. In a sense, then, we don't really care much about the score that was obtained on either test. What we really want to know is whether that student can do algebra problems or drive well in traffic. So we are interested in the score on the tests the student actually took only in so far as those scores represent what they know or can do in the domain in which we are interested. This study seeks to clarify the relationship between the score obtained on a high-stakes test and the domain knowledge that the test score represents.

If, as in some states, scores on the state test go up, it is proper to ask whether the scores are also going up on other measures of the same domain. That is precisely what a gain score on a state assessment should mean. Gain scores should be the indicators of increased competency in the domain that is assessed by the tests, and that is why transfer measures that assess the same domain are needed. (Note 81)

If the high-stakes testing of students really induces teachers to upgrade curricula and instruction or leads students to study harder or better, then scores should also increase on other independent assessments. (Note 82) So we used the ACT, SAT, NAEP and AP exams as the other independent assessments, as measures of transfer. We are not alone in using these four measures to assess transfer of learning. For example, one analyst of the Texas high-stakes program believes: "If Texas-style systemic reform is working as advertised, then the robust achievement gains that TAAS reports should also be showing up on other achievement tests such as the National Assessment of Educational Progress (NAEP), Advanced Placement exams and tests for college admission." (Note 83)

In addition, the RAND Corporation recently used this same logic to investigate the validity of impressive gains on Kentucky's high-stakes tests. The researchers compared the students' performance on Kentucky's state test with their performance on comparable tests such as the NAEP and the ACT. Gains on the state test did not match gains on the NAEP or ACT tests. They concluded the Kentucky state test scores were seemingly inflated and were not a meaningful indicator of increased student learning in Kentucky. (Note 84)

In assessing the effects of testing in Texas, other RAND researchers noted "Evidence regarding the validity of score gains on the TAAS can be obtained by investigating the degree to which these gains are also present on other measures of these same skills." (Note 85)

Because some test data from the states with high-stakes tests do not show evidence of learning on some of the transfer measures, journalist Peter Schrag noted that "...the unimpressive scores on other tests raise unavoidable questions about what the numbers really mean [on the high-stakes tests] and about the cost of their achievement." (Note 86)

The National Research Council also supports transfer measures of the type we use by

relying on such data in their own analysis. They note, with dismay, that "There is some evidence to indicate that improved scores on one test may not actually carry over when a new test of the same knowledge and skills is introduced." (Note 87)

Sampling concerns. In each state the ACT and SAT tests are designed to measure the achievements of various percentages of the 60–70 percent of the total high school students in a state who intend to go to college. Within each state these tests probably attract a broad sample of students intending to go to college, while the AP tests are probably given to a more restricted and higher achieving sample of students. But in all three cases the samples are *not* representative of the state's high school graduates. However, these are all high-stakes tests for the students, with each test influencing their future. Thus, their motivation to do well on the state's high-stakes test and these other indicators of achievement is likely to be similar. This leads to a conservative test of transfer of learning, because it ought to be easier to find indicators of transfer, if it occurs, among these generally higher ability, more motivated students, rather than in a sample that included all the students in a state.

Motivation to achieve well may be diminished in the case of the NAEP because no stakes are attached to those tests. But the NAEP state data is obtained from a random sample of the states' schools, and thus may provide the most representative sample among the four measures of transfer of learning we use. Nevertheless, even with NAEP there is a problem. At each randomly selected school it is the local school personnel who decide if individual students will participate in NAEP testing. As will become clear later, sometimes the participation rates in NAEP testing seem suspect, leading to concerns about the appropriateness of the NAEP sample, as well.

In each high-stakes state, from the year in which the first graduating class was required to pass a high school graduation examination, we asked: What happened to achievement in the domains assessed by the American College Test (ACT), in the domains assessed by the Scholastic Achievement Test (SAT), in the domains assessed by the National Assessment of Educational Progress (NAEP), (Note 88) and in the domains assessed by the Advanced Placement (AP) tests. We asked also how participation rates in these testing programs changed and might have affected interpretations of any effects found.

An archival time-series research design was chosen to examine the state-by-state and year-to year data on each transfer measure. Time-series studies are particularly suited for determining the degree to which large-scale social or governmental policies make an impact. (Note 89) In archival time-series designs strings of observations of the variables of interest are made before, and after, some policy is introduced. The effects of the policy, if any, are apparent in the rise and fall of scores on the variable of interest.

We may consider the implementation of the state policy to engage in high-stakes testing as the independent variable, or treatment, and the scores from year to year on the ACT, SAT, NAEP and AP tests, before and after the implementation of high-stakes testing, as four dependent variables of interest. Relationships between the treatments and effects (between independent and dependent variables) are demonstrated by studying the pattern in the trend lines before and after the intervention(s), that is, before and after it was mandatory to pass state tests. (Note 90) Table 3 presents the dates at which high school graduation requirements of this type were first introduced in the eighteen states under study.

Table 3
Years in Which High School Graduation Exams
Affected Each Graduating Class (Note 91)

	Graduating classes required to pass different
--	--

		graduation exams to receive a regular high school diploma.				
State	Year in which the state's 1st graduation exam policy was introduced	1st Exam Class of...	2nd Exam Class of...	3rd Exam Class of...	4th Exam Class of...	5th Exam Class of...
Alabama	1983	1985	1993	2001, 2002, 2003		
Florida	1976	1979	1990	1996	2003	
Georgia	1981	1984	1995	1997, 1998	Future (Note 92)	
Indiana	1996	2000				
Louisiana	1989	1991	2003, 2004			
Maryland	1981	1987	2007			
Minnesota	1996	2000				
Mississippi	1988	1989	2003, 2004, 2005, 2006			
Nevada	1979	1981	1985	1992	1999	2003
New Jersey	1981	1984	1987	1995	2003, 2004, 2006	
New Mexico	1988	1990				
New York	1960s (Note 93)	1985	1995	2000, 2001, 2002, 2003, 2004, 2005		
North Carolina	1977	1980	1998 (Note 94)	2005		
Ohio	1991	1991	1994	2007		
South Carolina	1986	1990	2005, 2006, 2007	Future		
Tennessee	1982	1986	1998	2005		
Texas	1980	1983 (Note 95)	1987	1992	2005	
Virginia	1983	1986	2004			

Two strategies were used to help evaluate the strength of the effects of the high-stakes testing policy, and our confidence in those effects. First, data points before the introduction of the tests provided baseline information. (Note 96) Whether changes in the transfer measure occurred was determined by comparing the post intervention data with the baseline or pre-intervention data. If there was a change in the trend line for the data, just after intervention occurred, it was concluded that the treatment had an effect.

Secondly, national trend lines were positioned alongside state trend lines to help control for normal fluctuations and extraneous influences on the data. (Note 97) The national group was used as a nonequivalent comparison group to help estimate how the dependent variable would have oscillated if there had been no treatment. (Note 98) The national trend lines controlled for whether effects at the state level were genuine or just