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

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 uncertainly 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 disquised 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).

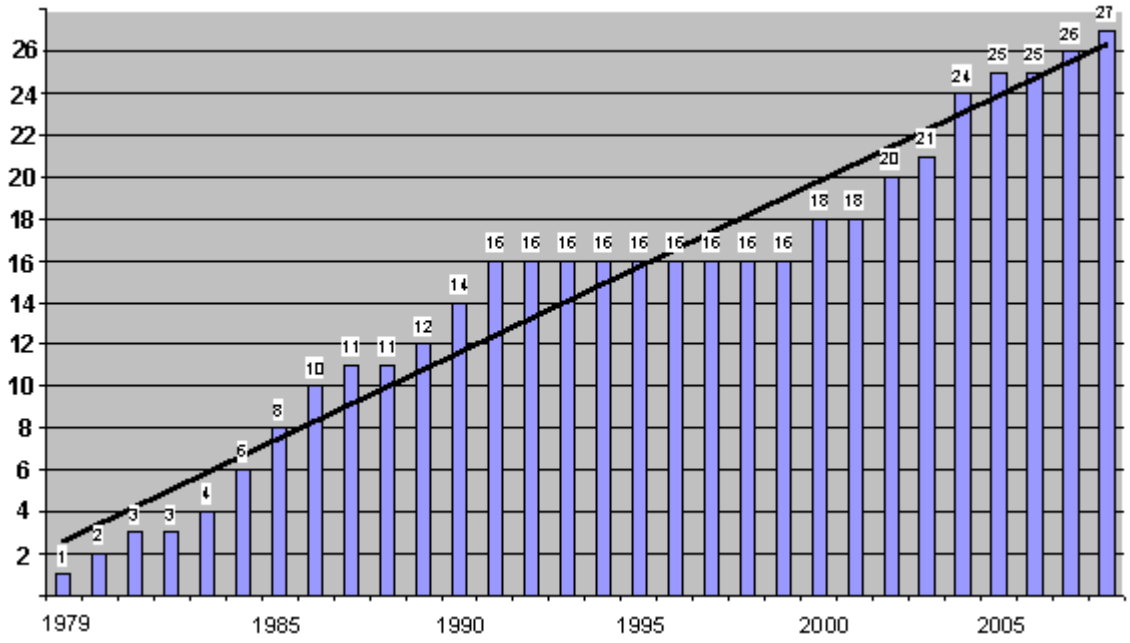

Figure 1. Number of states with high school graduation exams 1979–2008 (Note 29)

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 be 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. Gardener'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 else-where[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	X	X	X	X				X

			70)								
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			

[a]Graduation contingent on high school grad. exam.

[b]Grade promotion contingent on exam.

[c]State publishes annual school or district report cards.

[d]State rates or identifies low performing schools according to whether they meet state standards or improve each year.

[e]Monetary awards given to high performing or improving schools.

[f]Monetary awards can be used for "staff" bonuses.

[g]State has the authority to close, reconstitute, revoke a school's accred. or takeover low performing schools.

[h]State has the authority to replace school personnel due to low test scores.

[i]State permits students in failing schools to enroll elsewhere.

[j]Monetary 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	Year

		did not meet the graduation requirement (Note 77)	
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, 20061			
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 reflections of national trends. Figure 2, using actual data from the state of Alabama, and presented again in Appendix B, illustrates how the archival time series and our analyses of effects worked.

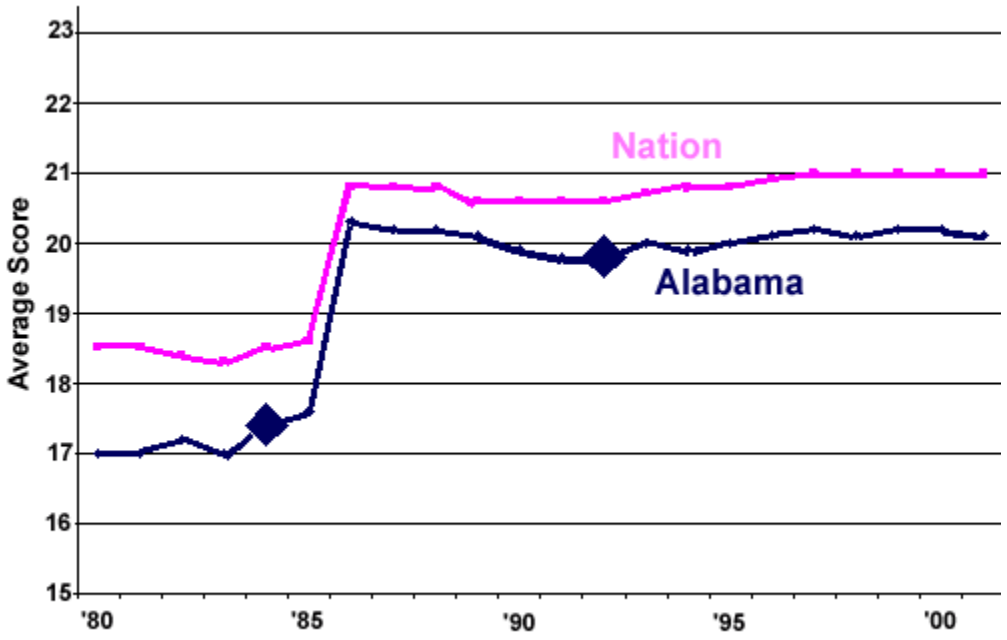

Figure 2. (From Appendix B) Analysis of the American College Test (ACT), Alabama (Note 99)

Alabama implemented its 1st high school graduation exam in 1983. It was a prerequisite for graduation that first affected the class of 1985. Alabama's 2nd exam first affected the class of 1993. The enlarged diamond shape signifies the year *before* the 1st graduating class was required to pass the exam. The policy intervention occurs in the year following the large diamond. From these data we conclude that:

- From 1984–1985 Alabama gained .1 point on the nation.
- From 1984–1992 Alabama gained .3 points on the nation.
- From 1992–1993 Alabama gained .1 point on the nation.
- From 1992–2001 Alabama lost .1 point to the nation.

To interpret these data, one inspects the state trend line and notes from the bold diamond shapes that there were two different points at which Alabama instituted high-stakes tests. After the first test was implemented, there was a score gain on the ACT in Alabama. (Note 100) After the second test there was an equally modest rise in Alabama's ACT scores. But in each case the national trend line showed similar effects, which moderates those conclusions. We can conclude from plotting the ACT scores each year that: 1) there were, indeed, small short term gains on the ACT in the year after new high-stakes tests were instituted; and 2) that the long term effects that may have occurred were substantial after the first test, but resulted in a small negative effect after the second high-stakes test was instituted. As can be seen, the national trend lines are quite important for interpreting the effects of a high-stakes testing policy on a measure of transfer.

A combined national trend line was used because the creation of a comparison group from the 32 states with no or low stakes attached to their tests was not feasible. Designation of which category a state was in changed from year to year so there were never clear cases of "states with high-stakes tests" and a comparison group made up of

"states without high-stakes tests" across the years. Using the combined national trend line was the best comparison group available, even though this trend line included each of the states that were under analysis and the other 17 states that we designated as high-stakes states and were also the object of study. Because of these factors there are some difficulties in the comparison of the state and national trend lines, perhaps introducing some bias toward or against finding learning effects when comparing state trend lines with the national trend lines. If such bias exists, we believe its effects would be minimal.

Sources of Data

In an archival time series analysis, effects of the independent variable were measured using historical records and data collected from agency and governmental archives (Note 101) and extensive telephone calls and emails to and from agency personnel and directors. The following state-level data archives were collected:

American College Test (ACT)

- ACT composite scores – 1980–2001
- ACT participation rates – 1994–2001

SAT

- SAT composite scores – 1977–2001
- SAT participation rates – 1991–2001

National Assessment of Educational Progress (NAEP)

- NAEP Grade 4 Mathematics composite scores – 1992, 1996, 2000
- NAEP Grade 8 Mathematics composite scores – 1990, 1992, 1996, 2000
- NAEP Grade 4 Reading composite scores – 1992, 1994, 1998
- NAEP Grade 8 Reading composite scores – 1998

Advance Placement (AP)

- Percentage of 11th / 12th graders who took AP exams 1991–2000
- Percentage of 11th / 12th graders receiving a 3 or above 1995–2000

State summaries for each of the 18 states with the highest stakes written into their K–12 testing policies were constructed to facilitate the time series analysis. These are presented in Appendix A. The summaries include contextual and historical information about each state's testing policies. Each summary should help readers gain more insight about each state's testing policies and the values each state attributes to high-stakes tests, beyond the information offered in Table 1. Most importantly, each summary includes background information regarding the key intervention points, or years in which graduating seniors were first required to pass different versions of high school graduation exams as summarized in Table 3. These intervention points were illustrated in each archival time series graph, and each interpretation of state data relied on what happened after these key points in time. The archival time series graphs for each of the transfer measures we used are included in the different Appendices. The data associated

with each of the transfer measures will now be described.

The American College Test (ACT) and the Scholastic Achievement Test (SAT)

The American College Test (ACT) (Note 102) and Scholastic Achievement Test (SAT) (Note 103) are the two predominant entrance exams taken by students prior to enrolling in higher education. College-bound students take the ACT or SAT to meet in-state or out-of-state university enrollment requirements. Scores on these tests are used by college admissions officers as indicators of ability and academic achievement and are used in decisions about whether an applicant has the minimum level of knowledge to enter into, and prosper, at the college to which they applied. Although many studies have been conducted questioning the usefulness of these tests in predicting a student's actual success after enrolling in college, they continue to be widely used by universities when accepting students into their institutions.. (Note 104).

Despite questions about their predictive validity, both ACT and SAT scores can be considered as sensible indicators of academic achievement in the domains that constitute the general high school curriculum of the United States. Averaged at the state level, both tests can be thought of as external and alternative indicators of achievement by the students of a particular state. Both of these tests can serve as measures of transfer of learning.

At this time, we know that the set of states without high-stakes tests perform better on the ACT and SAT. We do not know, however, how performance on the ACT and SAT tests changed after high school graduation exams were implemented in the 18 states that have introduced high-stakes testing policies. The objective of the first section of this inquiry is to answer this question.

There are, however, limitations to using these measures. For example, students who take the ACT and SAT are college-bound students and do not represent all students in a given state. But in 2001 38% and 45% of all graduating seniors took the ACT and the SAT tests, respectively. Although the sample of students is not representative, we can still use these scores to assess how high-stakes tests affected an average of approximately 2 out of every 5 students across the nation. Additionally, because participation rates vary by state we can use state participation rates to assess how in some states high-stakes tests affected the academic performance of more than 75% of graduating seniors.

It should be noted, as well, that some states are ACT states or states in which the majority of high school seniors take the ACT. Other states are SAT states or states in which the majority of high school seniors take the SAT. In Mississippi, for example, only 4% of high school seniors took the SAT in 2001 but in that same year 89% of high school seniors took the ACT. This would make Mississippi an ACT state. Whether states with high-stakes tests are ACT or SAT states should be taken into consideration to help us understand the sample of students who are taking the tests. If within Mississippi only 4% of high school seniors took the SAT it can be assumed that those students were probably among the brightest or most ambitious high school seniors in Mississippi. These students probably take the SAT because they were seeking out-of-state universities. Conversely, if 89% of high school seniors took the ACT, it can be assumed

that those students were probably a bit less talented or ambitious seniors, predominantly students trying to meet the requirements of the universities within the state of Mississippi. It is likely, however, that this sample also includes those seeking entrance to out of state universities that accept ACT scores. The participation rates for each test helps to decipher whether different samples of college bound students performed differently.

It should also be noted that the ACT and SAT tests are high-stakes tests. A student's score does influence to which colleges a student may apply and in which colleges a student may enroll. It seems likely, therefore, that students who take these tests are trying to achieve the highest scores possible. This would deflate arguments that students try harder on high school graduation exams than college entrance exams. If anything, the opposite might be true.

The purpose in the next two analyses is to assess how student learning changed in the domains represented by the ACT and SAT. Student scores and participation rates on these tests will be examined in each state after high-stakes high school graduation tests were implemented. Effects will be analyzed from the year in which the first graduating class was required to pass a high school graduation exam. It is also the purpose of the next two analyses to assess how high school seniors who are likely to be bound for out-of-state colleges, and seniors likely to be bound for in-state colleges, performed after high school graduation high-stakes exams were implemented.

American College Test (ACT)

The ACT data for each of the 18 states with high-stakes testing is included in Appendix B. Short-term, long-term, (Note 105) and overall achievement trends on the ACT were analyzed in the years following a states implementation of a high-stakes high school graduation exam. These analyses are summarized in Appendix B as well. The data and analysis for the state of Alabama, which we included as Figure 2, illustrated the way we examined each state's ACT data. A summary of those trends across the 18 states with the highest stakes is provided in Table 4.

Table 4
Results from the Analysis of ACT Scores (Note 106)

State	Effect after 1st HSGE		Effect after 2nd HSGE		Effect after 3rd HSGE		Effect after 4th HSGE		Overall Effects
	Short Term	Long Term	Short Term	Long Term	Short Term	Long Term	Short Term	Long Term	
Alabama	1984–'85 +0.1	1984–'92 +0.3	1992–'93 +0.1	1992–'01 –0.1					Positive
Florida	n/a	1980–'89 –0.4	1989–'90 –0.1	1989–'95 –0.2	1995–'96 –0.3 (+2%)	1995–'01 –0.6 (+5%)			Negative
Georgia	1983–'84 +0.2	1983–'94 –0.5	1994–'95 –0.1 (0%)	1994–'01 –0.6 (0%)					Negative

Indiana	1999–'00 +0.2 (–1%)	1999–'01 +0.2 (–1%)							Positive
Louisiana	1990–'91 0	1990–'01 –0.2							Negative
Maryland	1986–'87 +0.1	1986–'01 –0.6							Negative
Minnesota	1999–'00 –0.1 (0%)	1999–'01 0 (0%)							Negative
Mississippi	1988–'89 0	1988–'01 –0.4							Negative
Nevada	1980–'81 –0.1	1980–'84 +0.1	1984–'85 –0.3	1984–'91 +0.1	1991–'92 +0.2	1991–'98 +0.1	1998–'99 +0.1 (–1%)	1998–'01 –0.1 (–5%)	Positive
New Jersey	1983–'84 +0.3	1983–'86 –1.4	1986–'87 –0.2	1986–'94 –0.1	1994–'95 –0.5 (–1%)	1994–'01 –0.5 (–1%)			Negative
New Mexico	1989–'90 +0.1	1989–'01 –0.5							Negative
New York	1984–'85 –0.2	1984–'94 –0.5	1994–'95 +0.1 (–1%)	1994–'01 +0.4 (–6%)					Negative
North Carolina	1979–'80 n/a	1980–'97 –1.1	1997–'98 +0.1 (0%)	1997–'01 +0.4 (0%)					Negative
Ohio	1993–'94 +0.1	1993–'01 +0.1							Positive
South Carolina	1989–'90 +0.1	1989–'01 –0.5							Negative
Tennessee	1985–'86 +0.3	1985–'97 –0.3	1997–'98 +0.1 (–7%)	1997–'01 +0.3 (–6%)					Positive
Texas	1986–'87 +0.2	1986–'91 +0.7	1991–1'92 0	1991–'01 0					Positive
Virginia	1985–'86 –0.1	1985–'01 –1.3							Negative

From Table 4, looking at all the states simultaneously, and in comparison to the nation, we can evaluate short-term, long-term, and the overall effects of high stakes testing policies.

Short-term effects. In the short term, ACT gains were posted 1.6 times more often than losses after high school graduation exams were implemented. Short-term gains were evident sixteen times, losses were evident ten times, and no apparent effects were evident three times. But the gains and losses that occurred were partly artificial, because the states' short-term changes in scores were correlated ($-0.51 < r < 0.13$) (Note 107) to

the states short-term changes in participation rates. This modest negative correlation informs us that if the participation rate in ACT testing went down then the scores on the ACT went up, and vice versa. Under these circumstances it is hard to defend the thesis that there are reliable short-term gains from high-stakes tests.

Long-term effects. In the long term, and also in comparison to the nation, ACT losses were posted 1.9 times more often than gains after high school graduation exams were implemented. Long-term gains were evident ten times, losses were evident nineteen times, and no apparent effects were evident two times. These gains and losses were "real" given that the states' long-term changes in score were unrelated ($r = –0.18$) (Note 108) to the states' long-term changes in participation rates.

Overall effects. In comparison to the rest of the nation, negative ACT effects were displayed 2 times more often than positive effects after high-stakes high school graduation exams were implemented. Six states displayed overall positive effects, while twelve states displayed overall negative effects. In this data set overall losses or gains were unrelated to whether the percentage of students participating in the ACT increased or decreased.

Assuming that the ACT can serve as an alternative measure of the same or a similar domain as a state's high-stakes achievement tests, there is scant evidence of learning. Although states may demonstrate increases in scores on their own high-stakes tests, it appears that transfer of learning is not a typical outcome of their high-stakes testing policy. Sixty-seven percent of the states that use high school graduation exams posted *decreases* in ACT performance after high school graduation exams were implemented. These decreases were unrelated to whether participation rates increased or decreased at the same time. On average, the college-bound students in states with high school graduation exams decreased in levels of academic achievement as measured by the ACT.

One additional point about the ACT data needs to be made. In ACT states (states in which more than 50% of high school seniors took the ACT) students who are thought to be headed for in-state colleges were just slightly (1.3 times) more likely to post negative effects on the ACT. In SAT states (states in which less than 50% of high school seniors took the ACT) the students who are more likely bound for out-of-state colleges were 2.7 times more likely to post negative effects on the ACT. If anything, high school graduation exams hindered the performance of the brightest and most ambitious of the students bound for out-of-state colleges. Seventy-three percent of the states in which less than 50% of students take the ACT posted overall losses on the ACT.

Analysis of ACT Participation Rates. (Note 109) Just as ACT scores were used as indicators of academic achievement, ACT participation rates were used as indicators of the rates by which students in each state were planning to go to college. Arguably, if high school graduation exams increased academic achievement in some broad and general sense, an increase in the number of students pursuing a college degree would be noticed. An indicator of that trend would be increased ACT participation rates over time. So we examined changes in the rates by which students participated in ACT testing after the year in which the first graduating class was required to pass a high school graduation exam and for which data were available. These results are presented in Table 5.

Table 5

Results from the Analysis of ACT Participation Rates

State	Year in which students had to pass 1st HSGE to graduate	Change in % of students taking the ACT 1994–2001 as compared to the nation*	Overall Effects
Alabama	1985	+9%	Positive
Florida	1979	+4%	Positive
Georgia	1984	0%	Neutral
Indiana	2000	–1%	Negative
Louisiana	1991	+5%	Positive
Maryland	1987	–1%	Negative
Minnesota	2000	0%	Neutral
Mississippi	1989	+14%	Positive
Nevada	1981	–6%	Negative
New Jersey	1985	–1%	Negative
New Mexico	1990	0%	Neutral
New York	1985	–6%	Negative
North Carolina	1980	+2%	Positive
Ohio	1994	+2%	Positive

South Carolina	1990	+15%	Positive
Tennessee	1986	+10%	Positive
Texas	1987	–2%	Negative
Virginia	1986	+4%	Positive

*1999–2001 data were used for Indiana and Minnesota.

From this analysis we learn that from 1994–2001 ACT participation rates, as compared to the nation, increased in 50% of the states with high school graduation exams. When compared to the nation, participation rates increased in nine states, decreased in six states, and stayed the same in three states. Thus there is scant support for the belief that high-stakes testing policies within a state have an impact on the rate of college attendance.

The Scholastic Achievement Test (SAT)

The SAT data for each of the 18 states with high-stakes testing is included in Appendix C. Short-term, long-term, and overall achievement trends were analyzed following the states' implementation of their high-stakes high school graduation exam and these analyses are summarized in Appendix C, as well. The state of Florida was randomly chosen from this data set to illustrate what a time series for the SAT looks like. These data are provided in Figure 3. A summary of those trends across the 18 high-stakes testing states is provided in Table 6.

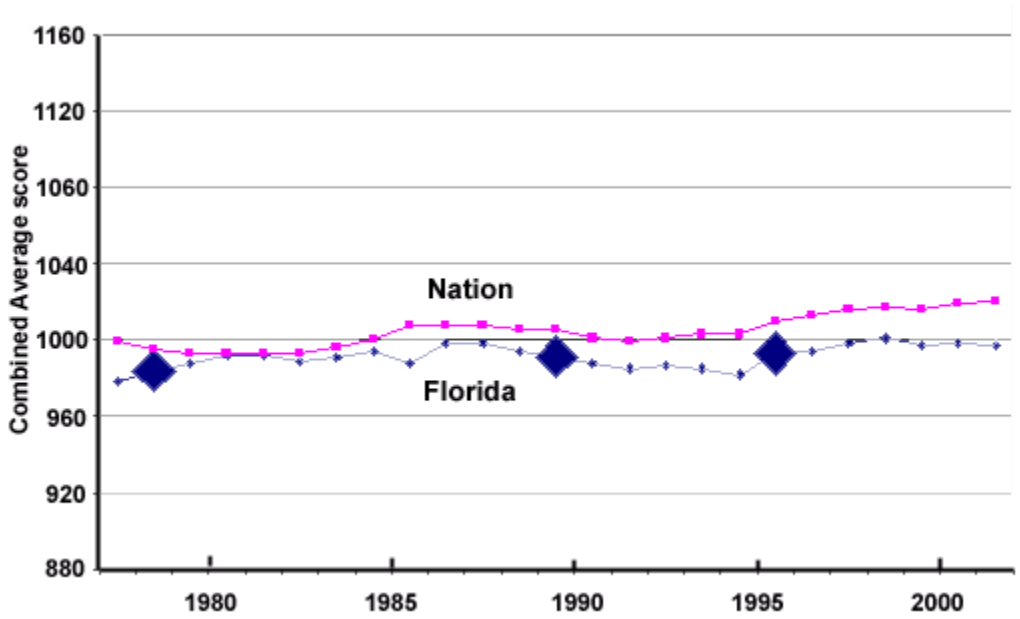

Figure 3. Florida: SAT scores

Florida implemented its 1st high school graduation exam in 1976. It was a prerequisite for graduation that first affected the class of 1979. Florida's 2nd exam first affected the class of 1990 and its 3rd exam the class of 1996 – see points of intervention (diamonds) enlarged to signify the year *before* the 1st graduating class was required to pass each exam:

- From 1978–1979 Florida gained 6 points on the nation.
- From 1978–1989 Florida lost 4 points to the nation.
- From 1989–1990 Florida gained 2 points on the nation.
- From 1989–1995 Florida lost 2 points to the nation.
- From 1995–1996 Florida lost 2 points to the nation.
- From 1995–2001 Florida lost 6 points to the nation.

Table 6
Results from the Analysis of SAT Scores Across the States (Note 110)

State	Effect after 1st HSGE		Effect after 2nd HSGE		Effect after 3rd HSGE		Effect after 4th HSGE		Overall Effects
	Short Term	Long Term	Short Term	Long Term	Short Term	Long Term	Short Term	Long Term	
Alabama	1984–85 +13	1984–92 +23	1992–93 +7 (0%)	1992–01 +4 (–2%)					Positive
Florida	1978–79 +6	1978–89 –4	1989–90 +2	1989–95 –2	1995–96 –2 (0%)	1995–01 –6 (+2%)			Negative
Georgia	1983–84 0	1983–94 +21	1994–95 –2 (+1%)	1994–2001 +10 (–5%)					Positive
Indiana	1999–00 +2	1999–01 +2							Positive
Louisiana	1990–91 +3	1990–01 +19							Positive
Maryland	1986–87 +3	1986–01 –6							Negative
Minnesota	1999–00 –12	1999–01 –19							Negative
Mississippi	1988–89 –13	1988–01 +7							Negative
Nevada	1980–81 +3	1980–84 –6	1984–85 –16	1984–91 –10	1991–92 +1 (+2%)	1991–98 –15	1998–99 +7	1998–01 –2	Negative
New Jersey	1983–84 –2	1983–86 +2	1986–87 +4	1986–94 +8	1994–95 –2 (0%)	1994–01 +1 (+7%)			Positive
New Mexico	1989–90 –3	1989–01 –29							Negative
New York	1984–85 –3	1984–94 –11	1994–95 –3 (–1%)	1994–2001 –6 (–2%)					Negative
North Carolina	1980–81 +7	1980–97 +32	1997–98 +3	1997–01 +10					Positive
Ohio	1993–94 +7	1993–01 –1							Positive
South Carolina	1989–90 +2	1989–01 +15							Positive
Tennessee	1985–86 –3	1985–97 +8	1997–98 +1	1997–01 –9 (–3%)					Negative
Texas	1986–87 –2	1986–91 +7	1991–92 –1 (0%)	1991–01 –8 (+6%)					Negative

Short–term effects. Looking across all the states simultaneously, and in comparison to the nation, we see that in the short term, SAT gains were posted 1.3 times more often than losses after high school graduation exams were implemented. Short-term gains were posted seventeen times, losses were posted thirteen times, and no apparent effects were posted once. But the gains and losses that occurred were partly artificial because the states' short-term changes in scores were related ($-0.60 < r < 0.38$) to the states short-term changes in participation rates. The negative correlations inform us that if the participation rate in SAT testing went down the scores on the SAT went up, and vice versa. The modest positive correlations inform us that in a few cases if the participation rate in SAT testing went down the scores on the SAT went down, and vice versa. Under these circumstances it is hard to defend the thesis that there are any reliable short-term gains on measures of general learning associated with high-stakes tests.

Long-term effects. In the long term, and also in comparison to the nation, SAT losses were posted 1.1 times more often than gains after high school graduation exams were implemented. Long-term gains were evident fifteen times, and losses were evident sixteen times. These gains and losses were partly artificial, however, given that the states' long-term changes in score were negatively correlated ($r = -0.41$) to the changes in participation rates for taking the SAT. The fewer students taking the test, the higher the SAT scores, and vice versa.

Overall effects. In comparison to the rest of the nation, negative SAT effects were posted 1.3 times more often than positive effects after high school graduation exams were implemented. Eight states displayed overall positive effects, while ten states displayed overall negative effects. But the gains or losses in score were related to increases and decreases in the percentage of students participating in the SAT. Thus it is hard to attribute any effects on the SAT to the implementation of high-stakes testing.

If we assume that the SAT is an alternative measure of the same or a similar domain as a state's own high-stakes achievement tests, then there is scant evidence of learning. Although states may demonstrate increases in scores on their own high-stakes tests, it appears that transfer of learning is not a typical outcome of their high-stakes testing policy. Fifty-six percent of the states that use high school graduation exams posted decreases in SAT performance after high school graduation exams were implemented. However, these decreases were slightly related to whether SAT participation rates increased or decreased at the same time. Thus, there is no reliable evidence that high-stakes high school graduation exams improve the performance of students who take the SAT. Gains and losses in SAT scores are more related to who participates in the SAT than the implementation of high school graduation exams.

One additional point about the SAT data needs to be made. In SAT states (states in which more than 50% of high school seniors took the SAT) students who are thought to be headed for in-state colleges were equally likely to post negative and positive effects on the SAT. In ACT states (states in which less than 50% of high school seniors took the SAT) the students who are more likely bound for out-of-state colleges were 1.7 times more likely to post negative effects on the SAT. If anything, high school graduation exams hindered the performance of the brightest and most ambitious of the students bound for out-of-state colleges. Sixty-three percent of the states in which less than 50% of students take the SAT posted overall losses on the SAT.

Analysis of SAT Participation Rates. Just as SAT scores were used as indicators of

academic achievement, SAT participation rates were used as indicators of the rates by which students in each state were planning to go to college. Arguably, if high school graduation exams increased academic achievement in some broad and general sense, an increase in the number of students pursuing a college degree would be noticed. An indicator of that trend would be increased SAT participation rates. So we examined changes in the rates by which students participated in SAT testing after the year in which the first graduating class was required to pass a high school graduation exam and for which data were available. These results are presented in Table 7.

Table 7
Results from the Analysis of SAT Participation Rates

State	Year students must pass 1st HSGE to graduate	Change in % of students taking the SAT 1991–2001 as compared to the nation*	Overall Effects
Alabama	1985	–2%	Negative
Florida	1979	+3%	Positive
Georgia	1984	–2%	Negative
Indiana	2000	–1%	Negative
Louisiana	1991	–5%	Negative
Maryland	1987	–2%	Negative
Minnesota	2000	–1%	Negative
Mississippi	1989	–3%	Negative
Nevada	1981	+5%	Positive
New Jersey	1985	+4%	Positive
New Mexico	1990	–2%	Negative

New York	1985	–1%	Negative
North Carolina	1980	+5%	Positive
Ohio	1994	+2%	Positive
South Carolina	1990	–4%	Negative
Tennessee	1986	–2%	Negative
Texas	1987	+6%	Positive
Virginia	1986	+5%	Positive

* 1993–2001 data were used for Ohio and 2000–2001 data were used for Indiana and Minnesota. Participation rates were not available for 1998 and 1999.

From this analysis we learn that from 1991–2001 (1993–2001 in Ohio, and 2000–2001 in Indiana and Minnesota) SAT participation rates, as compared to the nation, fell in 61% of the states with high school graduation exams. Participation rates in the SAT increased in seven states and decreased in eleven states. There is scant support for the belief that high-stakes testing policies will increase the rate of college attendance. Students did not participate in the SAT testing program at greater rates after high-stakes high school graduation exams were implemented.

National Assessment of Educational Progress (NAEP)

Some may argue that using ACT and SAT scores to assess the effects of high school graduation exams is illogical because high school graduation exams are specifically intended to raise the achievement levels of those students who are the most likely to fail – the poor, in general, and poor racial minorities, in particular. These students do not take the ACT or SAT in great numbers. But the effects of high-stakes policies on these particular populations can be assessed with data from the National Assessment of Educational Progress.(Note 111)

The National Assessment of Educational Progress (NAEP), commonly known as ‘the nation's report card," is the test administered by the federal government to monitor the condition of education in the nation's schools. NAEP began in 1969 as a national assessment of three different age or grade levels, for which students were randomly sampled and tested to provide information about the outcomes of the nation's various educational systems. In 1990 NAEP was expanded to provide information at the state level, allowing for the first time state-to-state comparisons.

States that volunteered to participate in NAEP could gauge how they performed in math

and reading in comparison to each other and to the nation, overall. This way states could assess the effects of the particular educational policies they had implemented. Under President Bush's national education policy, however, states are required to take the NAEP because it is believed to be the most robust and stable instrument the nation has to gauge learning and educational progress across all states.(Note 112) The federal government believes, as we do, that NAEP exams can be used to asses transfer, that NAEP is an alternate measure of the domains that are assessed by each of the states.

Weaknesses of the NAEP. It is proper to acknowledge that the NAEP has a number of weaknesses influencing interpretations of the data that we offer below. First, state level NAEP data pertain only to 4th and 8th grade achievement. The national student data set includes 12th grade data as well, and some additional subjects are tested, but at the state levels, only 4th and 8th grade achievement is measured. Given these circumstances it is not logical to attempt an assessment of the effects of implementing a high school graduation exam, or any other exam that is usually administered at the high school level, by analyzing NAEP tests given at the 4th or 8th grade. On the other hand, it is not illogical to make the assumption that other state reform policies went into effect at or around the same time as high-stakes high school graduation exams were put into place, *including the use of other high-stakes tests at lower grade levels*.(Note 113) The usefulness of the NAEP analyses that follow rest on the assumption that states' other K–12 high-stakes testing policies were implemented at or around the same time as each state's high school graduation exam. Table 1 describes these policies, and these policies are elaborated on in Appendix A. Other researchers who have used NAEP data to draw conclusions about the effects of high-stakes tests have used this logic and methodology as well.(Note 114)

Secondly, the NAEP does not have stakes attached to it. Students who are randomly selected to participate do not have to perform their best. However, because each student only takes small sections of the test, students appear to be motivated to do well and the scores appear to be trustworthy.(Note 115)

Third, states like North Carolina have aligned their state-administered exams with the NAEP, making for state-mandated tests that are very similar to the NAEP.(Note 116) In such cases gains in score on the NAEP may be related to similarities in test content rather than actual increases in school learning. States that align their tests with the NAEP have an unfair advantage over other states that aligned their tests with their state standards, but such imitative forms of testing occur. State tests that look much like the NAEP will probably become more common now that President Bush is attempting to attach stakes to the NAEP, and this will, of course, make the NAEP much less useful as a yardstick to assess if genuine learning of the domains of interest is taking place.

Finally, when analyzing NAEP data it is important to pay attention to who is actually tested. The NAEP sampling plan uses a multi-stage random sampling technique. In each participating state, school districts are randomly sampled. Then, schools within districts are randomly sampled. And then, students within schools are randomly sampled. Once the final list of participants is drawn, school personnel sift through the list and remove students who they have classified as Limited English Proficient (LEP) or who have Individualized Education Plans (IEPs) as part of their special education programs. Local personnel are required to follow "carefully defined criteria" in making determinations as to whether potential participants are "capable of participating."(Note 117) In short, although the NAEP uses random sampling techniques, not all students sampled are

actually tested. The exclusion of these students biases NAEP results.

Illusion from exclusion. Walter Haney found that exclusion rates explained gains in NAEP scores and vice versa. Texas, for example, was one state in which large gains in NAEP scores were heralded as proof that high-stakes tests do, indeed, improve student achievement. But Haney found that the percentages of students excluded from participating in the NAEP increased at the same time that large gains in scores were noted. Exclusion rates increased at both grade levels escalating from 8% to 11% at grade 4 and from 7% to 8% at grade 8 from 1992–1996. Meanwhile, in contrast, exclusion rates declined at both grade levels at the national level during this same time period, decreasing from 8% to 6% at grade 4 and from 7% to 5% at grade 8. Haney, therefore, termed the score gains in Texas an "illusion arising from exclusion."(Note 118)

Unfortunately, however, such illusions from exclusions hold true across the other states that use high-stakes tests. For example, North Carolina was the other state in which large gains in NAEP scores were heralded as proof that high-stakes testing programs improve student achievement. On the 4th grade NAEP math test North Carolina recorded an average composite score of 212 in 1992 and an average composite score of 232 in 2000. The nation's composite score increased from 218 to 226 over the same time period. North Carolina gained 20 points while the nation gained 8, making for what would seem to be a remarkable 12-point gain over the nation, the largest gain made by any state. But North Carolina excluded 4% of its LEP and IEP students in 1992 and 13% of its LEP and IEP students in 2000. Meanwhile, the nation's exclusion rate decreased from 8% to 7% over the same time period. North Carolina excluded 9% more of its LEP and IEP students while the nation excluded 1% less making for a 10% divergence between North Carolina's and the nation's exclusion rates from 1992–2000. North Carolina's grade 4 math 1992–2000 exclusion rates increased 325% while the nation's exclusion rate decreased. In addition, North Carolina's grade 8 math 1992–2000 exclusion rates increased 467% while the nation's exclusion rate stayed the same.

There is little doubt that the relative gains posted by North Carolina were partly, if not entirely, artificial given the enormous relative increase in the rates by which North Carolina excluded students from participating in the NAEP. The Heisenberg Uncertainty Principle appears to be at work in both Texas and North Carolina, leading to distortions and corruptions of the data, giving rise to uncertainty about the meaning of the scores on the NAEP tests.

North Carolina and Texas, however, are not the only states in which exclusionary trends were observed. In states with high-stakes tests, between 0%–49% of the gains in NAEP scores can be explained by increases in rates of exclusion. Similarly, 0%–49% of the losses in score can be explained by decreases in rates of exclusion over the same years.(Note 119) The more recent the data, the more the variance in NAEP scores can be explained by changes in exclusion rates. In short, states that are posting gains are increasingly excluding students from the assessment. This is happening with greater frequency as time passes from one NAEP test to the next. That is, as the stakes attached to the NAEP become higher, the Heisenberg Uncertainty Principle in assessment apparently is having its effects, with distortions and corruptions of the assessment system becoming more evident.

The state scores on the NAEP math and reading tests, at grades 4 and 8, will be used in our analysis to test the effects on learning from using high-stakes tests in states that have

implemented high-stakes high school graduation exams. Given that exclusion rates affect gains and losses in score, however, state exclusion rates will be presented along side the relative gains or losses posted by each state. In this way readers can make their own judgments about whether year-to-year gains in score are likely to be "true" or "artificial." The gains and losses in scores and exclusion rates have all been calculated in comparison to the pooled national data.

Analysis of NAEP Grade 4 Math Scores

For each state, after high-stakes tests were implemented, an analysis of NAEP mathematics achievement scores was conducted. The state of Georgia was randomly chosen to serve as an example of the analysis we did on the grade 4 NAEP math tests (see Figure 4). The logic of this analysis rests on two assumptions. First, that high-stakes tests and other reforms were implemented in all grades at or around the same time, or soon after high-stakes high school graduation exams were implemented. Second, that such high-stakes test programs and the reform efforts that accompany them should affect learning in the different mathematics domains that make up the K–4 curriculum. NAEP is a test derived from the K–4 mathematics domains.

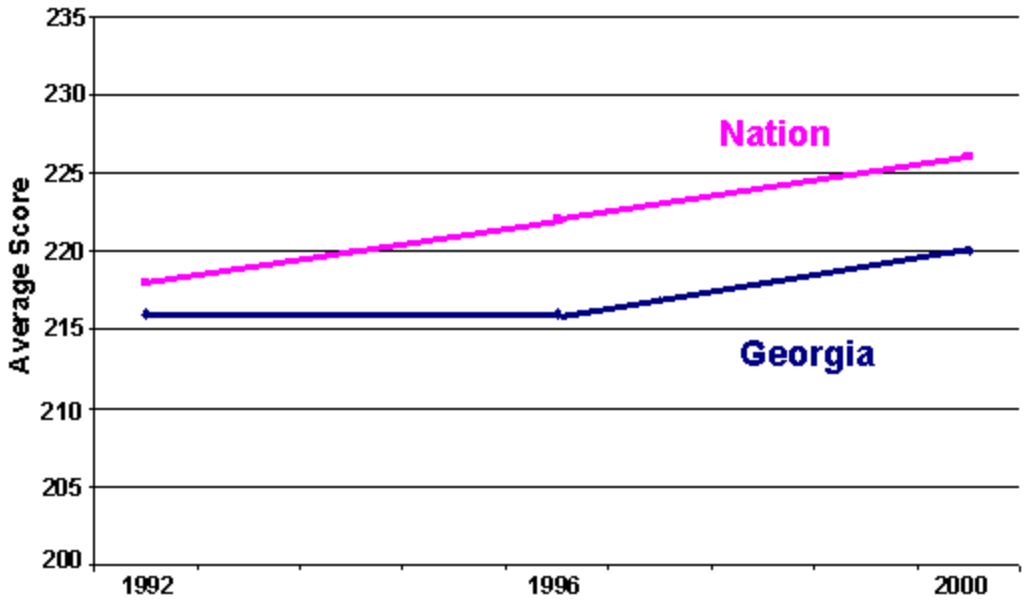

Figure 4. NAEP Math, Grade 4: Georgia

Trend lines and analytic comments for all the other states are included in Appendix D. A summary of these data across all 18 states is presented as Table 8.

Georgia implemented its 1st high school graduation exam in 1984. Assuming that other stakes attached to Georgia's K–8 tests (see Table 1) were attached at or around the same time or some time thereafter:

- From 1992–1996 Georgia lost 4 points to the nation.
- From 1996–2000 Georgia gained 4 points, as did the nation.
- From 1992–2000 Georgia lost 4 points to the nation.

Table 8
Results from the Analysis of NAEP Math Grade 4 Scores

State	Year in which students had to pass 1st HSGE to graduate	1992–1996 Change in score	1992–1996 Change in % excluded	1996–2000 Change in score	1996–2000 Change in % excluded	1992–2000 Change in score	1992–2000 Change in % excluded	Overall Effects
Alabama	1985	0	+3%	+2	–1%	+2	+2%	Positive
Florida	1979	–2	n/a	n/a	n/a	n/a	n/a	Negative
Georgia	1984	–4	+4%	0	–1%	–4	+3%	Negative
Indiana	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Louisiana	1991	+1	+6%	+5	–1%	+6	+5%	Positive
Maryland	1987	–1	+6%	–2	0%	–3	+6%	Negative
Minnesota	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Mississippi	1989	+2	+3%	–1	–3%	+1	0%	Positive
Nevada	1981	n/a	n/a	–1	0%	n/a	n/a	Negative
New Jersey	1984	–4	n/a	n/a	n/a	n/a	n/a	Negative
New Mexico	1990	–3	+7%	–4	–1%	–7	+6%	Negative
New York	1985	0	+5%	0	+3%	0	+8%	Neutral
North Carolina	1980	+8	+5%	+4	+5%	+12	+10%	Positive
Ohio	1994	+2	n/a	+2	n/a	+4	+5%	Positive
South Carolina	1990	–3	+3%	+3	0%	0	+3%	Neutral
Tennessee	1986	+4	+4%	–3	–3%	+1	+1%	Positive
Texas	1987	+7	+4%	0	+4%	+7	+8%	Positive

The time period 1992–1996. From Table 8, in comparison to the nation as a whole, we see that the states that implemented high-stakes tests 1 or more years before 1996 posted losses 1.2 times more often than gains on the 1992–1996 grade 4 NAEP math tests. Six states posted gains, seven states posted losses, and two states posted no changes, as compared to the nation. Thus, only 40% of the states with high-stakes tests posted gains from 1992–1996. These gains and losses may be considered "real" given that the states' 1992–1996 changes in score were unrelated (r = 0) to the states' 1992–1996 exclusion rates.

The time period 1996–2000. Table 8 also reveals that on the 1996–2000 grade 4 NAEP math tests the states that implemented high-stakes tests 1 or more years before 2000 posted gains 1.2 times more often than losses, as compared to the nation. Six states posted gains, five states posted losses, and three states posted no changes as compared to the nation. Thus, only 43% of the states with high-stakes tests posted gains from 1996–2000. These gains and losses, however, were partly artificial since the states' 1996–2000 changes in score were positively correlated (r = 0.45) with the states' 1996–2000 exclusion rates.

The time period 1992–2000. Table 8 also reveals that states that implemented high-stakes tests 1 or more years before 2000 posted gains 2.7 times more often than losses. Another way to look at these data is to note that these states were 1.6 times more likely to show gains rather than losses or no changes on the grade 4 NAEP math tests over the time period from 1992–2000. Eight states posted gains, three states posted losses, and two states posted no changes as compared to the nation. Thus, gains were posted by 62% of the states with high-stakes tests from 1992–2000. But these gains and losses were partly artificial given that the states' 1992–2000 changes in score were positively correlated (r = 0.39) to the states' 1992–2000 exclusion rates. The higher the percent of students excluded, the higher the NAEP scores obtained by a state. Because of the correlation we found between exclusion rates and scores on the NAEP, there is uncertainty about the meaning of those improved scores.

The overall data set. In the years for which data were available, across all time periods, the implementation of high-stakes tests resulted in positive effects 1.3 times more often than negative effects on the grade 4 NAEP tests in mathematics. Eight states displayed positive effects, six states displayed negative effects, and two states displayed neutral effects. Thus, in comparison to national trends, 50% of the states with high-stakes tests posted positive effects but these gains and losses were partly artificial, given that the overall positive or negative changes in score were related to changes in the overall state exclusion rates.

In short, when compared to the nation as a whole, high-stakes testing policies did not usually lead to improvement in the performance of students on the grade 4 NAEP math tests between 1992 and 2000. Gains and losses were more likely to be related to who was excluded from the NAEP than to the effects of high-stakes testing programs in a state. In the 1992–1996 time period, when participation rates were unrelated to gains and losses, the academic achievement of students may have even been thwarted in those states where high-stakes testing was implemented. High-stakes tests within states probably had a differential impact on students from racial minority and economically disadvantaged backgrounds.

Analysis of NAEP Grade 8 Math Scores

For each state, after high-stakes tests had been implemented, an analysis of NAEP mathematics achievement scores was conducted. The state of Mississippi was randomly chosen to serve as an example of the analysis we did on the grade 8 NAEP math tests (see Figure 5). The logic of this analysis rests on two assumptions. First, that high-stakes tests and other reforms were implemented in all grades at or around the same time, or soon after high-stakes high school graduation exams were implemented. Second, that such high-stakes test programs should affect learning in the different mathematics domains that make up the K–8 curriculum. NAEP is a test derived from the domains that make up the K–8 curriculum.

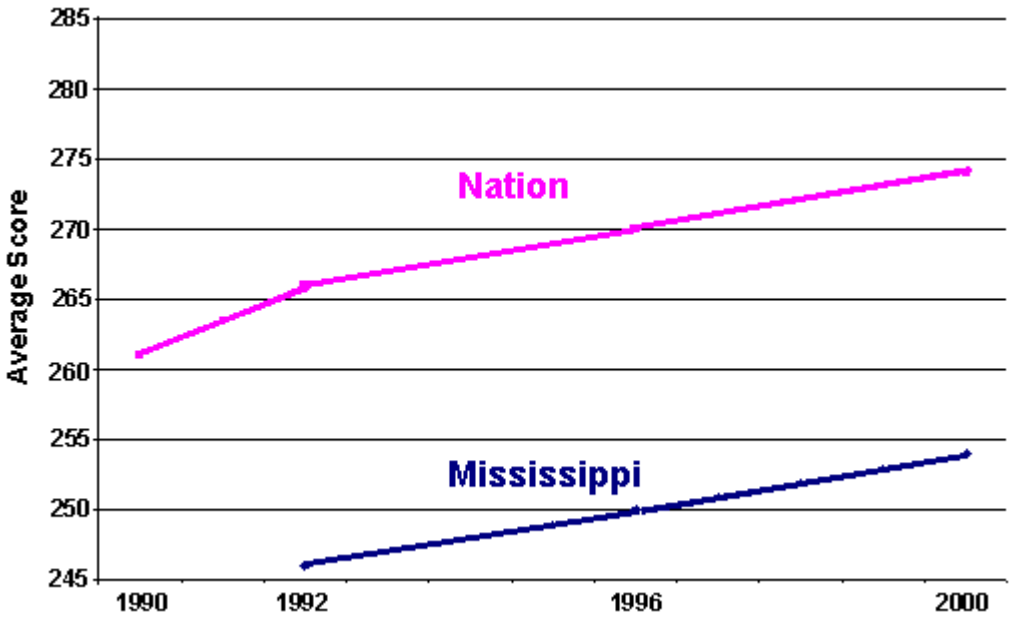

Figure 5. Mississippi – NAEP math grade 8

Mississippi implemented its 1st high school graduation exam in 1988. Assuming that the stakes attached to Mississippi's K–8 tests (see Table 1) were attached at or around the same time or some time thereafter. From 1990–1992 Mississippi data were not available. From 1992–1996 Mississippi gained 4 points, as did the nation. From 1996–2000 Mississippi gained 4 points, as did the nation. From 1990–2000 Mississippi NAEP data were not available.

All other states' trend lines and analytic comments are included in Appendix E. A summary of these data across all 18 states is presented as Table 9.

Table 9
Results from the Analysis of NAEP Math Grade 8 Scores

State	Year in which students	1990–92	1990–1992	1992–96	1992–96 Change	1996–00	1996–00 Change	1990–00	1990–00 Change	Overall Effects

	had to pass 1st HSGE to graduate	Change in score	Change in % excluded	Change in score	in % excluded	Change in score	in % excluded	Change in score	in % excluded	
Alabama	1985	–6	–2%	0	+4%	+2	–4%	–4	–2%	Negative
Florida	1979	–1	n/a	0	n/a	n/a	n/a	n/a	n/a	Negative
Georgia	1984	–5	0%	–1	+4%	0	–2%	–6	+2%	Negative
Indiana	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Louisiana	1991	–1	–2%	–2	+4%	+3	–2%	0	0	Neutral
Maryland	1987	–1	–2%	+1	+4%	+2	+2%	+2	+4%	Positive
Minnesota	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Mississippi	1989	0	n/a	0	+2%	n/a	n/a	n/a	n/a	Neutral
Nevada	1981	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
New Jersey	1984	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
New Mexico	1990	–2	–3%	–2	+5%	–6	+2%	–10	+4%	Negative
New York	1985	0	0%	0	+2%	+2	+3%	+2	+5%	Positive
North Carolina	1980	+3	–2%	+6	+3%	+8	+8%	+17	+9%	Positive
Ohio	1994	–1	–1%	+3.5	n/a	+3.5	n/a	+6	+2%	Positive
South Carolina	1990	n/a	n/a	–4	+2%	+2	–1%	n/a	n/a	Negative
Tennessee	1986	n/a	n/a	+1	+1%	–4	–1%	n/a	n/a	Negative
Texas	1987	+2	–1%	+1	+4%	+1	–1%	+4	+2%	Positive
Virginia	1986	–2	–2%	–2	+4%	+3	+1%	–1	+3%	Negative

The time period 1990–1992. Table 9 reveals that, in comparison to the nation as a whole, states that implemented high-stakes tests one or more years before 1992 posted losses 4 times more often than gains on the 1990–1992 grade 8 NAEP math tests. Compared to the nation, two states posted gains, eight states posted losses, and two

states posted no change. Over this time period gains on the NAEP tests were posted by 17% of the states with high-stakes tests. These gains and losses were "real" given that the states' 1990–1992 changes in score were unrelated (r = 0) to the states' 1990–1992 exclusion rates.

The time period 1992–1996. Table 9 also reveals that states that implemented high-stakes tests 1 or more years before 1996 were as likely to post gains as losses on the 1992–1996 grade 8 NAEP math tests. Five states posted gains, five states posted losses, and four states posted no changes as compared to the nation. Thus, from 1992–1996 only 36% of the states with high-stakes tests posted gains. These gains and losses were "real" given that the states' 1992–1996 changes in score were unrelated (r = 0) to the states' 1992–1996 exclusion rates.

The time period 1996–2000. Looking at the grade 8 NAEP math tests over the 1996–2000 time period we see that states that implemented high-stakes tests 1 or more years before 2000 posted gains 4.5 times more often than losses. Nine states posted gains, two states posted losses, and one state posted no changes, as compared to the nation. Thus, in the time period from 1996–2000 gains were posted by 75% of the states with high-stakes tests, but those NAEP scores were related to whether exclusion rates increased or decreased over the same time period, raising some uncertainty about the authenticity of these gains. Gains and losses during this time period must be considered partly artificial given that the states' 1996–2000 changes in score were positively related (r = 0.35) to the states' 1996–2000 exclusion rates.

The time period 1990–2000. Looking over the long term, states that implemented high-stakes tests one or more years before 2000 posted gains 1.3 times more often than losses on the 1990–2000 grade 8 NAEP math tests. Five states posted gains, four states posted losses, and one state posted no changes as compared to the nation. These gains and losses were partly artificial, however, given that the states' 1996–2000 changes in score were substantially related (r = 0.53) to the states' 1990–2000 exclusion rates.

Overall, across the years for which data were available, the states that had implemented high-stakes tests displayed negative effects 1.4 times more often than positive effects. Five states displayed positive effects, seven states displayed negative effects, and two states displayed neutral effects. Another way of interpreting these data is that 36% of the states with high-stakes tests posted positive effects from 1990–2000 on the grade 8 NAEP math examinations, while losses were posted by 50% of the states with high-stakes tests over this same time period. These gains and losses were partly artificial, however, given that the overall positive or negative changes in score were related to overall exclusion rates.

In short, there is no compelling evidence that high-stakes testing policies have improved the performance of students on the grade 8 NAEP math tests. Gains were more related to who was excluded from the NAEP than to whether there were high-stakes tests being used or not. If anything, the weight of the evidence suggests that high-stakes tests thwarted the academic achievement of students in these states.

Analysis of the Grade 4 NAEP Reading Scores

For each state, after high-stakes tests had been implemented, an analysis of NAEP reading achievement scores was conducted. The state of Virginia was randomly chosen

to serve as an example of the analysis we did on the grade 4 NAEP reading tests (see Figure 6). The logic of this analysis rests on two assumptions. First, that high-stakes tests and other reforms were implemented in all grades at or around the same time, or soon after high-stakes high school graduation exams were implemented. Second, that such high-stakes test programs should affect learning in the different domains of reading that make up the K–4 curricula. NAEP is a test derived from the various domains that constitute the K–4 reading curriculum.

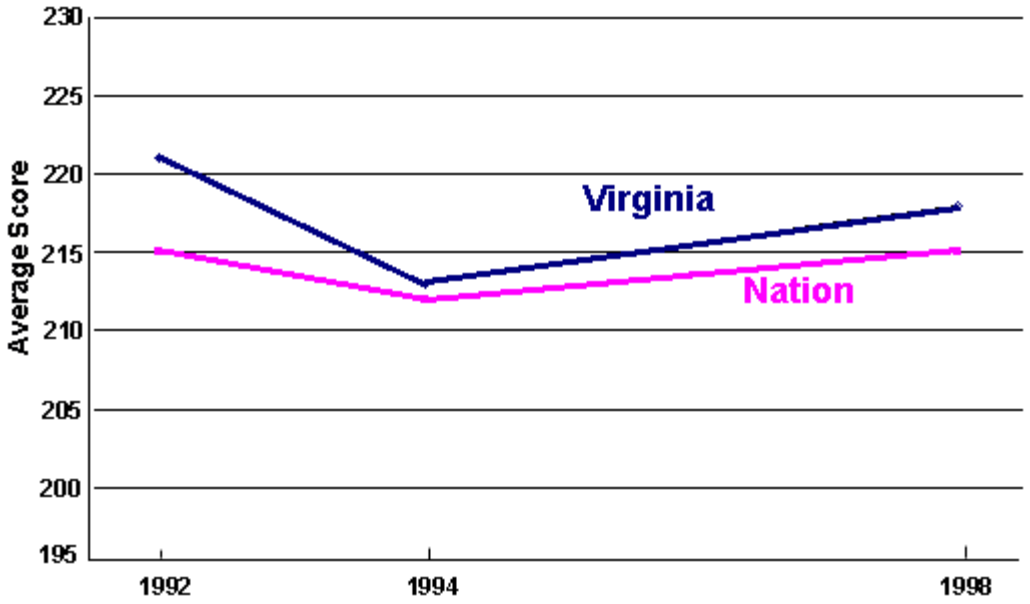

Figure 6. Virginia – NAEP reading grade 4

Virginia implemented its 1st high school graduation exam around 1981. Assuming that the stakes attached to Virginia's K–8 tests (see Table 1) were attached at or around the same time or some time thereafter 1) From 1992–1994 Virginia lost 5 points to the nation; 2) From 1994–1998 Virginia gained 2 points on the nation; 3) From 1992–1998 Virginia lost 3 points to the nation. Trend lines and analytic comments for all other states are included in Appendix F. A summary of these data across all 18 states is presented as Table 10.

Table 10
Results from the analysis of NAEP reading grade 4 scores

State	Year in which students had to pass 1st HSGE to graduate	1992–94 Change in score	1992–94 Change in % excluded	1994–98 Change in score	1994–98 Change in % excluded	1992–98 Change in score	1992–98 Change in % excluded	Overall Effects
Alabama	1985	+4	–1%	0	+4%	+4	+3%	Positive

Florida	1979	0	+1%	–1	–1%	–1	0%	Negative
Georgia	1984	–2	0%	0	+2%	–2	+2%	Negative
Indiana	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Louisiana	1991	–4	+2%	+4	+7%	0	+9%	Neutral
Maryland	1987	+2	0%	+2	+3%	+4	+3%	Positive
Minnesota	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Mississippi	1989	+6	+1%	–1	–2%	+5	–1%	Positive
Nevada	1981	n/a	n/a	n/a	n/a	n/a	n/a	n/a
New Jersey	1984	n/a	n/a	n/a	n/a	n/a	n/a	n/a
New Mexico	1990	–3	0%	–2	+3%	–5	+3%	Negative
New York	1985	0	+2%	+1	0%	+1	+2%	Positive
North Carolina	1980	+5	+1%	0	+6%	+5	+7%	Positive
Ohio	1994	n/a	n/a	n/a	n/a	n/a	n/a	n/a
South Carolina	1990	–4	+1%	+4	+5%	0	+6%	Neutral
Tennessee	1986	+4	+1%	–4	–1%	0	0%	Neutral
Texas	1987	+2	+3%	+2	+3%	+4	+6%	Positive
Virginia	1986	–5	+1%	+2	+2%	–3	+3%	Negative

The time period 1992–1994. We note in Table 10 that on the grade 4 reading test, during the time period 1992–1994, states that implemented high-stakes tests 1 or more years before 1994 posted gains 1.2 times more often than losses, in comparison to the nation. Compared to national trends six states posted gains, five states posted losses, and two states posted no changes at all. Thus, only 46% of the states with high-stakes tests posted gains from 1992–1994. These gains and losses were "real" given that the states' changes in score for the time period 1992–1994 were virtually unrelated ($r = –0.10$) to the states' exclusion rates.

The time period 1994–1998. Table 10 also reveals that those states implementing high-stakes tests 1 or more years before 1998 posted gains 1.5 times more often than losses when compared to national trends. Six states posted gains, four states posted

losses, and three states posted no changes when compared to the national trends. Thus, only 46% of the states with high-stakes tests from 1994–1998 posted gains. These gains and losses were partly artificial, however, given that the states' 1994–1998 changes in score were strongly correlated (r = 0.63) to the states' 1994–1998 exclusion rates.

The time period 1992–1998. Table 10 also informs us that states implementing high-stakes tests 1 or more years before 1998 posted gains 1.5 times more often than losses in comparison to the national trends during the time period 1992–1998. Six states posted gains, four states posted losses, and three states posted no changes in comparison to national trends. Thus, only 46% of the states with high-stakes tests posted positive effects from 1992–1998 on the NAEP grade 4 reading test. The gains and losses may be considered "real" given that the states' 1992–1998 changes in score were virtually unrelated (r = 0.11) to the states' changes in 1992–1998 exclusion rates.

In short, in comparison to the national trends, high-stakes tests did not improve the learning of students as judged by their performance on the NAEP grade 4 reading test. This was clearest in the time periods from 1992–1994 and from 1992–1998. The learning effects over these years were unrelated to the rates by which students were excluded from the NAEP. We note, however, that in 1998 75% of the states with high-stakes tests had 1998 exclusion rates that were higher than the nation. Given the typical positive (and substantial) correlation between increased exclusion rates and increased NAEP scores, states' gains and losses in score need to be carefully evaluated. If anything, in comparison to national trends, the academic achievement of students in states with high-stakes testing policies seemed to be lower, particularly for students from minority backgrounds.

NAEP Cohort Analyses

Another way of investigating growth in achievement on measures other than states' high-stakes tests is to look at each state's cohort trends on the NAEP.(Note 122) The NAEP analyses preceding this section gauged the achievement trends of different samples of students over time, for example, 4th graders in one year compared to a different group of 4th graders a few years later. There is a slight weakness with this approach because we must compare students in one year with a different set of students a few years later. We are unable to control for differences between the different groups or cohorts of students.(Note 123) To compensate for this we did a cohort analysis, an analysis of the growth in achievement made by "similar" groups of students over time.

This is possible because NAEP uses random samples of students. Thus the 4th graders in 1996 should be representative of the same population of 8th graders tested four years later. Random sampling techniques made the groups of students similar enough so that the achievement effects made by the "same" (statistically the same) students can be tracked over time.(Note 124) Analyzing cohort trends in the 18 states with high-stakes tests helped assess the degree to which students increased in achievement as they progressed through school systems that were exerting more pressures for school improvement, including the use of high-stakes tests. We examined the growth of these students by tracking the relative changes in math achievement of 4th graders in 1996 to 8th graders in 2000, and by looking at the reading achievement of 4th graders in 1994 to that of 8th graders in 1998. The changes we record for each state are all relative to the national trends on the respective NAEP tests.

Cohort Analysis of NAEP Mathematics Scores: Grade 4 (1996) to Grade 8 (2000)

The state of New York was randomly chosen to serve as an example of the analysis we did for the NAEP mathematics cohort over the years 1996 to 2000 (see Figure 7). The logic of this analysis rests on the same two assumptions as previous NAEP analyses. First, that high-stakes tests and other reforms were implemented in all grades at or around the same time, or soon after high-stakes high school graduation exams were implemented. Second, that such high-stakes test programs should affect learning in the different domains of mathematics from which NAEP is derived.

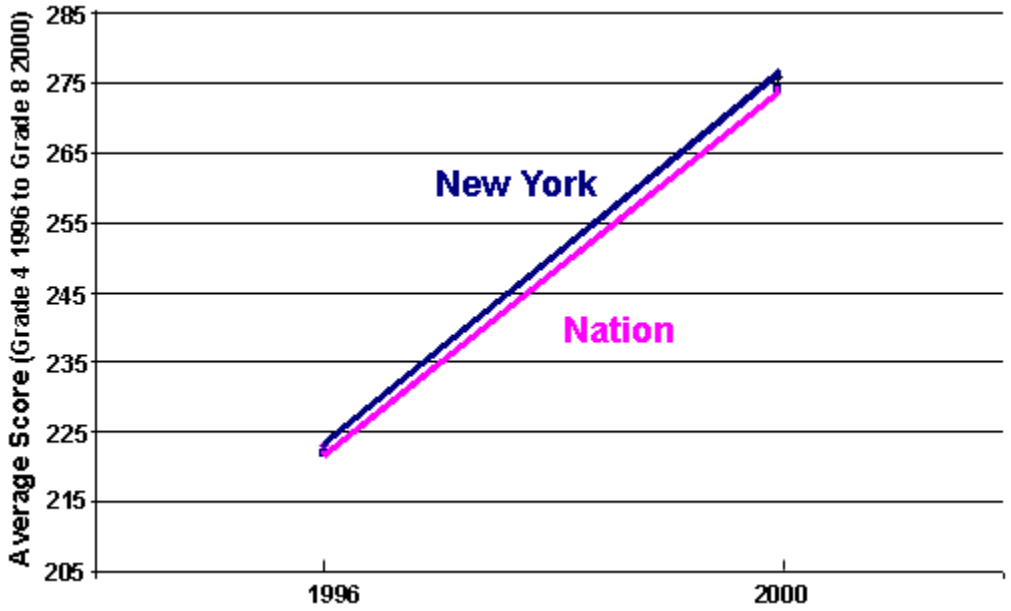

Figure 7. New York by cohort: NAEP math grade 4 1996 to grade 8 2000

New York implemented its 1st high school graduation exam around 1981. Assuming that the stakes attached to New York's K–8 tests (see Table 1) were attached at or around the same time or some time thereafter, from 4th grade in 1996 to 8th grade in 2000 New York gained 1 point on the nation. Trend lines and analytic comments for all other states are included in Appendix G. A summary of these data across all 18 states is presented as Table 11.

Table 11
Results from the Analysis of NAEP Math Cohort Trends

State	Year in which students had to pass 1st HSGE to graduate	Change in score from grade 4 1996 to grade 8 2000	Change in % excluded from grade 4 1996 to grade 8 2000	Overall Effects 1996–00
Alabama	1985	–2	–2%	Negative
Florida	1979	n/a	n/a	n/a

Georgia	1984	–2	–1%	Negative
Indiana	2000	n/a	n/a	n/a
Louisiana	1991	–2	–3%	Negative
Maryland	1987	+4	+2%	Positive
Minnesota	2000	n/a	n/a	n/a
Mississippi	1989	–6	0%	Negative
Nevada	1981	–1	0%	Negative
New Jersey	1984	n/a	n/a	n/a
New Mexico	1990	–6	–1%	Negative
New York	1985	+1	+4%	Positive
North Carolina	1980	+4	+6%	Positive
Ohio	1994	n/a	n/a	n/a
South Carolina	1990	+1	0%	Positive
Tennessee	1986	–8	–2%	Negative
Texas	1987	–6	–1%	Negative
Virginia	1986	+3	+2%	Positive

The 1996–2000 cohort. From 1996 to 2000 cohorts of students moving from 4th to 8th grade in states that had implemented high-stakes tests in the years before 2000 posted losses 1.6 times more often than gains. In comparison to the national trends five states posted gains, and eight states posted losses. Said differently, in comparison to the nation, 62% of the states with high-stakes tests posted losses as their students moved from the 4th grade 1996 NAEP to the 8th grade 2000 NAEP. These gains and losses, however, were partly artificial because gains and losses in score for the cohorts in the various states were strongly correlated ($r = 0.70$) with overall exclusion rates. This cohort analysis finds no evidence of gains in general learning as a result of high-stakes testing policies.

Cohort Analysis of NAEP Reading Scores: Grade 4 (1994) to Grade 8 (1998)

The state of Tennessee was randomly chosen to serve as an example of the analysis we did for NAEP reading cohort over the years 1994 to 1998 (see Figure 8). The logic of this analysis rests on the same two assumptions made in the previous NAEP analyses. First, that high-stakes tests and other reforms were implemented in all grades at or around the same time, or soon after high-stakes high school graduation exams were implemented. Second, that such high-stakes test programs should affect learning in the different domains of reading from which NAEP is derived.

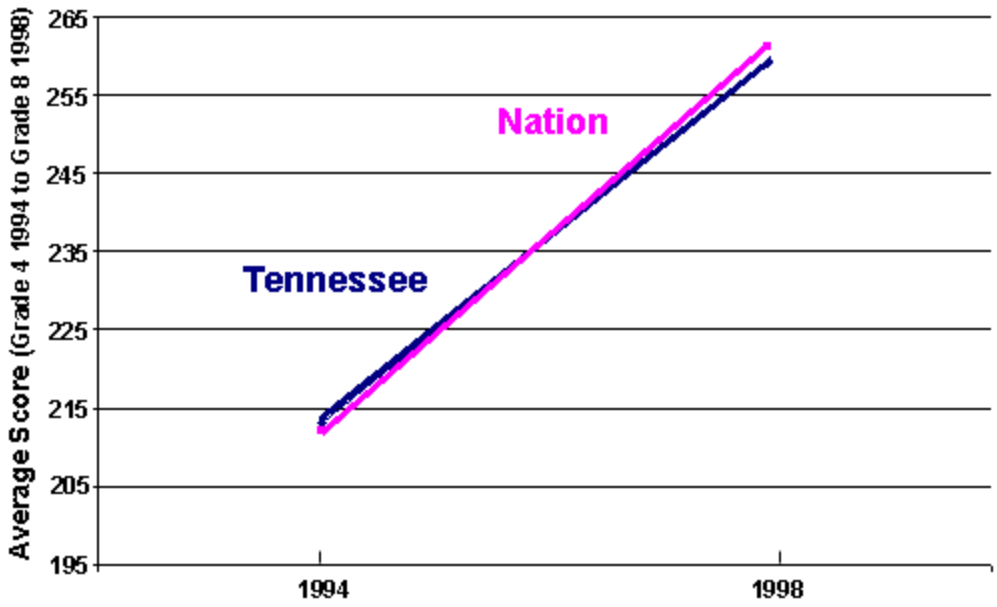

Figure 8. Tennessee by cohort: NAEP reading grade 4 1994 to grade 8 1998

Tennessee implemented its 1st high school graduation exam in 1982. Assuming that the stakes attached to Tennessee's K–8 tests (see Table 1) were attached at or around the same time or some time thereafter, from 4th grade in 1994 to 8th grade in 1998 Tennessee lost 3 points to the nation. Trend lines and analytic comments for all other states are included in Appendix H. A summary of these data across all 18 states is presented as Table 12.

Table 12
Results from the Analysis of NAEP Reading Cohort Trends

State	Year in which students had to pass 1st HSGE to graduate	Change in score from grade 4 1994 to grade 8 1998	Change in % excluded from grade 4 1994 to grade 8 1998	Overall Effects 1994–98
Alabama	1985	–2	+5%	Negative
Florida	1979	–1	–2%	Negative
Georgia	1984	+1	+4%	Positive
Indiana	2000	n/a	n/a	n/a

Louisiana	1991	+6	+6%	Positive
Maryland	1987	+3	+3%	Positive
Minnesota	2000	n/a	n/a	n/a
Mississippi	1989	–20	+4%	Negative
Nevada	1981	n/a	n/a	n/a
New Jersey	1984	n/a	n/a	n/a
New Mexico	1990	+4	+2%	Positive
New York	1985	+5	+4%	Positive
North Carolina	1980	+1	+7%	Positive
Ohio	1994	n/a	n/a	n/a
South Carolina	1990	+3	+2%	Positive
Tennessee	1986	–3	+1%	Negative
Texas	1987	+1	–1%	Positive
Virginia	1986	+4	+3%	Positive

The 1994–1998 cohort. In comparison to national trends, cohorts of students in states that implemented high-stakes tests in the years before 1998 posted gains 2.3 times more often than losses from the 4th to the 8th grade on the 1994 and the 1998 NAEP reading exams. Nine states posted gains, and four states posted losses. These gains and losses were "real" given that gains and losses in score were unrelated (r = 0) to overall exclusion rates.

Thus far in these analyses this is the only example we found of gains in achievement on a transfer measure that meet criteria of acceptability. As their students moved from the 4th grade in 1994 to the 8th grade in 1998, 69% of the states with high-stakes tests posted gains on the NAEP reading tests. Since these gains and losses were unrelated to increases and decreases in exclusion rates they appear to be "real" effects. To put these gains in context we note that in the states that showed increases in scores from 1994 to 1998, the average gain was 52 points. By any metric a 52-point gain is sizeable. But when these gains are compared to the national trends over the same time period, as shown in table12, we see that the gains in the states with high-stakes testing policies

was, on average, only 3 points over the national trend. On the other hand, although fewer in number, the states that posted losses in comparison to the nation fell an average of 6.5 points. This figure is skewed, however, by the fact that Mississippi lost 20 points more than the nation did on the 4th to 8th grade reading NAEP from 1994-1998. In sum, these gains in the reading scores in states with high-stakes testing policies seem real but modest given the losses shown by other states with high-stakes testing policies.

Advanced Placement (AP) Data Analysis

The Advanced Placement (AP) program offers high school students opportunities to take college courses in a variety of subjects and receive credits before actually entering college. We used the AP data(Note 125) as another indicator of the effects of high-stakes high school graduation exams on the general learning and motivation of high school students. Using the AP exams as transfer measures and the AP participation rates as indicators of increased student preparation and motivation for college, we could inquire whether, in fact, high school graduation exams increased learning in the knowledge domains that are the intended targets of high-stakes testing programs.

The participation rates and rates by which students passed AP exams that are used in the following analyses were calculated by the College Board,(Note 126) administrators of the AP program. Gains or losses were assessed after the most recent year in which a new high school graduation exam was implemented or after 1995 – the first year for which these AP data were available.

Table 13 presents for each state the percentages of students who passed AP examinations with a grade of 3 or better after high school graduation exams were implemented. As we worked, however, it became apparent that fluctuations in participation rates were related ($r = –0.30$) to the percent of students passing AP exams with a grade 3 or better. If participation rates in a state decreased, the percent of students who passed AP exams usually increased and vice versa. To judge the effect of this interaction, and in comparison to the nation, the percent change in students who passed the AP examination is presented along with the percent change in students who participated in AP exams during the time period 1995–2000. If an increase in one corresponded to a decrease in the other, caution in making judgments about the effects is required.

North Carolina was randomly chosen from the states we examine to be the example for the AP analysis. That data is presented in Figure 9. Trend lines and analytic comments for all other states are included in Appendix I and summarized in Table 13.

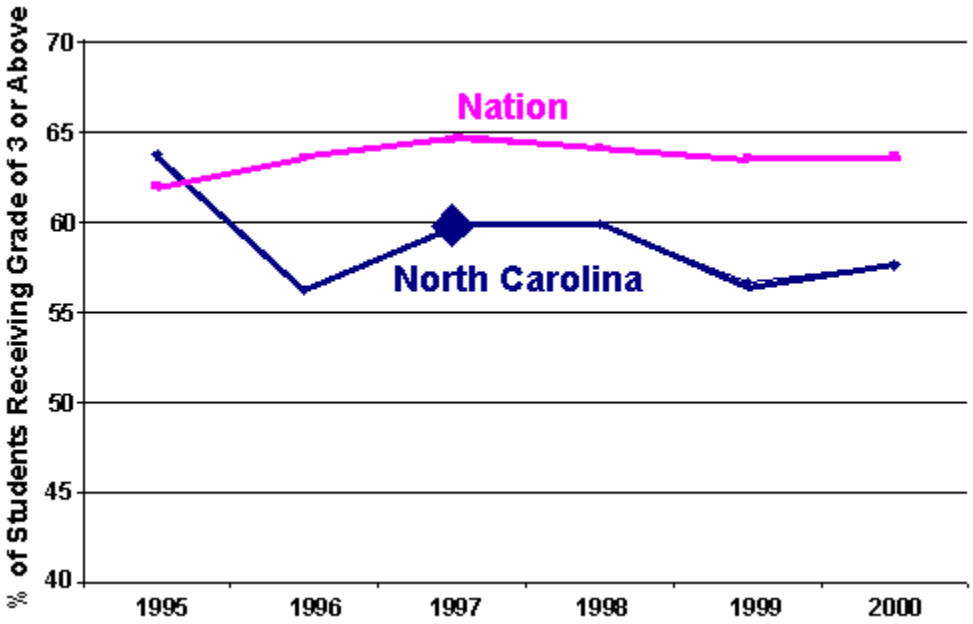

Figure 9. North Carolina: Percent passing AP examinations

North Carolina's 1st high school graduation exam first affected the class of 1980. North Carolina's second exam first affected the class of 1998. From 1995–2000 North Carolina lost 7.7 percentage points to the nation. From 1997–1998 North Carolina gained .5 percentage points on the nation. From 1997–2000 North Carolina lost 1.3 percentage points to the nation.

Table 13
Results from the Analysis of AP Scores and Participation Rates

State	Year in which students had to pass 1st HSGE to graduate	Change in % of students passing AP exams 1995–2000 as compared to the nation*	Change in % of students taking AP exams 1995–2000 as compared to the nation*	Overall Effects
Alabama	1985	+9.6%	–6.5%	Positive
Florida	1979	+3.9%	–0.5%	Positive
Georgia	1984	+6.8%	–1.4%	Positive
Indiana	2000	+1.9%	–0.4%	Positive
Louisiana	1991	+2.6%	–4.4%	Positive
Maryland	1987	+0.5%	+2.3%	Positive
Minnesota	2000	+0.6%	–1.6%	Positive

Mississippi	1989	–2.4%	–4.6%	Negative
Nevada	1981	+3.2%	–2.7%	Positive
New Jersey	1985	+1.7%	+2.0%	Positive
New Mexico	1990	–4.1%	–1.6%	Negative
New York	1985	+7.7%	+3.9%	Positive
North Carolina	1980	–7.7%	+0.9%	Negative
Ohio	1994	–3.3%	–2.6%	Negative
South Carolina	1990	–9.8%	–3.7%	Negative
Tennessee	1986	+5.8%	–1.8%	Positive
Texas	1987	–10.5%	+5.1%	Negative
Virginia	1986	–1.6%	+3.9%	Negative

*(Indiana and Minnesota, 1999–2000)

The time period 1995–2000. In comparison to national trends from 1995–2000, students in states with high school graduation exams posted gains 1.6 times more than losses in the percentage of students passing AP exams with a score of 3 or better. Eleven states posted gains, and seven states posted losses. These gains and losses were partly artificial, however, given that gains and losses in the percentage of students passing AP exams were negatively correlated (r = –0.30) with the rate in which students participated in the AP program. The greater the percentage of students who participated in the AP program, the lower the percentage of students passing AP exams, and vice versa.

Compared to the national average participation rates fell in 67% of the states with high school graduation exams since 1995 (and since 1999 for Indiana and Minnesota). In comparison to the nation participation rates increased in six states and decreased in twelve states in the time period from 1995–2000.

Overall, 61% of the states with high-stakes tests posted gains in the rate by which students passed AP exams with a grade of 3 or better from 1995–2000 (1999–2000 in Indiana and Minnesota). But those increases and decreases in the percent passing AP exams were negatively correlated (r = –0.30) to whether participation rates increased or decreased at the same time. If we look at only those states where the participation rates did not seem to influence the percent passing AP exams(Note 127) as the overall correlation suggests it typically does, only Maryland (+), Mississippi (–), New Jersey

(+), New Mexico (–), New York (+), Ohio (–), and South Carolina (–) posted "true" effects, 57% of which were negative.

The special case of Texas. Texas, as has been mentioned, received attention as one of two states in which high-stakes tests purportedly improve achievement. Dramatic gains in the rates of students enrolled in AP courses were among several state indicators of achievement provided by the state in support of their academic gains. But another educational policy was put into effect around the same time as the high-stakes testing program was implemented in that state.(Note 128) The Texas state legislature substantially reduced the cost of taking AP courses and the accompanying exams.(Note 129) This highly targeted policy may have helped increase enrollments in AP courses in Texas much more than their high-stakes testing program. So the substantial drop in the percent passing the test is difficult to assess since many more students took the AP tests. As we have seen, as a greater percentage of students in a state take the test the scores go down, and as a smaller percentage of students take the test scores go up. Inferences about the meaning of test scores become more uncertain when participation rates are not steady from one testing year to another.

In conclusion, when we use the national data on AP exams as a comparison for state AP data, and we use the percent of students passing the various AP exams as an indicator of learning in the domains of interest, we find no evidence of improvement associated with high-stakes high school graduation exams. When controlling for participation rates there even appeared to be a slight decrease in the percent of students who passed AP examinations. Further, in the states under study, high-stakes high school graduation exams did not result in an increase in the numbers of students preparing to go to college, as indicated by the percent of students who participated in AP programs from 1995–2000.

Conclusion

If we assume that the ACT, SAT, NAEP and AP tests are reasonable measures of the domains that a state's high-stakes testing program is intended to affect, then we have little evidence at the present time that such programs work. Although states may demonstrate increases in scores on their own high-stakes tests, transfer of learning is not a typical outcome of their high-stakes testing policy.

The ACT data. Sixty-seven percent of the states that use high school graduation exams posted *decreases* in ACT performance after high school graduation exams were implemented. These decreases were unrelated to whether participation rates increased or decreased at the same time. On average, as measured by the ACT, college-bound students in states with high school graduation exams decreased in levels of academic achievement. Moreover, participation rates in ACT testing, as compared to the nation, increased in nine states, decreased in six states, and stayed the same in three states. If participation rates in the ACT program serve as an indicator of motivation to attend college, then there is scant support for the belief that high-stakes testing policies within a state have such an impact.

The SAT data. Fifty six percent of the states that use high-stakes high school graduation exams posted decreases in SAT performance after those exams were implemented. However, these decreases were slightly related to whether SAT participation rates

increased or decreased over the same time period. Thus, there is no reliable evidence of high-stakes high school graduation exams improving the performance of students who take the SAT. Gains and losses in SAT scores are more strongly correlated to who participates in the SAT than to the implementation of high school graduation exams. Moreover, SAT participation rates, as compared to the nation, fell in 61% of the states with high school graduation exams. If these participation rates serve as an indicator for testing the belief that high-stakes testing policies will prepare more students or motivate more students to attend college, then there is scant support for such beliefs. Students did not participate in the SAT testing program at greater rates after high-stakes high school graduation exams were implemented.

The NAEP mathematics data. High-stakes testing policies did not usually improve the performance of students on the grade 4 NAEP math tests. Gains and losses were more related to who was excluded from the NAEP than the effects of high-stakes testing programs in a state. However, during the 1992–1996 time period, when exclusion rates were unrelated to gains and losses in scores, mathematics achievement decreased for students in states where high-stakes testing had been implemented. High-stakes testing policies did not consistently improve the performance of students on the grade 8 NAEP math tests. Gains were more strongly correlated to who was excluded from the NAEP than to whether or not high-stakes tests were used. If anything, the weight of the evidence suggests that students from states with high-stakes tests did not achieve as well on the grade 8 NAEP mathematics tests as students in other states.

The NAEP reading data. High-stakes testing policies did not consistently improve the general learning and competencies of students in reading as judged by their performance on the NAEP grade 4 reading test. This was clearest in the time periods from 1992–1994 and over the time span of from 1992–1998. The learning effects over these years were unrelated to the rates by which students were excluded from the NAEP. By 1998, however, 75% of the states with high-stakes tests had exclusion rates higher than the national average. These exclusionary policies were probably the reason for the apparent increases in achievement in several states. As the NAEP tests become more important in our national debates about school success and failure the effects of the Heisenberg Uncertainty Principle, as applied to the social sciences, seems to be evident. When these exclusion rates are taken into account, in comparison to national trends, the reading achievement of students in states with high-stakes testing policies appeared lower, particularly for students from minority backgrounds.

The NAEP cohort data. Sixty-two percent of the states with high-stakes tests posted losses on the NAEP mathematics exams as a cohort of their students moved from the 4th grade in 1996 to the 8th grade in the year 2000. These gains and losses, however, must be considered artificial to some extent because of the very strong relationship of overall exclusion rates to the gains and losses that were recorded. This cohort analysis finds no evidence of gains in general mathematics knowledge and skills as a result of high-stakes testing policies.

For the cohort of students moving from the 4th to the 8th grade and taking the 1994 and the 1998 NAEP reading exams, gains in scores were posted 2.3 times more often than losses in the states with high-stakes testing policies. Nine states (69%) posted gains, and four states (31%) posted losses. These gains and losses were "real" given that gains and losses in score were unrelated to overall NAEP exclusion rates. While not reflecting unequivocal support for high-stakes testing policies, this is the one case of gains in

achievement on a transfer measure among the many analyses we did for this report. It is also true that over this time period many reading curriculum initiatives were being implemented throughout the country, as reading debates became heated and sparked controversy. Because of that it is not easy to attribute the gains made for the NAEP reading cohort to high-stakes testing policies. Our guess is that the reading initiatives and the high-stakes testing polices are entangled in ways that make it impossible to learn about their independent effects.

The AP data. High-stakes high school graduation exams do not improve achievement as indicated by the percent of students passing the various AP exams. When participation rates were controlled there was a decrease in the percent of students who passed AP examinations. Further, in the states with high-stakes high school graduation exams there was no increase in the numbers of students preparing to go to college, as indicated by the percent of students who chose to participate in AP programs from 1995–2000.

Final thoughts

. What shall we make of all this? At the present time, there is no compelling evidence from a set of states with high-stakes testing policies that those policies result in transfer to the broader domains of knowledge and skill for which high-stakes test scores *must* be indicators. Because of this, the high-stakes tests being used today do not, as a general rule, appear valid as indicators of genuine learning, of the types of learning that approach the American ideal of what an educated person knows and can do. Moreover, as predicted by the Heisenberg Uncertainty Principle, data from high-stakes testing programs too often appear distorted and corrupted.

Both the uncertainty associated with high-stakes testing data, and the questionable validity of high-stakes tests as indicators of the domains they are intended to reflect, suggest that this is a failed policy initiative. High-stakes testing policies are not now and may never be policies that will accomplish what they intend. Could the hundreds of millions of dollars and the billions of person hours spent in these programs be used more wisely? Furthermore, if failure in attaining the goals for which the policy was created results in disproportionate negative affects on the life chances of America's poor and minority students, as it appears to do, then a high-stakes testing policy is more than a benign error in political judgment. It is an error in policy that results in structural and institutional mechanisms that discriminate against all of America's poor and many of America's minority students. It is now time to debate high-stakes testing policies more thoroughly and seek to change them if they do not do what was intended and have some unintended negative consequences, as well.

Notes

1. Haladyna, Nolen, & Haas, 1991.

2. Figlio & Lucas, 2000.

3. Kreitzer, Madaus, & Haney, 1989.

4. Bracey, 1995; Heubert & Hauser, 1999; and Kreitzer, Madaus, & Haney, 1989.

5. Linn, 2000 and Serow, 1984.

6. U.S. Department of Education, 1983 and Bracey, 1995.

7. U.S. Department of Education, 1983.

8. Berliner & Biddle, 1995.

9. Quality Counts, 2001.

10. McNeil, 2000; Orfield & Kornhaber, 2001; Paris, 2000; Sacks, 1999; and Sheldon & Biddle, 1998.

11. Madaus & Clarke, 2001 and Campbell, 1975.

12. All of the following statistics come from extensive interviews conducted with knowledgeable testing personnel throughout the United States, and Quality Counts, 2001.

13. Administrative bonuses, 2001.

14. Neufeld, 2000.

15. Folmar, 2001.

16. Commission on Instructionally Supportive Testing, 2001.

17. California, Delaware, Michigan, Missouri, Nevada, and Ohio give scholarships to students for high performance on state mandated exams. See Quality Counts, 2001.

18. Durbin, 2001 and Ross, 2001.

19. Thanks to Professor J. Ryan, Arizona State University, for suggesting we investigate this story.

20. National Governor's Association, 2000; "Civil rights coalition," 2000; and "Using tobacco settlement revenues," 1999.

21. Heller, 1999. See also Durbin, 2001; Ross, 2001; and Swope & Miner, 2000.

22. "Civil rights coalition," 2000.

23. Heller, 1999.

24. Delaware, Ohio, South Carolina, and Texas have plans to promote students using test scores by the year 2003. Interview data and Quality Counts, 2001.

25. Florida implemented its first minimum competency test for the class of 1979, North Carolina implemented its first minimum competency test for the class of 1980, and Nevada implemented its first minimum competency test for the class of 1981.

26. U.S. Department of Education, 1983.

27. States that currently use high school graduation exams to grant or withhold diplomas are Alabama, Florida, Georgia, Indiana, Louisiana, Maryland, Minnesota, Mississippi, Nevada, New Jersey, New Mexico, New York, North Carolina, Ohio, South Carolina, Tennessee, Texas, and Virginia. Hawaii used a test until 1999 and has plans to implement a different exam in 2007.

28. States that are developing high school exit exams are Alaska, Arizona, California, Delaware, Hawaii, Massachusetts, Utah, Washington, and Wisconsin.

29. Data illustrated in this chart were collected through telephone interviews and cross-checked with information provided in Quality Counts, 2001. States are counted the year the first graduating class was (or will be) affected by the state's first high school high-stakes graduation exam. For example, since the class of 1987 was the first class that had to pass the TEAMS in Texas, Texas was defined as a state with a high school exit exam in 1987.

30. Percentages were calculated using 1997 National Center for Education Statistics finance data available: http://nces.ed.gov/. . Data were adjusted for cost of living.

31. See Elazar's classification of state's governmental traditions of centralism and localism in Elazar, 1984. Hawaii and Alaska were not included in his analyses so were not included in these calculations.

32. These numbers were calculated using 2000 Census Bureau data available: http://www.census.gov/.

33. Ibid.

34. In the West, Nevada has a high school graduation exam and Alaska, California, Utah, and Washington have exams in progress (5/10 western states).

35. These numbers were calculated using 1999 Census Bureau data available: http://www.census.gov/.

36. New Mexico, Louisiana, California, Mississippi, New York, Alabama, Texas, Arizona, Georgia, South Carolina and Florida are among the 16 states with the highest degrees of poverty that have or have plans to implement high school graduation exams. For child poverty levels see 2001 Kids Count Data Online available: http://www.aecf.org/kidscount/kc2001/.

37. Ohanian, 1999.

38. Goodson & Foote, 2001.

39. McNeil, 2000.

40. Clarke, Haney, & Madaus, 2000.

41. The most influential research we found that substantiated the effectiveness of high-stakes testing policies came from Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S., 2000. Using NAEP data, researchers in this study recommended duplicating the high-stakes testing programs in North Carolina and Texas, although

concrete evidence that high-stakes testing programs caused the achievement gains noted in those states was lacking. Only a few other studies have substantiated the positive effects of high-stakes testing. See Carnoy, Loeb, & Smith, 2000; Muller & Schiller, 2000; Scheurich, Skrla, & Johson, 2000; and Schiller & Muller, 2000.

42. Sacks, 1999 and Kohn, 2000b.

43. The attachment of accountability measures to high academic standards has enjoyed a full measure of bipartisan support for the last decade or more. Eilperin, 2001 and Valencia, Valenzuela, Sloan & Foley, 2001.

44. Haney, 2001; Haney, 2000; Neill & Gayler, 1999; and Sacks, 1999.

45. Firestone, Camilli, Yurecko, Monfils & Mayrowetz, 2000; Goodson & Foote, 2001; Haney, 2000; Heubert & Hauser, 1999; Klein, Hamilton, McCaffrey, & Stecher, 2000; Kohn, 2000a; Kossan & González, 2000; Kreitzer, Madaus, & Haney, 1989; McNeil, 2000; McNeil & Valenzuela, 2001; Reardon, 1996; Sacks, 1999; Thomas & Bainbridge, 2001; and Urdan & Paris, 1994.

46. Chiu, 2000.

47. Robelen, 2000.

48. Sacks, 1999.

49. Salzer, 2000.

50. Kossan, 2000.

51. Domench, 2000.

52. Gardner, 1999, p. 16.

53. Shorris, 2000; and Shorris, 1997.

54. "high-stakes tests," 2000.

55. Ibid.

56. Ibid.

57. Heubert & Hauser, 1999.

58. Linn, 2000, pg. 14.

59. Heubert & Hauser, 1999, pg. 75.

60. Heubert & Hauser, 1999.

61. McNeil & Valenzuela, 2001, pg. 133.

62. McNeil & Valenzuela, 2001, pg. 134.

63. Wright, (Forthcoming).

64. This listing of "stakes" is not exhaustive. For example, local school districts and local schools may attach additional stakes to the consequences written into state test policies.

65. In 2004, grade promotion decisions in grades 3, 5, and 8 will be contingent upon student performance on Georgia's new Criterion-Referenced Competency Tests. Eventually, all Georgia students will have to pass promotion tests at each grade level.

66. Beginning in the fall of 2000, grade promotion became contingent on grades 4 and 8 performance on the new Louisiana Educational Assessment Program (LEAP 21) tests. Louisiana became the first state to retain students in grade using test scores.

67. Grade promotion in grade 8 is contingent on a combination of CTBS/5 test scores, student classroom performance, and classroom assessments.

68. A grade promotion "gateway" exists at grade 5. Beginning in 2002, promotion gateways will exist at grades 3 and 8.

69. Teachers in schools that perform poorly and are identified as low-performing by the state face the possibility of having to take a teacher competency test. Jones, Jones, Hardin, Chapman, Yarbrough, & Davis, 1999.

70. In 2002, promotion to the 5th grade will depend on a student's 4th grade reading score on the Ohio Proficiency Reading Test. Plans to make more grades promotion gateways are in progress.

71. In 2002, promotion to the 5th grade will depend on a student's 4th grade test scores. Plans to make more grades promotion gateways are in progress.

72. In 2003 students must pass the grade 3 reading test to be promoted to the 4th grade. In 2005 students must pass the grade 5 reading and math tests to be promoted to the 6th grade. In 2008 students must pass the grade 8 reading and math tests to be promoted to the 9th grade.

73. The state also uses student or school test results to evaluate teachers. Georgia has similar teacher accountability plans underway.

74. Quality Counts, 2001.

75. Information included has been pooled from the state department web sites and multiple telephone interviews with state testing personnel.

76. States with an asterisk (*) do not collect the percent of students who do not graduate or receive a regular high school diploma because they did not meet the graduation requirement at the state level. Almost 50% of the states do not collect this information. For these states, a rough percent was calculated by taking the number of 12th graders who on their last attempt before the intended graduation date did not meet the graduation requirement divided by the total 12th grade enrollment that year.

77. This percentage does not account for students who dropped out, who enrolled in

alternative or GED programs, or who transferred out of state.

78. Mississippi does not collect the percent of students who do not graduate or receive a regular high school diploma because they did not meet the graduation requirement and does not collect any data on the test after the test is first administered. The number of 12th graders who took, failed, or passed the test was therefore unavailable so rough estimates were impossible to calculate.

79. New York's graduation exam requirement consists of a series of end-of-course exams. Students take the end-of-course tests as early as the 7th grade or when students complete each Regents course.

80. Students in North Carolina take the competency tests only if they did not pass a series of similar, end-of-grade tests taken at the end of the 8th grade.

81. Mehrens, 1998.

82. Neill & Gayler, 2001, pg. 108.

83. Fisher, 2000, pg. 2.

84. As cited in Haney, 2000.

85. Klein, Hamilton, McCaffrey & Stecher, 2000.

86. Schrag, 2000.

87. Heubert & Hauser, 1999, pg. 132.

88. NAEP data are not collected at the high school level by state. As such, NAEP scores will not be used as direct indicators of how high school students have been affected by high school graduation exams. However, if a state has a high school graduation exam in place it has appropriately been defined as one of the states with high-stakes written into K–12 testing policies. Accordingly, gains over time as compared to the nation will indicate how more general high-stakes testing policies have improved each state's system of education.

89. Judd, Smith, & Kidder, 1991 and Smith & Glass, 1987.

90. Fraenkel & Wallen, 2000; Glass, 1988; and Smith & Glass, 1987.

91. Information included was pooled from the state department web sites, multiple telephone interviews with state testing personnel, and Quality Counts, 2001. State testing personnel in all states but Florida and Virginia verified the information before it was included in this chart.

92. In 39% (7/18) of the high-stakes states – Georgia, Maryland, Mississippi, New York, South Carolina, Tennessee, and Virginia – students will take end-of-course exams instead of high school graduation, criterion-referenced tests once they complete courses such as Algebra I, English 1, Physical Science, etc. ... End-of-course exams seem to be the new fad, replacing high school graduation exams.

93. Since the 1960s student performance on New York's Regents Exams determined the type of diploma students receive at graduation – a Local Diploma or a Regents Diploma.

94. The competency tests are only given to 9th graders who did not pass the end-of-grade tests at the end of the 8th grade.

95. In 1983 students did not have to pass the Texas Assessment of Basic Skills to receive a high school diploma.

96. Glass, 1988 and Smith & Glass, 1987.

97. Glass, 1988, pg. 445–446.

98. Campbell & Stanley, 1963; Glass, 1988; and Smith & Glass, 1987.

99. From 1959 to 1989 the original version of the ACT was used. In 1989 an enhanced ACT was implemented but only scores back to 1986 were equated to keep scores consistent across time. This explains the slight jumps from 1985–1986 that will be apparent across all states. Although scores from 1980–1985 have not been equated, the correlation between scores from the original and enhanced ACT assessments is high: r =.96

100. See footnote #99 to explain the large increase illustrated from 1985–1986.

101. Smith & Glass, 1987.

102. ACT composite scores (1980–2000) were available on–line at http://www.act.org. or were obtained through personal communications with Jim Maxey, Assistant Vice President for Applied Research at ACT. We are indebted to him for providing us with these data.

103. SAT composite scores (1977–2000) were available on-line at http://www.collegeboard.com. or were provided by personnel at the College Board. We thank those at the College Board who helped us in our pursuit of these data.

104. Kohn, 2000a.

105. Trends were defined in the short term, as defined by the difference in score one year after the point of implementation, and in the long term, as defined by the difference in score the number of years from one point of intervention to the next or 2001 as compared to the nation.

106. Changes in participation rates as compared to the nation (1994–2001) are listed in parentheses.

107. Correlation coefficients represent the relationship between changes in score and changes in participation or exclusion rates for participating states with high-stakes tests. Only states with high-stakes tests were included in the calculations of correlation coefficients hereafter. Coefficients were calculated separately from one year to the next for the years for which data and participation rates were available.

108. These correlation coefficients were calculated using changes in score and changes

in participation rates for the years in which data and participation rates were available.

109. Within states colleges may change their policies regarding which tests are required of enrolling students. This may affect participation and exclusion rates hereafter.

110. Changes in participation rates as compared to the nation (1991–1997 and 2000–2001) are listed in parentheses.

111. State NAEP composite scores (1990–2000) are available on–line at http://nces.ed.gov/nationsreportcard.

112. For more information on the NAEP, for example its design and methods of sampling see Johnson, 1992.

113. For further discussion see Neill & Gayler, 2001.

114. See Grissmer, Flanagan, Kawata, & Williamson, 2000 and Klein, Hamilton, McCaffrey & Stecher, 2000.

115. Johnson, 1992.

116. Neill & Gayler, 2001.

117. See the NAEP website at http://nces.ed.gov/nationsreportcard.

118. Haney, 2000.

119. This figure represents the r-square of each correlation coefficient that was calculated by squaring the correlations between change in score and change in exclusion rates year to year.

120. Changes in exclusion rates are listed next to changes in score hereafter. Scores and exclusion rates were calculated as compared to the nation.

121. The exclusion rate for the nation in 1990 was not available. The exclusion rate was imputed by calculating the average exclusion rate for all states that participated in the 1990 8th grade math NAEP.

122. For a similar study see Camilli, 2000. Camilli tested claims made by Grissmer et al., 2000, that large NAEP gains made by students from 1992 to 1996 in Texas were due to high-stakes tests. Camilli found, however, that the cohort of Texas students who took the NAEP math as 4th graders in 1992 and then again as 8th graders in 1996 were just average in gains. Camilli analyzed cohort gains in Texas on the NAEP 1992 and 1996 math assessment only, however. This section of the study will expand on Camilli's work to include all states with high-stakes tests. Further, randomly sampled cohorts of students who took the NAEP math as 4th graders in 1996 and as 8th graders in 2000 and cohorts of students who took the NAEP reading as 4th graders in 1994 and as 8th graders in 1998 will be examined.

123. Toenjes, Dworkin, Lorence & Hill, 2000.

124. Klein, Hamilton, McCaffrey & Stecher, 2000.

125. AP data (1995–2000) were available in the AP National Summary Reports available on-line at http://www.collegeboard.org/ap.

126. Participation rates were calculated by dividing the number of AP exams that were taken by students in the 11th and 12th grade by each state's total 11th and 12th grade population. Grades received on the exams were calculated by dividing the number of students who received a grade of 3 or above, a grade of 3 being the minimum grade required to receive college credit, by the total number of 11th and 12th grade participants.

127. "Controlling" for participation rates was possible only in this analysis. Years for which we had participation rates matched the years for which we had the percentages of students who passed AP exams.

128. "Fisher," 2000.

129. "Advanced placement," 2000.

References

Administrative bonuses: Tied to student performance in Oakland. (2001, February 8). *The National Education Goals Panel Weekly* [On-line]. Available: http://www.negp.gov/weekly.htm

Advanced placement: Growth in Texas. (2000, August 31). *The National Education Goals Panel Weekly* [On-line]. Available: http://www.negp.gov/weekly.htm

Berliner, D. C. & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley Publishing Company, Inc.

Bracey, G. W. (1995). Variance happens: Get over it! *Technos, 4* (3), 22–29.

Camilli, G. (2000). Texas gains on NAEP: Points of light? *Education Policy Analysis Archives, 8* (42) [On-line]. Available: http://epaa.asu.edu/epaa/v8n42.html

Campbell, D. T. (1975). On the conflicts between biological and social evolution and between psychology and moral tradition. *American Psychologist, 30* (12), 1103-1126.

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.) *Handbook of research on teaching*. Chicago, IL: Rand McNally & Company.

Carnoy, M., Loeb, S. & Smith, T.L. (2000, April). *Do higher test scores in Texas make for better high school outcomes*? Paper presented at the American Educational Research Association Annual Meeting, New Orleans, LA.

Chiu, L. (2000, October 3). Education issues fuel Capitol rally. *The Arizona Republic* [On-line]. Available: http://www.azcentral.com/news/education/1003AIMS03.html

Civil rights coalition sues over race discrimination in Michigan Merit Scholarship Program. American Civil Liberties Union Freedom Network, (2000, June 27).

[On-line]. Available: http://www.aclu.org/news/2000/n062700a.html

Clarke, M., Haney, W., & Madaus, G. (2000). *High stakes testing and high school completion*. The National Board on Educational Testing and Public Policy [On-line]. Available: http://www.nbetpp.bc.edu/reports.html

Commission on Instructionally Supportive Testing. (2001). *Building tests to support instruction and accountability: A guide for policymakers* [On-line]. Available: http://www.aasa.org/issues_and_insights/assessment/

Domench, D. A. (2000, December). *School administrator web edition* [On-line]. Available: http://www.aasa.org/publications/sa/2000_12/domenech.htm

Durbin, D. (2001, March 21). Merit awards fail minorities. *The Detroit News* [On-line]. Available: http://www.detnews.com/2001/schools/0103/21/c07d-202027.htm

Eilperin. J. (2001. May 23). House backs annual reading, math tests. *Washington Post* [On-line]. Available: http://washingtonpost.com/wp-dyn/education/A63017-2001May22.html

Elazar, D. J. (1984). *American federalism: A view from the states* (3rd ed.). New York: Harper & Row, Publishers.

Figlio, D. N. & Lucas, M. E. (2000). *What's in a grade? School report cards and house prices*. National Bureau of Economic Research [On-line]. Available: http://papers.nber.org/papers/w8019

Firestone, W. A., Camilli, G., Yurecko, M., Monfils, L., & Mayrowetz, D. (2000). State standards, socio-fiscal context and opportunity to learn in New Jersey. *Education Policy Analysis Archives, 8* (35) [On-line]. Available: http://olam.ed.asu.edu/epaa/v8n35/

Fisher, F. (2000). Tall tales? Texas testing moves from the Pecos to Wobegon. Unpublished manuscript.

Folmar, K. (2001, August 9). Lawsuit delays payment of state bonuses to teachers. *San Jose Mercury News* [On-line]. Available: http://www0.mercurycenter.com/premium/local/docs/award09.htm

Fraenkel, J. R. & Wallen, N. E. (2000). *How to design and evaluate research in education* (4th ed.). Boston, MA: McGraw Hill, Inc.

Gardner, H. (1999). *The disciplined mind: What all students should understand*. New York: Simon & Schuster.

Glass, G. V (1988). Quasi-experiments: The case of interrupted time series. In R. M. Jaeger (Ed.) *Complementary methods for research in education*. Washington, DC: American Educational Research Association.

Goodson, I. & Foote, M. (2001). Testing times: A school case study. *Education Policy Analysis Archives, 9* (2) [On-line]. Available: http://epaa.asu.edu/epaa/v9n2.html

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student*

achievement: What NAEP test scores tell us. Santa Monica, CA: RAND Corporation [On-line]. Available: http://www.rand.org/publications/MR/MR924/

Haladyna, T., Nolen, S. B. & Haas, N. S. (1991). Raising standardized test scores and the origins of test score pollution. *Educational Researcher, 20* (5), 2–7.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Analysis Policy Archives, 8* (41) [On-line]. Available: http://epaa.asu.edu/epaa/v8n41/

Haney, W. (2001). *Revisiting the myth of the Texas miracle in education: Lessons about dropout research and dropout prevention*. Paper prepared for the "Dropout Research: Accurate Counts and Positive Interventions" Conference sponsored by Achieve and the Harvard Civil Rights Project, Cambridge, MA [On-line]. Available: http://www.law.harvard.edu/groups/civilrights/publications/dropout/haney.pdf

Heller, D. E. (September 30, 1999). Misdirected money: Merit scholarships take resources from have-nots. *Detroit Free Press* [On-line]. Available: http://www.freep.com/voices/columnists/qehell30.htm

Heubert, J. P. & Hauser, R. M. (Eds.) (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press [On-line]. Available: http://www.nap.edu/html/highstakes/

High stakes tests: A harsh agenda for America's children. (2000, March 31). Remarks prepared for U.S. Senator Paul D. Wellstone. Teachers College, Columbia University [On-line]. Available: http://www.senate.gov/~wellstone/columbia.htm

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29* (2), 95–110.

Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T. & Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan, 81* (3), 199–203.

Judd, C. M., Smith, E. R. & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

Klein, S. P., Hamilton, L.S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives, 8* (49) [On-line]. Available: http://epaa.asu.edu/epaa/v8n49/

Kohn, A. (2000a). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth, N.H.: Heinemann.

Kohn, A. (2000b, September 27). Standardized testing and its victims. *Education Week* [On-line]. Available: http://www.edweek.org/ew/ewstory.cfm?slug=04kohn.h20

Kossan, P. & González, D. (2000, November 2). Minorities fail AIMS at high rate: Some backers talking change. *The Arizona Republic* [On-line]. Available: http://www.arizonarepublic.com/news/articles/1102aimsrace02.html

Kossan, P. (2000, November 27). By trying too much too quick, AIMS missed mark.

The Arizona Republic [On-line]. Available: http://www.arizonarepublic.com/news/articles/1127aims27.html

Kreitzer, A. E., Madaus, G. F., & Haney, W. (1989). Competency testing and dropouts. In L. Weis, E. Farrar, & H. G. Petrie (Eds.) *Dropouts from school: Issues, dilemmas, and solutions*. Albany, NY: State University of New York Press.

Linn, R. L. (2000). Assessments and accountability. *Education Researcher, 29* (2), 4–15 [On-line]. Available: http://www.aera.net/pubs/er/arts/29-02/linn01.htm

Madaus, G. & Clarke, M. (2001). The adverse impact of high stakes testing on minority students: Evidence from one hundred years of test data. In Orfield, G., & Kornhaber, M. L. (Eds.). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press.

McNeil, L. M. & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric. In Orfield, G., & Kornhaber, M.L. (Eds.). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press.

McNeil, L. M. (2000). *Contradictions of school reform*. New York, NY: Routledge.

Mehrens, W.A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6* (13) [On-line]. Available: http://olam.ed.asu.edu/epaa/v6n13.html

Muller, C & Schiller, K.S. (2000). Leveling the playing field? Students' educational attainment and states' performance testing. *Sociology of Education, 73* (3), 196–218.

National Governor's Association. (2000, January 25). *1999 state initiatives on spending tobacco settlement revenues*. [On-line]. Available: http://www.nga.org/Pubs/IssueBriefs/2000/000125Tobacco.asp#Summary

Neill, M., & Gayler, K. (2001). Do high-stakes graduation tests improve learning outcomes? Using state-level NAEP data to evaluate the effects of mandatory graduation tests. In Orfield, G., & Kornhaber, M. L. (Eds.). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press.

Neufeld, S. (2000, October 2). Backlash fermenting against school tests: Groups organize to complain about STAR. *San Jose Mercury News* [On-line]. Available: http://www.mercurycenter.com/premium/local/docs/backlash02.htm

Ohanian, S. (1999). *One size fits few: The folly of educational standards*. Portsmouth, NH: Heinemann.

Orfield, G., & Kornhaber, M. L. (Eds.) (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press.

Pallas, A.M., Natriello, G., & McDill, E.L. (1989). *The changing nature of the disadvantaged population: Current dimensions and future trends*. Center for Research

on Elementary and Middle Schools. (ERIC Document Reproduction Service No. ED 320 655).

Paris, S. G. (2000). Trojan horse in the schoolyard: The hidden threats in high-stakes testing. *Issues in Education, 6* (1, 2), 1-16.

Quality Counts 2001. (2001). *Education Week* [On-line]. Available: http://www.edweek.org/sreports/qc01/

Reardon, S. F. (1996). *Eighth grade minimum competency testing and early high school dropout patterns*. Paper presented at the Annual Meeting of the American Educational Research Association (ERIC Document Reproduction Service No. ED 400 273).

Robelen, E. W. (2000, October 11). Parents seek civil rights probe of high-stakes tests in La. *Education Week* [On-line]. Available: http://www.edweek.org/ew/ewstory.cfm?slug=06ocr.h20

Ross, J. (2001, April 2). High-testing kids grab MEAP cash: More students savor $2,500 reward to use for college education, expenses. *Detroit Free Press* [On-line]. Available: http://www.freep.com/news/education/maward2_20010402.htm

Sacks, P. (1999). *Standardized minds: The high price of America's testing culture and what we can do to change it*. Cambridge, MA: Perseus Books.

Salzer, J. (2001, April 29). Georgia dropout rate highest in nation: Years of reform achieve little, and some fear testing will only worsen problem. *Atlanta-Journal Constitution* [On-line]. Available: http://www.accessatlanta.com/partners/ajc/epaper/editions/sunday/news_a3be6ad4f17dd10f007c.hml

Scheurich, J. J., Skrla, L & Johnson, J. F. (2000). Thinking carefully about equity and accountability. *Phi Delta Kappan, 82* (4), 293–299.

Schrag, P. (2000, January 3). Too good to be true. *The American Prospect* [On-line]. Available: http://www.prospect.org/archives/V11-4/schrag-p.html

Serow, R. C. (1984). Effects of minimum competency testing for minority students: A review of expectations and outcomes. *The Urban Review, 16* (2), 67-75.

Sheldon, K. M. & Biddle, B. J. (1998). Standards accountability and school reform: Perils and pitfalls. *Teachers College Record, 100* (1), 164-180.

Shorris, E. (1997). *New American Blues*. New York: W. W. Norton.

Shorris, E. (2000). *Riches for the poor*. New York: W. W. Norton.

Smith, M. L. & Glass, G.V (1987). *Research and evaluation in education and the social sciences*. Needham Heights, MA: Allyn and Bacon.

Swope, K. and Miner, B. (Eds.). (2000). *Failing our kids: Why the testing craze won't fix our schools*. Milwaukee, WI: Rethinking Schools, Ltd.

Thomas, M. D. & Bainbridge, W. L. (2001). "All children can learn:" Facts and

fallacies. *Phi Delta Kappan, 82* (9), 660–662.

Toenjes, L. A., Dworkin, A. G., Lorence, J., & Hill, A. J. (2000). *The lone star gamble: High stakes testing, accountability, and student achievement in Texas and Houston* [On-line]. Available: http://www.brook.edu/gs/brown/bc_report/2000/Houston.PDF

U.S. Department of Education. (1983). *A nation at risk: The imperative for educational reform* [On-line]. Available: http://www.ed.gov/pubs/NatAtRisk/index.html

Urdan, T. C. & Paris, S. G. (1994). Teachers' perceptions of standardized achievement tests. *Educational Policy, 8* (2), 137–157.

Using tobacco settlement revenues for children's services: State opportunities and actions. (1999, October). National Conference of State Legislatures and The Finance Project. [On-line]. Available: http://www.financeproject.org/tobaccoattach.htm

Valencia, R. R., Valenzuela, A., Sloan, K., & Foley, D. E. (2001). Let's treat the cause, not the symptoms: Equity and accountability in Texas revisited. *Phi Delta Kappan, 83* (4), 318-321, 326.

Wright. W. E. (Forthcoming). The effects of high stakes testing on an inner-city elementary school: The curriculum, the teachers, and the English Language Learners. *Current Issues in Education* [On-line]. Available: http://cie.ed.asu.edu/

About the Authors

Audrey L. Amrein
College of Education
Arizona State University
Division of Educational Leadership and Policy Studies
PO Box 872411
Tempe, AZ 85287-2411

Email: audrey.beardsley@cox.net

Audrey L. Amrein is an Assistant Research Professional in the College of Education at Arizona State University in Tempe, Arizona. Her research interests include the study of large-scale educational policies and their effects on students from racial minority, language minority, and economically disadvantaged backgrounds. Specifically, she is interested in investigating the effects of high-stakes tests, bilingual education, and charter school policies as they pertain to issues of equity.

David C. Berliner
Regents' Professor of Education
College of Education
Arizona State University
Tempe, AZ 85287-2411

Email: berliner@asu.edu

David C. Berliner is Regents' Professor of Education at the College of Education of

Arizona State University, in Tempe, AZ. He received his Ph.D. in 1968 from Stanford University in educational psychology, and has worked also at the University of Massachusetts, WestEd, and the University of Arizona. He has served as president of the American Educational Research Association (AERA), president of the Division of Educational Psychology of the American Psychological Association, and as a fellow of the Center for Advanced Study in the Behavioral Sciences and a member of the National Academy of Education. Berliner's publications include *The Manufactured Crisis*, Addison-Wesley, 1995 (with B.J. Biddle) and *The Handbook of Educational Psychology*, Macmillan, 1996 (Edited with R.C. Calfee). Special awards include the Research into Practice Award of AERA, the National Association of Secondary School Principals Distinguished Service Award, and the Medal of Honor from the University of Helsinki. His scholarly interests include research on teaching and education policy analysis.

Appendices

Appendices are available in either html format or Rich Text Format, the latter being a word processor file. The RTF file is 3.5 megabytes in size.

- Appendices in html.
- Appendices in Rich Text Format

The World Wide Web address for the *Education Policy Analysis Archives* is **epaa.asu.edu**

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board

EPAA Spanish Language Editorial Board