DOCUMENT RESUME

ED 429 089 TM 029 621

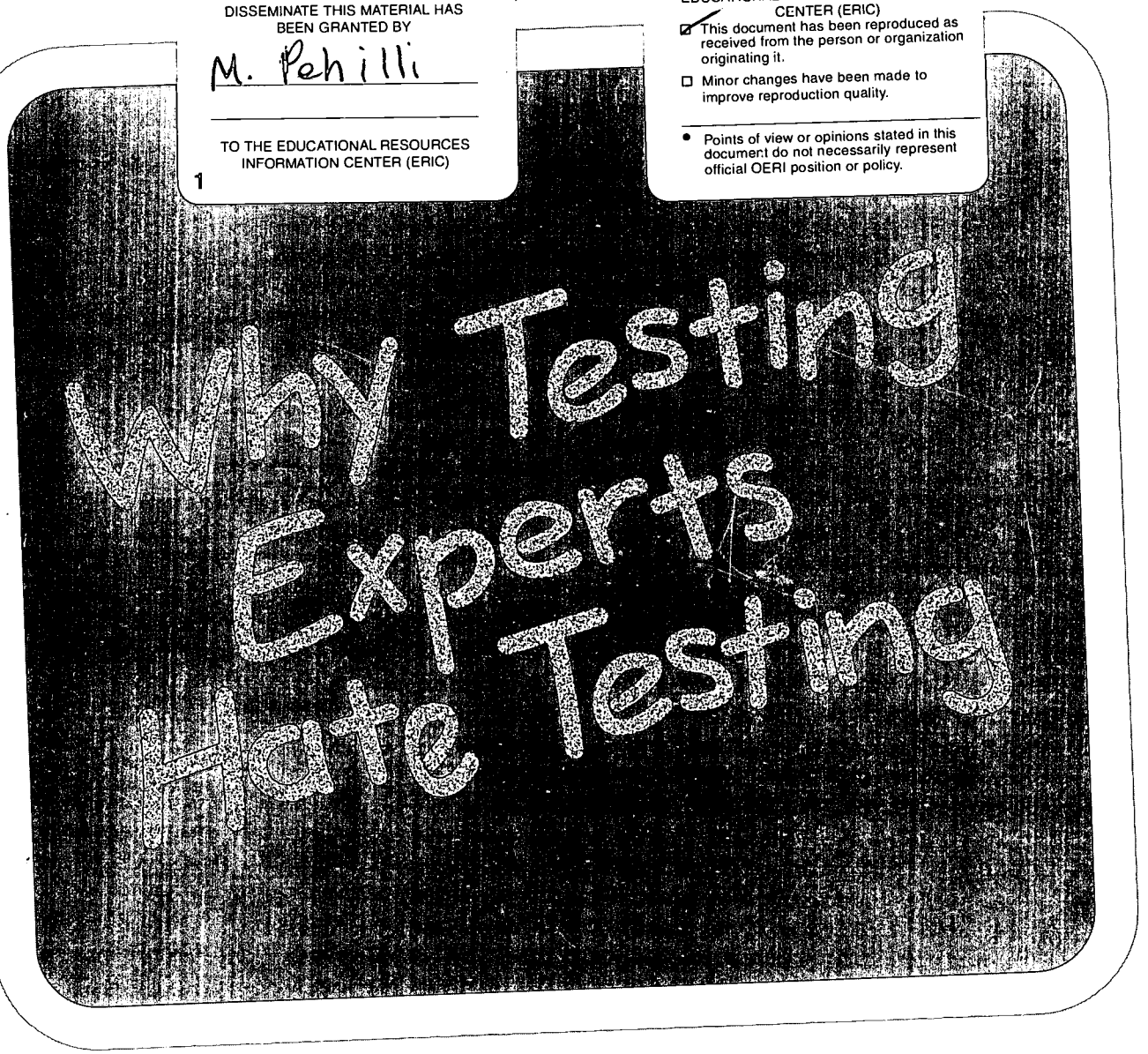
AUTHOR Phelps, Richard P.
TITLE Why Testing Experts Hate Testing.
INSTITUTION Thomas B. Fordham Foundation, Washington, DC.
PUB DATE 1999-01-00
NOTE 41p.
AVAILABLE FROM Thomas B. Fordham Foundation, 1627 K Street, NW, Suite 600, Washington, DC 2006 (single copies free).
PUB TYPE Collected Works - Serials (022)
JOURNAL CIT Fordham Report; v3 n1 Jan 1999
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Case Studies; Elementary Secondary Education; Minority Groups; Scores; *Standardized Tests; Test Bias; *Test Use; *Testing Programs
IDENTIFIERS *High Stakes Tests; National Assessment of Educational Progress; Scholastic Assessment Tests

ABSTRACT

The objections of testing experts to standardized testing are evaluated. The report begins with a foreword by Chester E. Finn, Jr., followed by an executive summary and an introduction. Four case studies include: (1) experts' opposition to high-stakes testing in Texas; (2) in North Carolina; (3) concerns raised in connection with the National Assessment of Educational Progress; and (4) in connection with the Scholastic Assessment Tests. Eight alleged harms of standardized testing are: (1) test score inflation; (2) curriculum narrowing; (3) emphasis on lower-level thinking; (4) declining achievement; (5) harm to women and minorities; (6) expense; (7) over use compared to other countries; and (8) opposition from parents, students, and teachers. Each of the claims is examined in detail, and a rebuttal is offered for each. The conclusion offers two views of testing and learning. (Contains 139 endnotes.) (SLD)

THOMAS B. FORDHAM FOUNDATION

OUTSIDE THE BOX



Why Testing Experts Like Testing

by Richard P. Phelps

JANUARY 1999

Fordham Report

Vol. 3, No. 1 January 1999

Why Testing Experts Hate Testing

by
Richard P. Phelps

THOMAS B. FORDHAM FOUNDATION
OUTSIDE THE BOX

Table of Contents

Foreword

No issue in U.S. education is more controversial than testing. Some people view it as the linchpin of serious reform and improvement, others as a menace to quality teaching and learning.

The public's view is clear: most Americans favor plenty of student testing for purposes of information, accountability and incentives. So do most policy makers. Hence statewide assessment systems are now the norm. Most of these systems rely in full or in part on "standardized" testing of some sort.

Particularly as the "stakes" that are attached to test results become more serious — with promotion and graduation now often hinging on attaining some minimum score — more attention is understandably being paid to the strengths and weaknesses of standardized testing and to ways of improving it. Many jurisdictions have been striving to make their assessments more sophisticated and sensitive. Some states supplement their "standardized" tests with more complex methods for gauging the performance of districts, schools and students.

Yet the critics are relentless. Even as the public and its elected officials want these testing regimens to become more consequential, many educators deplore them. In the widening use of such tests, they see a practice that distorts the curriculum, discourages higher-order thinking skills, and, ultimately, depresses student achievement.

The most curious aspect of this debate is the special animus that many testing "experts" hold for tests. Indeed, I have sometimes thought that the working definition of a testing expert is "a person with a Ph.D. who has the reputation of knowing something about testing but who has never met any test that he thinks should actually be used for any real purpose."

This is the third research report on testing from the Thomas B. Fordham Foundation. In July, 1998, we issued *A TIMSS Primer* by Harold W. Stevenson, and in October we published *Filling In the Blanks: Putting Standardized Tests to the Test* by Gregory J. Cizek.

Now we are pleased to present this critical essay by Richard P. Phelps, *Why Testing Experts Hate Testing*. In its pages, Phelps engages in a point-by-point analysis of eight arguments that testing experts commonly fling against standardized testing. He describes those who embrace this anti-testing canon and delves into the research they offer as proof. Finally, Phelps offers his estimation of the future of testing — and its critics.

Richard Phelps has conducted education research for over a dozen years for the Indiana Department of Education, the General Accounting Office and the American Institutes of Research. Recently, he has joined the staff of the Organisation for Economic Co-operation and Development in Paris, where he helps to coordinate OECD's World Education Indicators project. (This report was not written in

association with any of his aforementioned employers, however, and should not be construed to represent their views in any way.) Readers wishing to contact him directly may write to him at OECD ELS/SID 2, rue Andre Pascal, F-75775 Paris CEDEX 16 France or send e-mail to Richard.PHELPS@oecd.org.

The Thomas B. Fordham Foundation is a private foundation that supports research, publications, and action projects in elementary/secondary education reform at the national level and in the vicinity of Dayton, Ohio. Further information can be obtained from our web site (www.edexcellence.net) or by writing us at 1627 K Street, NW, Suite 600, Washington, DC 20006. (We can also be e-mailed through our web site.) This report is available in full on the Foundation's web site, and hard copies can be obtained by calling 1-888-TBF-7474 (single copies are free). The earlier testing reports by Messrs. Stevenson and Cizek may be obtained in the same ways. The Foundation is not connected to or sponsored by Fordham University.

Chester E. Finn, Jr., President
Thomas B. Fordham Foundation
Washington, DC
January 1999

Executive Summary

The American people have consistently advocated greater use of standardized student testing, and more states than ever now administer standardized tests on a regular basis. Yet such tests also evoke protests from many educators and testing experts. These critics oppose the concept of testing in general, and fervently denounce high-stakes tests (which carry rewards for success and consequences for failure) and multiple-choice tests in particular. Critics generally find standardized tests to be seriously flawed and allege that they even have pernicious effects.

Testing experts hurl an arsenal of arguments against a wide range of targets. This report appraises their objections, first by examining four case studies, then by scrutinizing eight of the principal arguments and the "research" that undergirds them.

The case studies include testing experts' opposition to high-stakes testing in Texas and North Carolina, where the assessments have been credited with producing marked gains in student achievement; objections to state-level reporting of scores on the National Assessment of Educational Progress (NAEP); and challenging the use of the SAT scores in college admissions decisions.

The basic argument made by testing critics is that the use of high-stakes standardized tests is counterproductive. Instead of leading to stronger academic achievement, it is said to interfere with good teaching and learning. In this contention, the critics embrace a sort of domino theory. Pressure to produce higher scores leads teachers to focus on material that will be covered by the tests and to exclude everything else. The curriculum is thereby narrowed, which means that some subjects are ignored. Within those that are taught, lower order thinking skills are emphasized. As a result, test scores get inflated while real learning suffers.

In addition to the alleged harms of 1) test score inflation, 2) curriculum narrowing, 3) emphasis on lower-order thinking, and 4) declining achievement, testing experts add a quartet of other arguments: 5) standardized tests hurt minorities and women, 6) the tests are too costly, 7) other countries don't test nearly as much as the U.S. does, and 8) parents, teachers and students in this country are all opposed to testing.

These eight claims are examined in detail and a rebuttal is offered to each. The arguments are found to be irrelevant, misplaced, overly simplistic or untrue. Many of the weaknesses attributed to standardized testing turn out to involve shortcomings of teaching and asssessment in general rather than standardized testing per se.

Introduction[1]

The public has often been asked how it feels about testing. Over several decades and in a variety of contexts, the American people have consistently advocated greater use of standardized student testing, preferably with consequences for failure (i.e., "high stakes"). The margins in favor have typically been huge, on the order of 70-point spreads between the percentage in favor of more testing and the percentage against.[2]

But the public may not get its way. Many educators and education "experts" oppose standardized testing and high stakes. This throng includes some school administrators who fear the fallout from poor test results, but also, and most notably, it includes most education school faculty. In a 1997 survey, a national sample of these faculty members voiced substantially less support for high-stakes standardized testing than did other groups. "[O]nly 49 percent believe raising the standards of promotion from grade school to junior high and letting kids move ahead only when they pass a test showing they've reached those standards, would do a great deal to improve academic achievement. In sharp contrast, the percentage reaches 70 percent among the general public (and 62 percent among teachers)."[3]

It frequently seems that experts on testing have never met an actual test that they like and want to see used.

The polling organization Public Agenda found that, "while supporting standards in concept, professors of education seem reluctant to put into place concrete, high-stakes tests that would signal when kids are meeting the standards."[4] They are especially opposed to multiple choice tests. "Fully 78 percent want less reliance on multiple-choice exams in the schools. [E]ducation professors...call for more reliance on portfolios and other authentic assessments."[5]

These faculty members don't think standardized tests demonstrate learning. "The fact is that all of the data say standardized tests don't predict what they are intended to. They just don't do it.... There is no standardized test that is good," a Boston professor told Public Agenda.[6] The professors recognize that the public has a different view of testing, however. Public Agenda reported that many faculty members expressed "disappointment and some exasperation that so much current educational research seems to be ignored or dismissed by the public."[7]

Several years ago, the American Educational Research Association (AERA), a group consisting primarily of education professors, hosted a press conference on student testing issues in Washington, DC as a "public service to build bridges between researchers and policy makers." Five prominent members of the group presented papers, such as "The Teacher, Standardized Testing, and Prospects of Revolution," in unanimous opposition to President Bush's then-pending proposal for national tests, as well as to "high-stakes" use of standardized tests, multiple-choice formats, "external" tests (i.e., tests not written by classroom teachers), and other features of student testing that they disliked.[8]

The reader may be struck by a paradox: it frequently seems that experts on testing have never met an actual test that they like and want to see used. What is it about testing that troubles them so? A now-familiar litany of assertions has been offered to explain how "research" shows that standardized testing is

bad. The anti-testing canon includes allegations that standardized tests, particularly those with "high stakes":

- induce "teaching to the test" which, in turn, leads to artificial inflation of scores;
- narrow the curriculum to a small domain of topics;
- tap only "lower-order thinking" and hence discourage innovative curricula and teaching strategies;
- cause student achievement to decline;
- are unfair to minorities and women;
- are costly in terms of money and time;
- are overused in the United States, especially in comparison with other countries; and
- are opposed by all those who truly care about children.

Not all testing experts hate testing, however. Hundreds of them work cheerfully for state and local testing agencies and for test developers. The opponents we hear the most from are a relatively small group of "testing policy" researchers, who are on the faculty of education schools or who work at organizations such as the federally funded Center for Research on Evaluation, Standards, and Student Testing (CRESST),[9] the Center for the Study of Testing, Evaluation, and Educational Policy (CSTEEP)[10] at Boston College, and an advocacy group known as the National Center for Fair and Open Testing (FairTest).[11] A brief excerpt from a FairTest publication entitled *Fallout from the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools* sums up the basic position of that organization.[12]

> Standardized tests often produce results that are inaccurate, inconsistent, and biased against minority, female, and low-income students. Such tests shift control and authority into the hands of the unregulated testing industry and can undermine school achievement by narrowing the curriculum, frustrating teachers, and driving students out of school. [13]

In this report, the arguments of the testing experts who hate testing will be held up for careful scrutiny. First we examine four case studies that suggest how these experts deploy their arguments in the real world. Then we appraise those arguments.

Case Studies

Case Study 1: The National Assessment of Educational Progress (NAEP)

Proposals for national testing systems, be they from George Bush or Bill Clinton, tend to attract a great deal of attention. To date, however, there is only one such test — the National Assessment of Educational Progress (NAEP). NAEP is an assessment based on samples of schools, and no individual student information is made available to anyone. It is a "no-stakes" test.

For decades, NAEP samples were exclusively national and so were NAEP scores. In the 1980s, however, many people pressed for state-representative NAEP samples (a.k.a., State NAEP). Almost half the states had instituted their own testing programs, many of them high-stakes "minimum-competency" graduation requirements. Some state leaders wanted to gauge their students' levels of achievement or the progress of their states' education reform efforts against an external benchmark, and the scores of state-representative samples of schools and students on NAEP seemed the perfect candidate to be that benchmark.

But what sounds like an obvious idea drew strong opposition from testing experts. Daniel Koretz, a researcher with CRESST and the RAND Corporation, made three separate arguments against releasing state-by-state NAEP scores. First, he argued that the public cannot be trusted with such information. Koretz wrote that "[S]ome differences among states would be too fragile — too dependent on the specifics of the test — to warrant the simple interpretations that they will receive."[14] Second, Koretz argued, academic success is predicted primarily by the socioeconomic background of students, so state-level NAEP will just show once again that richer states do better:[15] "To infer that a difference between two states on the NAEP reflects specific policies or practices, one needs to be able to reject with reasonable confidence other plausible explanations, such as economic or demographic difference."[16] (Other opponents of State NAEP have made these arguments even more forcefully.[17]) Third, Koretz insisted that, because State NAEP provides only cross-sectional data, it cannot show improvements in achievement that may coincide with education reform programs: "NAEP is purely cross-sectional, which eliminates a large number of the designs that could be used to draw causal inferences."[18]

The essence of these objections is that state-level NAEP results would be used to judge states and these judgments would inevitably be unfair. Since people who don't understand what the scores really mean would use this information to evaluate the states, we shouldn't gather the information at all.

In a counter to Koretz, Gary W. Phillips of the U.S. Education Department's National Center for Education Statistics noted that, although a single administration of State NAEP might not allow us to evaluate the impact of reforms, a system needed to be established that could be used to appraise such changes in the future. We had to start somewhere.[19] The National Academy of Education, assigned to review the efficacy of State NAEP, recommended implementation and reiterated that recommendation in a 1996 review.[20]

With several administrations of state-level NAEP now behind us, we have time-series data with which to gauge the progress (or lack thereof) each state's youngsters are making in mathematics, reading, and (soon) science. We can

thus begin to see where state education policies are effective, with background factors controlled. The utility of NAEP scores as markers for monitoring state education reforms will be seen in the following two case studies.

The 1988 legislation establishing state-level NAEP also permitted "standards-based" reporting of scores. Historically, NAEP results were reported only according to abstract "scale scores" that were not anchored to any standards. But the National Assessment Governing Board — to the continuing dismay of many testing experts — judged that NAEP results would be far more useful, particularly in tracking progress toward the national education goals that the President and governors set in 1989, if they showed how U.S. children were doing academically in relation to how well they *ought* to be doing. The Governing Board established three performance levels, which it termed "basic, proficient, and advanced," and accompanied each with descriptions written in plain English about the specific skills and abilities represented by each level. Like State NAEP, the performance level concept and the method for setting the levels have drawn controversy, with some testing critics favoring the old scale scores' aloof abstractness and many policymakers desiring more useful and understandable measures.[21]

It is clear that Americans want the kind of information about the achievement of the nation's students currently provided by NAEP summary scores and achievement-level results.

It would appear, however, that the performance levels are here to stay. The National Assessment Governing Board has remained steadfast. And a recent National Research Council review of NAEP, while agreeing with the critics on a number of specific points, also concedes that

> It is clear that Americans want the kind of information about the achievement of the nation's students currently provided by NAEP summary scores and achievement-level results.[22]

Case Study 2: Texas

Perhaps no state testing program has aroused the ire of testing critics more than the Texas Assessment of Academic Skills (TAAS), for over a decade the backbone of the Lone Star State's education accountability program. In its ratings of all state testing programs, FairTest rates the TAAS at 2 on a scale of 1 to 5, with 1 being the worst score possible.

FairTest explains its dim view of TAAS as follows:

> The Texas assessment system needs many major changes. It relies almost entirely on multiple-choice items, except for a writing prompt, and has a high-stakes graduation test. On most of the other standards, however, the state does very well. It has strong bias review procedures, provides solid public information, accords parents substantial rights, and has a thorough and continuing review system. Professional development appears fairly extensive.[23]

Observe that FairTest gives the state's testing program the second-lowest possible rating for only two reasons: the use of high-stakes and multiple-choice tests. According to Monty Neill of FairTest, "When you have high stakes and then add an exit exam, that jacks up the system so that the test becomes the curriculum.

One should not be using scores on tests to make serious educational decisions."[24]

Responding to evidence that pupil achievement in Texas has improved markedly since TAAS was introduced, Neill "concedes that the improvements are impressive," reports *Education Week*, "but he says that an enriched curriculum, not test preparation, is behind the shifts."[25] There may be a contradiction here. According to Neill, the test has *become* the curriculum in Texas and the improvement in student achievement is due to an enriched curriculum. Still, he declines to see the improvement as linked to the testing.

In addition to the FairTest criticisms of TAAS for its high-stakes and multiple-choice formats, the Texas testing program has been the subject of two separate lawsuits. The NAACP asserted that it was biased against blacks since they performed worse than whites on the test.[26] The Mexican-American Legal Defense Fund followed with a suit using the same logic.[27] Both cases were heard by the U.S. Education Department's Office for Civil Rights and both were dismissed.[28]

Perhaps no state testing program has aroused the ire of testing critics more than the Texas Assessment of Academic Skills (TAAS).

Through the clouds of flack, however, the citizens of Texas have remained on course, retaining and expanding TAAS. Moreover, the results do appear to be positive. Texas students' average state test scores have shown achievement gains year after year. That Texas students have also made gains well above the national average on NAEP throughout the past decade would seem to corroborate the improvement.[29]

Other benefits have also followed. Observers of Texas report

- a greater focus on academic learning;
- a culture of high expectations and enthusiasm toward reaching standards;
- generous and immediate remediation efforts offered to poorly performing students, both because a system is in place to identify their problems early and because, with high-stakes tests, students' problems are not just passed along to the next grade, where they become compounded;
- greater interest among teachers in academic strategies and more cooperation with each other to learn which ones work best, and how;
- with a regular system of assessment, quicker feedback for school faculty on which instructional systems work best; and
- the develoment in Texas of a school-specific information system on the World Wide Web for all parents to see, helping them understand their schools better.[30]

Though always intended to match Texas's curriculum and performance standards, the state's student testing program first took aim at basic skills and minimum competency. It has now been expanded to cover more grades (now 3 to 10) and purposes (statewide end-of-course examinations, for example). It has achieved better integration with the curriculum, professional development, and program planning, as well as student evaluation, and is today a key component of one of the most comprehensive accountability systems in the country.[31]

TAAS has also received strong political support from both parties. Both Republican and Democratic governors have resisted attempts to soften its requirements, even in the face of sus-

tained criticism. Indeed, gubernatorial opponents in the 1994 election attempted to outdo each other in their support for still higher standards and tougher requirements.[32] In 1998, it is not even an issue.

Case Study 3: North Carolina

A similar story can be told about North Carolina, a state that, like Texas, ranked near the bottom on NAEP but has improved its student achievement dramatically after instituting a comprehensive, integrated, high-stakes testing program, and sticking with it despite serious opposition.[33]

North Carolina ranked near the bottom on NAEP but has improved its student achievement dramatically after instituting a comprehensive, integrated, high-stakes testing program.

The North Carolina Education Department rates schools based on their results on state tests. It is a "value-added" rating system in which adjustments are made to the expected performance of each school for socioeconomic and other background factors. Teachers at schools rated "exemplary" are rewarded monetarily. But poorly performing schools are not abandoned. The department assembles teams of three to five experts in curriculum and instruction who work with those schools for an entire year. These teams help school staff align their curricula with state academic standards, demonstrate effective teaching techniques, and try to locate additional resources for the schools.[34] Fifteen schools designated for "mandatory assistance" at the end of the 1996-1997 school year finished 1997-1998 by achieving "exemplary" ratings for improving their performance by more than 10 percent.[35] In 1998-99, state "assistance teams" are visiting forty-six public and seven charter schools, eleven of them under a "mandatory assistance" provision for the worst-performing schools in the state, the rest under voluntary arrangements.[36]

The sixty-odd schools visited by state assistance teams represent less than 3 percent of the state's schools, however. Most schools either develop their own reform programs or rely on assistance from their school district. In Hoke County, the poorest county in the state, students who fail a test are offered a round of after-school classes, and are then allowed to retake the test without penalty. [37] A Southern Regional Education Board study of the Hoke County Schools' reform program found that

> [T]he percentage of students who now meet the state's algebra proficiency standard has doubled. Twenty percent more now meet the history standard. And the high school's overall Scholastic Assessment Test (SAT) scores are up 11 percent over three years. Also employers are more welcoming of graduates now.[38]

The whole process of reform in Hoke County was set in motion by its initial poor showing in the state testing program, which identified the district's academic problems.

Still, holding students, teachers, or schools to fixed standards means that some will do less well than others and some students may be held back. Johnson County, North Carolina, for example, passed a student-accountability policy of its own in 1996. The policy called for intensive remediation, but also for retention in grade of students who did not score at a proficient level on state exams.[39]

District officials claim the accountability program has boosted student performance.

According to Johnson County officials, more than one-third of students performed below grade level on the tests four years ago, yet just 1 percent of students were held back.[40] This year, less than one-fourth of the students performed below grade level and 8.8 percent of students were retained under the policy — and for other reasons, such as absenteeism. The other 16 percent were promoted "based on the grades they earned and other academic factors."[41]

Not everyone liked the new policy. Fourteen parents filed suit against Johnson County on behalf of children who were held back. They argued that the tests were intended by the state to rate districts and schools, not individual students, and thus were "not valid for measuring individual performance."[42]

Walter Haney, a researcher at CSTEEP at Boston College, agreed.

> It is a prime example of a test that was developed for one purpose...and applied for a purpose that is totally inappropriate and unintended.... The North Carolina end-of-grade tests were designed to hold schools and districts accountable. There is considerable potential for people trying to use [a] national test for similar decisions without stopping to examine whether, in fact, the content parallels the local curriculum.[43]

The North Carolina tests do match state curriculum standards, however, and cover a -representative sample of it. Because the state uses the tests to evaluate districts and schools, individual students usually see only one-third of each subject-area exam; by sampling this way, the state can cut testing time and costs. Had Johnson County held students back for poor performance on a test that covered only one-third of the curriculum, that would have been unfair. Instead, the district put the three separate pieces of the exam together to form complete exams that covered the entire curriculum.[44]

A U.S. District Court judge last year rejected the plaintiffs' request for an injunction to prevent another year of student retention.[45] The plaintiffs later dropped the case.[46]

In spite of great advances in student achievement linked to the testing system in North Carolina, FairTest gives the state's system its lowest rating of 1 (on a scale of 1 to 5).

Richard Jaeger, a professor at the University of North Carolina who was chosen by the AERA to speak at their press conference on testing, criticized his state's testing program in general, but particularly its high-stakes, minimum-competency element. That test is geared toward a relatively low level of basic skills and students have several chances to pass, starting in 10th grade. They may not graduate from high school until they pass it.

Jaeger argued that the costs to society of denying students diplomas might be too high. "As a determinant of a student's life chances in American society, possessing a high school diploma is far more important than scoring well on a basic skills competency test." [47] He cited statistics showing that high school dropouts are more likely to have blighted lives and argued that "the use of such tests jeopardizes the future of those young people denied a high school diploma by limiting their employability, reducing their quality of life, and diminishing their opportunity to contribute to society through the productive applications of their abilities."[48] Jaeger also presented evidence purporting to show that meeting higher standards

and passing high-stakes tests do not improve students' economic prospects. He implied that if North Carolina just gave poorly performing students their diplomas with no impediments, the state would enjoy less crime, fewer out-of-wedlock births, and shorter welfare rolls.[49]

George Madaus of Boston College and CSTEEP and Lorrie Shepard of the University of Colorado and CRESST have also accused high-stakes tests of increasing the dropout rate.[50] Their evidence, however, is spotty. Most U.S. dropouts leave school when they reach the limit of the compulsory attendance law, not when they fail an exam.[51] When students in the large-scale *Indiana Youth Poll* explained why some dropped out, either disinterest in school or non-academic-related problems (such as pregnancy or family problems) were cited more than four times more often than academic failure.[52]

A careful examination of the dropout issue by Griffen and Heidorn, using data from Florida from the early 1990s (when a test similar to the one used in North Carolina was in place), found that

> [F]ailure on a [minimum-competency test] provided a statistically significant increase in the likelihood of leaving school, but only for students who were doing well academically. Students with poorer academic records did not appear to be affected by MCT [minimum-competency test] failure; similarly, minority students did not demonstrate an increased likelihood of leaving school as a result of failing an MCT.[53]

Speaking about the same high-stakes exit exam in Florida, the psychologist and attorney Barbara Lerner explained as follows:

> On the first few tries, 80 to 90 percent of Florida's students failed the test. But they were not crushed, as the experts predicted, and they did not give up and drop out in droves without diplomas. They kept trying, and their teachers did too, working hard to help them learn from failure and, ultimately, to master the skills they needed to graduate. By the fifth try, better than 90 percent of them did just that. They left school not just with a piece of paper, but with basic skills that prepared them better for life.[54]

In spite of great advances in student achievement linked to the testing system in North Carolina, however, FairTest gives the state's system its lowest rating of 1 (on a scale of 1 to 5).

> North Carolina's assessment program needs a complete overhaul. It relies far too heavily on multiple-choice tests, tests too often, and has a graduation exam. It should reduce the grades tested, drop the graduation requirement, ensure districts do not rely on the tests for grade promotion decisions, and implement a performance assessment system based on the state standards.[55]

Overall, North Carolina showed the most improvement of any state on NAEP in the 1990s.[56]

Case Study 4: SAT

The test attracting the loudest and most sustained opprobrium from critics over the years is the SAT, used by almost two-thirds of U.S. colleges in making admissions decisions.[57]

One of the primary sustaining causes of FairTest is its crusade to convince colleges to cease using SAT scores in admissions decisions. If one read only FairTest's literature, one might well conclude that the group's campaign against the SAT has been very successful.[58] According to FairTest, 240 colleges now have optional or limited SAT requirements.[59]

Those colleges that offer the possibility of admissions *sans* test scores may, however, require additional proof of ability, such as a graded writing sample or on-campus interview. Moreover, even if not required for admission, the absence of a test score may still bias an application negatively.[60]

Still, the SAT's impact is often overstated. The overwhelming majority of colleges are not selective, so a low SAT score will rarely keep a student out of college. Even at the most selective colleges, the SAT is seldom used alone by college admissions staff to make decisions. Typically, it is one of many factors that include a student's high school grade point average, extra-curricular activities, recommendations, essays, and so on.[61] When surveyed, however, admissions counselors rate the SAT score as a more reliable measure than these other indicators.[62]

Admissions counselors rate the SAT score as a more reliable measure than other indicators.

The primary argument of SAT critics pertains to the test's "predictive validity"; it only explains 6 to 8 percent of the variation in first-year college grades, after other predictive factors are accounted for.[63] If that's all the good it does, why bother with it, they ask?[64] As Haney of CSTEEP says:

> Which is more accurate? Does a person's height more accurately predict a person's weight? Or do national college entrance exams more accurately predict a student's success in college?
>
> The answer: Height is a better predictor of weight. And there might be some crude relationship between height and weight. But it ain't real good.[65]

To a college admissions counselor, however, 6 to 8 percent is a lot of predictive power, and the SAT only costs about $20.[66] It costs society about $25,000 to educate a high school student. For an incremental cost of 0.08 percent over the cost of a high school education, the SAT score provides a college admissions counselor a 16 percent increase in information over what is provided by a student's high school record.[67] The incremental cost-benefit ratio for the SAT is 194:1 over the high school record. The "break even" value of the SAT is over $3,900 per student; at $20, it's a bargain.

The SAT is a nationally *standardized* measure; a grade point average is not. One student can achieve a high grade point average by working extremely hard in difficult courses in a high school with exacting standards, while another can get by choosing easy courses at a high school with low standards.

Ultimately, the makers of the SAT do not determine its success; its customers do. Those customers are thousands of college admissions officers throughout the United States who are doing their best to select students they believe can handle the level of academic rigor at their institution.

College admissions officers are not deaf and blind. They hear and read the arguments against use of the SAT. Nor are they elitist conspirators opposed to fair admission policies. Moreover, they are not required to use the SAT

(or ACT). They use such tests because they believe, based on personal experience, that they are valuable — so valuable that they consider test scores to be the second most important criterion in making admissions decisions, higher than grade point averages or class ranks, and second only to grades and test scores from Advanced Placement courses (which relatively few students take), the only other nationally standardized measure of achievement commonly available to them.[68]

Appraising the Criticisms

The basic argument made by testing critics is that the use of high-stakes standardized tests is counterproductive. Instead of leading to stronger academic achievement, it actually interferes with good teaching and learning. Testing experts embrace a sort of domino theory. Pressure to produce higher scores leads teachers to focus on material that will be covered by the tests and to exclude everything else.[69] The curriculum is thereby narrowed, which means that some subjects are ignored. Within those that are taught, lower-order thinking skills are emphasized since these are what the tests tap. As a result of teachers teaching to the tests, subsequent test scores are inflated while real learning suffers.

In addition to the alleged harms of 1) test score inflation, 2) curriculum narrowing, 3) emphasis on lower-order thinking, and 4) declining achievement, testing experts add a quartet of other arguments against testing — that: 5) standardized tests hurt minorities and women, 6) tests are too costly, 7) other countries don't test nearly as much as we do, and 8) parents, teachers, and students in this country are all opposed to testing. These eight claims will be examined in detail in the section that follows, and a rebuttal will be offered to each.

What testing experts particularly do not like are high-stakes, multiple-choice, external tests. They excoriate these tests with bad-sounding words (e.g., "lower-order thinking," "factory model of education," "uncreative," "rote recall," and so on), but the terms are seldom well explained. The root of most of the objections can be traced to the dominant worldview of testing experts (and many other educators).

The education philosophy driving many of these criticisms is constructivism, the view that every student and teacher constructs his or her own meanings from classroom activities, books, etc. Hence no construction is wrong or bad. We all know that there is often more than one way to get to a right answer. We all think differently, using different combinations of several different kinds of intelligence. Moreover, we all know that a student can process much of a problem well but still get the "wrong" answer in the end because of a fairly minor error, such as misplacing a decimal point.

The education philosophy driving many of these criticisms is constructivism, the view that every student and teacher constructs his or her own meanings.

As test critic (and constructivist) Mary Lee Smith of CRESST and Arizona State University describes it:

> Constructivist theory assumes that students construct their own knowledge (rather than passively receiving knowledge transmitted by school) out of intentional transactions with materials, teachers, and other pupils. Learning is more likely to happen when students can choose and become actively engaged in the tasks and materials, and when they can make their own connections across subject matter on tasks that are authentic and organized around themes. According to this theory, literacy is whole, embodying reading authentic texts and writing as a way of unifying all the subjects. For

example, to be literate is to be able to explain the reasoning one uses to discover and solve math problems. Explicit in constructivist theory is the rejection of the pedagogy of worksheets and the exclusive reliance on phonics, spelling out of context, computation, isolated subject matter and the like.[70]

Constructivists oppose school practices that they think "fix" behavior. They see standardizing curricula and instructional practice as restricting teacher behavior and multiple-choice standardized tests as shackling student responses to problems.

For constructivists, the more open-ended the assessment the better, and portfolios are the most open-ended of all. They involve no standardized, mandated, pre-set responses and not necessarily even a standardized question to impede any student's unique understanding of the problem, creative solution, and personal construction of the work.[71] This constructivist worldview will be seen to underlie most of the arguments marshaled by testing experts against testing.[72]

1) Test Score Inflation

An initial set of harms ascribed to standardized testing falls under the rubric of "teaching to the test." A CRESST paper entitled "The Effects of High-Stakes Testing on Achievement," by Daniel Koretz, Lorrie Shepard, and others, purports to demonstrate that high-stakes tests in fact cause teaching to the test.[73] The researchers compared student performance in math and reading from one commercial test given under high-stakes conditions in one school district to student performance on a different commercial test with no stakes. Student performance on the high-stakes test improved over time, according to the researchers, as the teachers adapted their instruction to the curriculum implicit in the test. Student performance on the other test, administered solely for the purpose of the study, did not improve over time. The difference in student performance between the two tests is offered by the CRESST researchers as proof that high-stakes tests "narrow the curriculum" and induce "teaching to the test." Test critics would describe the second set of scores on the high-stakes test as artificially inflated, "polluted," or "corrupted."

The idea behind score inflation is that, as teachers become more familiar with test content, they spend more time teaching that test content and less time teaching other material. So, over time, as familiarity grows, scores climb on the test while real learning suffers.

In the early 1980s, a West Virginia physician named John J. Cannell investigated a statistical anomaly that he had discovered: statewide average scores for students on some widely used test batteries were above the national average in every state in which they were given.[74] It was dubbed the "Lake Wobegon Effect" after the fictional community where "all the children are above average."

Response:

The Lake Wobegon anomaly might have been caused — observed Cannell and a group of test experts — by a number of factors, including schools reusing old tests year after year and growing familiar with their specific content, and test publishers waiting years before "renorming" the reference scales. Other factors could have included the "non-representativeness" of the norming samples;[75] the choice by school districts of the one test, from among various test versions, that most closely aligned with their curriculum and on which their pupils would likely perform best; and the fact that student achievement really was improving throughout the 1980s, as verified by

independent testing, such as that for SAT, ACT, and NAEP exams. There may also have been some statistical anomalies in Dr. Cannell's calculations.[76]

The Lake Wobegon controversy led to calls for more state government control over test content and administration and less local discretion. In most states, those calls were answered. Today most school districts are aware of the problem of test score inflation, and do not use tests with the exact same questions year after year. Many jurisdictions now either use tests that are custom-built to their state standards and curricula or that are adapted from commercial publishers' test item banks. A simple way to prevent score inflation is to use different tests or test forms from year to year without announcing in advance which one will be used. Indeed, most of the likeliest sources of the Lake Wobegon effect are fairly easily avoided.[77]

The larger argument about teaching to the test has several components, which will now be addressed.

Educators and policymakers are signalling that, in a world of tough choices among competing priorities, some subjects must in fact take a backseat to others.

2) Curriculum Narrowing

We might suppose that preparing youngsters to do well on tests would find favor with testing experts, yet many of them condemn all forms of "teaching to the test." These arguments tend to come in several forms. One is that valuable subjects that are not tested (e.g., art and music, maybe even social studies or science) will be ignored or slighted by test-obsessed teachers and school systems. Lorrie Shepard of CRESST and the University of Colorado has asserted:

> Although critics may originally have feared that testing would take instructional time away from 'frills,' such as art and citizenship, the evidence now shows that social studies and science are neglected because of the importance of raising test scores in the basic skills. [78]

A variation on this theme holds that, even within a subject that *is* taught, content coverage will be narrowed (or curricular depth made shallow) in order to conform to the content or style of the test.

Response:

There is only so much instructional time available, and choices must be made as to how it is used. (Of course, some new school designs also extend the school day or year to ameliorate this problem.) If non-tested subjects are being dropped, either they, too, should be tested or, perhaps, educators and policymakers are signalling that, in a world of tough choices among competing priorities, some subjects must in fact take a backseat to others. A state or school system could easily add high-stakes tests in art, music, language, and civics, or any other subjects. Attaching high-stakes to tests in some subjects and not others would be interpreted by most as a signal that the former subjects are considered to be more important. Perhaps that's actually true. Especially where students are sorely deficient in basic skills and need extra instruction in them, it is likely that few parents would object to such priorities. Survey results show clearly that the public wants students to master the basics skills first, before they go on to explore the rest of the possible curriculum.[79] If that means spending more time on "the basics," so

be it. As for subject content being narrowed or made shallow in anticipation of a test, a better response than eliminating the test might be to replace it with one that probes deeper or more broadly.

3a) Emphasis on Lower-Order Thinking (in Instruction)

Lorrie Shepard has also asserted:

> High-stakes testing misdirects instruction even for the basic skills. Under pressure, classroom instruction is increasingly dominated by tasks that resemble tests....Even in the early grades, students practice finding mistakes rather than do real writing, and they learn to guess by eliminating wrong answers...
>
> In an extensive 18-month observational study, for example, Mary Lee Smith and her colleagues found that, because of external tests, elementary teachers had given up on reading real books, writing, and undertaking long-term projects and were filling all available time with word recognition; recognition of errors in spelling, language usage, and punctuation; and arithmetic operations...[80]

Testing critics cannot abide the notion that preparing students for a standardized test could be considered instruction, because it is not the kind of instruction that they favor.

Response:

Critics like Smith and Shepard say that intensive instruction in basic skills denies the slow students instruction in the "the neat stuff" in favor of "lower-order thinking."[81] They argue that time for preparing students for high-stakes tests reduces "ordinary instruction." They cannot abide the notion that preparing students for a standardized test could be considered instruction, because it is not the kind of instruction that they favor.[82]

Instruction to which teachers may resort to help students improve their scores on standardized tests tends not to be constructivist. It is the type of instruction, however, that teachers feel works best for knowledge and skill acquisition. Teachers in high-stakes testing situations do not deliberately use instructional practices that impede learning: they use those that they find to be most successful.

These testing critics idealize the concept of teachers as individual craftspersons, responding to the unique needs of their unique pupils in unique ways with "creative and innovative" curriculum and instruction.[83] But the most difficult jobs in the world are those that must be created anew every day without any consistent structure, and performed in isolation without collaboration or advice. In Public Agenda's research, "teachers routinely complained that teaching is an isolated and isolating experience."[84]

By contrast, teachers in other countries are commonly held to more narrowly prescribed curricula and teaching methods. Furthermore, because their curricula and instructional methods are standardized, they can more easily and productively work together and learn from each other. They seem not to suffer from a loss of "creativity and innovation"; indeed, when adjusted for a country's wealth, teachers in other nations are commonly paid more, and usually have greater prestige.[85]

Critics like Shepard and Smith cannot accept that some teachers may *want* to conform to sys-

temwide standards for curriculum, instruction, and testing. Standardization brings the security, convenience, camaraderie, and common professional development that accompany a shared work experience.[86]

3b) Emphasis on Lower-Order Thinking (in Test Content)

One CSTEEP study, funded by the National Science Foundation, analyzed whether several widely used commercial (and mostly multiple-choice) tests required "higher-" or "lower-order" thinking. A press account boasted, "In the most comprehensive study of its kind yet conducted, researchers from Boston College have found evidence to confirm the widespread view that standardized and textbook tests emphasize low-level thinking and knowledge and that they exert a profound, mostly negative effect on classroom interaction."[87]

Researcher Maryellen Harmon told a reporter, "None of [the test content] calls for high-order thinking that requires that they go in-depth into the concept, that they use math skills in nonconventional contexts, or pull together concepts from geometry and algebra."[88] Project Director George Madaus was quoted as saying that the findings present a "depressing picture....If this doesn't change, an inordinate amount of time, attention, and preparation will be given to the wrong domains in math and science, domains that are not reflecting the outcomes we want."[89]

Response:

Many readers would be astonished, as I still am, by the vehemence of some critics' ire toward something as seemingly dull and innocuous as item response format. Yet many of the accusations leveled at multiple-choice items have little substance. For example, you can often find in CSTEEP and FairTest publications assertions that multiple-choice items demand only factual recall and "lower-order" thinking, while "performance-based" test do neither. Both claims are without merit. It is the structure of the *question*, not the response format, that determines the character of the cognitive processing necessary to reach a correct answer.

Test items can be banal and simplistic or intricately complex and, either way, their response format can be multiple-choice or open-ended. There is no necessary correlation between the difficulty of a problem and its response format. Even huge, integrative tasks that require fifty minutes to classify, assemble, organize, calculate, and analyze can, in the end, present the test-taker with a multiple-choice response format. Just because the answer to the question is among those provided, it is not necessarily easy or obvious how to get from the question to the answer.

Anyone who still thinks that multiple-choice items demand only factual recall should take a trip to the bookstore and look at some SAT or ACT help books. I purchased a copy of the Cliffs Notes SAT prep book and randomly picked a page. It was in the math section and four items are posed. Here's one: "What is the maximum number of milk cartons, each 2" wide by 3" long by 4" tall, that can fit into a cardboard box with inside dimensions of 16" wide by 9" long by 8" tall?" Five possible answers are provided, but the correct one, obviously, cannot just be "recalled." Calculations are required. My solution was to calculate the volume, in cubic inches, of a carton and the box, by multiplying the three dimensions in each case, and then to divide the former volume into the latter. I used pen and paper for two of the calculations and figured the other in my head. Interestingly, the Cliffs Notes book solves the problem graphically, by sketching a three-dimensional box and subdividing it with line segments along each dimension.[90]

Indeed, much of the Cliffs Notes book is devoted to convincing the student that there is usually more than one way to "construct" a response to a problem. The book contains sections that illustrate different approaches to solving similar problems. It's a very "constructivist" book; any student following its advice would make ample use — in taking the SAT — of pen, paper, calculator, formulas, diagrams, sketches, lateral thinking, meta-analysis, and other devices that constructivists hold dear. Students armed with multiple methods for solving problems, of course, will hit more correct answers on the SAT than students with fewer methods, other factors held equal. So, higher SAT scores should be taken as evidence of more "higher-order" thinking.

All the optical scanner will read in the end, however, is a sheet of circles, some filled in with pencil and others not. Moreover, all the computer will score in the end is the number of correct filled-in circles. The calculations, sketches, and diagrams the student used to solve the problems are left behind in the test booklet, on scratch paper, or in the student's head. Just because the optical scanner and computer do not see the "process" evidence of "higher-order" thinking, however, does not mean it did not take place.[91] That is, however, what the critics assume.

The most essential point for the critics in applying the "lower-order" label to multiple-choice and the "higher-order" label to performance tests, seems to be that, with open-ended questions, a student *shows* (some of) her work in the test-response booklet itself, and a scorer can see (some of) how the test-taker approached the problem through the exposition of the answer. This is undoubtedly helpful to teachers but far less necessary for purposes of informing parents, policymakers, admissions counselors, etc.

CSTEEP's researchers defined "higher-order thinking skills" as having three characteristics: problem solving (the ability to formulate problems, use a variety of problem solving strategies in non-routine situations, verify and interpret results); reasoning (the ability to infer, analyze, use logic); and communicating (the ability to speak, write, depict, or demonstrate ideas in prose, graphs, models, equations, and to describe, explain, or argue a position).[92]

The first two characteristics are typically found in definitions of "higher-order thinking."[93] The third was added by CSTEEP for the purposes of their study. The CSTEEP researchers crafted a definition of "higher-order thinking" that multiple-choice tests would invariably fail. According to them, one is not "communicating" when filling in a bubble for a multiple-choice item, no matter what mental or physical processes may have been used in getting a student to that point, but one is "communicating" when writing a textual response to an open-ended prompt.[94] If the scorer cannot see the work, the work does not exist. Obviously, if one can define "higher- order thinking skills" any way one wishes, as these CSTEEP researchers did, one can define any type of testing one dislikes as embodying only "lower-order thinking."

Even defined without the "communicating" component, is "higher-order thinking" always a superior form of thinking, as testing critics imply? Consider the type of thinking surgeons do. They are highly paid and well-respected professionals. Their study, however, consists of a considerable amount of rote memorization, and their work entails a considerable amount of routine and factual recall (all "lower-order thinking"). Moreover, the medical college admissions test is largely multiple-choice, and tests administered during medical training largely elicit the recall of discrete facts.

If you were about to go under the knife, which kind of surgeon would you want? Perhaps one who used only "higher-order thinking," only "creative and innovative" techniques, and "constructed her own meaning" from every operation she performed?

Or, would you prefer a surgeon who had passed her "lower-order thinking" exams — on the difference, say, between a spleen and a kidney — and used tried-and-true methods with a history of success: methods that other surgeons had used successfully? Certainly, there would be some situations where one could benefit from an innovative surgeon. If *no aspect whatsoever* of the study or practice of surgery were standardized, however, there would be nothing to teach in medical school, and your regular barber or beautician would be as well qualified to "creatively" excise your appendix as anyone else. Ideally, most of us would want a surgeon who possesses both "lower" and "higher" abilities.[95]

The surgery analogy also addresses another of the testing critics' arguments. They say that multiple-choice tests limit students to the "one correct answer" when there may really be more than one valid answer and more than one way to get to each. Moreover, they say, students should not get an entire exercise counted wrong if they analyze most of the problem correctly, but make one careless error.

Students from states, provinces, or countries with medium or high-stakes testing programs score better on neutral, common tests and earn higher salaries after graduation than do their counterparts from states, provinces, or countries with no- or low-stakes tests.

Most of us would sympathize with this sentiment, but we should remember that there are countless examples in real life where there *is* just one right answer or where one careless error can have devastating consequences — in brain surgery, for example.

4) Declining Achievement

Testing experts claim that high-stakes tests actually interfere with learning and student achievement in states that use them. In "High Stakes Tests Do Not Improve Student Learning," FairTest asserted that states with high-stakes graduation exams tend to score lower on NAEP. According to FairTest, this "contradicts the... common assumption of standards and tests-based school reform...that high-stakes testing...will produce improved learning outcomes."[96]

The FairTest solution is to restrict testing to occasional no-stakes monitoring with samples of students using the types of response formats that FairTest favors (no multiple-choice!). Scores on "portfolios" of each student's best work would track individual student progress.[97] Indeed, the only state testing program to garner the highest rating from FairTest was Vermont, which has a statewide portfolio program and no high-stakes or multiple-choice standardized testing.[98]

Response:

The claim that high-stakes tests inhibit learning is a weak argument supported by dubious research. The FairTest report provides a good example of just how simplistic that research can be. FairTest argues that states with high-stakes minimum-competency test graduation requirements tend to have lower average test scores on NAEP. They make no effort, however, to control for other factors that influence test performance, and the relationship between cause and effect is just assumed to run in the direction FairTest wants.[99] Most honest observers would assume the direction of cause and effect to be just the opposite — poorly performing states initiate high-stakes testing programs in an effort to

improve academic performance while high performing states do not feel the need to.

The work of the Cornell labor economist John Bishop does not get the press attention bestowed on FairTest. Yet in a series of solid studies conducted over a decade, Bishop has shown that, when other factors that influence academic achievement are controlled for, students from states, provinces, or countries with medium or high-stakes testing programs score better on neutral, common tests and earn higher salaries after graduation than do their counterparts from states, provinces, or countries with no- or low-stakes tests.[100]

Bishop recently turned his attention to the very same relationship that FairTest studied, only he looked at it in depth. He and his colleagues used individual-level data from the National Education Longitudinal Study (NELS:88), which began in 1988, and High School and Beyond (HSB), another longitudinal study that ran from 1980 to 1992. They controlled for socioeconomic status, grades, and other important factors, while comparing the earnings of graduates from "minimum-competency" testing states to those from non-testing states.[101] They found that test-taking students earned an average of 3 percent to 5 percent more per hour than their counterparts from schools with no minimum-competency tests. And the differences were greater for women, with as much as 6 percent higher earnings for those who had taken the tests. Other evidence of the success of high-stakes state testing programs continues to surface.[102]

5) Testing Hurts Women and Minorities

As mentioned in the case study of high-stakes testing in Texas, the NAACP and the Mexican-American Legal Defense Fund both argued that the Texas Assessment of Academic Skills was biased against minorities.

The brunt of FairTest's attack on the SAT involves alleged bias as well. The argument is straightforward: on average, girls score worse on the SAT than boys, despite getting better grades in school. Therefore the SAT is gender-biased. Blacks and Hispanics score lower than whites. Therefore the SAT is race-biased.[103] FairTest argues that this bias against minorities depresses minority college admissions.

Response:

After investigating why girls score worse on the SAT than boys, despite getting better grades in school, the Educational Testing Service (ETS), the SAT's developer, concluded that the gender difference in SAT scores was almost entirely explained by high school course selection (e.g., girls took fewer math and science courses than boys, and so got lower SAT math scores).[104] FairTest called the ETS explanation a "smokescreen."[105] Yet similar evidence is available for blacks and Hispanics: almost all the SAT math score differences between them and their white counterparts disappear when they take as much algebra and geometry in high school as white students do.[106]

The charge that the use of SATs in college admissions artificially depresses minority admissions is also misguided. As David W. Murray writes in "The War on Testing":

> Nor is it even clear that relying more exclusively on grades would bump up the enrollment numbers of blacks and Hispanics, as many seem to think. While it is true that more minority students would thereby become eligible for admission, so would other students whose grade point averages (GPAs) outstripped their test scores. A state commission in California, considering the adoption of such a scheme, discovered that in order to pick

students from this larger pool for the limited number of places in the state university system, the schools would have to raise their GPA cut-off point. As a result, the percentage of eligible Hispanics would have remained the same, and black eligibility actually would have dropped.[107]

There is a double sadness to the focus of some minority spokespersons on the messenger instead of the message. Black and Hispanic students in the United States generally receive an education inferior to that received by white students. This is a shame and a disgrace. By blaming standardized tests instead of the schools that are responsible for their students' poor achievement, however, these advocacy groups waste efforts that would be better expended reforming bad schools.

A recent Public Agenda survey of parents on education issues pertaining to race implies that the NAACP actions in Texas and other states against high-stakes standardized testing may not even reflect what most African Americans want. "Most African-American parents do not think standardized tests are culturally biased," reports Public Agenda, "and very few want race to be a factor when choosing the best teachers for their children..."[108] When asked why, on average, black students don't do as well as whites on standardized achievement tests, only 28 percent say it is mostly because "the tests are culturally biased against black students." Forty-four percent of black parents say "the tests measure real differences in educational achievement," and 18 percent say the reason for this difference is a failure of expectations.[109]

By blaming standardized tests instead of the schools that are responsible for their students' poor achievement, advocacy groups waste efforts that would be better expended reforming bad schools.

6) Excessive Cost

Some experts have criticized standardized tests, particularly those which include performance-based items, as too costly. Daniel Koretz, of CRESST, appeared before a congressional committee to argue against a national testing proposal, and stressed cost as a major negative. He claimed that the costs of performance-based national tests would be well over $100 per student, perhaps as high as $325 per student.

Another study of the extent and cost of testing by Walter Haney and George Madaus of CSTEEP calculated a "high" estimate of $22.7 billion spent on standardized testing in a year.[110] U.S. schools, the CSTEEP report claimed, suffered from "too much standardized testing" that amounted to "a complete and utter waste of resources."[111] Their estimate breaks down to about $575 per student per year for standardized testing.

A recent CRESST report by Larry Picus which counted cost components in much the same way as the CSTEEP study estimated costs of a certain state test at between $848 and $1,792 per student tested.[112]

Response:

Several years ago, the U.S. General Accounting Office (GAO) surveyed a national sample of state and local testing directors and administrators to appraise the costs of then-current statewide and districtwide tests, many of which contained some performance-based items. GAO found that eleven state tests ranging from 20% to 100% performance-based cost an average of $33 per student, including

the salary time of teachers and other staff engaged in test-related activity, as well as the purchase of test materials and services. GAO estimated that slightly over $500 million was spent by U.S. school systems on systemwide testing in a year, or about 0.2% of all spending on elementary and secondary schools.[113]

The GAO estimate of $33 per student contrasts with CRESST and CSTEEP estimates of $575 to $1792. The GAO estimate of about $500 million for the total national cost of systemwide testing contrasts with a CSTEEP estimate 45 times higher.

Testing critics estimate standardized tests' costs so much higher because they count the costs of any activities "related to" a test as costs *of* a test. In the CRESST study of Kentucky's performance-based testing program, for example, teachers were asked to count the number of hours they spent "preparing materials related to the assessment program for classroom use." In an instructional program with the intention of unifying all instruction and assessment into a "seamless" web, where the curriculum and the test mutually determine each other, *all* instruction throughout the entire school year will be "related to" the assessment.

Testing critics estimate standardized tests' costs so much higher because they count the costs of any activities "related to" a test as costs of a test.

Furthermore, the Kentucky Instructional Results Information System (KIRIS) is a comprehensive program that includes changes in curriculum, instruction, and evaluation. Assessment is just one component. All the changes were implemented at the same time, and some survey respondents could consider any or all KIRIS costs as "related to" the assessment. Given the manner in which it posed its questions, CRESST cannot discern which are costs of the test and which are costs of other parts of the KIRIS program.

The CSTEEP study counted even more cost items, such as student time. Walter Haney and the other CSTEEP researchers assumed that there is no instructional value whatsoever to student time preparing for or taking a test. (I would guess that students probably learn *more* while preparing for or taking a test.) Then they calculated the present discounted value of that "lost" instructional time against future earnings, assuming all future earnings to be the direct outcome of school instruction. The CSTEEP researchers also counted building overhead (maintenance and capital costs) for the time spent testing, even though those costs are constant (i.e., "sunk") and not affected by the existence of a test. In sum, CSTEEP counts any and all costs incurred simultaneously with tests, not just those caused by testing.

7) Other Countries Don't Test as Much

CSTEEP's Madaus has claimed that "American students [were] already the most heavily tested in the world."[114] He has also asserted that the trend in other developed countries is toward less standardized testing. He reasoned that other countries are dropping large-scale external tests because they no longer need them as selection devices since places in upper secondary programs are being made available to everyone and access to higher education programs has widened. Thus, he argued, a worldwide trend toward less external testing could be found at all levels of education "even at the postsecondary level" and it was unidirectional — large-scale, external tests were being "abolished."[115]

Response:

Are U.S. students the "most heavily tested in the world"? No. U.S. students actually spend less time taking high-stakes standardized tests than do students in other developed countries. A 1991 survey for the Organisation for Economic Co-operation and Development (OECD) revealed that "U.S. students face fewer hours and fewer numbers of high-stakes standardized tests than their counterparts in every one of the 13 other countries and states participating in the survey and fewer hours of state-mandated tests than their counterparts in 12 of the 13 other countries and states."[116]

What of Madaus's assertion of a trend toward less standardized testing in other countries?[117]

The primary trend appears to be toward more testing, with a variety of new test types used for a variety of purposes. In a study I conducted, I found twenty-seven countries and provinces had increased or planned to increase testing over the period 1974–1999, while only two decreased it. Altogether, fifty-three tests were added and only three dropped.[118]

8) All Those Who Really Care About Children Oppose Testing

Testing experts who hate testing imply that they speak on behalf of teachers and students, defending them against politicians, mean-spirited conservatives, and the greedy testing industry.[119] The critics claim that those who care about teachers and students see testing for what it really is, and oppose it.

Regarding teachers, for example, Robert Stake, who was chosen by the AERA to speak at their press conference on testing, said "[teachers] have essentially no confidence in testing as the basis of the reform of schooling in America."[120]

The laundry list of costs attributed to *students* from the use of standardized tests ranged from a change in instruction away from the "neat stuff" in the curriculum toward "lower-order thinking," to an increase in grade retention and dropout rates from the use of standardized tests in high-stakes situations. A CRESST study by Mary Lee Smith, referred to at the conference on the "unintended consequences of external testing," claimed to find "stress, frustration, burnout, fatigue, physical illness, misbehavior and fighting, and psychological distress" among the effects of testing on young students.[121]

Majorities of the general public favor more testing and higher stakes in testing.

Response

To learn the true attitudes toward testing among teachers, students, parents, and the public, I attempted to gather all relevant U.S. poll and survey items on student testing by collecting many surveys myself and searching the Roper Center archives. I discovered 200 items from seventy-five surveys over three decades.[122]

The results are fairly decisive. Majorities of the general public favor more testing and higher stakes in testing. The majorities have been large, often very large, and fairly consistent over the years, across polls and surveys, and even across respondent groups. Parents, students, employers, state education administrators, and even teachers (who exhibit more guarded opinions and sometimes fear being blamed if their students score badly on tests) consistently favor more student testing and higher stakes.

Twenty-seven polls taken between 1970 and the present asked specific respondents whether they thought education would improve if there

were higher (student) stakes in school testing. In twenty-six of the twenty-seven polls the answer was yes, in most cases by huge margins.

Which was the twenty-seventh study, the one claiming that respondents want lower stakes in student testing? It was a survey conducted by CSTEEP and funded by the National Science Foundation.[123] Its contrary conclusions may have a lot to do with its convoluted design. First, respondents were chosen selectively from urban, "high-minority" public school districts. High school teachers in the sample were limited to those with classes of "average and below average" students.[124] Moreover, the specific interview question that elicited opinions on the effects of mandated tests was, in my judgment, biased in a way that would generate negative answers. The question was: "Do you have any particular concerns or opinions about any of these standardized tests?" "Concerns" doesn't equal "criticisms" in meaning but, in this context, it's pretty close.[125] Then the CSTEEP researchers classified as "negative" responses those that others might classify as neutral or positive. For example, if a teacher said that her students "didn't test well," it was interpreted by the researchers as a "major source of invalidity" and a "negative" comment, even though students can test poorly for dozens of reasons, including not studying or not paying attention in class.[126]

The debate seems unbalanced only because one side is often missing from it, that of the pro-testing advocates who cannot speak out.

Do these CSTEEP researchers and other testing opponents at least represent other "education establishment" organizations in opposing high-stakes standardized testing?

Far from it. The National Association of State Boards of Education has come out strongly in favor of a greater use of high-stakes standardized testing.[127] So have state superintendents and governors. The American Federation of Teachers (AFT) has been the nation's most forceful and vocal advocate for greater use of high-stakes standardized student testing. The other large teachers' union, the National Education Association, has recently moved closer to the AFT position.

Nationwide polls of teachers conducted over three decades by the Carnegie Foundation for the Advancement of Teaching, the Metropolitan Life Insurance Co., the American Federation of Teachers, the *Phi Delta Kappan* magazine, and Public Agenda show strong teacher support for high-stakes standardized tests.[128]

Despite this widespread support for testing, press coverage of testing issues often seems one-sided *against* testing. It typically features a FairTest spokesperson as the anti-testing alternative to some sincere, beleaguered state or local testing director just trying to do her job.

I telephoned a few newspaper reporters to try to understand why their stories on testing were set up this way. They replied that they do not know of any advocacy group on the other side of the issue that could balance FairTest's perspective. They added that FairTest is also very reliable: they keep up with the issues and they return telephone calls promptly. In his review of SAT critiques, Gregory Cizek expresses disappointment that "the measurement profession has made no corresponding, popular, accessible, public defense of its mission or of testing."[129]

While one sees only a handful of education-researcher experts speaking out in favor of high-stakes standardized tests, there are in fact hundreds of qualified testing experts working for national, state, or local agencies (not to mention the experts working for organizations that develop tests under contract to governments) who are legally and ethically restricted

from expressing their views regarding testing policy. The debate seems unbalanced only because one side is often missing from it, that of the pro-testing advocates who cannot speak out.

For a reporter who arrives at the office in the morning with no story and who cannot leave in the evening without one, FairTest is a godsend. The majority that favors testing has no comparable voice.

Testing in Perspective

The fact that tests and test results can be misused is beyond dispute. Human beings are responsible for administering them and interpreting their results, and humans are imperfect creatures. There is also no denying that tests are imperfect measurement devices. If the items in the anti-testing canon were also beyond dispute, one might well be disposed to give up on high-stakes standardized testing. But that would be an enormous mistake.

The critics would have us believe that all problems with high-stakes and standardized testing must always be with us, i.e., that nothing can be changed or improved. They're wrong. Some of the alleged problems — that they hurt learning and are expensive, for instance — are really not problems at all, as shown above. Other problems apply equally to the alternatives to testing. Still others are solvable and are being or have been solved by state, local, or national testing directors.

Probably the single most important recent innovation in relation to the quality and fairness of testing in the United States has been the addition of managerial and technical expertise in state education agencies. At that level, it is possible to retain an adequate group of technically proficient testing experts, adept at screening, evaluating, administering, and interpreting tests, who are not "controlled by commercial publishers" or naïve about test results. They, along with governors and legislatures, are currently calling the shots in standardized testing. Some of the most important decisions affecting the design and content of the tests, the character of the testing industry, and the nature of its work, are today being made by state testing directors.

They can, for example, deploy a number of relatively simple solutions to the problems of score inflation, curricular compression and teaching to the test, including not revealing the contents of tests beforehand; not using the same test twice; adding items on the test that sample broadly from the whole domain of the curriculum tested; requiring that non-tested subjects also get taught (or testing them, too); and maintaining strict precautions against cheating during test administrations.

In Texas and North Carolina, they do something else about score inflation: they keep raising the bar. As instruction and learning improve and scores rise, they boost their standards.[130] Their students' dramatic improvements on the independent NAEP offer evidence that the achievement gains are real, not a result of score inflation caused by "narrowing the curriculum" and "teaching to the test."

In Texas, some Canadian provinces, and some other states and countries, they use "blended" or "moderated" scores for high-stakes decisions. The "blends" combine test scores with other measures, such as classroom grades and attendance records, so that instructional efforts will not focus exclusively on the standardized test and so that high-stakes decisions will not be based solely on single, or even multiple, attempts at passing a test.

One final argument against testing — that using test results to evaluate schools leads to unfair comparisons between rich districts with highly educated parents and poor districts with less-well educated parents — can also be effectively challenged. There are at least two solutions to this problem. The first is to set targets for schools based on their own past performance. The second is to calculate "value-added" test scores, as North Carolina does. This method estimates how much value a school adds to the level of achievement that would have been predicted (given the background and prior attainment of students),

and then adjusts a school's or district's test scores accordingly. Like any other system, "value-added" scoring can be abused. There's a particular danger that it can be used to let poorly managed school systems with lots of poor and minority children off the hook. "Value-added" scores can also be tricky to calculate. But many able and earnest analysts throughout the country are striving to make value-added systems work.

While some of the "problems" with standardized testing turn out not to be problems, and others turn out to be solvable, a third set of problems is inherent and inevitable — but similar problems are also present in the alternatives to standardized tests.

The critics unfairly compare high-stakes standardized testing to their own notion of perfection. Administration of high-stakes tests will never be perfect. There will always be some teachers and pupils who cheat. There will always be some students who are better prepared to take a test than others, and so on.

Without standardized tests, no one outside the classroom can reliably gauge student progress.

Perfection, however, is not a reasonable standard of comparison for standardized testing. Too often, the alternative is a system of social promotion with many levels of (nominally) the same subject matter being taught, ranging from classes for self-motivated kids to those for youngsters who quit trying years before, and whom the system has ignored ever since.[131] Too often, the result is a system that graduates functional illiterates.

If *none* of the curriculum is tested, we cannot know if any of it has been learned. Without standardized tests, no one outside the classroom can reliably gauge student progress. No district or state superintendent. No governor. No taxpayer. No parent. No student. Each has to accept whatever the teacher says and, without standardized tests, no teacher has any point of comparison, either.

While it is unfair to test what has not been taught, no such claim can be made about testing what has been taught. And if what is tested is the curriculum, then attacks on "teaching to the test" seem silly, since teachers are teaching what they should be teaching.

Eliminating *high-stakes* standardized testing would necessarily increase our reliance on teacher grading and testing. Are teacher evaluations free from all the complaints of the anti-testing canon? Not exactly. Individual teachers can also narrow the curriculum to that which they prefer. Grades are susceptible to inflation with ordinary teachers, as students get to know a teacher better and learn his idiosyncrasies. A teacher's grades and test scores are far more likely to be idiosyncratic and non-generalizable than any standardized tests'. [132]

Moreover, teacher-made tests are not necessarily any better supplied with "higher-order thinking" than are standardized tests. Yet many test critics would bar all high-stakes standardized tests and have us rely solely on teacher evaluations of student performance. How reliable are those evaluations? Not very. There are a number of problems with teacher evaluations, according to research on the topic. Teachers tend to consider "nearly everything" when assigning marks, including student class participation, perceived effort, progress over the period of the course, and comportment, according to Gregory Cizek. Actual achievement vis-à-vis the subject matter is just one factor. Indeed, many teachers express a clear preference for non-cognitive outcomes such as "group interaction, effort, and participation" as more important than averaging tests and quiz scores.[133] It's not so much what you know, it's how you act in class. Being enthusiastic and

group-oriented gets you into the audience for TV game shows and, apparently, also gets you better grades in school.

One study of teacher grading practices discovered that 66 percent of teachers feel that their perception of a student's ability should be taken into consideration in awarding the final grade.[134] Parents of students who assume that their children's grades represent subject matter mastery might well be surprised.

Conclusion: Two Views of Testing and Learning

There is perhaps no more concise exposition of the general philosophy undergirding opposition to standardized testing among education experts than that revealed in the Public Agenda survey of education school professors, *Different Drummers*.[135] Among the reasons most dislike standardized tests are their preferences for "process over content"; "facilitating learning" rather than teaching; and "partnership and collaboration" over imparting knowledge.[136]

A large majority of education school professors surveyed felt that it was more important that "kids struggle with the process of trying to find the right answers" (86 percent) than that "kids end up knowing the right answers to the questions or problems" (12 percent): "It is the process, not the content, of learning that most engages the passion and energy of teacher educators. If students learn how to learn, the content will naturally follow."[137]

The role of teachers, then, in this education worldview, should be that of "facilitator," not "sage on the stage." When asked which statement was "closer to their own philosophy of the role of teachers," 92 percent of the education professors agreed that "Teachers should see themselves as facilitators of learning who enable their students to learn on their own." Only 7 percent felt that "Teachers should see themselves as conveyors of knowledge who enlighten their students with what they know."[138]

Many testing experts share an ideological orientation that makes any type of standardized test impossible to swallow.

The constructivist criticism of any teaching or testing that fixes the manner of solving a problem and penalizes students for careless or "minor" errors is not shared by the public or even by students. In *Getting By*, Public Agenda reported that 79 percent of teens say "most students would learn more if their schools routinely assured that kids were on time and completed their homework.... [Sixty-one percent said] having their class work checked regularly and being forced to redo it until it was correct would get them to learn a lot more. When interviewed in focus groups, teens often remembered "tough" teachers with fondness: "I had a math teacher [who] was like a drill sergeant. She was nice but she was really strict. Now I don't have her this year, and looking back, I learned so much."[139]

In the real world, testing will continue. Testing experts have much to contribute to efforts to ensure that testing is done well. Unfortunately many of them share an ideological orientation that makes any type of standardized test impossible to swallow. Until these experts reexamine their most fundamental beliefs about teaching and learning, all the hard work of improving standardized tests will have to be done without them.

Notes

[1] The author would like to thank the Thomas B. Fordham Foundation for its support; Chester E. Finn, Jr., Diane Ravitch, Marci Kanstoroom, Steve Ferrara, and Mike Petrilli for their voluminous and very helpful edits; Scott Oppler, Chris Sager, and Deb Wetzel for providing background information; and James Causby, Karen Davis, Kathleen Kennedy Manzo, Vanessa Jeter, and Janet Byrd for supplying important information about North Carolina's testing program. The author retains all responsibility for errors.

[2] See Richard P. Phelps, "The Demand for Standardized Student Testing," *Educational Measurement: Issues and Practice*, 17(3), Fall 1998.

[3] Steve Farkas, Jean Johnson, and Ann Duffett, *Different Drummers: How Teachers of Teachers View Public Education*, New York: Public Agenda, 1997, pp.20,36.

[4] Ibid., p.20.

[5] Ibid., pp.13-14.

[6] Ibid., p.13.

[7] Ibid., p.14.

[8] See transcripts of the conference papers printed, along with an introduction, in "Accountability as a Reform Strategy," *Phi Delta Kappan*, November, 1991, pp.219-251.

[9] CRESST is headquartered at UCLA's education school, but associates with "partners" at the education schools of the Universities of Colorado and Southern California and Arizona State University, and the RAND Corporation. CRESST publishes dozens of reports every year that are objective, often concentrating on psychometric methods. Some of the best psychometric research in the country is produced by CRESST. The research that several of CRESST's affiliated scholars publish that relates to *testing policy*, however, typically subscribes to the canon.

[10] Not everyone associated with CSTEEP opposes testing. Indeed, the Third International Mathematics and Science Study (TIMSS), perhaps the standardized test most reviled by testing critics, was headquartered at CSTEEP. TIMSS showed U.S. students performing more and more poorly in comparison with their international counterparts as grade levels advanced to the last year of high school. But another group of researchers at CSTEEP, not associated with TIMSS, devote themselves almost exclusively to anti-testing research.

[11] FairTest receives much of its financial support from the Ford Foundation, a great deal of exposure in the media, and a seat at the table with study commissions on testing policy as an interested "stakeholder," for example, on the U.S. Congress's Office of Technology Assessment Advisory Panel for the report *Testing in American Schools: Asking the Right Questions*.

[12] The actual number of standardized tests administered annually in the public schools is around 40 million, not 100 million. Forty million tests for about 40 million students calculates to one test per student per year, and only one in four of those is for high-stakes. See Richard P. Phelps, "The Extent and Character of System-Wide Student Testing in the United States," *Educational Assessment*, 4(2), 89-121.

[13] Noe Medina and Monty Neill, *Fallout from the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools*. Cambridge: FairTest, 1990.

[14] Daniel M. Koretz, "State Comparisons Using NAEP: Large Costs, Disappointing Benefits," *Educational Researcher*, 20(3), April 1991, p.19.

[15] Ibid., pp.19-21.

[16] Ibid., p.20.

[17] See also Richard M. Wolf, "What Can We Learn from State NAEP?" *Educational Measurement: Issues and Practice*; Bruce J. Biddle, "Foolishness, Dangerous Nonsense, and Real Correlates of State Differences in Achievement," *Phi Delta Kappan*. Bloomington, IN: Phi Delta Kappa Online Article, August 8, 1998.

[18] Koretz, "State Comparisons Using NAEP," p.20.

[19] Gary Phillips, "Benefits of State-by-State Comparisons," *Educational Researcher*, 20(3), April 1991, pp.17-19.

[20] Ibid., p.17. See also, George Bohrnstedt, Project Director, *The Trial State Assessment: Prospects and Realities*, Stanford: National Academy of Education, 1993.

[21] James W. Pellegrino, Lee R. Jones, and Karen J. Mitchell, *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress*, Washington: National Academy Press, 1998, p.2.

[22] Ibid., p.2.

[23] FairTest, "How the States Scored," Cambridge: FairTest, Summer 1997.

[24] Robert C. Johnston, "In Texas, the Arrival of Spring Means the Focus is on Testing," *Education Week*. Vol.17, No.33, April 29, 1998, p.20.

[25] Ibid.

[26] Lonnie Harp, "OCR Probes Bias Complaint Against Texas Exit Test," *Education Week on the Web*, Feb. 7, 1996.

27 Linda Jacobson, "State Graduation Tests Raise Questions, Stakes," *Education Week on the Web*. June 24, 1998.

28 "ED Clears Texas Tests," in "News in Brief," *Education Week on the Web*, August 6, 1997.

29 Johnston, "In Texas, the Arrival of Spring Means the Focus is on Testing," p.21.

30 Ibid., pp.1,20,21; "Pass or Fail," *Teacher Magazine: Education Week on the Web*. September, 1994; Lonnie Harp, "Final Exam," *Teacher Magazine: Education Week on the Web*. September, 1994; Lonnie Harp, "Texas Politicians Wrangle Over School Rankings," *Education Week on the Web*. September 14, 1994; Robert C. Johnston, "Texas Governor Has Social Promotions in His Sights," *Education Week on the Web*. February 11, 1998.

31 See "Pass or Fail, " Harp, "Final Exam," Harp, "Texas Politicians Wrangle Over School Rankings"; Johnston, "Texas Governor Has Social Promotions in His Sights".

32 They have eased one requirement, however. The "no pass, no play" rule, originally recommended by an education reform commission chaired by Ross Perot, barred students failing courses from participating in team sports for six weeks. That's been reduced to 3 weeks as other, broader requirements have been put in place. See Lonnie Harp, "Texas Politicians Wrangle Over School Rankings," September 14, 1994; Harp, "Texas Lawmakers Reach Accord on Overhaul of Education Laws," *Education Week on the Web*, May 31, 1995.

33 See Kathleen Kennedy Manzo, "N.C. Consensus Pushes for New Set of Reforms," *Education Week on the Web*. April 9, 1997; "Quality Counts '98: North Carolina Summary" *Education Week on the Web*; "High Stakes: Test Truth or Consequences," *Education Week*, Vol.17, No.8, October 22, 1997; "N.C. Gets First School-by-School Performance Results," *Education Week*, September 3, 1997, p.26; "Struggling N.C. Schools Buoyed by State Teams," *Education Daily*, July 10, 1998, p.4.

34 "Struggling N.C. Schools Buoyed by State Teams," p.4.

35 Ibid.

36 Telephone conversations with Vanessa Jeter and Janet Byrd of the North Carolina Department of Public Instruction, October 19, 1998.

37 David Molpus, "Improving High School Education," *National Public Radio Morning Edition*, Sept. 15, 1998.

38 Ibid.

39 Kathleen Kennedy Manzo, "High Stakes: Test Truths or Consequences," *Education Week on the Web*, October 22, 1997, pp.1-2.

40 Ibid., p.3.

41 Ibid.

42 Ibid., p.2.

43 Ibid., p.1.

44 Ibid., p.2.

45 Ibid., p.4.

46 Telephone conversations with James Causby, superintendent of the Johnson County Schools, September 24, 1998.

47 Richard M. Jaeger, "Legislative Perspectives on Statewide Testing," in "Accountability As a Reform Strategy," *Phi Delta Kappan*, November, 1991, p.242.

48 Ibid.

49 Ibid.

50 George F. Madaus, "The Effects of Important Tests on Students: Implications for a National Examination System," *Phi Delta Kappan*. November 1991, p.228; Lorrie A. Shepard, "Will National Tests Improve Student Learning?" *Phi Delta Kappan*. November 1991, p.234.

51 See Indicator C3 in Organization for Economic Co-operation and Development, *Education at a Glance: OECD Indicators 1997*. Paris: author, 1997. At age 16, the United States has a lower percentage of students enrolled than do Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, Germany, the Netherlands, Norway, New Zealand, and Sweden, all countries with high-stakes secondary-level exit exams. Rates in Hungary and Ireland are similar to ours. Switzerland and the United Kingdom are the only countries, among those included, with high-stakes exit exams and lower enrollment rates than the U.S. at age 16. Comparisons at age 17 are similar. The conclusion: students drop out in the United States for reasons other than not passing an exit exam.

52 J.B. Erickson, *Indiana Youth Poll: Youths' Views of High School Life*. Indianapolis: Indiana Youth Institute, 1991, p.33.

53 Bryan W. Griffin and Mark H. Heidorn. "An Examination of the Relationship Between Minimum Competency Test Performance and Dropping Out of High School," *Educational Evaluation and Policy Analysis*, Vol.18, No.3, Fall, 1996, pp.243–252.

54 C. Boyden Gray and Evan J. Kemp, Jr., "Flunking Testing: Is Too Much Fairness Unfair to School Kids?" *Washington Post*. September 19, 1993, p.C3.

55 "Testing Our Children: North Carolina," FairTest. Cambridge: author, Summer 1997.

56 See Indicators 8 and 9 in U.S. Department of Education, National Center for Education Statistics. *State Indicators in Education 1997*, NCES 97-376, by Richard P. Phelps, Andrew Cullen, Jack C. Easton, and Clayton M. Best. Project Officer, Claire Geddes. Washington, D.C., 1997.

57 While the SAT tends to be more visible, about half of state colleges and well over one-third of all U.S. colleges use the competing American College Test (ACT). Some colleges allow applicants to submit either test.

58 FairTest, "FairTest Fact Sheet: The SAT," Cambridge: author, p.2.

[59] They don't mention, however, the continual *growth* in the number of colleges *using* the SAT — 102 added since 1990 — bringing the total to 1,450 four-year institutions. See Charles A. Kiesler, "On SAT Cause and Effect," *Education Week*, May 13, 1998, p.43.

[60] See Debbie Goldberg, "Putting the SAT to the Test," *Washington Post Education Review*, October 27, 1996, pp.20-21; Many colleges have complained to the National Association for College Admission Counseling (NACAC) about their presence on "The List" of colleges that FairTest claims waive the SAT requirement. Many colleges FairTest includes waive the requirement only for a few students under extraordinary circumstances (e.g., disabilities, remote foreign locations, etc.) Telephone conversation with NACAC officials, Aug.14, 1998.

[61] See, for example, Joyce Slayton Mitchell, "A Word to High School Seniors — SATs Don't Get You In," *Education Week*, May 29, 1998, p.33; or National Association for College Admission Counseling, "Members Assess 1996 Recruitment Cycle in Eighth Annual NACAC Admission Trends Survey," *News from National Association for College Admission Counseling*, October 28, 1996, pp.2, 4.

[62] National Association for College Admission Counseling, "Members Assess 1996 Recruitment Cycle" in pp.2, 4.

[63] See Warren W. Willingham, et.al., *Predicting College Grades*, New York: The College Board, 1990, chapters 5 and 12; or Thomas F. Donlon, ed. *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*, New York: The College Board, 1984, Chapter VIII.

[64] Actually, one could make the same argument about high school grade point averages (GPAs). After all other predictive factors are accounted for, including SAT or ACT scores, high school GPA only explains another several percent of the variation in first-year college grades. So, why bother with the GPA? See also Gerald W. Bracey, "The $150 Million Redundancy," *Phi Delta Kappan*, 70(9), May, 1989, pp.698-702; Lucy May, "Tests don't have all the answers"; Edward B. Fiske, "Questioning an American Rite of Passage: How valuable is the SAT?" *New York Times*, January 18, 1989, p.B10. Walter Haney of Boston College is often quoted on this issue. See, for example, Lucy May, "Test Don't Have All the Answers to How Ky. Kids Rank," *The Lexington Herald-Leader*, July 6, 1995; or Debbie Goldberg, "Putting the SAT to the Test," *Washington Post Education Review*, Oct.27, 1996, p.21; or Peter Sacks, "Standardized Testing: Meritocracy's Crooked Yardstick," *Change*, March/April, 1997, p.26.

[65] May, "Tests don't have all the answers to how Kentucky kids rank," pp.44-45.

[66] Even 6 to 8 percent underestimates the predictive power of the SAT or ACT. Because colleges publish the mean and range of the admissions test scores of their first-year class, a high school senior can pick potential colleges where his test score fits in the range, and he is more likely to be admitted. Likewise, colleges can focus their recruiting efforts where they are likely to find attractive applicants who can succeed on their campuses and who might be willing to come. This represents an added benefit of the SAT — applicants and colleges don't waste time chasing after poor matches. Technicians call this benefit "allocative efficiency." Allocative efficiency is very difficult to estimate, but the Educational Testing Service has calculated that, just for the colleges alone, it must add at least another two percentage points of predictive power to the additional 6 to 8 percent already accounted for by SAT scores.

[67] At best, a student's high school record explains only 44 percent of the variation in first year college grades. SAT scores alone explain 42 percent of the variation. To a large extent, however, high school record and SAT scores represent the same thing, mastery of academic subject matter. Thus, when both high school record and SAT scores are used together in equations to predict students' first-year college grades, the two predictive factors overlap. If SAT scores are put in the equation first, high school record adds only a comparatively smaller amount of predictive power. Likewise, if high school record is put in the equation first, SAT scores add only a comparatively smaller amount of predictive power. After subtracting the proportion of predictive power that the two predictive factors share in common, SAT scores predict an additional 6 to 8 percent of the variation in first-year college grades. This 6 to 8 percent predicted by the SAT (averaged to 7 percent) represents 16 percent of the variation in first-year college grades explained by high school record alone (which was 44 percent). Thus if we put high school record in the prediction equation first, SAT scores represent a 16 percent incremental increase in predictive power when added to the equation. See Willingham, et.al., *Predicting College Grades*, chapters 5 and 12.

[68] Last year's annual survey by the National Association for College Admission Counseling shows that their members consider the following criteria the most important in determining admission (by percentage mentioning each criterion of considerable or moderate importance): grades in college prep courses, such as Advanced Placement courses (90); admission test scores, such as the SAT or ACT (82); grades in all subjects (79); class rank (71); essay or writing sample (53); counselor recommendation (66); and teacher recommendation (55). National Association for College Admission Counseling, "Members Assess 1996 Recruitment Cycle," pp.2,4.

[69] To some extent, the criticisms are tautological. A CSTEEP study of several commercially available math and science tests, managed by George Madaus and funded by the National Science Foundation, concluded that the tests promoted "test preparation" practices. Eighty-one percent of math teachers and 53 percent of science teachers engaged in some form of "test preparation," according to CSTEEP. However, the researchers "coded 'test preparation' as 'present' when the teacher or administrator made an explicit link between a particular activity and test scores, or gave such evidence in spite of denying test preparation." Thus, if a teacher taught an ordinary math or science lesson and hoped that it would improve students' performance on a test, that's "test preparation."

Mary Maxwell West and Katherine A. Viator, *The Influence of Testing on Teaching Math and Science in Grades 4-12: Appendix D: Testing and Teaching in Six Urban Sites*, Boston: CSTEEP, October 1992, pp.27-28.

70 Mary Lee Smith, et.al., *Reforming Schools by Reforming Assessment: Consequences of the Arizona Student Assessment Program (ASAP): Equity and Teacher Capacity Building*. CSE Technical Report 425, Los Angeles: CRESST, March 1997, p.2.

71 The history of the large-scale, standardized use of portfolios is spare and brief. (FairTest, "Testing Our Children: Introduction," Cambridge: author, 1998, p.2.) There appear to be many problems with a sole reliance on portfolios to measure student progress: they're far more susceptible to cheating, coaching, gaming, and outright plagiarism than are standardized tests. (See "Test Violations Uncovered," in "News in Brief", Education Week on the Web, August 6, 1997, p.5; Maryl Gearhart and Joan L. Herman, "Portfolio Assessment: Whose Work Is IT?" *Evaluation Comment*, CSE, CRESST, Winter 1996; and Daniel M. Koretz, "Sometimes a Cigar is Only a Cigar" in Diane Ravitch, ed. *Debating the Future of American Education*. Washington: Brookings Institution, 1995 p.160-162.) Moreover, they reward occasional, exceptional brilliance and not steady competence; and they are difficult to score with consistency. (Daniel Koretz, et.al., *The Reliability of Scores from the 1992 Vermont Portfolio Assessment Program*. Technical Report No.355, Los Angeles: CRESST, December, 1992.) These sound like the same tenor of criticisms FairTest, the most prominent advocate for the exclusive use of portfolios, makes of standardized tests.

72 One can also find elements of other theories and philosophies in the critics' rhetoric — that of "multiple intelligences," popularized by Howard Gardner, and what E.D. Hirsch labels "romantic progressivism," for example. As this article is not meant to focus on philosophy, however, I have kept this digression spare.

73 Daniel M. Koretz, Robert L. Linn, Stephen B. Dunbar, and Lorrie S. Shepard, "The Effects of High-Stakes Testing on Achievement: Preliminary Findings about Generalization across Tests," paper presented at the 1991 Annual Meeting of the AERA, Chicago, April 3-7. (ERIC ED340730)

74 See a discussion of the phenomenon that includes the physician, John Jacob Cannell, and many others in full-issue coverage, in *Educational Measurement: Issues and Practice*, Summer 1988.

75 Test publishers make economic and logistical trade-offs by using convenient samples, such as Chapter 1 students they are already testing to meet Chapter 1 requirements, as norming samples.

76 See Gary W. Phillips and Chester E. Finn, Jr. "The Lake Wobegon Effect: A Skeleton in the Testing Closet?" *Educational Measurement: Issues and Practice*. Summer 1988, pp.10-12.

77 Ibid.

78 Shepard, "Will National Tests Improve Student Learning?" pp.233-234.

79 Farkas, Johnson, and Duffet, *Different Drummers*, p. 7; and Jean Johnson and John Immerwahr, *First Things First: What Americans Expect from the Public Schools*. New York: Public Agenda, 1994.

80 Shepard, "Will National Tests Improve Student Learning?" pp.233-234.

81 See, for example, Mary Lee Smith, "Put to the Test: The Effects of External Testing on Teachers," *Educational Researcher*, Vol.20, No.5, June, 1991; "Meanings of Test Preparation," *American Educational Research Journal*, Vol.28, No.3, Fall 1991; "The Role of Testing in Elementary Schools," CSE Technical Report 321, Los Angeles, UCLA, May, 1991; and Lorrie Shepard, et.al. "Effects of High-Stakes Testing on Instruction," paper presented at the Annual Meeting of the AERA, Chicago, April 1991.

82 See, for example, Lorrie A. Shepard, "Will National Tests Improve Student Learning," pp.233-234.

83 See, for example, the example on the first two pages of Mary Lee Smith, "Put to the Test; The Effects of External Testing on Teachers."

84 Farkas, Johnson, and Duffett, *Different Drummers*, p.12.

85 See OECD, *Education at a Glance*, 1997, p.200 for the salary figures. See also John H. Bishop, "Impacts of School Organization and Signaling on Incentives to Learn in France, The Netherlands, England, Scotland, and the United States," Working Paper 94-30, Center for Advanced Human Resource Studies, New York State School of Industrial and Labor Relations, Cornell University, Ithaca, New York, December, 1994; and "Incentives for Learning: Why American High School Students Compare So Poorly to Their Counterparts Overseas," Working Paper No.89-09. Cornell University School of Industrial and Labor Relations, 1989, for discussions of the relationship of external tests and teacher status.

86 See Richard P. Phelps, "Benchmarking to the World's Best in Mathematics," paper presented at the Annual Meeting of the AERA, Chicago, April, 1997; Linda Ann Bond and Darla A. Cohen, "The Early Impact of Indiana Statewide Testing for Educational Progress on Local Education Agencies," in Rita G. O'Sullivan and Robert E. Stake, eds. *Advances in Program Evaluation*. Vol.1, Part B, 1991, pp.87-88; or see George Stigler's work comparing time use in U.S., German, and Japanese lower secondary mathematics and science classes, in George W. Stigler and James Hiebert, "Understanding and Improving Classroom Mathematics Instruction: An Overview of the TIMSS Video Study." *Phi Delta Kappan*. Bloomington: Phi Delta Kappa, September, 1997, pp.14-21.

87 Robert Rothman, "Study Confirms 'Fears' Regarding Commercial Tests," *Education Week*. Vol.12, No.7, October 21, 1992, p.1,13.

88 Malcolm Gladwell, "NSF Faults Science and Math Testing," *Washington Post*. October 16, 1992. pp.A1,A4.

89 Ibid., p. A1.

90 Jerry Bobrow, *Cliffs SAT I Preparation Guide*, Lincoln, NE: Cliffs Notes, 1994, p.63.

91 They are made explicit, however, in the cognitive laboratory testing that some multiple-choice tests undergo when they are developed.

92 Maryellen C. Harman and Claudette Fong-Kong Mungal, *The Influence of Testing on Teaching Math and Science in Grades 4-12: Appendix B: An Analysis of Standardized and Text-Embedded Tests in Mathematics.* Boston: CSTEEP, October 1992, p.5.

93 Here's another definition of "higher-order thinking" that can be used as a point of comparison: "Students engage in purposeful, extended lines of thought during which they: identify the task or problem type; define and clarify essential elements and terms; judge and connect relevant information; and evaluate the adequacy of information and procedures for drawing conclusions and/or solving problems. In addition, students become self-conscious about their thinking and develop self-monitoring problem-solving strategies. Commonly specified higher-order reasoning processes are: cognitive: analyze, compare, infer/interpret, evaluate; metacognitive: plan, monitor, review/revise." (Edys S. Quellmalz, "Needed: Better Methods for Testing Higher-Order Thinking Skills," *Educational Leadership*. Vol.43, No.2, October 1985, p.30.)

94 See George F. Madaus, Mary Maxwell West, Maryellen C. Harmon, Richard G. Lomax, and Katherine A. Viator, *The Influence of Testing on Teaching Math and Science in Grades 4-12: Executive Summary*. Chestnut Hill: CSTEEP, Boston College, October, 1992; and Mary Maxwell West and Katherine A. Viator, *The Influence of Testing on Teaching Math and Science in Grades 4-12: Appendix D: Testing and Teaching in Six Urban Sites.* Chestnut Hill: CSTEEP, Boston College, October 1992.

95 E.D. Hirsch, of course, makes a more detailed and eloquent argument for the acceptance of both process and content as necessary components of intelligence. See E.D. Hirsch, Jr.. *The Schools We Need and Why We Don't Have Them*. New York: Doubleday, 1996.

96 Monty Neill, *High Stakes Tests Do Not Improve Student Learning*, Cambridge: FairTest, 1998.

97 FairTest, "Testing Our Children: Introduction," Cambridge: author, 1998, p.2.

98 FairTest, "How the States Scored," and "Vermont," *FairTest Examiner*. Summer 1997, p.1 and pp.1-3,

99 Monty Neill, *High Stakes Tests Do Not Improve Student Learning*. Cambridge: FairTest, January 1998.

100 Bishop, John H., "Education Quality and the Economy," paper presented at the Seventh International Conference on Socio-Economics of the Society for the Advancement of Socio-Economics, Arlington, VA, April 8, 1995; Bishop, John H., "Do Curriculum-based External Exam Systems Enhance Student Achievement?" Working Paper 97-28, Center for Advanced Human Resource Studies, New York State School of Industrial and Labor Relations, Cornell University, Ithaca, New York, December, 1997; and Bishop, John H., "The Effect of Curriculum-Based Exit Exam Systems on Student Achievement," Working Paper #97-15, Cornell University, School of Industrial and Labor Relations, Center for Human Resource Studies, 1997.

101 Bishop, John H., "Diplomas for Learning, Not Seat Time: The Impacts of New York Regents Examinations," Working Paper #97-31, Cornell University, School of Industrial and Labor Relations, Center for Advanced Human Resource Studies, 1997, pp.11-17.

102 Debra Viadero, "Assessment Payoff," *Education Week*, September 10, 1997, p.32. For example, a group of University of Maryland researchers studying Maryland's State School Performance Assessment Program, which bears consequences for poorly performing *schools* (while another state test bears consequences for poorly performing students), found large gains in student achievement in schools that embraced the entire programs.

103 FairTest, "FairTest Fact Sheet: The SAT" and "SAT, ACT Bias Persist" *FairTest Examiner*, Fall, 1995, Cambridge: author.

104 Nancy Cole and Warren Willingham, *Gender and Fair Assessment*, Princeton, ETS, 1997.

105 "ETS Gender Bias Report a 'Smokescreen', *FairTest Examiner*, Cambridge: author, Fall, 1997.

106 Sol Pelavin and Michael Kane, *Changing the Odds*. New York: The College Board, 1990.

107 David W. Murray, "The War on Testing," *Commentary*, September, 1998, pp.34-37; see also Jessica L Sandham, "Ending SAT May Hurt Minorities, Study Says," *Education Week*. January 14, 1998, p. 5.

108 "Diversity Takes Back Seat to Standards in New Poll," *Education Daily*, July 30, 1998, pp.3,4.

109 Steve Farkas, Jean Johnson, Stephen Immerwahr, and Joanna McHugh, *Time to Move On: African-American and White Parents Set an Agenda for Public Schools*. New York: Public Agenda, 1998, pp.16-17.

110 Walter M. Haney, George F. Madaus, and Robert Lyons, *The Fractured Marketplace for Standardized Testing*. Boston: Kluwer, 1993, p.119.

111 Ibid., p.122.

112 Lawrence O. Picus and Alisha Tralli, *Alternative Assessment Programs: What are the True Costs?* CSE Technical Report 441, Los Angeles: CRESST, February 1998, p.47.

113 U.S. General Accounting Office, *Student Testing: Current Extent and Expenditures, with Cost Estimates for a National Examination*, Report GAO/PEMD-93-8. Washington, DC: author, 1993, p.66; and U.S. Department of Education, National Center for Education Statistics, *Digest of Education Statistics 1997*, by Thomas D. Snyder and Charlene M. Hoffman, Washington: U.S.GPO, 1997, Table 33, p.36.

114 George F. Madaus, "The Effects of Important Test on Students," *Phi Delta Kappan*. November 1991, p.227.

[115] George F. Madaus and Thomas Kellaghan, "Student Examination Systems in the European Community: Lessons for the United States," contractor report submitted to the Office of Technology Assessment, June 1991.

[116] See Richard P. Phelps, "Are U.S. Students the Most Heavily Tested on Earth?" *Educational Measurement*, Vol.15, No.3, Fall, 1996.

[117] George F. Madaus and Thomas Kellaghan, "Student Examination Systems in the European Community: Lessons for the United States. "

[118] I looked at OECD countries, plus Russia and China. I scoured any source documents I could find that noted instances of countries adding or dropping systemwide tests. The documents consisted of several editions of international encyclopedias of education with chapters on each country written by resident experts; several, and some ongoing, comparative surveys of countries' testing programs by the OECD, the U.S. General Accounting Office, the U.S. Education Department's National Center for Education Statistics (NCES) and Office of Policy and Planning (OPP), the National Center on Education and the Economy, and the Statistical Office of the European Union; and the scholars Eckstein and Noah, Bishop, and Stevenson and Lee.

[119] On the latter point, one speaker, Linda Darling-Hammond, then of Columbia University Teachers College, now at Stanford University, said, "In contrast to testing in most other countries, testing in the U.S. is primarily controlled by commercial publishers and nonschool agencies that produce norm-referenced, multiple-choice instruments designed to rank students cheaply and efficiently." Linda Darling-Hammond, "The Implications of Testing Policy for Quality and Equality," *Phi Delta Kappan*. November 1991, p.220.

[120] Robert E. Stake, "The Teacher, Standardized Testing, and Prospects of Revolution," *Phi Delta Kappan*. November 1991, p.246.

[121] Mary Lee Smith and Claire Rottenberg, "Unintended Consequences of External Testing in Elementary Schools," *Educational Measurement: Issues and Practice*. Winter 1991, pp.10-11.

[122] See Richard P. Phelps, "The Demand for Standardized Student Testing," *Educational Measurement: Issues and Practice*, 17(3), Fall, 1998.

[123] Mary Maxwell West and Katherine A. Viator, *Teachers' and Administrators' Views of Mandated Testing Programs*. Boston: CSTEEP, October 1992, Table 3.

[124] Ibid., p.2.

[125] Ibid., p.6.

[126] Ibid., pp.39-40. Other sources of "test invalidity" included "kids are not on grade level," even though a student can be so because he doesn't study or pay attention in class; "kids don't try on tests" even though it can be the fault of the student that he doesn't try; or "tests have weird words, content unfamiliar to the students (language/culture bias)," even though words can be "weird" and "content unfamiliar" because a student doesn't do his reading, study, or pay attention in class.

[127] National Association of State Boards of Education, *The Full Measure: Report of the NASBE Study Group on Statewide Assessment Systems*, Alexandria, VA: author, 1997; and Millicent Lawton, "State Boards' Leaders Call for Assessments Bearing Consequences," *Education Week on the Web*, October 22, 1997.

[128] For an interesting study of the positive opinions of teachers and administrators toward one state test, see Linda Ann Bond and Darla A. Cohen, "The Early Impact of Indiana Statewide Testing for Educational Progress on Local Education Agencies," in Rita G. O'Sullivan and Robert E. Stake, eds. *Advances in Program Evaluation*. Vol.1, Part B, 1991, pp.78-79, 87-88

[129] Gregory J. Cizek, "The Case Against the SAT," book review, *Educational and Psychological Measurement*. 50(3), Autumn, 1990, p.705.

[130] For example, see "State News Roundup," *Education Week on the Web*, June 8, 1994, p.1.

[131] According to Jeff Moss, the Associate School Superintendent for the Hoke County, North Carolina schools, before the accountability reforms, "We had seven levels of instruction for a subject matter, such as seven levels of biology, seven levels of English One, which ranged from remedial to honors or college preparatory. So the teacher expectation was such that if I labeled you a basic student I needed to put you in basic English and not require much from you." See David Molpus, "Improving High School Education," *National Public Radio Morning Edition*, September 15, 1998.

[132] For a comprehensive overview of the quality and reliability of teacher evaluations of student achievement, see Richard J. Stiggins and Nancy Faires Conklin, *In Teachers' Hands: Investigating the Practices of Classroom Assessment*, New York: SUNY Press, 1992.

[133] Gregory J. Cizek, "Grades: The Final Frontier in Assessment Reform," *NASSP Bulletin*, December, 1996.

[134] Robert B. Frary, et.al., "Testing and Grading Practices and Opinions of Secondary School Teachers of Academic Subjects: Implications for Instruction in Measurement," *Educational Measurement: Issues and Practice*, Fall, 1998, pp.23-30.

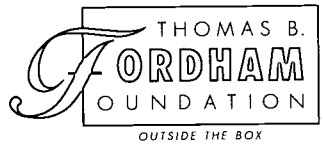
[135] Farkas, Johnson, and Duffett, *Different Drummers*.

[136] Ibid., pp.10-12.

[137] Ibid., pp.10-11.

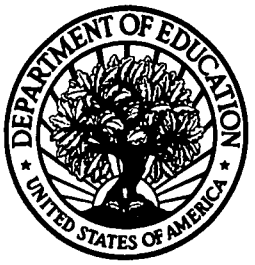
[138] Ibid., p.11.

[139] Ibid., pp.15-16.



The Thomas B. Fordham Foundation
1627 K Street, N.W. ∘ Suite 600 ∘ Washington, D.C. 20006
Telephone: (202) 223-5452 ∘ FAX: (202) 223-9226
http://www.edexcellence.net

To order publications: 1-888-TBF-7474 (single copies are free)



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").