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. It is contended that, as a whole, standardized tests do not measure much. Current standardized multiple-choice tests are severely flawed, based on assumptions about human ability that cannot be proven, with inadequate content and construct validity. An agenda for test reform calls for the use of "authentic" or "performance-based" evaluation, using properly constructed, validated, and administered tests. Schools, test-takers, and independent researchers should have access to data needed to verify test publishers' claims about their tests. Both test developers and test users have the obligation to promote proper, reasonable, and limited use of standardized tests as one of a series of assessment mechanisms. Appendix A presents results of a survey of officials from all states and from 56 sample school districts in 38 states focusing on the use of standardized tests. Appendix B discusses "authentic evaluation." A 72-item annotated bibliography is included. (SLD)
Fallout From the Testing Explosion:
How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools

by Noe Medina and D. Monty Neill

with the staff of the
National Center for Fair & Open Testing
(FairTest)

Third Edition (Revised)
March, 1990

BEST COPY AVAILABLE
Contents

Introduction 3
Test Use in U.S. Schools 5
Problems with Standardized Tests 8
Impact on the Public Schools 24
Reform Agenda 33
Notes 37
Appendix A: Survey Results 47
   Table 1: State-Mandated Tests 53
   Table 2: District-Mandated Tests 54
   Table 3: Test Use Calculations 57
Appendix B: Authentic Evaluation 58
Annotated Bibliography 61
FALLOUT FROM THE TESTING EXPLOSION:
HOW 100 MILLION STANDARDIZED EXAMS UNDERMINE EQUITY AND EXCELLENCE IN AMERICA’S PUBLIC SCHOOLS

"Standardized testing seems to have become the coin of the educational realm. In recent years, it seems that the aims of education and the business of our schools are addressed not so much in terms of curriculum...as in terms of what gets tested."
- George Madaus & Walt Haney

Standardized tests dominate the educational landscape in contemporary America. Based on a recent study, the National Center for Fair & Open Testing (FairTest) conservatively estimates that public schools in the United States administered at least 100 million standardized tests to their 39.8 million students during the 1986-87 school year— an average of more than two and one-half tests per student per year. An estimate of 200 million—an average of five per year—is probably not extreme.

Standardized test results have become the major criteria for a wide range of school decisions. Test scores limit the programs students enter or dictate the ones in which they are placed. Some tests are used to decide who will be promoted and who will be retained in grade; others determine which students graduate from high school. Test results also are used to assess the quality of teachers, administrators, schools and whole school systems.

Test proponents, of course, applaud these trends. They see tests as “objective” mechanisms to inject accountability into public schools and thereby improve student achievement, staff competence and educational quality. They see standardized exams as essential elements of the “School Reform Movement.”

In fact, experience with standardized test use paints quite a different picture. Rather than being “objective” instruments, standardized tests often produce results that are inaccurate, inconsistent, and biased against minority, female and low-income students. Rather than promoting accountability, tests shift control and authority into the hands of an unregulated testing industry. By narrowing the curriculum, frustrating teachers, and driving students out of school, tests undermine school improvement rather than advance its cause.

FairTest concurs with the National Academy of Education that “the nation has a right to know what students achieve, what schools
Fallout from the Testing Explosion

Relying on standardized tests will lead to a weaker, not stronger, educational system.

are doing, and what more should be done." Standardized tests, even when properly constructed, validated, administered and used, can only play a limited role in this effort. Too often, however, standardized tests do not meet these basic standards. Moreover, the essential nature of these types of instruments makes them inadequate tools for assessing what is most valuable in education. As a result, relying on standardized tests as the primary criterion for making various school decisions will lead to a worse, not better, public understanding of the schools and a weaker, not stronger, educational system.
I. TEST USE IN U.S. SCHOOLS

During the 1986-87 school year, American educators reported that at least 93 to 105 million standardized tests or test batteries were administered to 39.8 million elementary and secondary public school students. This includes:

- 38.9 million standardized achievement, competency and basic skills tests administered to fulfill local testing mandates;

- 16.8 million standardized achievement, competency and basic skills tests administered in 42 states and the District of Columbia to fulfill state testing mandates;

- between 30 and 40 million standardized tests administered to compensatory and special education students;

- between 1.5 and 1.75 million screening tests for kindergarten and pre-kindergarten students; and

- between 6 and 7 million standardized tests administered to identify or place gifted or limited-English proficient students (for which there are no reliable figures). Nor does it include tests administered by private and parochial schools to their students.

This data was gathered by FairTest staff through a series of telephone interviews with officials from all 50 state departments of education, the District of Columbia school district and 56 sample school districts in 38 states. Additional information was gathered by examining recent surveys documenting the use of other standardized exams, including IQ tests, behavioral tests, readiness tests for young children and placement tests [see Appendix A].

This estimate of 93-105 million tests is a conservative one. FairTest counted each test battery as one test. However, batteries usually include a number of separate exams: at least one each for math and reading, and frequently social studies, science and other subjects. If we had counted each exam within a battery as a separate test, the local- and state-mandated totals would easily double and the special education total would potentially double. The total also does not include tests administered to identify or place gifted or limited-English proficient students (for which there are no reliable figures). Nor does it include tests administered by private and parochial schools to their students.
Fallout from the Testing Explosion

Test use may also have been underreported by some school officials. For example, the Milwaukee Public Schools reported 64,500 tests administered in 1986-87. However, a detailed investigation conducted by the Milwaukee Assessment Task Force revealed a total of 484,956 tests administered in 1988-89 (each test in a battery was counted). Thus, rather than being tested at a rate of .67 tests per student per year, the rate is 5 per student per year. Milwaukee students in special programs were found to be tested far more heavily than students in regular classes.5

Given factors such as these, total standardized testing across the U.S. could exceed 200 million per year. Thus, in 12 years of schooling (plus kindergarten), a child could take, on average, some 60 standardized tests, or five per year.

In addition, the FairTest survey revealed that the number of states which mandate testing has greatly increased in recent years. Compared to the findings of a 50-state survey conducted by Education Week in 1985:

• the number of states requiring students to pass a standardized test for high school graduation increased from 15 in 1985 to 24 in 1987;

• the number of states employing standardized tests to determine whether students should be promoted to the next grade increased from 8 in 1985 to 12 in 1987; and

• the number of states using standardized tests as part of a state assessment program increased from 37 in 1985 to 42 in 1987.6

The survey also revealed three significant patterns of standardized test use in public schools. First, ten Southern states (Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina and Tennessee) administer standardized tests to fulfill state testing mandates at a rate more than twice that of schools in the remainder of the nation. In fact, the four states mandating the most tests per pupil (Kentucky, North Carolina, Alabama, and Georgia) were all located in the South.7

Second, seven states (Iowa, Minnesota, Montana, North Dakota, Ohio, Vermont, and Wyoming) which did not have a state testing mandate all have relatively small minority enrollments in their public schools (ranging from a low of 1% in Vermont to a high of only 15% in Ohio, compared to a national average of about 25%). Alaska is the only state without a testing mandate which also has a
substantial minority student population (primarily Native Americans). However, Alaska planned to institute a state testing program in the 1988-89 school year.\[8\]

Finally, schools in larger school districts are likely to administer standardized tests to fulfill local mandates at a higher rate than schools in smaller districts. The largest school districts (those with student enrollments exceeding 100,000) administered tests at a rate one and one-quarter times that of medium-sized districts (those with enrollments between 25,000 and 100,000) and one and one-half times that of the smallest districts (those with enrollments less than 25,000). Within these overall trends, the rate of test administration...
Fallout from the Testing Explosion

**II. PROBLEMS WITH STANDARDIZED TESTS**

"The importance of understanding what it is that tests can and cannot tell us is critical. Not all tests are accurate measures of the skills and knowledge they purport to measure and even the more accurate tests are at best approximations." - Congressional Budget Office

Standardized tests are consistently sold as scientifically developed instruments which simply, objectively and reliably measure student achievement, abilities or skills. In reality, there are serious problems in the construction, validation, administration and use of standardized tests and their results.

Standardized tests are constructed in ways that often guarantee biased results against minorities, females and low-income students. Test results are evaluated and scored in ways that are often at odds with modern theories of intelligence and child development. The test validation process is often inadequate and far from objective. Many tests are administered in an environment that undermines any claims they may have to being “standardized.” Even those that adhere to “standard” administration practices may be biasing the results against minorities, low-income students and females by using examiners who are unfamiliar to the test-takers and by using tightly timed tests.

These flaws undermine testmakers’ claims of objectivity and produce test results that are inaccurate, unreliable or biased. Ultimately, many tests fail to effectively measure test-takers’ achievement, abilities or skills.

**Test Bias**

Because most standardized tests are written by and for the middle- to upper-class white population, their results often fail to accurately measure the performance of those who do not fit this category. Testmakers claim that the lower test scores of racial and ethnic minorities and of low-income students simply reflect the bias and inequity that exists in American schools and society. While such biases and inequities certainly exist, standardized tests do not just reflect their impact, they compound them.

Joseph Gannon provided documentation for this conclusion in a 1980 study for the National Conference of Black Lawyers. Gannon...
college seniors from the same universities and with comparable undergraduate grade point averages. Even after controlling for these characteristics a gap of 100 points remained when black and Hispanic scores were compared with those of white students even though they had demonstrated equal academic ability in college.10

Researchers have identified several characteristics of standardized tests which could bias results against minority and low income students. Each characteristic reflects the middle- to upper-class white focus of such tests. As such, test results are as much a measure of race/ethnicity or income as they are of achievement, ability or skill.

Some of these characteristics also lead to gender bias in standardized tests, which affects both males and females. Among very young children, some tests appear to be biased against boys. On the other hand, among older children and adolescents, most bias affects girls.11

Several of the characteristics that bias test results relate to language. To communicate their level of achievement, ability or skills, test-takers must understand the language of the test. Obviously, tests written in English cannot effectively assess the achievement, skills or abilities of students who primarily speak Spanish or some other language.

Many groups of English-speakers are affected by a similar, but more subtle, form of language bias. Most standardized tests are written in an elaborated, stylized language rather than the simple and common vernacular. Researchers have discovered that the use of such forms of English prevents tests from accurately measuring achievement, ability or skills of students who use nonstandard English dialects. This includes speakers of African-American, Hispanic, white southern, Appalachian, and working-class dialects.12

A related type of bias stems from stylistic or interpretive differences related to culture, income or gender. For instance, the word "environment" is often associated by black students with terms such as "home" or "people." White students tend to associate the word with "air," "clean" or "earth." Neither usage is wrong; one simply centers on the social environment while the other centers on the natural one. Unfortunately, on a standardized test only one of these two usages, generally the one reflecting the white perspective, will be acceptable.13

Similarly, researchers have discovered that individuals exhibit "different ways of knowing and problem-solving" which reflect different styles, not different abilities. These differences are often
Fallout from the Testing Explosion

Standardized tests assume that all individuals perceive information and solve problems in the same way.

correlated with race/ethnicity, income level and gender. Yet standardized tests assume that all individuals perceive information and solve problems in the same way.14

For example, middle class whites are more apt to be trained, simply through cultural immersion, to respond to questions removed from any specific context and to repeat information the test-taker knows the questioner already possesses. Heath found that working class black children, in their communities, were rarely asked the sorts of questions in which the questioner already knew the answer, the sort found on tests.15 Again, assumptions about the universal application of one particular style limits the validity of test results.

Another cause of test bias is apparent in the content of questions which assume a cultural experience and perspective which not all children share. “Correct” answers to such questions usually reflect the experiences, perspectives and knowledge of children and adults from a white middle- to upper-class background. Answers which draw upon the different experiences, perspectives and knowledge of racial/ethnic minorities, children from poor families, and children from inner city or rural backgrounds are ignored. Although some answers may be correct in these different geographical or cultural contexts, they are generally counted “incorrect” on the test.

The WISC-R IQ test, for example, asks “What is the thing to do when you cut your finger?” The best response, according to the test, is to “put a bandaid on it.” Partial credit is also given for a response of “go to the doctor (hospital).” No credit is awarded for responses of “cry,” “bleed” or “suck on it.” Minority children usually perform poorly on this item. A few years ago a Baltimore, Maryland sociologist asked several inner-city youths why they answered the question the way they did. She found that many of these kids answered “go to the hospital” because they thought that “cut” meant a big cut. When the children were told that “cut” meant “little cut,” almost all then responded, “Put a bandaid on it.”16

Finally, students tend to perform better on tests when they identify with the subjects of the test questions. Research on Mexican-Americans, African-Americans, and girls all reveal that “items with content reference of special interest” to each group seem to improve their test scores.17 Unfortunately, standardized tests remain dominated by questions about and for white middle- to upper-class males.18
Nonetheless, test companies maintain that they effectively screen out biased questions. Though they subject items to review by experts who can supposedly detect bias, such screening is of low accuracy, and many biased items cannot be detected in this way. Though most major test-makers also apply some form of statistical bias-detection procedure, even when bias is found items are not necessarily removed. Moreover, the procedures themselves are problematic.

Bias-reduction techniques, such as Differential Item Functioning (DIF), generally attempt to match test-takers by “ability” and then locate those items that test-takers who are of similar “ability” but differ by race, ethnicity, or gender answer correctly at different rates. The items which exhibit large race, ethnicity or gender differentials are then flagged for further review, but not necessarily discarded from the test even if the differential answer rate is very large. However, the procedure used to match test-takers by “ability” is to match them by their overall scores on the test. Since the overall test score is nothing more than the sum of all test items, this procedure is obviously circular (as the Educational Testing Service has admitted) and inadequate for detecting any systematic bias in the test as a whole. The fundamental problem is that the procedure assumes the very thing it ought to be checking for.20

Given the role of knowledge and language in creating culture, there is no reason to believe a totally “culture-free” test can be constructed. Tests however, can be better designed so as to reduce their discriminatory impact by careful selection of content that is less likely to be unfamiliar to minorities or by use of bias reduction techniques such as the “Golden Rule” procedures.21

It is important to note that the use of standardized tests is often defended on the grounds of “objectivity.” But all “objective” really means is that the test can be scored without human subjectivity, by machines. As Banesh Hoffman noted in The Tyranny of Testing, “the term ‘objective test’ is a misnomer. The objectivity resides not in the test as a whole but merely in the fact that no subjective element enters the process of grading once the key is decided upon.”22

Bias can still creep into the questions themselves. What content and which items to include on the test, the wording and content of the items, and the determination of the correct answer, as well as how the test is administered and the uses made of test scores, are all decisions made subjectively by human beings. In fact, the purported objectivity of tests is often no more than the standardization of bias.
Fallout from the Testing Explosion

The replacement of the potential bias of individual judgement with the numerical bias of an “objective” test is not progress.

Test Construction

The ability of standardized tests to accurately report students’ knowledge, abilities, or skills is limited by assumptions that these attributes can be isolated, sorted to fit on a linear scale, and reported in the form of a single score. Gould labels these the fallacies of reification (treating “intelligence” as though it were a separable unitary thing underlying the complexity of human mental activity) and ranking (“our propensity for ordering complex variation as a gradual ascending scale” using a number). He concludes, “Thus, the common style embodying both fallacies of thought has been quantification, or the measurement of intelligence as a single number for each person.”

Many of the assumptions and structures of achievement tests are based on IQ tests and operate in the same way. For example, assumptions regarding the linear sorting of students are common to both.

The use of a linear scale not only creates false differences, it also can mask real differences. Assume, for example, that one student can compute using addition, subtraction, multiplication or division, but is unable to apply those concepts to fractions. Meanwhile, another student can compute with either whole numbers or fractions, but has difficulty with multiplication and division. If a mathematics test included four questions on whole numbers and four on fractions and each question required the student to employ a different type of
score (50%). Yet, their identical scores would mask real differences in skills.

The simple fact is that, while our knowledge of thinking, learning, teaching, and child development has grown over recent years, many standardized tests have not. Despite claims that testing is now more advanced and scientific, Oscar Buros noted that “little progress has been made in the past fifty years — in fact in some areas, we are not doing as well. Except for the tremendous advances in electronic scoring, analysis and reporting of test results, we don’t have a great deal to show for fifty years’ work. Essentially, achievement tests are being constructed today in the same manner they were fifty years ago...”

The same can be said for I.Q. and school readiness testing. The WISC-R “has remained virtually unchanged since its inception in 1949. . . Developments in the fields of cognitive psychology and neuroscience have revolutionized our thinking about thinking, but the WISC-R remains the same.”

Resnick and Resnick point out that behaviorist and associationist psychological theories from early in the twentieth century underlie standardized tests. These theories assume that knowledge can be decomposed or broken into separable bits, taken out of any context, and learned as an accumulation of these bits. The bits can then be tested one by one and without context. This is sometimes called the “banking” or “empty vessel” theory of learning. While few psychologists hold to these theories today, these assumptions remain built into standardized testing.

Test constructors not only erroneously presume that the knowledge, skill or ability being measured is one-dimensional and decomposable, but also “that it tends to be distributed according to the statistical “normal” bell-shaped curve. The bell-shaped curve is used for statistical convenience, not because any form of knowledge or ability is actually distributed in this manner.

Again, modern theories emphasize the complexity of human intelligence. Researchers have observed that knowledge, learning and thinking have multiple facets, and that a high level of development in one area does not necessarily indicate a high level of development in others. Unitary test scores and linear scaling of scores ignore the true complexity and thus provide a deceptive picture of individual achievement, ability or skills. This is a fundamental problem underlying all standardized tests in education.
Creation of a norm-referenced curve exaggerates the differences among people.

In order to construct a “normal curve,” test-makers must separate test-takers along a continuum. Thus, a norm-referenced test must not have many questions all test-takers can answer and must have some that almost none can answer. Creation of the curve exaggerates the differences among people. In general, test-makers discard those questions on which low-scoring test-takers do well but high-scorers do poorly. As a result, an item on which African-Americans do particularly well but whites do not is likely to be discarded for the compound reason that African-Americans are a minority group in the U.S. and generally score low. Even if minorities are included in the test companies' samples consistent with their portion of the overall population, at least three quarters of the sample is white. Moreover, minorities are disproportionately among the low-scoring group. Thus, questions that might favor minorities are apt to be excluded for not fitting the “required” statistical properties of the test.

For standardized achievement tests, problems also arise when test publishers use national test score averages (“norms”) as reference points for interpreting student performance. Using these norms, schools can determine, for example, that a certain test score ostensibly represents performance at the 65th national percentile, i.e. higher than 65% of all other students.

Test norms are developed by administering the test to a group of students which, in theory, represents the national student population. However, some widely used tests have been normed on small, unrepresentative populations. The popular Gesell Preschool Test, for example, employed a normative sample “composed of 40 girls and 40 boys at each 6-month age level from 2 through 6 years, for a total sample of 320 girls and 320 boys. However, nearly all were Caucasians and all resided in the state of Connecticut.”

The WISC-R was normed using a population of 100 boys and 100 girls for each of its age levels from 6-1/2 to 16-1/2. According to the test publisher, the norming group was representative of the national population as of 1970 with respect to race, geographic region, occupation of head of household, and urban-rural residence. Aside from the fact that the national population has changed considerably since 1970, this means that the test is normed for African-Americans using only 10 boys and 10 girls at each grade level and for Hispanics using only 5 or 6 each of boys and girls.

Norms also may be distorted even if samples of a thousand or more students for each level are used (as is the case with major
standardized achievement batteries) because school systems volunteer to participate in the norming process. The quality of the norming sample therefore depends on how accurately the self-selected schools reflect national performance. Test publishers typically fail to report the extent of this potential variance.37

A recent study by Friends for Education raises additional doubts about the validity of the national norms. That organization "discovered that no state is below average at elementary level on any of the six major nationally normed, commercially available tests" and blamed inaccurate initial norms and teaching to the test for the inflated scores.38

Critics have noted that the report released by Friends for Education contains inaccuracies and that many of the tests' statistical procedures were interpreted incorrectly. However, these flaws do not undermine the major finding of the report (dubbed the Lake Wobegon Phenomenon after Garrison Keillor's mythical community in which "all the children are above average"). In a recent analysis, Daniel Koretz reaffirmed the report's finding and conclusion that outdated norms are in part to blame for overstating student performance.39 A subsequent study conducted for the Center for Research on Evaluation, Standards and Student Testing (CRESST) concluded that the "Lake Wobegon effect" does exist and argued that "teaching to the test" is a primary cause.40

Test Reliability

Publishers' claims that standardized tests exhibit a high level of reliability do not necessarily mean that test results will be similar in successive administrations. Rather, for the testing industry, "reliability" is a technical term encompassing several different concepts.

The term can mean consistency of test scores over time, the popular perception of reliability. However, it can also refer to the consistency between scores on the entire test and responses to individual test questions, between two halves of the test, or between different forms of the same test. Despite sharing the same term, the concepts are certainly not interchangeable, even though each measures some aspects of the "range of fluctuation likely to occur in a single individual's score as a result of irrelevant, chance factors."41

Psychometricians have developed a variety of statistical approaches to measure these different types of reliability. These approaches all report their results as "reliability coefficients" (numbers generally ranging from 0 to 1). For most standardized tests, these re-
Fallout from the Testing Explosion

The publisher of the Gesell does not report any reliability data for its tests.

Arbitrary use of test scores will assign some children to a remedial program who do not belong there, while excluding others who should be in the program.

For most standardized tests, these reliability coefficients are very high — often exceeding .8 and .9. Test companies often report internal or inter-form reliability rather than consistency over time (which usually has a lower reliability), even where the latter may be more important than the former.

Nevertheless, significant scoring differences can still result. Anne Anastasi presents an example in her text, *Psychological Testing*, of an IQ test with a reliability coefficient of .89 and a standard deviation of 15. A student administered that test twice would be likely to score up to 13 points higher or lower on a retest. (There is also a chance a student could score even higher or lower). Thus a school system could deny entry into a “Gifted & Talented” Program requiring a score of 130 on this test to a student scoring 117 when the same student could, within the bounds of reliability, readily score 130 on a readadministration of the same test.

Similarly, Kaufman and Kaufman found that the reliability of the Gesell School Readiness Test meant that “a child measured to be four and one-half years old developmentally and unready for school could very likely be five and fully ready.” The publisher of the Gesell, in fact, does not report any reliability data for its tests.

The effect of imperfect reliability on decision making is expressed in various ways. The “error of measurement” means that decisions made using test scores with imperfect reliability will include both false positives (those who will be included but who should not be) and false negatives (those who will be excluded but should have been included). For example, arbitrary use of test scores will assign some children to a remedial program who do not belong there, while excluding others who should be in the program. Setting the “cut score,” the point at which to include or exclude, high will lead to more false negatives, while setting it low will lead to more false positives. The setting of the cut score is an ultimately subjective act by test makers or test users.

Test-makers also can compute a test’s “margin of error” and its “standard error of difference.” The Scholastic Aptitude Test (SAT), for example, has a “standard error of measurement” of 67 points on the combined test (which has a scale that runs from 400 to 1600). The error of difference is 144 points. That is, two test-takers’ scores must differ by at least that much before it can be said their abilities (as measured by the test) differ. Decisions on college entrance, however, are sometimes based on SAT score differences as small as 10 points.
Even with tests which report overall high reliability, the reliability of test sub-sections may be much lower. On many tests, reliability is also lower for children below the third grade. Thus the chance for error increases when such tests are used for placing young children or when decisions are made based on sub-test scores. Given the reality of test reliability, it is not surprising that the Congressional Budget Office noted, “One indication of the limitations of standardized results is often marked disparities in the results they yield.”

Analysis of test reliability leads to one inescapable conclusion: No test has reliability high enough that it can be used as the sole or primary basis for making decisions about students.

**Test Administration**

Educators, researchers and members of the public generally assume that standardized tests are administered in a standardized context under relatively uniform conditions. Anastasi emphasizes the importance of such a controlled setting: “Even apparently minor aspects of the testing situation may appreciably alter performance... In general, children are more susceptible to examiner and situational influences than are adults; in the examination of preschool children, the role of the examiner is especially crucial.”

In fact, recent research has demonstrated that tests are administered in far from “standard” conditions. One study concluded that “the actual context (of test administration) often includes confusion, anxiety, behavioral resistance, negative attitudes toward testing on the part of staff and students, lack of properly trained test examiners, developmentally or educationally immature children, and other institutional problems that are endemic to many schools.” Because reliability coefficients do not take into account the possibility that tests are administered under such variable situations, the reliability of test scores, particularly for tests administered to very young children, is likely to be lower than estimates computed from controlled studies and reported by test publishers. Moreover, many of the ideal “standard” conditions called for by test developers may actually place certain groups of students at a disadvantage. For example, the use of unfamiliar test examiners reduces test scores of low socio-economic status (SES) and black students, but does not affect the scores of high SES students. This factor alone can account for half of the difference in I.Q. test scores between low and high SES students. Similarly, the time limitations associated with most standardized tests harm minorities and women. Generally, these groups appear to cope less effectively with the pressures inherent in a time-limited test than do white males.
Fallout from the Testing Explosion

The validity of any standardized test depends entirely on its context.

Test Validity

"A test," write Airasian and Madaus, "is a sample of behaviors from a domain about which a user wishes to make inferences.... Test validity involves an evaluation of the correctness of the inferences about the larger domain of interest." Thus, the validity of a standardized test tells us whether the test measures what it claims to measure, how well it measures it, and what can be inferred from that measurement. Test validity cannot be measured in the abstract. It can only be determined in the context of the specific uses to which a test's results will be put. Thus, information and conclusions regarding a test's validity in one context may not be relevant and applicable in different contexts.

Like reliability, the term "validity" encompasses several concepts:

- **Content-related validity** determines whether the test questions relate to the trait or traits the test purports to measure.

- **Criterion-related validity** compares test performance (for example, on a reading test) against a standard that independently measures the trait (such as reading ability) the test purports to measure. Criterion validity takes two forms, concurrent and predictive. Concurrent validity describes how well test results coincide with another measure of the same trait made at the same time. Predictive validity examines how well the test forecasts performance on another measure of the tested ability when assessed at a later date.

- **Construct-related validity** examines how well a test actually correlates with the underlying theoretical characteristics of the trait it purports to measure. For example, does the test accurately measure "academic ability" or "competence." This form of validity is rarely reported by test makers even though it is essential for assessing how useful and accurate a test will be in practice.

The types of evidence required to demonstrate a test's validity will differ depending on how test results are to be used. For example, an achievement test used to determine how well students have learned math would require content-related validity. If used to predict students' future math performance, the same test would also require criterion-related predictive validity. A test should only be used for purposes for which it has been adequately validated. Too often, test are used invalidly.
To be content valid at the level of sophistication of the appropriate domain, the test must adequately include what the domain covers. This is so difficult to do within the multiple-choice format that it is, essentially, not done (see discussion below, pp. 21-22).

Lack of adequate content validity can have wide-ranging effects. For example, if a U.S. history test only measures factual recall and the test is used to guide curriculum (as is increasingly the case), then not only will most of the real content of history not be measured, it will be excised from the curriculum.

Test developers (both commercial institutions and governmental agencies) generally validate a test’s use by asking subject area experts to make qualitative judgments about the relationship between individual test items and the trait(s) the test seeks to measure. The selection of test items typically is done by panels of experts who review textbooks for content, draft items and then review them—a method occasionally referred to as BOGSAT (Bunch Of Guys Sitting Around a Table). Essentially, the subjective views of individuals are aggregated to design a test whose content is labelled “objective” and comprehensive.

Because each test item must be one that reasonably should be on the exam, experts are asked whether the item should be included. This is a simple, affirmative format. However, what content validity studies should examine are disconfirming hypotheses: What is not included? Is the overall balance of the items adequate to cover the content? Given the limited number of questions, is the content range a fair approximation of the domain? Such disconfirming questions generally are not asked during test development.57

Many test developers do not go beyond content-related validity.58 For example, the widely-used and highly-respected Iowa Test of Basic Skills “is somewhat lacking when it moves beyond content validity into other validity realms.”59 Similar comments are made by reviewers of other standardized achievement tests.60

Test developers who do go beyond content-related validity generally rely upon other tests to demonstrate criterion-related and construct-related validity. For example, Mitchell demonstrated the predictive validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis by correlating scores on those tests with scores on the Stanford Achievement Test. However, she failed to explain what the Stanford Achievement Test measured and how validly it did so.61
Fallout from the Testing Explosion

Another approach to demonstrating criterion-related validity relies upon comparisons of test scores with teachers' grades. This, however, undermines a major selling point of standardized tests— that they are an objective substitute for overly subjective teacher judgments.62

If test scores and grades agree completely, then why have the tests at all? If they differ significantly, which is better and how do we know? The simple claim of test “objectivity” is insufficient. The question, then, is whether the test is more reliable and valid than are teachers' judgments or some other plausible measure of ability or achievement.63 This last point is important because test-makers will argue that even with low validity, tests can improve decision-making as compared with pure chance. However, teacher judgments and other high-quality alternatives are not decisions equivalent to pure chance.

Validity, like reliability, can be measured by statistical methods which produce numbers called validity coefficients. For many standardized multiple-choice tests, validity coefficients are quite low, and even high coefficients can still result in significant margins of error. “Although various readiness tests are correlated with later school performance, predictive validities for all available tests are low enough that 30 to 50 percent or more of children said to be unready [for first grade] will be falsely identified.” The predictive validity of many developmental screening tests, exams that purport to indicate children’s possible disabilities, is also low and often determined by comparing one test with another.65

Another major concern about predictive validity is whether the performance predicted is, in fact, created through a self-fulfilling prophecy. If a child scores low on a test and is then placed in a program in which he or she is not challenged and does not progress, is a low score on a subsequent achievement test a measure of the accuracy of the first test or only the result of placement in a slow track? Predictive validity studies do not address this question.

The constructs underlying tests are often entirely outdated, as is the case with IQ, readiness and achievement tests that are based on early twentieth-century psychology. Often, a test will purport to measure one thing when, in fact, it measures another. Deborah Meier, Principal of Central Park East Secondary School in Manhattan, argues that reading tests do not measure reading but rather measure “reading skills,” such as phonics decoding, which is not the same thing.66 That is, the tests are based on a faulty understanding of reading and learning to read. As Jerome Harste puts it, “Standard-
ized testing is hopelessly out of date given what we know about reading."

What is true of reading is true across the board: the tests do not explain how people learn nor what they know beyond a very narrow and limited area. The tests also misidentify the content and the construct they purport to measure. As a result, the real knowledge and abilities of our students are not measured.

This is true not only when standardized exams are used to test individuals, but also when used to assess programs. As Airasian and Madaus write, "Are traditional standardized achievement tests construct valid in terms of inferences about school or program effectiveness? In general, the answer is no." The lack of construct validity has a direct impact on teaching when curriculum becomes dominated by testing.

In the work of leading psychometric theoreticians, construct validity has become the essential core of validity, subsuming content and criterion validity. In large part, this is because questions about the meaning of the content or the effects of the prediction enter the realm of underlying hypotheses, theories and assumptions. Tests are not constructed and used independent of theories of knowledge, ability and performance, as well as theories about the domain to be measured. (For example, the domain of history must be conceptualized to provide a construct that can be measured.) The relationships among theories, tests and test use should be examined as part of construct validity studies. Typically, as indicated above, either the constructs are not considered at all or they are woefully inadequate or outdated.

Messick, among others, has argued that the construct validity of a test cannot be considered outside of social or educational values and the consequences of its use. Though Messick himself maintains that we must distinguish between those adverse social consequences of test use that are attributable to the test and those that are not, this expansion of the concept of construct validity opens the entire enterprise of testing to serious question. If the general social results of testing are harmful, and the harm can be traced to the nature and structure of the tests, then testing should, by its own terms, be rejected as lacking in validity.

From this perspective, reliance on the multiple-choice format itself limits the construct validity of many tests. For example, this format appears fundamentally incapable of adequately measuring what are now commonly referred to as "higher order thinking..."
Fallout from the Testing Explosion

The multiple-choice format itself limits the construct validity of many tests. For example, this format appears fundamentally incapable of adequately measuring what are now commonly referred to as "higher order thinking skills." In part, this is because the format requires that there be only one right answer. Some questions on a test may have, by accident, more than one correct answer. But more fundamentally, the multiple-choice format precludes the possibility of designing problems that have more than one correct response or more than one method of solution. Yet problems of these types are necessary to reflect the very essence of complex thinking. Further, multiple-choice questions are carefully constructed and defined, while real-world problems are usually messy and ill-structured. Nor are multiple-choice tests useful for explaining how students think as they solve problems. Multiple-choice testing is not based on evaluating real work done under real conditions; that is, it is not based on students' authentic performances.

In general, the content, criterion and construct validity of standardized tests is so limited as to make most test use invalid. To statisticians or test developers, the reliability and validity of their tests may seem adequate. However, the degree of error remains so high that any decision made about a child based on a single instrument risks harmful consequences. The content of the tests tends to be extremely narrow, due in large part to the multiple-choice format. The predictive validity of the tests often resides in circular logic or self-fulfilling prophecy. And test construct validity is generally absent or based on outdated theories of learning and knowledge. Because of the enormous educational damage caused by reliance on standardized tests (discussed below), it should now be up to the test-makers to prove that any extensive use of their instruments is valid.

Finally, Johnston argues that the philosophy of science underlying the very concept of validity presumes a model of education in which the student and the teacher are both objects. This model, he charges, disempowers student and teacher, with detrimental effects to both as well as to education and society. What is needed, he concludes, is a different conception of science connected to a fundamentally different educational practice—different values, different consequences, and a different conception of what is valid.
Summary

Even a cursory examination of current standardized tests reveals many flaws:

- Race, class, linguistic and gender bias may lead to incorrect scores and thus incorrect treatment of many groups and individuals.

- The assumptions about "intelligence" and how people learn are outdated and designed to fit statistical procedures, not reality.

- Norming is often inadequate or misleading and the norms are frequently out-of-date.

- The reliability of many tests is too low for accurate educational decision-making, particularly for young children.

- Variations in test administration can reduce test reliability, particularly for minorities and low-income children.

- The multiple-choice format is incapable of measuring higher order skills, thinking or creativity.

- Although proper test use requires different types of validity studies, test-makers generally perform only a limited content-related validity study. Validity theory itself now calls into question the validity of the entire testing enterprise.

Taken as a whole, standardized tests do not measure much, and what they do measure, they measure inadequately.

Taken as a whole, standardized tests do not measure much, and what they do measure, they measure inadequately. Thus, tests should be used only with great caution. Unfortunately, in too many American schools that is simply not the case.
III. THE IMPACT OF TESTING ON THE PUBLIC SCHOOLS

"The adverse impact of an educational system based on tests may be far worse than anticipated. A lesson is to be learned from an English experiment in 1863 which utilized tests as a basis for accountability... What transpired was nothing short of disaster... Almost all courses that were not addressed in the test were dropped and reading materials were limited to those that appeared on the tests... Some teachers quickly left the system and prospective teachers were reluctant to enter a profession that operated in this manner. Needless to say 'payment by results' was abandoned in fairly short order."78

—Mary Dilworth

Traditionally, standardized tests have been one of several educational tools used to assess student achievement and to diagnose their academic strengths and weaknesses. In recent years, American public schools, like their British counterparts more than a century ago, have begun to treat standardized tests as the all-purpose answer for promoting educational improvement and ensuring school accountability. In the process, standardized tests have become the primary or sole criterion used by public schools for making a number of decisions affecting students, teachers and schools. In many schools, standardized tests serve as gatekeepers for:

- assignment to special education or remedial programs;
- admission to gifted and talented or accelerated programs;
- grade promotions;
- high school graduation;
- merit pay awards to teachers;
- teacher certification and recertification;
- allocation of funds to schools or school systems; and
- school system certification and decertification.

Reliance on standardized tests as educational gatekeepers is growing. In just two years, 1986-88, the number of states using tests to determine student promotions increased from 8 to 12. Similarly, the number of states using tests to determine eligibility for high school graduation increased from 15 to 24.
Given the limited range of skills and knowledge measured by standardized tests, the impact of race, ethnicity, income and gender on test results, and questions regarding their proper construction, validation and administration, the use of standardized tests as the primary or sole criteria for making any “high stakes” decision is reckless. Moreover, as standardized tests have become the all-powerful gatekeepers of American education, they have affected educational goals and curriculum, student progress and achievement and local control — and created a new set of problems in each area.

Impact on Educational Goals and Curriculum

Children go to school not just to learn basic academic skills, but also to develop the personal, intellectual and social skills to become happy, productive members of a democratic society. Unfortunately, the current emphasis on standardized tests threatens to undermine this educational diversity by forcing schools and teachers to focus on quantifiable skills at the expense of less easily quantifiable academic and non-academic abilities.

This is particularly true for young children. As the National Association for the Education of Young Children (NAEYC) recently noted: “Many of the important skills that children need to acquire in early childhood — self-esteem, social competence, desire to learn, self-discipline — are not easily measured by standardized tests. As a result, social, emotional, moral, and physical development and learning are virtually ignored or given minor importance in schools with mandated testing programs.”

Many schools have embarked on a single-minded quest for higher test scores even though this severely narrows their curriculum:

- Deborah Meier noted that when synonyms and antonyms were dropped from the New York City test for word meaning, teachers promptly dropped academic material that stressed them. She also noted that students read “dozens of little paragraphs about which they then answer multiple-choice questions” — an approach that duplicates the form of the standardized tests the students take in the spring.

- Gerald Bracey, former Director of Research, Evaluation and Testing in the Virginia Department of Education, noted that some teachers did not teach their students how to add and subtract fractions because the state’s minimum competency test included ques-
Fallout from the Testing Explosion

Test preparation can begin months prior to the test.

"The curriculum falls in line with the test, and the test becomes the curriculum."
--P.S. Hlebowitsch

- In one Georgia school, the goal in essay writing is to produce "a five-paragraph argumentative essay written under a time limit on a topic about which the author may or may not have knowledge, ideas, or personal opinions." Not surprisingly, this exercise exactly matches the requirement for the Georgia Regents' Test essay exam.

Sometimes, the curriculum is narrowed simply because "testing takes time, and preparing students for testing takes even more time. And all this time is time taken away from real teaching." In fact, test preparation can begin months prior to the test. Susan Harman of the American Reading Council has observed classrooms discarding all other curriculum in January to prepare for April testing in New York City. The study by Smith, et al., shows schools in Arizona focusing on test preparation in January for April testing. A separate study for the Arizona Department of Education showed the results of the focus on the multiple-choice tests: three-fourths of the state curriculum remained untaught because it was not covered by the tests. The essential reason appeared to be the pressure teachers felt to raise test scores.

Unfortunately, a closer link between tests and curriculum has become a very conscious goal for some school systems. School systems in at least 13 states and the District of Columbia are seeking to "align" their curriculum so that students do not spend hours studying materials upon which they will never be tested regardless of the value or benefits which could be derived from that effort. As a result, curriculum alignment "subordinates the process of curriculum development to external testing priorities... Thus, the curriculum falls in line with the test, and, for all intents and purposes, the test becomes the curriculum." The psychometric approach, rooted in outmoded psychology, thus comes to control education.

The educational price paid for allowing tests to dictate the curriculum can be a high one. Julia R. Palmer, Executive Director of the American Reading Council, recently wrote, "[T]he major barrier to teaching reading in a common-sense and pleasurable way is the nationally normed standardized second grade reading test." Ms. Palmer explained that the test questions force teachers and students to focus on "reading readiness" exercises and workbooks in their early grades and not on reading. As a result, many students become disenchanted with reading because they rarely get a chance to participate in it or to read anything of real interest to them.
Mathematics instruction has also been harmed by the emphasis on testing. Constance Kamii reports that the tests are unable to distinguish between students who understand underlying math concepts and those who are only able to perform procedures by rote and are thus unable to apply them to new situations. Focusing instruction on teaching to the test, therefore, precludes teaching so that children grasp the deeper logic. The National Council of Teachers of Mathematics has concluded that unless assessment is changed, the teaching of math cannot improve.

In general, test control over the curriculum narrows education, "dummies-down" learning materials and instruction, and makes school less interesting to both students and teachers. This narrowing of curriculum is a virtually unavoidable by-product of emphasizing instruments of limited construct validity that utilize a multiple-choice format. As teaching becomes test coaching, real learning and real thinking are crowded out in too many schools.

Just as curricula have been narrowed, so too have textbooks. Diane Ravitch argues that "textbooks full of good literature began to disappear from American classrooms in the 1920's, when standardized tests were introduced. Appreciation of good literature gave way to emphasis on the 'mechanics' of reading." Similarly, a recent report by the Council for Basic Education concluded that the emphasis on standardized tests and curriculum alignment are among the main causes of the increasingly poor quality of textbooks. The report noted that "instead of designing a book from the standpoint of its subject or its capacity to capture the children's imagination, editors are increasingly organizing elementary reading series around the content and time of standardized tests... As a result, much of what is in the textbooks is incomprehensible." As Goodman, et al. point out, many textbooks have end-of-chapter or end-of-section exams that are badly made standardized, multiple-choice tests, and these exams define the content of the text and control how the material is taught. These tests may be even more prevalent than regular standardized tests.

Finally, by narrowing the curriculum, standardized tests are undermining many of the most important goals of the current school improvement movement. Recent education reform efforts have sought to promote "higher-order thinking skills," imagination and creativity in American students. Yet standardized tests focus on basic skills, not critical thinking, reasoning or problem-solving. They emphasize the quick recognition of isolated facts, not the more profound integration of information and generation of ideas.
Fallout from the Testing Explosion

Teachers, in general, are not pleased with the massive increases in testing.

Several studies have demonstrated that "teaching behaviors that are effective in raising scores on tests of lower-level cognitive skills are nearly the opposite of those behaviors that are effective in developing complex cognitive learning, problem-solving ability, and creativity." As Linda Darling-Hammond of the Rand Corporation concluded, "It's testing for the TV generation — superficial and passive. We don't ask if students can synthesize information, solve problems or think independently. We measure what they can recognize." 100

Teachers, in general, are not pleased with the massive increases in testing. Research has shown that teachers believe that too much time is spent on testing and that tests don't measure student achievement very comprehensively, are not accurate for minority students, are not necessary for placement decisions, and are not instructionally useful. Further, most teachers believe that teaching to the test is educationally wrong, but has become necessary due to the pressure to raise test scores. 101

Impact on Student Progress and Achievement

By controlling or compelling student placement in various educational programs, standardized tests perpetuate and even exacerbate existing inequities in educational services, particularly for minority and low-income students. Thus, standardized test results lead to larger numbers of racial and ethnic minorities being placed in special education and remedial programs. For example, national data shows that African-American children are three times as likely as white children to be placed in special education programs. 102 A number of sources have reported to FairTest that school districts are placing ever more children in special education programs so that their test scores are not counted in the school or district averages. 103

Standardized tests also perpetuate the domination of white upper-middle class students in "advanced" classes. In New York City for example, IQ tests are used in some districts to place children in "gifted and talented" programs, creating white, upper-middle class enclaves in districts whose enrollment is dominated by racial and ethnic minorities. 104 Similarly, test results assign boys to advanced math and science programs and keep girls out.

At the same time, standardized tests, particularly when used as promotional gates, can act as a powerful exclusionary device — again aimed disproportionately at minority and low-income students. In the end, they both narrow the educational opportunities available to many segments of our student population and maintain the isolation of different racial and social groups and classes. 105 For
example, academic research has demonstrated that, for a student who has repeated a grade, the probability of dropping out prior to graduation increases by 20 to 40%. Thus, students who are not promoted because they have failed an often unreliable, invalid and biased standardized test are more likely to become high school dropouts.

The impact of standardized tests is particularly devastating when used to determine readiness for first grade. These tests are among the least valid and reliable and are among the most difficult to administer under relatively uniform conditions. Moreover, Shepard and Smith, after examining 14 controlled studies on the effects of kindergarten retention, concluded that retention provided no increase in subsequent academic achievement while imposing a significant social stigma on the retained students.

Nor does the use of standardized tests affect only low-achieving students. High-achieving students are likely to be frustrated by a narrowed curriculum, which has been “dumbed down” in response to standardized tests, particularly minimum competency tests. These students, too, are likely to drop out.

One of the most insidious effects of the overuse of standardized tests is on teachers’ perceptions of their students. The existence of a “Pygmalion effect” as it relates to test results has long been a source of controversy. However, a 1984 study by Stephen Raudenbush has carefully documented its existence for students entering a new school (in this case, 7th grade students entering junior high school). Where teachers have little information on students, conclusions about student knowledge, skills, and abilities based on often inaccurate and unreliable test results can become self-fulfilling prophesies.

Standardized tests are also used to determine pupil placements within regular programs, a practice often termed “tracking.” Low-scoring students are placed in slower tracks while high-scoring students are placed in faster tracks. In the slower tracks, students receive a watered-down curriculum at a slower pace. In lower tracks, students also are far more apt to receive instruction that focuses on rote and drill. Allington has documented the differences in reading instruction in various groups in which children in lower groups do not read much, and similar problems have been noted in other subject areas.

Once tracked into a slow group, a student usually falls further and further behind those in higher groups, in terms both of content and skills learned and of test scores.
Fallout from the Testing Explosion

The negative effect testing has on the curriculum damages low-income and minority students most of all.

and skills learned and of test scores. This problem is even more acute for those labeled “learning disabled.” One consequence is that in the effort to increase test scores in the short run, students in lower tracks receive ever more instruction that resembles test-taking. That is, their schooling becomes reduced to test coaching.

Students in the slower tracks are disproportionately from low-income or minority-group backgrounds. This has led to segregation in many schools.\textsuperscript{112} Minority and low-income students are administered biased and invalid tests, tracked into special education or slow programs, taught what often amounts to no more than test-preparation, and fall further behind their peers. The negative effect testing has on the curriculum damages low-income and minority students most of all. Their subsequent low test scores “confirm” the predictions made by earlier tests, and the growing test-score differentials are used to justify policies of tracking and retention.

Increasing evidence has shown that tracking, like retention, is rarely helpful. Students in the slower groups are clearly harmed by the process, but students in faster groups do not necessarily gain. Multi-ability classrooms would be preferable for low-achievers and not harm high achievers.\textsuperscript{113}

Impact on Local Control

Because standardized tests increasingly determine what is taught in the classroom, parents and other citizens are losing their traditional control over the public schools. This shift of power from local communities to state and national government reduces the level of input and influence available to both parents and teachers in the management of the schools. This, in turn, reduces “the responsiveness of schools to their clientele and so reduces the quality of education” available in those schools.\textsuperscript{114}

Local control over the schools is also being lost to private organizations, namely the test developers. Because of the influence of testing on curricula and instruction, test-makers are effectively dictating the content and form of education. Test publishers and textbook publishers are often divisions of the same company, and both are increasingly owned by international corporations. One result is “greater centralization, tighter control ... [and] less accountability ... as the channels of intellectual thought are controlled by fewer and fewer publishers.”\textsuperscript{115}

Despite the significant and growing role their products play in educational decisions, test manufacturers face little government
regulation or supervision. Unlike other businesses, such as communications, food & drugs, transportation, and securities, there are virtually no regulatory structures at either the federal or state level governing the billion dollar a year testing industry.

States and school districts have neither the expertise nor the resources to independently develop and validate the standardized tests that they need. Instead they turn to private testing companies, who design and market a tremendous variety of products. Even here, states and school systems have neither the skills nor the funds to adequately investigate claims by test developers regarding test validation or to review the test validation process.116

Even if the expertise and resources did exist, the secrecy which is rampant in the testing industry would likely prevent any effective outside evaluation. As the late Dr. Oscar K. Buros (editor of the Mental Measurement Yearbook) lamented, "It is practically impossible for a competent test technician or test consumer to make a thorough appraisal of the construction, validation, and use of standardized tests...because of the limited amount of trustworthy information supplied by the test publishers."117

Testing: An Invalid Enterprise

In sum, current standardized, multiple-choice tests are severely flawed instruments. Their overuse and misuse cause substantial individual and social harm. Many factors contribute to these problems:

• Test-makers make assumptions about human ability that cannot be proven but that lead directly to harmful educational assumptions and practices.

• No test is sufficiently reliable to be used as the sole or primary criteria for decision-making, but such decisions are made constantly.

• The content validity of tests is inadequate because the tests cannot measure the complex material contained in most learning or performance domains.

• Predictive criterion validity is too low to use tests as the sole or primary criteria for decision-making. The limited degree of validity that does exist often results from self-fulfilling prophecies.

• The construct validity of tests is also inadequate: tests often do not measure the traits they claim to measure or do so only poorly.

Fallout from the Testing Explosion

There are virtually no regulatory structures at either the Federal or State level governing the billion dollar a year testing industry.

"It is practically impossible for a competent test technician or test consumer to make a thorough appraisal...because of the limited amount of trustworthy information supplied by the test publishers."

—Oscar K. Buros
Most standardized, multiple-choice-type testing is invalid.

- Standardized exams often fail to accurately measure persons from atypical backgrounds, and test results are used to segregate and devalue persons from minority groups.

- The effects of testing not only cause irreparable harm to many individuals, they also are destructive to the educational process as a whole. Low-income and minority-group students are disproportionately subjected to the poorest, narrowest, most rigidly test-driven curriculum and instruction.

If, as some of testing's foremost theoreticians suggest, the validity of testing is inseparable from its social consequences, then most standardized, multiple-choice-type testing (including I.Q., readiness and most developmental screening testing) is invalid. Continued reliance on standardized testing will prevent necessary school reform. The ongoing domination of testing means that millions of students, predominantly those most in need of improved education, will be dumped into dead end tracks and pushed out of school. To prevent damage and to allow needed reforms, standardized testing must become an occasional adjunct, used for attaining basic but limited information about educational policies, and not a controlling factor over students or curricula.
IV. AN AGENDA FOR TESTING REFORM

FairTest concurs with the National Academy of Education that "information on student progress, wisely interpreted, is of obvious value to the public, to educators and to policy-makers at all levels of government." If properly constructed, validated, administered and used, standardized tests could serve as one limited tool in this effort.

Unfortunately, it has become all too obvious that standardized tests are not properly constructed, validated and administered. Moreover, their widespread use is creating serious problems for students, teachers and the schools themselves. Reliance on standardized tests is now a roadblock to significant reform in curriculum and pedagogy. The question arises then: What should be done to reform tests and test use in the public schools?

In response to the misuse of standardized tests in U.S. society, FairTest has developed an agenda to answer this question. Our Testing Reform Agenda is guided by two essential principles: standardized, multiple-choice testing should be supplanted in most instances by authentic forms of evaluation, and standardized tests that remain must be markedly improved in their construction and use. The FairTest Reform Agenda has four parts:

- It is time to use new methods of measuring student achievement as part of a major reform in schooling as a whole. Standardized multiple-choice tests can only measure a very limited range of student knowledge, abilities and skills. Current standardized tests should become no more than occasional complements to performance-based evaluations. Their use should largely be limited to matrix survey-samples, which will provide any programmatically useful information while limiting their effect on curriculum and preventing their being used to make decisions about individuals.

  Emerging methods, commonly referred to as "authentic" and "performance-based" evaluation (See Appendix B, "What Is Authentic Evaluation?"), provide new opportunities to expand society’s capability to more fully and accurately evaluate a greater range of knowledge, abilities and skills. Authentic evaluations can and should be used to diagnose the strengths and weaknesses of students in order to help them learn, rather than to sort, stratify or segregate them. Further, good assessment can encourage the teaching of challenging and comprehensive curricula in ways that spur
students to serious thinking and greater accomplishments. Today, “teaching to the test” is a synonym for narrowing education; with authentic assessments, “teaching to the test” can become a method of enriching and expanding education.

Testing should be undertaken when it can be directly helpful to student learning; any other reasons must be carefully justified and not allowed to negatively affect instruction. Tests and other forms of educational evaluation, whether performance-based or multiple-choice, should measure meaningful and important knowledge and capabilities possessed by students. Questions must be relevant to the knowledge, abilities or skills being tested. Test items and instructions should be written clearly and accurately. The tests themselves should take into account the diversity of language, experience and perspective embodied in the test-taking population. If necessary, different assessments should be used for different population groups to ensure the elimination of bias. At the same time, questions and scoring procedures should acknowledge the complexity and diversity of intelligence and individual development.

• Test validation should ensure that the content of the test matches the content of what is taught, but test developers cannot stop at content validation. They must document assumptions about the relationship between test results and future performance, and do so in ways that are more than documentation of a self-fulfilling prophecy. At the same time, they must demonstrate that test results are accurately related to the underlying knowledge, skills and abilities the test claims to measure, and that the theoretical conceptions of knowledge and learning used to develop test constructs are accurate. Test companies should not provide sub-scores that lack adequate reliability and validity, nor make recommendations for instructional practice when the tests have not been validated as diagnostic instruments.

Those who develop and use tests must ensure that the testing environment is both consistent for and supportive of all test-takers. Where the environment cannot be made standard and supportive, the only alternative is to refrain from testing. Moreover, the standard environment must not be constructed in a manner that creates disadvantages for particular students through artificial distractions or pressures.

• Public schools, test-takers and independent researchers should have access to the descriptive and statistical data needed to verify test publishers’ claims regarding test construction and validation. This should include the release of questions used on previous tests.
as well as data on test results, identified by race/ethnicity, gender, socio-economic status, geographical residence, and other distinctions.

Test publishers have long argued against the release of old test questions. They have claimed that any large-scale release would require the development of a massive number of new items, thus increasing test development cost. This may not be the case. One study found that the release of old test questions did not affect test scores. Nor has this been a problem in college admissions testing, where tests have been disclosed since 1980. Thus the release of old test questions may not require such a large scale development of new questions. On the other hand, some test publishers also publish test coaching books that contain material nearly identical to the tests. Some critics have called the use of these materials "cheating." Test publishers should not sell coaching materials that undermine the validity of their tests.

Test users or independent public agencies should also fully investigate the claims of test publishers regarding the construction and validity of their tests. Test company materials as well as data to facilitate independent analysis must be made available by test companies. At the same time, test administrators and users should disclose and monitor their own processes for test administration.

• Both test developers and test users should work to ensure that test results are properly interpreted and employed by educators, policymakers, test-takers and the general public. At a minimum, test scores should never be used as the sole or primary factor in "high stakes" educational decisions.

At the same time, test developers and test users must recognize that standardized tests are only limited measures of educational progress. Used alone, they present distorted pictures of what they seek to measure and their use often undermines the quality of education offered in our public schools, particularly the education offered to low-income and minority-group students. Both test developers and test users have the affirmative obligation to promote a proper, reasonable and limited use of standardized tests as one of a series of assessment mechanisms.

Although FairTest believes that those institutions that develop and use standardized tests have a primary obligation to reform tests and test use, government also has a major role to play. By establishing guidelines for the testing industry, requiring information on standardized tests to be made public, and analyzing test results to
Fallout from the Testing Explosion

The price which has been paid by our schools and our children for this infatuation with tests is high.

guard against bias, the government can go a long way toward improving the quality of tests and test use. More importantly, public agencies can set the standard for intelligent and proper use of tests.

Unfortunately, too many policymakers and educators have ignored the complexities of testing issues and the obvious limitations they should place upon standardized test use. Instead, they have been seduced by the promise of simplicity and objectivity. The price which has been paid by our schools and our children for this infatuation with tests is high. Unless Americans act now to limit and reform the use of standardized tests in the schools, that price will continue to increase.
NOTES


5 Peterson, B. "Half Million Standardized Tests Given to MPS Students," Rethinking Schools (May-June, 1989), and personal communication from Peterson to author Neill. Limited-English Proficiency students were not included in the total. The Task Force also counted 860,000 end-of-unit basal reader tests; if these are included as standardized tests, then Milwaukee administers nearly 1,350,000 tests per year, or about 14 per student per year. As K. Goodman, et al., report, tests in basal readers are, in general, less well developed but essentially the same thing as standardized multiple-choice achievement tests (Report Card on Basal Readers, Katonah, NY: Richard C. Owen Publishers, 1988).


7 Southern states may be taking the lead in reducing the amount of standardized testing. Over the past several years, North Carolina banned the use of standardized achievement tests in grades one and two and is implementing developmentally appropriate assessments in those grades; Mississippi stopped requiring the testing of young children; Georgia dropped, after one year, a required test for admission to first grade; Kentucky has lifted its requirement for annual testing in every grade and may be turning to performance-based assessments (see Appendix B for discussion of performance-based assessment).

8 Since this was written in 1988, Ohio has also adopted state-wide testing. Vermont is considering performance-based assessments through portfolios, but may also use other methods of testing on a statewide basis.


Fallout from the Testing Explosion


18 The mathematics section of the WISC-R test, for example, includes 8 questions about 13 boys or men who save money on purchases, trade fairly, cleverly divide their efforts and money, and work at jobs, compared to only one question featuring a girl—who loses her hair ribbon. Wechsler, D. *Wechsler Intelligence Scale for Children-Revised* (Psychological Corporation, 1974).

Fallout from the Testing Explosion


26 Meier, D. “Why Reading Tests Don’t Test Reading,” Dissent (Winter, 1982-83). Meier notes that questions correctly answered by “low-performing” students, but not by “high-performing” students are removed from test use as inaccurate measures of ability or achievement. This ignores the possibility that these questions reflect differing language or cognitive styles possessed by “low-performing” students. See also A. Martin “Screening, Early Intervention, and Remediation: Obscuring Children's Potential,” Harvard Educational Review (Vol. 58, No. 4, Nov., 1988) pp, 488-501; in which she points out that the testing and screening processes produce an overemphasis on children’s weaknesses, treat variation as a problem, and disempower teachers.


Fallout from the Testing Explosion


37 Ninth Mental Measurement Yearbook (1985). See reviews of the California Achievement Test, Comprehensive Tests of Basic Skills, Iowa Tests of Basic Skills, Metropolitan Achievement Test, Stanford Achievement Test, and SRA Achievement Series.

38 Cannell, J.J. Nationally Normed Elementary Achievement Testing in America's Public Schools (Daniels, WV: Friends for Education, 1987) p. 6. This report examined the same six tests listed in the previous footnote.


40 Rothman, R. "Physician's Test Study was 'Clearly Right,'" Education Week (April 5, 1989).


Congressional Budget Office (1986) p. 10. A second report by CBO, *Educational Achievement: Explanations and Implications of Recent Trends* (Washington, D.C.: Government Printing Office, August, 1987) p. 10-11, concluded that "one of the most serious mistakes made by some analysts attempting to explain recent achievement trends...has been to assume that patterns evident in the scores of one test will appear in others as well."

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing* (Washington, D.C.: APA, 1985), Standard 8.12 (p.54): "In elementary or secondary education, a decision or characterization that will have a major impact on a test taker should not automatically be made on the basis of a single test score."


For discussions of this problem as regards the NTE (National Teacher Exam), see: Horner, B. and J. Sammons. *The Test That Fails: An Analysis of the National Teachers Examination in New York* (New York: NYPRG, 1987) pp. 4-6; L. Darling-Hammond. "Teaching Knowledge: How Do We Test It?" *American Educator* (Fall, 1986) p. 88. FairTest recently received a questionnaire circulated by ETS asking respondents to indicate whether certain content areas ought to be on the NTE. Not only did the questionnaire demonstrate continued use of the affirmative
hypothesis, but also it demonstrated continuing reliance on using separable bits of "knowledge" that can be reduced to multiple-choice questions.


69 Airasian and Madaus (1983).


72 Messick (1989).


76 Resnick and Resnick (1989).


Fallout from the Testing Explosion


84 Wise, A.E. "Legislated Learning Revisited," Phi Delta Kappan (January, 1988) p. 330. See also, D.W. Dorr-Bremme and J.L. Herman. Assessing Student Achievement: A Profile of Classroom Practices (Los Angeles, CA: Center for the Study of Evaluation, UCLA, 1986) which concluded, based upon a nationally representative sample of 114 school districts, that "preparations for a test can begin days or even weeks before the test is given." They also observed that "for each hour that students spend taking tests, teachers seem to spend two to three more" hours in preparation.

85 Oral communication to author Neill (Jan., 1990).


87 Bishop, C.D. Presentation to the Joint Legislative Committee on Goals for Educational Excellence (Arizona Department of Education, Nov. 9, 1989).


90 Palmer, J.R. Letter to the editor, New York Times (December 14, 1987). See also, J.T. Guthrie Indicators of Reading Education (New Brunswick, NJ: Center for Policy Research in Education, 1988), which concludes that the strengthening of students’ reading skills goes hand-in-hand with finding better ways to measure reading achievement. The primary shortcoming of reading tests is that they don’t reflect the complexity of the reading process. See also Note 66.

91 Kamii (1989).


97 See note 5.
Fallout from the Testing Explosion


103 The city of Boston, for example, does not test some 30 percent of its students. City Wide Education Coalition “CWEC Fact Sheet on Boston Public Schools and Student Achievement” (Boston: CWEC, Nov., 1986).


107 Shepard and Smith (1988) p. 34. See also, Shepard and Smith (1989) and Smith & Shepard (1987).


Fallout from the Testing Explosion


APPENDIX A

DESCRIPTION OF FAIRTEST SURVEY AND RESULTS

Study Methodology

In mid-1987, FairTest staff conducted a series of telephone interviews with officials from all 50 state departments of education, from the District of Columbia school district and from 56 sample school districts in 36 states. Interviews of state officials focused on standardized tests administered by the public schools to fulfill testing mandates established by the state, while interviews of school district officials focused on tests administered to fulfill testing mandates established by the district. All interviews sought responses to three basic questions:

• How many tests were administered by the public schools to fulfill the state or local testing mandates?

• Which standardized tests were used to fulfill the state or local testing mandates?

• For what purposes did the state or school district mandate standardized tests?

The responses collected through these interviews related entirely to the use of standardized achievement, competency or basic skills tests. Public schools also use many other standardized exams, including IQ tests, behavioral tests, readiness tests for young children, and placement tests for special education, remedial education and bilingual education programs. However, the use of these tests varies considerably among schools, even within the same districts, and records of their use appear to be unreliable or nonexistent. Thus the study results reflect only a portion of the standardized tests actually administered to students by the public schools.

Study Results

State-Level Testing Mandates. During the 1986-87 school year, schools in 42 states and the District of Columbia administered nearly 17 million standardized achievement, competency and basic skills tests to almost 36.3 million students to fulfill state testing mandates—a rate of almost one test for every two students. FairTest counted each test battery as one test administration. (If individual tests within test batteries are counted, this number would easily double,
Fallout from the Testing Explosion

to 34 million tests.) This rate varied considerably among the states however. A detailed listing of the number of standardized tests administered by the public schools in each state to fulfill state testing mandates is presented in Table 1.

On average, schools in the South administered standardized tests to fulfill state mandates at much higher rates than schools in the remainder of the nation. Schools in the 11 Southern states (Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee and Virginia) administered more than 6 million tests to over 9.3 million students — a rate of one test for every 1.5 students. Schools in the remaining 31 states administered tests at about half that rate — one test for every 2.5 students. In fact, schools in all but two Southern states (Florida and Virginia) administered tests at a rate higher than the average for the remainder of the nation. Moreover, the four states with the highest rates of test administration were all in the South. Kentucky, North Carolina, Alabama, and Georgia all administered about one test per student per year.

Outside the South, no clear patterns of test use to fulfill state mandates emerged. Although schools in New Mexico, a state with a minority population in the public schools above the national average, reported a high rate of standardized test use, so did schools in Utah, a state with a relatively low minority population. Conversely, schools in states like Texas and California, with relatively high minority populations in the public schools, reported rates almost equal to Wisconsin and Kansas, states with relatively low minority populations.

A clear pattern did emerge among the eight states (Alaska, Iowa, Minnesota, Montana, North Dakota, Ohio, Vermont and Wyoming) which did not have any state testing mandates. Seven of these eight states have minority student populations significantly below the national average. While minorities make up more than one-quarter of public school students nationwide, the minority student populations in these seven states range from a high of 15% in Ohio to a low of 1% in Vermont. Most in fact have minority populations that are less than 10% of the total student population. The one exception is Alaska, with a large Native American population, and an overall minority population just above the national average. However, Alaska will also implement a state testing mandate beginning in the 1987-88 school year.
District-Level Testing

During the 1986-87 school year, schools in the 56 sample school districts administered more than 7.8 million standardized tests to over 5.7 million students to meet local testing mandates — a rate of one and one-third tests for every student. (Again, FairTest counted only whole batteries, not tests within batteries.) However, the rate of test administration among districts varies even more than it did among different states. A detailed listing of the number of standardized tests administered by the public schools in each of the sample school districts to fulfill local testing mandates is presented in Table 2.

The schools in Newark, New Jersey are excluded from the discussion in the following paragraph due to their extremely high reliance on standardized tests. During the 1986-87 school year, schools in Newark administered over 600,000 standardized tests to about 67,000 students — a rate of more than 9 tests per student. This rate was more than three times that of the next highest school system.

Excluding the Newark school system, the rate of test administration ranges from a high of almost 3 tests per student in Cleveland to a low of only one test per 12.5 students in Fairfax County, Virginia. Overall, six districts reported administering more than 2 tests per students per year, while seven others reported administering less than 1 test for every 2 students.

To analyze the variations among the different rates of tests administered by the different districts, districts were categorized by size. Three categories were created: large districts (with student populations exceeding 100,000); medium-sized districts (with populations between 25,000 and 100,000); and small districts (with populations below 25,000). The results from Newark were excluded from these calculations. On average, large districts administered standardized tests to fulfill local testing mandates at a rate (1.38 tests per student) 25% higher than that of medium-sized districts (1.11 tests per student). The rate of medium-sized districts, in turn, exceeded that of small districts (0.88 tests per student) by almost the same proportion.

The estimate that schools administered almost 38.9 million standardized achievement, competency and basic skills tests to 39.8 million students during the 1986-87 school year to fulfill local testing mandates is based upon these average variations and the distribution of students among school districts of different sizes.
(See Table 3 for the detailed computations.) Combining this figure with the tests administered to fulfill state testing mandates produces an estimate of 55.7 million standardized achievement-type tests administered by the public schools during the 1986-87 school year, or 1.4 tests for every public school student. (If we counted each test, not each battery, the total could approach three per student per year.)

**Additional Surveys on Test Use**

Although the FairTest survey focused on the use of standardized achievement, competency and basic skills tests in the public schools, public schools also use standardized tests for a variety of other purposes. These include:

- screening and readiness tests administered to kindergarten and pre-kindergarten students;

- tests administered to gifted, disadvantaged, handicapped and limited-English proficient students for placement into or graduation from gifted & talented, compensatory education, special education and bilingual education programs;

- tests administered to randomly selected samples of students as part of the U.S. Department of Education's National Assessment of Educational Progress.

- admissions tests administered to students seeing to enroll in particular secondary schools;

- college admissions tests administered to high school juniors and seniors;

- GED (General Education Development) tests administered to individuals who did not complete high school;

In addition, some school systems continue to administer IQ tests to their entire student populations, although most school districts administer IQ tests only to some individuals.

As noted previously, information on the number of other standardized tests administered was neither as specific nor as reliable as the data gathered for achievement, competency and basic skills tests. As a result, it is not possible to compute specific totals on the use of these tests. From a variety of sources, however, general figures can be obtained:
Fallout from the Testing Explosion

- Test publishers have reported that the total number of college and secondary school admissions, GED and NAEP tests administered to elementary and secondary school students was between 6 and 7 million.

- A survey conducted in the early 1980's indicated that students in compensatory education and special education programs were tested two to three times as often as their peers in mainstream education program. Since the mainstream student is administered 1.4 standardized tests per year (according to our survey), compensatory and special education students are administered between 3 and 4 standardized tests per year. Given that over 10 million students participate in federally-funded compensatory and special education programs, an estimated 30 to 40 million additional standardized tests are administered to these students. (If individual tests rather than tests within batteries are counted, the numbers could double to 6-8 tests per students per year for a total of 60-80 million tests.)

- A survey conducted in 1985 concluded that almost half of the kindergarten and pre-kindergarten students in the public schools were administered screening tests. Based upon the 1985 kindergarten enrollment, this means that between 1.5 and 1.75 million tests were administered to these children. This total (which still excludes tests administered to gifted and limited-English proficient students and some proportion of I.Q. testing) yields an additional 37.5 to 48.75 million tests.

Summary

The results of the FairTest survey revealed that at least 55.7 million standardized achievement-type tests were administered to public school students in the 1986-87 school year. An additional 37.5 to 48.75 million standardized tests were administered by the public schools to their students for other purposes during that year. Based on these two estimates, between 93 million and 105 million standardized tests were administered to 39.8 million students in the public schools. If individual tests within batteries are counted, the total would increase dramatically. Looked at in this way, and including the uncounted IQ and other tests, the number of standardized tests administered yearly could approach 200 million, close to 5 per year per student.
NOTES ON SURVEY

1 Harris J., and J. Sammons. *Failing our Children* (New York: New York Public Interest Research Group, 1989). This survey revealed that tests validated for one purpose are often used for other purposes for which the publisher claims no validity; e.g., achievement tests may be used to determine school readiness or even to screen for disabilities.

2 See Chapter I, "Test Use in U.S. Schools," for some changes that have occurred since 1987.

3 In fact, recent evidence indicates that Newark may not be unusual—see Note 5 to the main text.


<table>
<thead>
<tr>
<th>STATE</th>
<th>NUMBER OF TESTS</th>
<th>SCHOOL ENROLLMENT</th>
<th>STATE-MANDATED TESTS PER STUDENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>720,000</td>
<td>733,735</td>
<td>0.98</td>
</tr>
<tr>
<td>Alaska</td>
<td>0</td>
<td>107,973</td>
<td>0.00</td>
</tr>
<tr>
<td>Arizona</td>
<td>480,000</td>
<td>534,538</td>
<td>0.90</td>
</tr>
<tr>
<td>Arkansas</td>
<td>179,000</td>
<td>437,438</td>
<td>0.41</td>
</tr>
<tr>
<td>California</td>
<td>1,420,000</td>
<td>4,377,989</td>
<td>0.32</td>
</tr>
<tr>
<td>Colorado</td>
<td>208,800</td>
<td>558,415</td>
<td>0.37</td>
</tr>
<tr>
<td>Connecticut</td>
<td>99,000</td>
<td>468,847</td>
<td>0.21</td>
</tr>
<tr>
<td>Delaware</td>
<td>60,000</td>
<td>94,410</td>
<td>0.64</td>
</tr>
<tr>
<td>District of Columbia</td>
<td>119,000</td>
<td>85,612</td>
<td>1.39</td>
</tr>
<tr>
<td>Florida</td>
<td>437,352</td>
<td>1,607,320</td>
<td>0.27</td>
</tr>
<tr>
<td>Georgia</td>
<td>1,020,000</td>
<td>1,096,425</td>
<td>0.93</td>
</tr>
<tr>
<td>Hawaii</td>
<td>75,000</td>
<td>164,640</td>
<td>0.46</td>
</tr>
<tr>
<td>Idaho</td>
<td>30,000</td>
<td>208,391</td>
<td>0.14</td>
</tr>
<tr>
<td>Illinois</td>
<td>418,000</td>
<td>1,825,185</td>
<td>0.23</td>
</tr>
<tr>
<td>Indiana</td>
<td>211,000</td>
<td>966,780</td>
<td>0.22</td>
</tr>
<tr>
<td>Iowa</td>
<td>0</td>
<td>481,286</td>
<td>0.00</td>
</tr>
<tr>
<td>Kansas</td>
<td>135,000</td>
<td>416,091</td>
<td>0.32</td>
</tr>
<tr>
<td>Kentucky</td>
<td>650,000</td>
<td>642,778</td>
<td>1.01</td>
</tr>
<tr>
<td>Louisiana</td>
<td>390,000</td>
<td>795,188</td>
<td>0.49</td>
</tr>
<tr>
<td>Maine</td>
<td>48,000</td>
<td>211,752</td>
<td>0.23</td>
</tr>
<tr>
<td>Maryland</td>
<td>168,000</td>
<td>675,747</td>
<td>0.25</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>410,000</td>
<td>833,918</td>
<td>0.49</td>
</tr>
<tr>
<td>Michigan</td>
<td>326,285</td>
<td>1,681,880</td>
<td>0.19</td>
</tr>
<tr>
<td>Minnesota</td>
<td>0</td>
<td>711,134</td>
<td>0.00</td>
</tr>
<tr>
<td>Mississippi</td>
<td>240,000</td>
<td>498,639</td>
<td>0.48</td>
</tr>
<tr>
<td>Missouri</td>
<td>664,000</td>
<td>800,606</td>
<td>0.83</td>
</tr>
<tr>
<td>Montana</td>
<td>0</td>
<td>153,327</td>
<td>0.00</td>
</tr>
<tr>
<td>Nebraska</td>
<td>20,508</td>
<td>267,139</td>
<td>0.08</td>
</tr>
<tr>
<td>Nevada</td>
<td>55,000</td>
<td>161,239</td>
<td>0.34</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>30,000</td>
<td>163,717</td>
<td>0.18</td>
</tr>
<tr>
<td>New Jersey</td>
<td>630,000</td>
<td>1,107,467</td>
<td>0.57</td>
</tr>
<tr>
<td>New Mexico</td>
<td>80,000</td>
<td>281,943</td>
<td>0.28</td>
</tr>
<tr>
<td>New York</td>
<td>1,691,000</td>
<td>2,607,719</td>
<td>0.65</td>
</tr>
<tr>
<td>North Carolina</td>
<td>1,085,000</td>
<td>1,085,248</td>
<td>1.00</td>
</tr>
<tr>
<td>North Dakota</td>
<td>0</td>
<td>118,703</td>
<td>0.00</td>
</tr>
<tr>
<td>Ohio</td>
<td>0</td>
<td>1,793,508</td>
<td>0.00</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>234,000</td>
<td>593,183</td>
<td>0.39</td>
</tr>
<tr>
<td>Oregon</td>
<td>15,000</td>
<td>449,307</td>
<td>0.03</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>538,212</td>
<td>1,674,161</td>
<td>0.32</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>40,000</td>
<td>134,126</td>
<td>0.30</td>
</tr>
<tr>
<td>South Carolina</td>
<td>450,000</td>
<td>611,629</td>
<td>0.74</td>
</tr>
<tr>
<td>South Dakota</td>
<td>5,000</td>
<td>125,458</td>
<td>0.40</td>
</tr>
<tr>
<td>Tennessee</td>
<td>503,000</td>
<td>818,073</td>
<td>0.61</td>
</tr>
<tr>
<td>Texas</td>
<td>1,500,000</td>
<td>3,209,515</td>
<td>0.47</td>
</tr>
<tr>
<td>Utah</td>
<td>258,000</td>
<td>415,994</td>
<td>0.62</td>
</tr>
<tr>
<td>Vermont</td>
<td>0</td>
<td>92,112</td>
<td>0.00</td>
</tr>
<tr>
<td>Virginia</td>
<td>377,000</td>
<td>975,135</td>
<td>0.39</td>
</tr>
<tr>
<td>Washington</td>
<td>182,000</td>
<td>761,428</td>
<td>0.24</td>
</tr>
<tr>
<td>West Virginia</td>
<td>200,000</td>
<td>351,837</td>
<td>0.57</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>309,000</td>
<td>767,819</td>
<td>0.40</td>
</tr>
<tr>
<td>Wyoming</td>
<td>0</td>
<td>100,955</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>UNITED STATES</strong></td>
<td>16,753,157</td>
<td>39,837,459</td>
<td>0.42</td>
</tr>
</tbody>
</table>
### Table 2. Standardized Tests Administered in the Public Schools To Fulfill Local Mandates (1986-1987 School Year).

<table>
<thead>
<tr>
<th>SCHOOL DISTRICT</th>
<th>NUMBER OF TESTS</th>
<th>SCHOOL ENROLLMENT</th>
<th>DISTRICT-MANDATED TESTS PER STUDENT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CALIFORNIA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Los Angeles</td>
<td>949,899</td>
<td>540,903</td>
<td>1.76</td>
</tr>
<tr>
<td>San Juan Unified</td>
<td>33,000</td>
<td>44,186</td>
<td>0.75</td>
</tr>
<tr>
<td>San Francisco</td>
<td>21,000</td>
<td>58,378</td>
<td>0.36</td>
</tr>
<tr>
<td>San Diego City Unified</td>
<td>34,040</td>
<td>110,631</td>
<td>0.31</td>
</tr>
<tr>
<td><strong>COLORADO</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jefferson County</td>
<td>52,500</td>
<td>76,000</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>CONNECTICUT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greenwich</td>
<td>5,720</td>
<td>6,772</td>
<td>0.84</td>
</tr>
<tr>
<td><strong>FLORIDA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hillsborough County (Tampa)</td>
<td>98,000</td>
<td>115,323</td>
<td>0.85</td>
</tr>
<tr>
<td>Pinellas County (Clearwater)</td>
<td>112,000</td>
<td>85,339</td>
<td>1.31</td>
</tr>
<tr>
<td>Duvall County (Jacksonville)</td>
<td>190,000</td>
<td>99,512</td>
<td>1.91</td>
</tr>
<tr>
<td>Dade County (Miami)</td>
<td>230,000</td>
<td>250,000</td>
<td>0.92</td>
</tr>
<tr>
<td>Broward Cty. (Ft. Lauderdale)</td>
<td>336,000</td>
<td>129,478</td>
<td>2.60</td>
</tr>
<tr>
<td><strong>GEORGIA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fulton County (Atlanta)</td>
<td>13,000</td>
<td>35,523</td>
<td>0.37</td>
</tr>
<tr>
<td>DeKalb County (Decatur)</td>
<td>15,000</td>
<td>66,000</td>
<td>0.23</td>
</tr>
<tr>
<td><strong>ILLINOIS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicago</td>
<td>468,544</td>
<td>435,000</td>
<td>1.08</td>
</tr>
<tr>
<td><strong>INDIANA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indianapolis</td>
<td>66,600</td>
<td>50,600</td>
<td>1.32</td>
</tr>
<tr>
<td><strong>IOWA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Des Moines Independent</td>
<td>44,500</td>
<td>30,000</td>
<td>1.48</td>
</tr>
<tr>
<td><strong>KANSAS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wichita</td>
<td>65,729</td>
<td>44,729</td>
<td>1.47</td>
</tr>
<tr>
<td><strong>KENTUCKY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jefferson County (Louisville)</td>
<td>47,000</td>
<td>95,020</td>
<td>0.49</td>
</tr>
<tr>
<td><strong>LOUISIANA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orleans Parish (New Orleans)</td>
<td>84,000</td>
<td>84,000</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>MARYLAND</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baltimore City</td>
<td>275,800</td>
<td>120,000</td>
<td>2.30</td>
</tr>
<tr>
<td>Prince Georges County</td>
<td>175,000</td>
<td>103,000</td>
<td>1.70</td>
</tr>
<tr>
<td><strong>MASSACHUSETTS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boston</td>
<td>60,000</td>
<td>55,000</td>
<td>1.09</td>
</tr>
<tr>
<td>Brookline</td>
<td>3,500</td>
<td>5,400</td>
<td>0.65</td>
</tr>
<tr>
<td><strong>MICHIGAN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detroit City</td>
<td>269,000</td>
<td>200,000</td>
<td>1.35</td>
</tr>
<tr>
<td>SCHOOL DISTRICT</td>
<td>NUMBER OF TESTS</td>
<td>DISTRICT-MANDATED SCHOOL ENROLLMENT PER STUDENT</td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------------</td>
<td>-------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>MINNESOTA Minneapolis</td>
<td>39,712</td>
<td>1.23</td>
<td></td>
</tr>
<tr>
<td>MISSOURI St. Louis</td>
<td>119,327</td>
<td>2.45</td>
<td></td>
</tr>
<tr>
<td>MONTANA Missoula</td>
<td>6,990</td>
<td>1.24</td>
<td></td>
</tr>
<tr>
<td>NEVADA Las Vegas</td>
<td>87,684</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>NEW HAMPSHIRE Concord</td>
<td>7,296</td>
<td>1.46</td>
<td></td>
</tr>
<tr>
<td>NEW JERSEY Newark</td>
<td>603,000</td>
<td>9.00</td>
<td></td>
</tr>
<tr>
<td>NEW MEXICO Albuquerque</td>
<td>133,320</td>
<td>1.67</td>
<td></td>
</tr>
<tr>
<td>NEW YORK New York City</td>
<td>1,133,000</td>
<td>1.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rochester</td>
<td>35,000                                        1.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Buffalo City</td>
<td>63,000                                        1.41</td>
<td></td>
</tr>
<tr>
<td>NORTH CAROLINA Mecklenburg Cty. (Charlotte)</td>
<td>18,000</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wake County (Raleigh)</td>
<td>40,500</td>
<td>0.70</td>
</tr>
<tr>
<td>NORTH DAKOTA Fargo</td>
<td>6,000</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>OHIO Cincinnati</td>
<td>123,800</td>
<td>2.34</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Akron</td>
<td>20,520                                        0.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cleveland</td>
<td>219,000                                       2.92</td>
<td></td>
</tr>
<tr>
<td>OKLAHOMA Oklahoma City</td>
<td>31,300</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>OREGON Portland</td>
<td>30,000</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>PENNSYLVANIA Philadelphia</td>
<td>400,000</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>SOUTH CAROLINA Greenville County</td>
<td>37,000</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>TENNESSEE Memphis City</td>
<td>188,200</td>
<td>1.76</td>
<td></td>
</tr>
</tbody>
</table>

55
<table>
<thead>
<tr>
<th>SCHOOL DISTRICT</th>
<th>NUMBER OF TESTS</th>
<th>DISTRICT-MANDATED TESTS ENROLLMENT PER STUDENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEXAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dallas Independent</td>
<td>217,584</td>
<td>127,584</td>
</tr>
<tr>
<td>Houston Independent</td>
<td>250,000</td>
<td>193,702</td>
</tr>
<tr>
<td>UTAH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salt Lake City</td>
<td>56,000</td>
<td>72,000</td>
</tr>
<tr>
<td>VERMONT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burlington</td>
<td>2,212</td>
<td>3,800</td>
</tr>
<tr>
<td>VIRGINIA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fairfax County</td>
<td>10,000</td>
<td>124,631</td>
</tr>
<tr>
<td>Virginia Beach City</td>
<td>0</td>
<td>54,870</td>
</tr>
<tr>
<td>Prince William County</td>
<td>105,000</td>
<td>57,213</td>
</tr>
<tr>
<td>WASHINGTON</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seattle</td>
<td>41,500</td>
<td>44,000</td>
</tr>
<tr>
<td>WEST VIRGINIA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kanawha County (Charleston)</td>
<td>20,300</td>
<td>37,399</td>
</tr>
<tr>
<td>WISCONSIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Milwaukee</td>
<td>64,500</td>
<td>96,387</td>
</tr>
<tr>
<td>WYOMING</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laramie County (Cheyenne)</td>
<td>11,000</td>
<td>13,000</td>
</tr>
</tbody>
</table>
TABLE 3. COMPUTATION OF ESTIMATE OF STANDARDIZED TESTS ADMINISTERED BY PUBLIC SCHOOLS TO FULFILL STATE OR LOCAL TESTING MANDATES (1986-87 SCHOOL YEAR).

<table>
<thead>
<tr>
<th>TYPE OF SYSTEM</th>
<th>SURVEYED</th>
<th>ESTIMATED</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STUDENTS</td>
<td>TESTS</td>
<td>RATE</td>
</tr>
<tr>
<td>LARGE (OVER 100,000)</td>
<td>3,680,375</td>
<td>5,035,067</td>
<td>1.37</td>
</tr>
<tr>
<td>MEDIUM (25,000 TO 100,000)</td>
<td>2,026,008</td>
<td>2,790,792</td>
<td>1.11</td>
</tr>
<tr>
<td>SMALL (LESS THAN 25,000)</td>
<td>48,812</td>
<td>42,718</td>
<td>0.88</td>
</tr>
</tbody>
</table>

| NOTE | The computation of the testing rate in medium-sized school systems excludes the unusually high testing rate of Newark, New Jersey (which is three times higher than the next highest system.) The estimated number of tests administered in medium and small school systems was computed using the actual rate administered for medium and small systems based on survey results.
Appendix B
What Is Authentic Evaluation?

Authentic evaluation of educational achievement directly measures actual performance in the subject area. Standardized multiple-choice tests, on the other hand, measure test-taking skills directly, and everything else either indirectly or not at all.

Also called “performance,” “appropriate,” “alternative,” or “direct” assessments, authentic evaluation includes a wide variety of techniques, such as written products, solutions to problems, experiments, exhibitions, performances, portfolios of work and teacher observations, checklists and inventories, and cooperative group projects. They may be the evaluation of regular classroom activity or take the form of tests or special projects.

Authentic evaluations indicate what we value by directing instruction toward what we want the student to know and be able to do. They are appropriate to the student’s age and level of learning and the subject being measured, and are useful to both teachers and students.

All forms of authentic assessment can be summarized numerically or put on a scale. Therefore, individual results can be combined to provide a variety of information about aggregate performance at the classroom, school, district, state and national levels. Thus, state and federal requirements for comparable quantitative data can be met.

Authentic assessment was developed in the arts and in apprenticeship systems, where assessment has always been based on performance. It is impossible to imagine evaluating a musician’s ability without hearing her or him sing or play an instrument, or judging a woodworker’s craft without seeing the table or cabinet. It is also impossible to help a student improve as a woodworker or musician unless the instructor observes the student in the process of working on something real, provides feedback, monitors the student’s use of the feedback, and adjusts instruction and evaluation accordingly. Authentic assessment extends this principle of evaluating real work to all areas of the curriculum.

The most widely used form of authentic assessment in education today is in writing. For example, twenty-eight states, the National Assessment of Educational Progress (NAEP), and many other na-
tions ask students to write on assigned topics. The essays and stories are graded by teams of readers (usually teachers) who assign grades according to standard guidelines. The readers are trained and retrained throughout the process to maintain reliable standards, a process that produces a high degree of agreement among judges. As with all the examples, this methodology can be used to evaluate classroom work that has been collected in a portfolio; it only has to be adjusted for subject area and student age.

Similar procedures are now being followed with open-ended mathematics questions. These ask students to write their own response (not just the answer) to a problem. There is no single way to find a "right answer" because the question is designed to see how a student thinks through a problem, thereby indicating her or his ability to use math. The answers are scored by groups of teacher-readers, again following a standard grading procedure. Two-sevenths of the NAEP math questions will be open-ended in 1990.

Performance assessments in science ask students to plan or perform experiments or use scientific apparatus, as is done in New York State. Science assessments can be graded by observation (where a teacher or other observer uses a checklist) or by scoring the students' written answers to the questions. These assessments can be developed to indicate understanding of basic scientific concepts and methods.

History/social studies assessments frequently require group projects, such as preparing a history of the neighborhood or discovering how a group of people changed a law or policy, tasks which allow students to demonstrate that they grasp important concepts about history and about democratic processes. Foreign language assessments ask students to use the language in a real-life situation, orally and in print.

For young students, reading is best evaluated by having a student read aloud from material of varying levels of difficulty, while keeping a record of "misreads" that reveal the reader's strengths and weaknesses and the strategies used to solve problems. The reading can be taped and reviewed by teacher and student for further analysis and to monitor progress. For older and younger students, the material can be discussed to evaluate comprehension and critical thinking. A writing assignment responding to the ideas of the reading passage can reveal the student's proficiency and thinking in both reading and writing.
All these assessments can be designed to closely follow the curriculum. They provide continuous, qualitative data that can be used by teachers to help instruction. They can be used by students, who learn to assume responsibility for their portfolios and records and thereby engage in regular self-analysis of their work and progress. They provide a direct measure of achievement and therefore are worth the time spent preparing for and doing them. They also encourage an intelligent, rich curriculum rather than the dumbed-down, narrow curriculum fostered by teaching to and coaching for multiple-choice tests.

Teachers can and should be assisted in the evaluation process by community groups, parents, administrators, and university faculty. Outside participation can ensure that racial or cultural bias does not distort the assessment process. For example, a team can examine student portfolios and then compare their evaluations with those of the teacher. These teams should also be helpful in strengthening the evaluation capabilities of teachers by providing feedback.

Authentic evaluation will provide far more information than any multiple-choice test possibly could. The costs of teacher involvement in designing, administering, and scoring new assessments can be counted as part of professional and curriculum development, since no other activity involves teachers more deeply in thinking about their teaching, its objectives, methods and results. Schools and communities will see that authentic assessments are promoting the thinking curriculum everyone wants for our children, and thereby providing genuine accountability.
1. MATERIALS FROM FAIRTEST

FairTest

FairTest Examiner.

Quarterly newsletter surveys developments in testing and testing reform, including pre-school, grade school, IQ and related tests. (Available from FairTest, $15/yr individuals; $25/yr institutions.)

Neill, D.M. and Medina, N.J.


Similar to Fallout from Testing Explosion in content, though shorter, no tables on extent of test use and no annotated bibliography. Included in a special section in this Kappan on testing and alternatives to testing (see 5. Authentic Evaluation, below).

FairTest has the following fact sheets available (send SASE with request for fact sheet): "What's Wrong with Standardized Tests," "How Standardized Testing Harms Education," "What Is Authentic Assessment?" "Testing Reform in Grades 1 and 2, North Carolina."

For other FairTest publications, see the order form at the back of this report.

2. GENERAL DISCUSSIONS OF TESTING

(Many of these articles are also relevant to the subsections that follow.)

Airasian, P. W.


Under some conditions, measurement driven instruction may corrupt the measurement process.


The testing profession's latest revision of its guidelines for the proper construction and use of standardized tests. This work is a useful tool in reform efforts because much test construction and use fails to meet even these weak guidelines.

Bastian, A., et.al.


Concludes that democratizing schools to meet the needs of all students ought to be the focus of school reform efforts. Standardized tests are shown to be one of the barriers to equal educational opportunity and educational quality.

Congressional Budget Office


The first of two CBO reports on educational achievement. Includes general discussion of the ways in which standardized test results are used to measure educational achievement and the problems with such uses. Warns against overreliance on test results in evaluating trends in educational achievement.

Fiske, E.


Contains general overview of the growth of standardized testing in America's public schools and the resulting problems. Discusses various uses made of standardized tests, efforts by schools to improve test scores, and criticisms and limitations of standardized tests. It also suggests that standardized tests will undermine the goals of the current school reform movement.

Gould, S.J.


Criticizes the idea that there is a single unitary thing that can be called "intelligence" and the tendency to reduce information to a single number. Provides a history of "mismeasures," including "I.Q." tests. Highly readable, valuable critique.

Haney, W.


Provides a detailed history of the use of standardized tests in America (divided into three eras: pre-WWI; WWI to 1950; and 1950
to present). In particular, chronicles the recent growth in test use and test criticism. Includes brief discussion of the intensity of use of standardized tests in the public schools and a longer discussion of the types of uses. Concludes by suggesting avenues for further research and development in testing.

Haney, W. & G. Madaus


Documents the increasing attention paid to testing compared with curriculum in educational literature. Discusses four broad issues which determine the impact of testing on education and lists seven major problems regarding the impact of tests on individuals and schooling. Authors suggest efforts that can minimize the negative impact and maximize the possible benefits of testing. Two sections specifically discuss NAEP and its future.

Madaus, G. F.


An excellent, comprehensive look at the effects of testing on the curriculum. After describing the various types of tests and testing programs, Madaus poses seven general principles describing the impact of testing on the curriculum. He argues that tests can become the "ferocious master" instead of being the "compliant servant," and details the evidence of their effects on programs, teachers and students.

Madaus, G. & D. Pullin


Presented in question-and-answer format; focuses on use of standardized tests in "high stakes" educational situations (i.e. where significant sanctions or rewards are associated with the test). Lists several "high stakes" uses of tests, discusses ways in which tests are designed or selected. Examines potential structure of testing programs, possible conclusions that will be drawn from their results, and their likely impact on students and schools. (NCAS, 100 Boylston St., Boston, MA 02116; free.)

National Elementary Principal

(March/April 1975 and July/August 1975).

The first issue, "IQ: The Myth of Measurability," contains a variety of criticisms of IQ tests. The second, "The Scoring of Children: Standardized Testing in America," examines the testing industry, test construction, and achievement tests.
Negro Educational Review


Book-length issue contains numerous articles on testing, including discussions of: psychometric, language and cultural biases against blacks (particularly working class blacks) and other minorities; IQ testing; and alternatives to standardized tests. Several specific articles are noted below. (Available from NER, Box 2895, Jacksonville, FL 32203; $20). The July-October 1977 issue of this journal, “Testing Black Students,” also includes many excellent articles.

Ninth Mental Measurement Yearbook

(Lincoln, Nebraska: Buros Institute of Mental Measurement, 1985).

Contains one or more reviews plus bibliography for each of hundreds of tests; older tests may also be in previous editions of the Yearbook. The primary source for test reviews.

Phi Delta Kappan

“What is the Proper Role of Testing?” (Special Section). (May 1985) pp. 599 - 639.

Nine articles focus on the use, problems and impact of standardized testing in the public schools. Topics covered include: misuse of SAT scores in assessing the quality of American education; popularity of standardized tests; impact of testing on pedagogy and instruction; problems with teacher testing; impact of testing on educational equity; and alternatives to multiple-choice writing tests. One article describes the use of standardized tests to “drive the curriculum” in three states and one school district.

Pipho, C.


One in a series of articles discussing the state education reform efforts of 1983 - 1985. Lists state uses of standardized tests.

Rethinking Schools

(Jan/Feb 1989).

“Focus on Testing” section includes critiques of testing and its effects, and a discussion of high school alternatives. (P.O. Box 93371, Milwaukee, WI 53202; $2.00.)
3. SPECIFIC PROBLEMS WITH STANDARDIZED TESTS

A. TESTING YOUNG CHILDREN

Bredekamp, S. and L. Shepard


Contains excellent critique of the misuses of standardized testing on young children as well as discussion of appropriate use. Adds to NAEYC publications noted below. Includes helpful bibliography.

Kamii, C., ed.


This book summarizes the problems caused by reliance on multiple-choice achievement testing, has chapters on specific problems for systems, principals, teachers and others, and examines the problems testing causes for teaching literacy and math. Appropriate assessment in math and literacy is also presented.

Martin, A.


Finds disability testing and special needs intervention frequently produce an overemphasis on children's weaknesses and promote a view of children as having deficits to be corrected, rather than having individual differences and strengths on which to build. The same factors disempower classroom teachers.

Meisels, S.J.


Looks at different tests and their limitations, urges caution in their use. Specifically examines the Gesell instruments. The issue contains a response from Gesell and a rebuttal by Meisels. A major source for the NAEYC position. For a specific discussion of screening tests and a review of a number of tests (many of which lack adequate validity), see Meisels, S.J. Developmental Screening in Early Childhood: A Guide (NAEYC, Third Edition, 1989).

Reviews appropriate and inappropriate test use. Urges caution in the use of tests. Basis for pamphlet (below). See also, "NAEYC Position Statement on Developmentally Appropriate Practice in the Primary Grades, Serving 5-Through 8-Year-Olds," Young Children (Jan. 1988) p. 64-84, which includes section on testing and has a comprehensive bibliography.

Publishers Association for the Education of Young Children


Pamphlet discusses the potentially harmful impact on young children of standardized testing. Describes proper uses of standardized tests and suggests how schools can "help ensure that all children get off to a sound start in kindergarten, first, and second grade." (Available from NAEYC, 1834 Connecticut Avenue, N.W., Washington, D.C. 20009; $.50 each or 100 for $10.)


Chapters 4 and 5 explore the harm of retaining young children in grade or placing them in "transitional programs," often done through the use of test results. See also by the same authors, "Flunking Kindergarten," American Educator (Summer 1988) pp. 34-38, and "What Doesn't Work: Explaining Policies of Retention in the Early Grades," Phi Delta Kappan (October 1987).

B. TEST BIAS

First, J.M. and J. Willshire Carrera


Includes discussion of the harmful effects of standardized testing on new immigrant students. (NCAS, 100 Boylston St., Boston, MA 02116; $11.95).

Hilliard, A.


A good discussion of the basis of I.Q. testing.
Hoover, M.R., R.L. Politzer & O. Taylor


Details language-related bias in standardized tests against speakers of non-standard English, including phonological (sound), syntactical (structural), and lexical (word choice and vocabulary) biases. Consequences of bias include school program misplacement and tracking resulting in inadequate education for students who are not white middle- to upper-class. Eliminating these biases is important for reducing educational and societal biases against working class and minority children.

Levidow, L.


Summarizes a number of problems with I.Q. tests, including bias in the tests, and argues they are harmful social constructs. Urges caution in using other methods of assessment as they can be as damaging as tests.

Sosa, A.S.


Wide-ranging document contains abstracts of two major papers plus testimony to the Commission. One paper discusses secondary education; seven witnesses address elementary and secondary education.

Taylor, O. & D.L. Lee


Contains detailed discussion of sources and kinds of cultural and language bias in standardized tests. These biases cause African-Americans (particularly working-class blacks) and minorities to be invalidly assessed: "At times... the results fail to accurately represent actual abilities." In conclusion, "... the very assumptions and paradigms upon which most standardized tests are based need to be revised."

Willie, C.V.


Discusses disproportionate impact of test use on minority students. Argues that tests ignore the diversity among various American ethnic groups and implicitly undercut the value and legitimacy of this diversity.
C. TEST VALIDITY & RELIABILITY

Cannell, J.J.  
*Nationally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States Are Above the National Average* (Friends for Education: Daniels, W.Va., 1987).

Describes the results of a nationwide survey on achievement test scores: no state average score at the elementary level was below the national norm on any of the six most popular achievement tests. Concludes that this results from improper norming of the tests and teaching to the tests. (For confirming studies, see Notes 39 and 40 at end of main body of *Fallout*, and Koretz, below).

Congressional Budget Office  

Second of two CBO reports on educational achievement, includes a general discussion of the problems and limitations of standardized tests, particularly the problems around validity and reliability.

Johnston, P.  

Criticizes the search for “objectivity” as both futile and educationally destructive. Points out that the core of validity is construct validity, which is a “socially negotiated variable within a changing pluralistic society.” Concludes that “the overwhelming concern for objective, valid measurement as a means to improve teaching and learning is not very helpful.” Discusses the need for a “different view of science” and for improving the quality of teachers. Concludes with various concrete examples of teaching and evaluating. (For a different view, calling for expanding validity studies to include teacher and learner use of tests, see C.K. Tittle, “Validity: Whose Construction Is it in the Teaching and Learning Context,” *Educational Measurement*, Spring 1989, pp. 5-13, 34.) The two articles also provide excellent bibliographies of current validity discussions.

Koretz, D.  

Analyzes and substantially confirms Cannell’s report (above) and essentially answers “Yes” to the questions it poses in the title. Calls for changing tests and re-focusing on the broader goals of education, not just testing.

The most recent in a series on validity by Messick. Argues for construct validity as the essence of validity, then expands the definition of construct validity to include value implications and social consequences; e.g., if use of a test in a particular situation has harmful consequences, the use of the instrument could lack construct validity if the harm is caused in part by the nature of the test itself. See also by Messick: "The Once and Future Issues of Validity," in Howard Wainer and Henry I. Braun, eds., *Test Validity* (Hillsdale, NJ: Lawrence Erlbaum, 1988); and his lengthy, technical article in Robert Linn, ed., *Educational Measurement* (New York: MacMillan, 1989).

D. TEST ADMINISTRATION


Authors analyze data from 22 controlled studies involving 1489 subjects and discover that familiarity with the test examiner had different impacts on test-takers. Use of unfamiliar examiners reduces test scores for low-income and black students, but does not affect the scores of upper-income students.


Study examines the administration of standardized tests in eight kindergartens, finding that testing practices in five of the eight kindergartens were so nonstandardized as to render their test scores incomparable and quite possibly unreliable as well." Concludes that using results from such administrations is likely to lead to unsound educational decisions.

4. THE IMPACT OF TESTING ON EDUCATION

Section 2 (above) also includes items relevant to this topic.


Provides very brief history of testing in American public schools, suggests reasons for the increased use of tests in the schools. Describes
various problems with test use including: loss of authority for teachers' professional judgments; loss of local control over education; narrowing the educational curriculum; teaching to the test; disproportionate impact on disadvantaged students; and elevation of coaching over teaching.

A. IMPACT ON CURRICULUM

Allington, R.L.


Finds that good readers are taught differently from poor readers. Poor readers experience lots of oral reading (with regular interruptions), decoding and word emphasis, with absolute test score gains as the measure. Good readers experience silent reading and comprehension discussions, both of which correlate with quicker and stronger comprehension improvement (decoding emphasis leads to increases in decoding on tests). Recommends more silent reading, summary and discussion for poor readers, plus more research, as this report is often tentative. See also by Allington: "Poor Readers Don't Get to Read Much in Reading Groups," Language Arts (Nov.-Dec. 1980), and "Shattered Hopes: Why Two Federal Reading Programs Have Failed to Correct Reading Failure," Learning (July-Aug. 1987).

Bussis, A. M.


"Skills" instruction displaces real reading in classrooms despite lack of evidence that "any identifiable group of subskills was essential to reading," (according to IRA study in 1973 and NIE study in 1975). "Only after educators began to realize that many children could master decoding skills and still fail to read effectively did 'the focus of research begin to expand'." Presents constructionist psychological view, distinguishes information from meaning. The skill of reading is complex, but not an assemblage of subskills. Different children learn differently. Concludes with teaching suggestions focused on real reading and increasingly observant teachers.

Bussis, A.M. and E A. Chittendon.


Excellent summary of research indicating "little correspondence between contemporary theories of the reading process and assumptions implicit in the tests."

Based on a six-year study of how children learn to read, compares test item types to children’s differing strategies in learning to read, finds much incompatibility. Connects this research to more general incompatibility between tests and current cognitive, developmental and learning theory.


Reports on a nation-wide survey of teachers and principals in 114 school districts. Seeks to identify the amount of time devoted to testing, uses of test results, teachers’ and principals’ perceptions about testing, and issues of equity as a result of standardized test use. Concerns about impact of standardized test use include the fact that most teachers pay more attention to standardized test results for low SES students than for high SES students. Tests also can reduce time spent on other educational goals and narrow curriculum. Calls for more “rational” relationship between teacher-designed tests, external texts and the curriculum, but emphasizes that the curriculum must drive the tests and not vice-versa.


From abstract: “There is evidence that tests do influence teacher and student performance and that multiple-choice tests tend not to measure the more complex cognitive abilities.” Describes several experiments showing that open-ended (free response, multiple-level) problem-solving tests measure different, higher and more complex, cognitive abilities than do multiple-choice tests, even multiple-choice tests directly derived from the open-ended tests. Most of the problems in the world are ill-structured, unlike multiple-choice problems which dominate tests used for accountability. “We need a much broader conception of what a test is if we are to use test information in improving educational outcomes.”


Prepared for the National Council of Teachers of English. Critical analysis of the history and content of basal readers and how they hinder the teaching of reading. Includes discussion of tests contained in basals and the relationship between basals and standardized tests.
Kamii, C.  *Young Children Continue to Reinvent Arithmetic, 2nd Grade* (New York: Teachers College Press, forthcoming).

Chapter 10, on evaluation, indicates that standardized tests cannot reveal whether students understand math concepts and reasoning. Teaching to the tests makes it less likely that students develop quantitative reasoning concepts and skills.


"It is untenable to agree that achievement is the product, and that test scores are its measure, and then assert, 'Please pay too much attention to the scores.'" Discusses test use in Pittsburgh, where the focus is diagnostic testing that is relevant, timely (quick turn-around, frequent), short, used by teachers to inform instruction. Pittsburgh uses regular achievement tests for evaluative, not diagnostic purposes, and only uses samples of students. Testing is still mostly multiple-choice.


Short discussion of the effects of testing includes the argument, with references, that test-based methods used to teach basic skills (rote, drill, retention) are counter productive to teaching and learning higher order thinking skills.


Part of a series by McNeil, "Contradictions of Control." Explores the destructive effects of mandated testing on an exceptionally good school program.


Shows that reading instruction aimed at increasing standardized test scores hinders learning to read, since higher test scores do not necessarily indicate improved reading ability. Argues that reading and other tests are biased against minority and low-income youth. Tests cannot measure the utility or effect of school reforms: "... testing not only fails to be helpful but sabotages good education."
National Academy of Education


Although this review discusses the recommendations of the Alexander/James Study Group to dramatically expand the National Assessment of Educational Progress (NAEP), it also provides some general discussions of the problems of testing. In particular, it notes how testing can narrow the curriculum, threaten educational equity, and undermine educational goals that are not easily quantifiable.

Resnick, L.B. and D.P. Resnick


Shows how tests are based on outmoded behaviorist and associationist psychological theories from late 19th century in which knowledge is reduced to isolated bits and learning is defined as passively absorbing the bits; current theory indicates knowledge is whole and contextual, learning is active, contextual, and constructed in people's minds. Current tests are incompatible with needed educational reforms. New forms of assessment must be based on accurate learning theory and help school reform.

Salganik, L.H.


Links the growing use of standardized tests in the schools with a loss of public confidence in teachers. Because increased reliance on testing undermines the authority of teachers' judgments, a cycle of declining professional authority and declining public confidence is created. Policy issues, such as educational equity, the goals of schooling and control over school decisions, are masked by the emphasis on technical questions regarding testing.

Smith, F.


Excellent, readable book on learning, reading, cognitive theory and how testing and much reading instruction are insulting to the intelligence of children (and teachers), who then turn off to reading and schooling. Also has much on good methods of teaching and how to reform education.

Suhor, C.

"Objective Tests and Writing Samples: How Do They Affect Instruction in Composition?" Phi Delta Kappan (May 1985) pp. 635-639.

Criticizes the use of multiple-choice standardized writing tests.
because they undercut "real" writing instruction. Urges use of computerized "writing sample" tests as an alternative.


Discusses the quality of textbooks in the public schools. Concludes that the increasing emphasis on testing has been a major contributor to the declining quality of American textbooks. Indicts the current "curriculum alignment" movement which has affected school districts in 22 states. Encourages the use of a more diverse curriculum and set of student assessment mechanisms. (CBE, 725 15th St., N.W., Washington, D.C. 20005; $13.)

B. IMPACT ON STUDENT PROGRESS AND ACHIEVEMENT


Summarizes the harmful effects of tracking on black students. For the effects of tracking on segregation, see also J.L. Epstein, "After the Bus Arrives: Resegregation in Desegregated Schools," Journal of Social Issues (Vol. 41, No. 3, 1985) pp. 23-43.


The key work on the harmful effects of tracking in the public schools, often done on the basis of test scores. Tracking is most harmful to low-income and minority-group children because of their isolation from other children and the watered-down curriculum they receive. See also Oakes' article, "Keeping Track," Phi Delta Kappan (Sept. & Oct., 1986).


This report discusses: increased use of standardized tests in the public schools; their use as barriers to educational opportunity for at-risk children; misclassification of minority students; impact of tests on handicapped students; and the general measurement limitations of tests. (NCAS, 100 Boylston St., Boston, MA 02116; free.)
Raudenbush, S.


Reanalyzes the "Pygmalion Effect," originally identified and described by Robert Rosenthal and Lenore Jackson in 1965 in *Pygmalion in the Classroom*. Discovers "Effect" prevalent among students entering seventh grade, not those in grades three to six. "Effect" most likely to occur when teachers know little about their students beyond the statistical information they are provided. Where that information (including test scores) under- or overestimates students' abilities, teacher behavior affects student achievement (regardless of previous student achievement levels).

Shepard, L. A. and M.I. Smith, eds.

*Flunking Grades* (cited above, 3A Young Children).

While Chs. 4 & 5 discuss the effects of retention on young children, the book as a whole surveys the extent of retention, why it happens—including test-based retention—and alternatives to retention.

C. IMPACT ON LOCAL CONTROL

Wise, A.E.


Demonstrates how increasing use of standardized testing undermines local control over the public schools and increases state and national control over education. Also discusses other impacts such as narrowed curriculum, loss of teaching time, lower teacher morale, and loss of teacher authority.

5. AUTHENTIC EVALUATION AND REDUCED TESTING

The following materials are introductions to the concepts and essential elements of the practice of authentic, direct, performance-based evaluation. Detailed descriptions of possible ways to create systems of evaluation that are performance-based seem not yet to be generally available, excepting the North Carolina materials noted below, but hopefully will be available soon.

Archbold, D. and F. Newman


Discusses "what is authentic academic achievement," how to assess it and educational programs, and implementing assessment
programs. Considers secondary school and college-level assessment alternatives currently in place. Appendix includes a critique of standardized tests.

Edelsky, C. & S. Harman


Provides excellent summary of many problems with standardized tests, suggests appropriate assessment procedures to meet the different needs of parents, teachers and students; the public and elected officials; and researchers.

Educational Leadership

(Vol. 46, #7, April 1989) “Redirecting Assessment.”

Includes 17 articles on developing authentic assessments, provides a wide range of materials useful for understanding alternatives theoretically and in practice. The best introduction to the range of “alternatives” being developed, most of which are performance-based.

Gardner, H.


Detailed discussion of process and product alternatives rooted in students' classroom work and recent scientific understandings. Gardner has recently published a number of other articles on assessment.

Haney, W.


Describes the limited educational utility of standardized tests. Examines three efforts to make educational use of testing (Portland, OR; Orange County, FL; Pittsburgh, PA) and one school that uses no tests but relies on alternative evaluations (Prospect School, North Bennington, VT).

Johnston, P.


Discusses the fact that teachers can and do evaluate and urges that teachers need to be helped so that they can become evaluation experts.
Grades 1 and 2 Assessment

N.C. has instituted developmentally appropriate assessment statewide in grades one and two, in Communication Skills and Mathematics. They have a packet of material available for $50.00, including a thick notebook, two videos, and material from staff development sessions. (116 W. Edenton St., Raleigh, N.C. 27603.)

Wiggins, G.


Criticizes standardized, multiple-choice norm and criterion referenced tests as inadequate, inauthentic and harmful. Argues that authentic tests (as sometimes different from assessment in general) can be valuable for learning. Defines authentic tests as evidencing and providing a means to judge knowledge and gives examples. States that authentic tests enable interaction with the student and therefore enhance equity. Establishes criteria for authentic exams and for grading them. Links new forms of testing and assessment to re-structuring schools. (There are several articles on alternative assessments in this issue of Kappan.)
**FairTest Publications**

- **FairTest Examiner**, published quarterly... $15.00/year (Individually), $25.00 (Institutions)
- Back Issues of **FairTest Examiner**, $4.00 each: Vol... No... (All 11 for $32.00)
- **Standing Up to The SAT**, by John Weiss, Barbara Beckwith and Bob Schaeffer (192 pp.)... $6.95
- **Fallout From the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools**, 3rd ed. by Noe Medina and Monty Neill (77 pp.)... $8.95
- **Sex Bias in College Admissions Tests: Why Women Lose Out**, by Phyllis Rosser and FairTest staff, 3rd ed. (60 pp.)... $7.95
- **The Reign of ETS**, by Allan Naim and Ralph Nader (550 pp.)... $30
- **Beyond Standardized Tests: Admissions Alternatives That Work**, by Amy Allina with FairTest staff (18 pp.)... $5.50
- **None of the Above, Behind the Myth of Scholastic Aptitude**, by David Owen (327 pp.)... $5.95

**Contributions to FairTest**

I agree there's no better way to be part of the testing reform movement than supporting FairTest! Enclosed is my tax-deductible contribution!*

- **FairTest Associate** ($25.00--includes a free one-year subscription to the **FairTest Examiner**)
- **FairTest Sponsor** ($50.00--includes Examiner subscription and **Standing Up to the SAT**)
- **FairTest Sustainer** ($100.00 or more--Includes above plus a 20% discount on publications)
- **Other**

---

Name ____________________________ ____________________________ ____________________________

Address ____________________________ ____________________________ ____________________________

City ________________ State ________________ ZIP ________________

*Contributions to FairTest are fully tax-deductible, except for the value of publications received.

Please make checks payable to FairTest and mail to: FairTest, 342 Broadway, Cambridge MA 02139