THINKING ABOUT TESTS AND TESTING:

A Short Primer in "Assessment Literacy"

Gerald W. Bracey

American Youth Policy Forum
in cooperation with the
National Conference of State Legislatures

AMERICAN YOUTH POLICY FORUM

The American Youth Policy Forum (AYPF) is a non-profit professional development organization based in Washington, DC. AYPF provides nonpartisan learning opportunities for individuals working on youth policy issues at the local, state and national levels. Participants in our learning activities include: Government employees—Congressional staff, policymakers and Executive Branch aides; officers of professional and national associations; Washington-based state office staff; researchers and evaluators; education and public affairs media.

Our goal is to enable policymakers and their aides to be more effective in their professional duties and of greater service—to Congress, the Administration, state legislatures, governors and national organizations—in the development, enactment, and implementation of sound policies affecting our nation's young people. We believe that knowing more about youth issues—both intellectually and experientially—will help them formulate better policies and do their work more effectively. AYPF does not lobby or take positions on pending legislation. We work to develop better communication, greater understanding and enhanced trust among these professionals, and to create a climate that will result in constructive action.

Each year AYPF conducts 35 to 45 learning events (forums, discussion groups and study tours) and develops policy reports disseminated nationally. For more information about these activities and other publications, contact our web site at www.aypf.org.

This publication is not copyrighted and may be freely quoted without permission, provided the source is identified as: *Thinking About Tests and Testing: A Short Primer in "Assessment Literacy"* by Gerald W. Bracey. Published in 2000 by the American Youth Policy Forum, Washington, DC. Reproduction of any portion of this for commercial sale or profit is prohibited.

AYPF events and policy reports are made possible by the support of a consortium of philanthropic foundations: Ford Foundation, Ford Motor Fund, General Electric Fund, William T. Grant Foundation, George Gund Foundation, James Irvine Foundation, Walter S. Johnson Foundation, W.K. Kellogg Foundation, McKnight Foundation, Charles S. Mott Foundation, NEC Foundation of America, Wallace-Reader's Digest Fund, and others. The views reflected in this publication are those of the author and do not reflect the views of the funders.

American Youth Policy Forum
1836 Jefferson Place, NW
Washington, DC 20036-2505

Phone: 202-775-9731
Fax: 202-775-9733
E-Mail: aypf@aypf.org
Web Site: www.aypf.org

ABOUT THE AUTHOR

A prolific writer on American public education, **Gerald W. Bracey** earned his Ph.D. in psychology from Stanford University. His career includes senior posts at the Early Childhood Education Research Group of the Educational Testing Service, Institute for Child Study at Indiana University, Virginia Department of Education, and Agency for Instructional Technology. For the past 16 years, he has written monthly columns on education and psychological research for *Phi Delta Kappan* which, in 1997, published his *The Truth About America's Schools: The Bracey Reports, 1991-1997.* Among Bracey's other books and numerous articles are: *Final Exam: A Study of the Perpetual Scrutiny of American Education* (1995), *Transforming America's Schools* (1994), *Setting the Record Straight: Responses to Misconceptions About Public Education in America* (1997), and *Bail Me Out!: Handeling Difficult Data and Tough Questions About Public Schools* (2000). Bracey, a native of Williamsburg, Virginia, now lives in Alexandria, Virginia.

Editors at the American Youth Policy Forum include Samuel Halperin, Betsy Brand, Glenda Partee, and Donna Walker James. Sarah Pearson designed the covers. Rafael Chargel formatted the document.

CONTENTS

INTRODUCTION: THE NEED FOR "ASSESSMENT LITERACY" 1

PART I: ESSENTIAL STATISTICAL TERMS 2

1. WHAT IS A MEAN? WHAT IS A MEDIAN? WHAT IS A MODE? 2
2. WHAT DOES IT MEAN TO SAY "NO MEASURE OF CENTRAL TENDENCY WITHOUT A MEASURE OF DISPERSION?" 3
3. WHAT IS A NORMAL DISTRIBUTION? 4
4. WHAT IS STATISTICAL SIGNIFICANCE? 4
5. WHY DO WE NEED TESTS OF STATISTICAL SIGNIFICANCE? 5
6. HOW DOES STATISTICAL SIGNIFICANCE RELATE TO PRACTICAL SIGNIFICANCE? 5
7. WHAT IS A CORRELATION COEFFICIENT? 6

PART II: THE TERMS OF TESTING: A GLOSSARY 7

1. WHAT IS STANDARDIZED ABOUT A STANDARDIZED TEST? 7
2. WHAT IS A NORM? WHAT IS A NORM-REFERENCED TEST? 7
3. WHAT IS A CRITERION-REFERENCED TEST? 8
4. HOW ARE NORM-REFERENCED AND CRITERION-REFERENCED TESTS DEVELOPED? 9
5. WHAT IS RELIABILITY IN A TEST? 10
6. WHAT IS VALIDITY IN A TEST? 11
7. WHAT IS A PERCENTILE RANK? A GRADE EQUIVALENT? A SCALED SCORE? A STANINE? 12
8. WHAT ARE MULTIPLE-CHOICE QUESTIONS? 13
9. WHAT DO MULTIPLE-CHOICE TESTS TEST? 14
10. WHAT IS "AUTHENTIC" ASSESSMENT? 15
11. WHAT ARE PERFORMANCE TESTS? 15
12. WHAT ARE PORTFOLIOS? 16
13. WHAT IS A "HIGH STAKES" TEST? 16
14. WHAT IS AN IQ TEST? 16
15. WHAT IS THE DIFFERENCE BETWEEN AN ABILITY OR APTITUDE TEST AND AN ACHIEVEMENT TEST? 17
16. WHAT ARE THE ITBS, ITED, TAP, STANFORD-9, METRO, CTBS AND TERRA NOVA? 18

17. WHAT IS A MINIMUM COMPETENCY TEST? 18
18. WHAT ARE ADVANCED PLACEMENT TESTS? 19
19. WHAT IS THE INTERNATIONAL BACCALAUREATE? 19
20. WHAT IS THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS? 19
21. WHAT IS THE NATIONAL ASSESSMENT GOVERNING BOARD? 20
22. WHAT IS THE THIRD INTERNATIONAL MATHEMATICS AND SCIENCE STUDY (TIMSS)? 20
23. WHAT IS "HOW IN THE WORLD DO STUDENTS READ?" 22
24. WHAT IS THE COLLEGE BOARD? 22
25. WHAT IS THE EDUCATIONAL TESTING SERVICE? 22
26. WHAT IS THE SAT? 22
27. WHAT IS THE PSAT? 23
28. WHAT IS THE NATIONAL MERIT SCHOLARSHIP CORPORATION? 23
29. WHAT IS THE ACT? 23
30. WHAT IS FAIRTEST? 23
31. WHAT IS A STANDARD? 24
32. WHAT IS A CONTENT STANDARD? WHAT IS A PERFORMANCE STANDARD? 24
33. WHAT IS ALIGNMENT? 24
34. WHAT IS CREDENTIALING? 24

PART III: SOME ISSUES IN TESTING 25

1. WHY IS TEACHING TO THE TEST A PROBLEM IN EDUCATIONAL SETTINGS, BUT NOT ATHLETIC SETTINGS? 25
2. WHO DEVELOPS TESTS? 25
3. WHAT AGENCIES OVERSEE THE PROPER USE OF TESTS? 26
4. WHY DO CORRELATION COEFFICIENTS CAUSE SO MUCH MISCHIEF? 26
5. WHY IS THERE NO MEANINGFUL NATIONAL AVERAGE FOR THE SAT OR ACT? 26
6. WHY DID THE SAT AVERAGE SCORE DECLINE? 27
7. WHY WAS THE SAT "RECENTERED?" 27
8. DO THE SAT AND ACT "WORK? " 28
9. DO COLLEGES OVER RELY ON THE SAT? 28

WHY "ASSESSMENT LITERACY"? 30

INTRODUCTION: THE NEED FOR "ASSESSMENT LITERACY"

Tests in education gradually entered public consciousness beginning around 1960. Forty years ago, people didn't pay much attention to tests. Few states operated state testing programs. The National Assessment of Educational Progress (NAEP) would not exist for another decades. SAT (Scholastic Aptitude, later Assessment, Test) scores had not begun their two decade-long decline. Guidance counselors, admissions officers and the minority of students wishing to go to college paid attention to these SAT scores, but few others did. There were no international studies testing students in different countries. Only Denver had a "minimum competency" test as a requirement of high school graduation.

Now, tests are everywhere. Thousands of students in New York City attended summer school in an attempt to raise their test scores enough to be promoted to the fourth grade. Because of the pressure on test scores, a number of schools in New York City were found to be cheating in a variety of ways. Experts are debating whether or not Chicago's policy of retaining students who don't score high enough is a success or failure. The State Board of Education in Massachusetts has been criticized for setting too low a passing score on the Massachusetts state tests. The Virginia Board of Education is wrestling with how to lower Virginia's excessively high cut score without looking like they're also lowering standards. Arizona failed 89% of its students in the first round of its new testing program. Tests are being widely used – and misused – to evaluate students, teachers, principals and administrators.

Unfortunately, tests are easy to misinterpret. Some of the inferences made by politicians, employers, the media and the general public about recent testing outcomes are not valid. In order to avoid misinterpretations, it is important that informed citizens and policymakers understand what the terms of testing really mean. The American Youth Policy Forum hopes this glossary provides such basic knowledge.

This short primer is organized into three parts. Part I introduces some *statistics* that are essential to *understanding testing concepts* and for talking intelligently about tests. Those who are familiar with statistical terms can skip Part I and go straight to the discussion of current test terms. Part II presents some *fundamental terms of testing*. Both Parts I and II deal with "what": What is a median, a percentile rank, a norm-referenced test, etc? Part III fleshes out Parts I and II with discussions *about testing issues*. These are more "who" and "why" questions. Together, these three parts have the potential of raising public understanding about what is, far too often, a source of political mischief and needless educational acrimony.

— American Youth Policy Forum

PART I
ESSENTIAL STATISTICAL TERMS

1. WHAT IS A MEAN? WHAT IS A MEDIAN? WHAT IS A MODE?

These are the three words that people call something "average." The most common term in both testing and the general culture is the mean, which is simply the sum of all scores divided by the number of scores. If you have the heights of eleven people, to calculate the mean you add all eleven heights together and divide by eleven.

The median, another common statistic, is the point above which half the scores fall and below which half fall. If you have the heights of eleven people, you arrange them in ascending or descending order and whatever value you find for the sixth score is the median (five will be above it, five below).

Means and medians can differ in how well they represent "average" because means are affected by extreme values and medians are not. Medians only involve counting to the middle of the distribution of whatever it is you're counting. If you are averaging the worth of eleven people and one of them is Bill Gates, the mean salary will be in the billions even if the other ten people are living below the poverty level. In calculating the median, Bill is just another guy, and to find the median you need only find the person whose score splits the group in half.

The third statistic that is labeled an "average" is called the mode. It is simply the most commonly occurring score in a set of scores. Suppose you have the weights of eleven people. If four of them weigh 150 pounds and no more than three fall at any other weight, the mode is 150 pounds. Modes are not much seen in discussions of testing because the mean and median are usually more descriptive. In the preceding weight example, for instance, 150 pounds would be the mode even if it were the lowest or highest weight recorded.

To illustrate the different averages, consider this list as the wealth of residents in Redmond, Washington (which, for our purposes, contains only 11 citizens).

$10,000	$10,000	$20,000	$20,000
$20,000	$50,000	$60,000	$70,000
$75,000	$125,000	$70 billion	

Mean wealth = $6.4 billion

Median wealth = $50,000

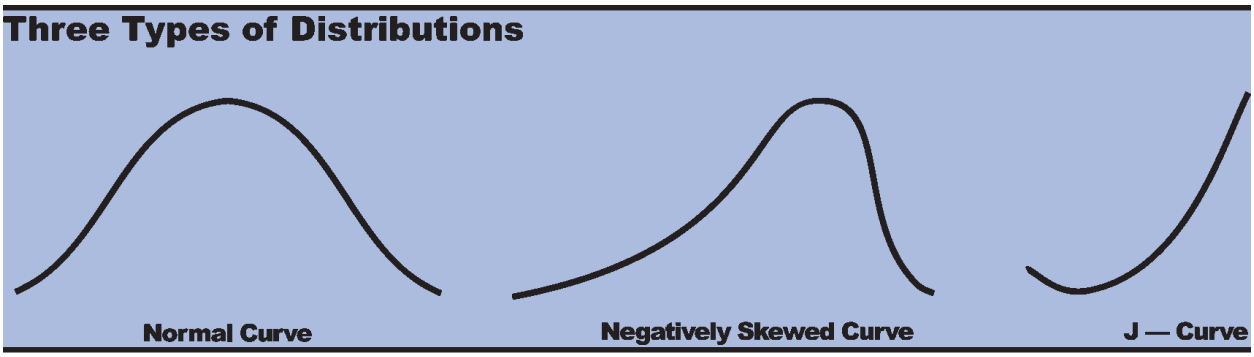
Modal wealth = $20,000

The seventy billion was roughly Bill Gates' net worth as of late 1999. When we calculate the mean, that wealth gets figured in and all the inhabitants look like billionaires, with the average (mean) wealth in excess of $6 billion.

When we calculate the median, we look for the score that divides the group in half. In the example, this is $50,000: five people are worth more than $50K and five are worth less. Gates' billions don't matter because we are just looking for the mid-point of the distribution.

In the Redmond of our example, three people have wealth equal to $20,000, so this is the most frequently occurring number and is, therefore, the mode.

Many distributions of statistics in education fall in a bell-shaped curve, also called a "normal distribution." In a normal distribution of scores, the mean, median and mode are identical.



Modes become useful when the shape of the distribution is not normal and has two or more values where scores clump. Thus, if you gave a test and the most frequent score was 100, that would be the mode, but if there was also another cluster of scores around, say, 50, it would be most descriptive to refer to the distributions as "bi-modal."

The curve on the left is normal. That in the middle is skewed, with many scores piling up at the upper end. This could happen because either the test was easy for the people who took it or because instruction had been effective and most people learned most of what they needed to know for the test.

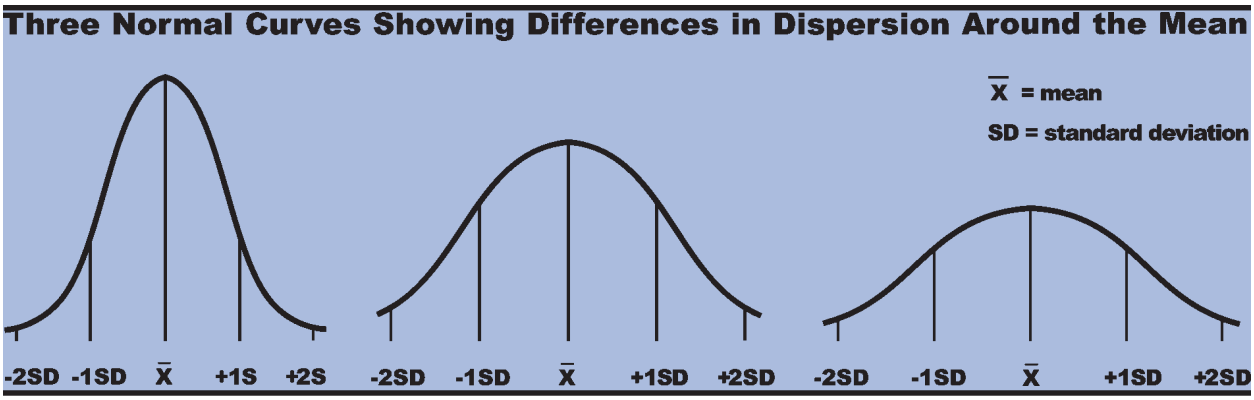
When constructing a "norm-referenced test," test makers *impose* a normal distribution of scores by the way in which items are selected for the test. When it comes to "criterion-referenced" tests, a bell-curve would be irrelevant. We are usually looking to make a yes-no decision about people: did they meet the criterion or not? Or, are we looking to place them in categories such as "basic," "proficient" and "advanced?" Noted educator Benjamin Bloom argued that in education the existence of a bell-curve was an admission of failure: it would show that most people learned an average amount, a few learned a lot and a few learned a little. The goal of education, Bloom argued should be a curve somewhat shaped like a slanted "j", the curve on the right. This would indicate that most people had learned a lot and only a few learned a little.

2. WHAT DOES IT MEAN TO SAY "NO MEASURE OF CENTRAL TENDENCY WITHOUT A MEASURE OF DISPERSION?" AND WHY WOULD ANYONE EVER SAY THIS?

Mean, median and mode are all measures of average or what statisticians call "measures of central tendency." We need a measure of how the scores are distributed around this average. Does everyone get nearly the same score or are the scores widely distributed?

One way of reporting dispersion is the range: the difference between the highest and lowest score. The problem with the range is that, like the mean, it can be affected by extreme scores.

The most common measure of dispersion is called the "standard deviation." In the world of statistics, the difference between the average score and any particular score is called a "deviation." The standard deviation tells us how large these deviations are on average. Statisticians use the standard deviation a lot because it has useful and important mathematical properties, particularly when the scores are distributed in a normal, bell-shaped curve.



Three different distributions and their standard deviations are shown above. Note that these are all bell curves. They differ in how much the scores are spread out around the average. Despite these differences, some things are the same. For instance, the distance between the mean and + 1 or - 1 standard deviation always contains 34% of the scores. Another 14% will fall between + or - one and + or - two standard deviations. A person who scores one standard deviation above the mean always scores at the 84th percentile—there are 34% of the scores between the mean and +1 standard deviation and then there are another 50% that are below the mean. (Please see SCALED SCORES on p. 13 for an example using SAT and IQ scores.)

Merely reporting averages often obscures important differences that might have important policy implications. For instance, in the Third International Mathematics and Science Study, the average 8th grade math and science scores for the United States were quite close to the average of the 41 nations in the study. As a nation, we looked average. However, the highest scoring states in the United States outscored virtually every nation while the lowest scoring states outscored only three of the 41 nations. The average obscured how much the scores varied among the 50 states.

3. WHAT IS A NORMAL DISTRIBUTION?

For statisticians, a "normal" distribution of test scores is the bell curve. There is nothing "magical" about bell curves, the title of a famous book notwithstanding (see note on p. 17). It happens, though, that many human characteristics are distributed in bell-curve fashion, such as height and weight. Grades and test scores have been traditionally expressed in bell-curve fashion.

4. WHAT IS STATISTICAL SIGNIFICANCE?

Tests of "statistical significance" allow researchers to judge whether or not their results are "real" or could have happened by chance. Educational researchers can be heard saying things like "the mean difference between the two groups was significant at the point oh (.0) one level." What on earth do they mean? They mean that the difference between the average scores of the two groups probably didn't happen by chance. More precisely, the chances that it *did* happen by chance are less than one in one hundred. This is written as $p < .01$. The "p" stands for "probability"—the probability that the results could have happened by chance.

5. WHY DO WE NEED TESTS OF STATISTICAL SIGNIFICANCE?

Because we use samples, not total populations.

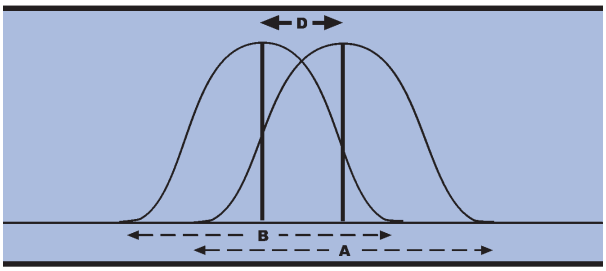
Let's take the simplest case where we are comparing only two groups. Let's say one group of students is taught to read with whole language, another with phonics. At the end of the year we administer a reading test and find that the two groups differ. Is it likely or unlikely that that difference occurred by chance? That's what a test of statistical significance tells us.

You might well ask, if the two groups actually had the same average score, why did we find any difference in the first place? The answer is that we are dealing with samples, not populations. If you give the test to everyone (the total population), whatever difference you find is real, no matter how large or small it is (presuming, for the moment, there is no measurement error in the test). But any given sample might not be representative of the population. This is particularly true in educational research that often must use "samples of convenience," that is, the kids in nearby schools. If you compared two different samples, you might get a different result. If you compared phonics against whole language in another school, you might get somewhat different scores, and it is unlikely that the difference would be *exactly* as you found it in the earlier comparison.

6. HOW DOES STATISTICAL SIGNIFICANCE RELATE TO PRACTICAL SIGNIFICANCE?

It doesn't. The results from an experiment like our example above can be highly significant statistically, but still have no practical import. Conversely, statistically insignificant results can be immensely important in practical terms. To repeat, statistical significance is only a statement of odds: "How likely was it that the differences we observed occurred by chance?" It's important to keep this in mind because many researchers have been trained in the use of statistical significance and act as if statistical and practical significance are the same thing. The chances of finding a statistically significant result increase as the sample becomes larger. The most common statistical tests were designed for small samples, about the size of a classroom. If the samples are large, tiny differences can become significant. As samples grow in size we become more confident that we're getting a representative sample, a sample that accurately represents the whole population.

The decision about practical significance must be weighed in other terms. For instance, can we find collateral evidence that students who are taught reading by whole language differ from students who are taught with phonics? Do teachers report that the kids taught with one method or the other *like* reading more? Do the two groups differ in how much the kids read at home? How much do the two programs cost? Do the benefits of either program justify those costs?



Let's take an example. Suppose that the two distributions above represent the scores of students who had learned to read with two different instructional programs. Their average scores differ by the amount, D. A test of statistical significance will tell us how likely it was that a D that large could have occurred by chance if, in the whole population D=0.

Now what? Well, it looks like we should consider A over B. But that decision cannot be based solely from the statistical results. The calculation of an "effect size" (described in the next section) will give some idea of how big the difference is in practical terms, but it alone will not lead to a decision. We need to determine for certain that a test was equally fair to both programs. In one study that compared phonics against whole language, students in both programs scored about the same on a standardized test. Students in the phonics program, however, scored poorly on a test about character, plot, and setting – aspects of reading treated by whole language, but not phonics.

If we think the statistical result is valid, then we can ask questions like: How do the teachers feel about the two programs? How do the students feel? Does one program cost much more than the other? How much additional teacher training is required for teachers to become competent in the two programs? Do students in one program spend more time voluntarily reading than students in the other? A "programmed text" used to teach B.F. Skinner's notions about learning was used in undergraduate psychology programs in the 1960s. It was touted as insuring that students would master the concepts. They did. But the format of the book made it simultaneously difficult to read and boring. Students came away hating both Skinner and programmed texts.

7. WHAT IS A CORRELATION COEFFICIENT?

"Correlation coefficients" show how changes in one variable are related to changes in another. One example used several times in this document is the correlation between SAT scores and college freshman grades. People who get higher scores on the SAT tend to have higher college grades in their freshman year. This is an indication of a positive correlation: as test scores get higher, grades tend to increase as well.

The important word in the last sentence is "tend." Not all people who score well on the SAT will do well in college. If the relationship between scores and grades were perfect and positive, then the correlation would be at its highest possible value, +1.00. If the relationship between test scores and grades were perfect and negative, the correlation coefficient would be -1.00. This would describe a peculiar situation in which people with the highest test scores received the lowest grades.

All this statistical terminology is important when reading and interpreting test and test scores, the subject of Part II.

PART II
THE TERMS OF TESTING

1. WHAT IS STANDARDIZED ABOUT A STANDARDIZED TEST?

Virtually everything. The questions are the same for all test takers. They are in the same format for all takers (usually, but not exclusively, the multiple-choice format). The instructions are the same for all students (some exceptions exist for students with certain handicaps). The time limits are the same (some exceptions exist for students with certain handicaps). The scoring is the same for all test takers, and there is no room for interpretation. The way scores are reported to parents or school staff are the same for all takers. The procedures for creating the test itself are quite standardized. The statistics used to analyze the test are standardized.

Where interpretations of open-ended responses are possible, as in some individually administered IQ tests, the administrators themselves are quite standardized. That is, they are trained in how to give the test, what answer variations to accept and what to refuse (this is especially important when testing young children who are anything but standardized), and how, generally, to behave in the test setting. It would not do to have an IQ score jump from 100 to 130 or fall to 70 based on who was giving the child the test.

2. WHAT IS A NORM? WHAT IS A NORM-REFERENCED TEST?

The norm is a particular median, the median of a norm-referenced, standardized test. It and other medians are also referred to as the 50th percentile. Whatever score divides testtakers into two groups with 50% of the scores above and 50% below that score, that is the norm.

Test publishers refer to the median of their tests as "the national norm." If the test has been properly constructed, the average student in the nation would score at the national norm.

Unlike internal body temperature, there is nothing evaluative about the norm of a test. Ninety-eight point six degrees Fahrenheit (98.6^{o} F) is the norm for body temperature. It is one indicator of health and departures from this norm are bad. The norm in test scores, though, merely denotes a place in the *middle* of the distribution of scores. (Yet, some administrators place students in remedial classes or Gifted & Talented programs solely on the basis of the students' relations to this norm.)

Once the norm has been determined, all other scores are described in reference to this norm, hence the term "norm-referenced test." The Iowa Tests of Basic Skills and other commercial tests of achievement, the SAT, and IQ tests, are all examples of norm-referenced tests.

The idea of establishing national norms in this way disturbs some people because, by definition, half of all people who take the test are below average. They argue that it might hurt children to think they are below average when they are actually doing quite well.

How can one be doing quite well and still be below average? Because a norm-referenced test tells you nothing about how well *anyone* is doing. If you score at the 75th percentile on such a test, you know you did better than 75% of other test

takers. That's all. Maybe *everyone* who took the test did poorly. You just happen to be better than 75% of the group. On the other hand, if you score at the 30th percentile of people taking the Graduate Record Examination (GRE), you are "below average" but still in a fairly elite group. If you bothered to take the GRE, chances are you will complete four years or more of college, something accomplished by only a quarter of all adults in the country, and by only 50% of those who begin college today.

This is important to keep in mind: *scores from a norm-referenced test are always relative, never absolute*.[1] If you visit Africa and rank your height with a group of Watusis, chances are you'll be below average; if you visit pygmies and perform the same measurements, you might be at the 99th percentile. Your absolute height never changed, but the nature of the reference group did.

Moreover, about every five years, test publishers re-norm their commercial achievement tests. Curricula change to reflect changes in knowledge or changes in instructional emphasis. The old tests might not measure the contents of the new curricula. So test publishers must renorm every so often to keep the tests current. There is overwhelming evidence that educational achievement has fluctuated up and down over the last 40 years so that the "50th percentile" reflects different amounts of achievement at different times.

To get away from the relativism of norm-referenced tests, people have sought to develop tests that have "criterion-referenced scores."

[1]*Until 1996, the SAT was an exception to this rule. Its norm was established in 1941 and was a fixed norm until the College Board "recentered" the SAT in 1996. Recentering is the same as "renorming," something that commercial achievement test publishers do about every five years.*

3. WHAT IS A CRITERION-REFERENCED TEST?

In theory, for any task, we can imagine achievement on a continuum from total lack of skill to conspicuous excellence. Any level of achievement along that continuum can be referenced to specific performance criteria. For instance, if the skill were ice-skating, the continuum might run from "Cannot Stand Alone on Ice" to "Lands Triple Axel." Professional baseball uses a criterion-referenced system. The major leagues represent "conspicuous excellence" whereas the various levels of farm teams represent different points of achievement on the continuum. We can train judges to agree almost unanimously about the quality of performance.

Unfortunately, the educational domains are not nearly so specific as those found in athletics. The "criteria" of criterion-referenced tests are generally limited to establishing a cut score on some test. Many current tests that are called criterion-referenced would be better referred to as "content-referenced." Thus in Virginia's Standards of Learning Program, the Commonwealth of Virginia described certain content that students should strive to learn. Tests were then developed to measure how well the students have mastered the material specified in the standard.

These tests have cut scores, scores that determine whether a student passes or fails. This cut score is often referred to as the "criterion." As a consequence, these tests are often referred to as criterion-referenced tests, but the phrase is not used in the original sense outlined in the first paragraph above. The "criterion" is simply attaining a score above the designated cut score in order to graduate from high school. If the cut score is, say 70, all that matters is getting a 70 or better. A pass-fail decision is based on the score,

nothing else. A true criterion-referenced test would have criteria associated with scores above 70 and with the lower scores as well.

In most states, the test for a driver's license is partly a content-referenced test with a "criterion" and also a true criterion-referenced test. The paper-and-pencil test covers specific content and applicants must get a certain number correct to pass. In addition, there is a behind-the-wheel test with true criteria. For instance, the applicant must parallel park the car within a certain distance of the curb and without knocking over the poles that represent other cars.

4. HOW ARE NORM-REFERENCED AND CRITERION-REFERENCED TESTS DEVELOPED?

The procedures for the two tests are quite different. In norm-referenced tests, the test publishers examine the curriculum materials produced by the various textbook and workbook publishers. Then item writers construct items to measure the skills and topics most commonly reflected in these books. These items are then judged by panels of experts for their "content validity." Content validity is an index of whether or not a test measures what it says it measures (considered in more detail in the section on test validity). A test that claims to be a measure of reading skills but which consists only of vocabulary items would not have high content validity.

After that, the items must be tried out to see if they "behave" properly. Proper behavior in an item is a statistical concept. If too many people get the item right or too many people get it wrong, the item does not behave properly. Most items included on norm-referenced tests are those that between 30% and 70% of the students get right in the tryouts. The test maker will also eliminate questions that people with overall high scores get wrong and people with overall low scores get right. The theory is that when that happens, there is something peculiar about the item.

Test makers choose items falling in the 30-70% correct range because of how norm-referenced tests are generally used. They are used to make differential predictions (e.g., who will succeed in college) or to allot rewards differentially (e.g., who gets admitted to gifted and talented programs). If everyone gets items right or if everyone gets items wrong, everyone would have the same score and no differential predictions would be possible. Keep in mind that a principal use of norm-referenced tests is to make such predictions.

For norm-referenced tests, vocabulary must be restricted to words that everyone can be expected to know except, of course, on a vocabulary test. Terms that were taken from specialized areas such as art or music, for example, would be novel to many students who then might pick a wrong (or a right) answer for the wrong reason. A teacher-made test, on the other hand, can incorporate words that have recently been used in instruction, whether or not they are commonly familiar to most people.

As a small digression, we observe that building a test with "words that everyone can be expected to know" is not as simple as one might initially think. In a polyglot nation such as the United States, different subcultures use different words. A small war was waged over the word "regatta" which appeared in some editions of the SAT. People argued that students from low-income families would be much less likely to encounter "regatta" or similar words that reflected activities only of the affluent.

The process of developing a criterion-referenced test is quite different. For most such tests, a set of objectives and perhaps even an entire curriculum is specified and the goal of the test is

to determine how well the students have mastered the objectives or curriculum. As with teacher-made tests, a criterion-referenced test can contain words that are unusual or rare in everyday speech and reading, as long as they occur in the curriculum and as long as the students have had an opportunity to learn them.

With a criterion-referenced test, we are not much interested in differentiating students by their scores. Indeed, the goal of some such tests, such as for a driver's license, is to have everyone attain a passing mark. When criterion-referenced tests do differentiate among students it is usually to place them into categories—such as basic, proficient and advanced—rather than to line students up by percentile ranks.

Historically, most of the tests used in the United States have been norm-referenced: standardized achievement tests, the SAT and ACT, IQ tests, etc. Recently developed tests' state standards are criterion-referenced in the sense of having a cut score.

Both norm-referenced and criterion-referenced tests must be evaluated in terms of two technical qualities, reliability and validity, considered next.

5. WHAT IS RELIABILITY IN A TEST?

In testing, reliability is a measure of consistency. That is, if a group of people took a test on two different occasions, they should get pretty much the same scores both times (we assume that no memory of the first occasion carries over to the second). If people scored high at time one and low at time two, we wouldn't have any basis for interpreting what the test means.

Initially, the most common means of determining reliability was to have a person take the same test twice or to take alternate forms of a test. The scores of the two administrations of the test would be correlated. Generally, one would hope for a correlation between the two administrations to reach .85 or higher, approaching the maximum a correlation can be, +1.00. (See WHAT IS A CORRELATION COFFICIENT? for an explanation of what values it can take.)

Testing people twice is often inconvenient. There is also the problem of timing: if the second administration comes too close to the first, the memory of the first testing might affect the second. If the interval between tests is too long, many things in a person's cognitive makeup can change and might lower the correlation. An alternative to test-retest reliability is called split-half reliability. This means treating each half of the test as an independent test and correlating the two halves. Usually the odd-numbered questions are correlated with the even-numbered ones.

6. WHAT IS VALIDITY IN A TEST?

Reliability is the *sine qua non* of a test: if it's not reliable, it has to be jettisoned. However, a test can be reliable without being valid. If a target shooter fires ten rounds that all hit at the "two o-clock" position of the target, but a foot away from the bull's eye, we could say that the shooter was reliable—he hits the same place each time—but not valid since the goal is the bull's eye.

Validity is somewhat more complicated than reliability. There are several terms that can be used preceding the word validity: content, criterion, construct, consequential, and face. A test has content validity if it measures what it says it is measuring. This requires people to analyze the test content in relation to what the test is supposed to measure. This might require, in the case of criterion-referenced tests, holding the test up against the contents of a syllabus.

Criterion-related validity, also called predictive validity, occurs if a test predicts something that we are interested in predicting. The SAT was developed to predict freshman grades in college. To see if it does, we correlate the two scores on the test with grades. If the test has predictive validity, those who score high on it will also tend to get better grades than those who score low.

Determining whether or not a test has *sufficient* predictive validity to justify its continuance is a matter of judgment or cost-benefit analysis. Few if any colleges would require the SAT if they had to pay for it. (Students now pay the costs.) The predictions from high schools and rank-in-class would be high enough. The SAT adds little to the accuracy of the predictions and would cost colleges millions of dollars if they, rather than the applicants, bore the cost.

Construct validity is a more abstract concept. It is a bit like content validity in that we are trying to determine if a test measures what it says it does, but this time we are not interested in content, such as arithmetic or history, but in psychological constructs such as intelligence, anxiety or self-esteem. Construct validity is of interest mostly to other professionals working in the field of the construct. They would try to determine if a new test of, say anxiety, yielded better information for purposes of treatment or if it fit better with other constructs in the field.

Consequential validity refers to a test's consequences and whether or not we approve of those consequences. It also refers to inferences made from the test results. For instance, once a test is known, teachers often spend more time teaching material that is on the test than material that is not. Is that a good thing? The answer depends on how we judge what is being emphasized and what is being left out. It might be that the test is doing a good job of focusing teachers' attention on important material, but it might be that the test is causing teachers to slight other, equally important material and to narrow their teaching too much. Numerous states have developed tests to determine if students have mastered certain content and skills. On the first administration of these tests, many students failed. Some inferred that teachers were not teaching the proper material or were not teaching well. Others inferred that the students weren't learning well. Others inferred that the cut scores on the tests were set too high. And some said the tests were simply no good. These were all consequences of using the test.

Researchers have differed on the importance of "face validity." Face validity has to do with how the test appears to the test taker. If the content of the test appears inappropriate or irrelevant, the test taker's cooperation with the test is compromised, possibly disturbing the other kinds of validity as well.

7. WHAT IS A PERCENTILE RANK? A GRADE EQUIVALENT? A SCALED SCORE? A STANINE?

These terms are all metrics that are used to report test results. The first two are the most common while stanine is seldom used any more. It stands for "standard nine" and was a means of collapsing percentile ranks into nine categories. This was important at the time it was invented because data were processed in computers by means of 80-column punch cards and space on the cards was at a premium. By condensing the 99 percentile ranks into 9 stanines, testing results would occupy only one column.

Percentile ranks, grade equivalents, and normal curve equivalents pertain to norm-referenced tests only. Scaled scores are used for both norm-referenced and criterion-referenced tests.

Percentile ranks. Percentile ranks provide information in terms of how a given child, class, school, or district performed in relation to other children, classes, schools, or districts. A student in the first percentile is outranked by everyone, a student in the 99th percentile outranks everyone and a student at the 50th percentile is at the national average.

It is important to note that percentiles are ranks, not scores. From rankings alone you cannot tell anything about performance. When the final eight sprinters run the100 meter dash in the Olympics, someone *must* rank last. This person is still the 8th fastest human being on the planet that day. Percentile ranks are usually reported in relation to some nationally representative group, but they can be tailored to "local norms."

Large cities often compare themselves to other large cities in order to avoid the national rankings that include scores from suburbs. Suburbs seldom compare themselves to other suburbs because they look better when compared to national samples that contain students from large cities and poor rural areas.

Grade equivalents. Grade equivalents also rate students in reference to the performance of the average student. A grade equivalent of 3.6 would be assigned to the student who received an average score on a test given in the sixth month of the third grade. If a student in the fourth month of the fourth grade receives a grade equivalent of 4.4 on a test, that student is said to be "at grade level." This manner of conceptualizing grade level creates a great deal of confusion.

Newspapers sometimes start scandals by reporting half of the students in some school are "not reading at grade level." There is no scandal. We have defined "grade level" as the score of the average student. Therefore, nationally, half of all students are *always* below grade level. By definition.

We don't have to define grade level this way. We could give grade level a criterion-referenced interpretation and hope that all children achieve it, but it is not usually defined with a criterion-referenced meaning.

The concept of grade level also creates confusion when students score above or below their grade level. Parents of fourth graders whose children are reading at, say, the seventh grade level will wonder why their child isn't in the seventh grade, at least for reading. But a fourth grader receiving a grade equivalent of seven on a test is not reading like a seventh grader. This is the grade equivalent that the average seventh grader would obtain *reading fourth grade material*. It is unlikely—but not impossible—that a fourth grader reading at seventh grade level could actually cope with seventh grade reading material.

A "**scaled score**" is hard to explain without getting into a lot of statistical detail. Conceptually, scaling converts raw scores into a metric in terms of the standard deviation. (Please see chart showing standard deviation on p. 4). Suppose one test was 100 items long and another only 50. A raw score of 43 would likely mean very different things on the two tests. But both tests can be converted to a scale in terms of their standard deviations.

Converting to scaled score from a raw score produces a scale with an average of 0.0 and a standard deviation of 1.0 (the average score minus the average score = 0, and 0 divided by anything is zero). Statisticians early on decided that such a scale didn't look very pretty. As it happens, you can add a constant to all scaled scores or multiply them all by a constant without changing their relationships to each other. A distribution of scaled scores has a mean of 0.0 and a standard deviation of 1.0. If we multiply all the scaled scores by 100 and add 15 we get the common IQ scale – a mean of 100 and a standard deviation of 15. If we multiply them by 100 and add 500 we get the scale of the SAT – a mean of 500 and a standard deviation of 100.

Scaled scores also permit normative comparison of scores across different scales: an IQ score of 115 is the "same" as an SAT verbal score of 600 because both are one standard deviation above the average, which are 100 and 500, respectively. A person with an IQ score of 115 and an SAT verbal score of 600 has scored at the 84th percentile on both tests. If the scores on the two tests had been different, we would want to explore whether or not the tests were measuring different constructs. (In this actual case, the contructs are highly correlated. When the SAT was invented in 1926 it was referred to as an intelligence test). Scaled scores can only be meaningfully interpreted if the scores fall into a normal, or bell-shaped curve, or a close approximation thereof.

8. WHAT ARE MULTIPLE CHOICE QUESTIONS?

They are questions in which one reads some material, then picks an answer from a list of pre-selected answers, usually four or five. Invented in 1914 by Frederick J. Kelly at the University of Kansas, multiple choice questions made possible the mass testing of military recruits in World War I to assess differential talents and abilities. Everyone could take the test at the same time, in the same format, in a short period of time, and the answers could be scored quickly and cheaply.

These qualities still figure in why multiple-choice tests are often favored today. Since World War I, the principal changes in multiple-choice technology have been developments in the scoring technology. Computers scan and score thousands of answer sheets in an hour. The chief disadvantage of multiple-choice questions is that they usually test small samples of knowledge out of context. Multiple-choice tests that tap higher-order thinking can be built, but one rarely sees them except in graduate schools.

9. WHAT DO MULTIPLE CHOICE TESTS TEST?

This vague question could have a variety of answers, but as used here it refers to the "level" of knowledge or skill that various tests test. Many people object to multiple-choice tests on the grounds that they can only test "factoids" or small bits of decontextualized knowledge. Others contend that multiple-choice tests can test reasoning and higher-order thinking as well as any other kind of test. The resolution of this dispute would appear to lie in the word "can." Multiple-choice questions *can* test all kinds of analytical skills, but they rarely *do*.

The use of multiple-choice tests in testing higher order thinking is largely seen in graduate schools, not in tests used in elementary or secondary school or anywhere else on a large scale. These tests might describe an experiment in, say, psychology. The exposition of the experiment might take a full page or more—a far longer "stem" than seen in other tests. The questions would then ask the students to draw conclusions about what the results of the experiment showed. A complete test might consist of only two or three such passages with four-to-eight questions built around each.

By contrast, most tests used in schools require students to answer many questions in a short period of time. Students who stop to think about a question are in trouble: they won't finish the test. Indeed, one piece of test taking advice the College Board gives to students practicing for the SAT is, "Keep Moving." The SAT, more complex than most achievement tests, contains questions like the following:

Rib Cage: Lung:

a) skull:brain
b) appendix:organ
c) sock:foot
d) skeleton:body
e) hair:scalp

Or, If the product of five integers is negative, at most how many of the five can be negative?

a) 1
b) 2
c) 3
d) 4
e) 5

In the first question, the person must evaluate each alternative answer to see which one describes a relationship most like that in the stem of the question. This analogical thinking is important in the real world, but seldom is such thinking so constrained by the format of an item. Nor does it take place in so brief a time.

In the second question, the students need only recall that the product of two negative integers is positive and that the product of a positive and a negative is negative. Therefore, any group of integers that contains an odd number of integers can all be negative (even: -1 x-1 x -1 x-1 = +1; multiplying by another -1 = -1 and so on). Those who advocate performance assessment or "authentic" assessment are generally trying to go beyond the limits of multiple-choice tests.

10. WHAT IS "AUTHENTIC" ASSESSMENT?

Authentic assessment is an attempt to measure performance directly in a "real life" setting. In a multiple-choice test it is usually impossible to know why a student chooses a particular answer. Rote memory? Lucky guess? Guess after eliminating two wrong answers? A well reasoned line of thought to a right answer? A well reasoned line of thought to a wrong answer?

In tests of arithmetic, clues can be garnered occasionally because the incorrect responses can be chosen to reflect particular kinds of misunderstandings. This sort of diagnostic use of tests is quite difficult in other subject areas. For these reasons and others, some people have become interested in constructing tests that assess performance in more "authentic" settings.

The word "authentic" is not an especially good choice because it implies that any other kind of assessment is "inauthentic." Perhaps a better word would be "direct" assessment. All tasks that have been given the name "authentic" are some form of direct assessment. Advocates contend that we can't determine how well someone knows a skill or a body of knowledge unless we have an opportunity to directly observe a performance. We can't tell much about one's writing skills by using multiple-choice tests. In such tests, a person reads a few sentences where some parts are underlined. The person then picks one of four or five choices about which underlined part contains a misspelling, a grammatical error or improper syntax. To learn about students' writing skills, we must observe them *perform*—they must write! We must have the students perform by making the edits themselves.

Beyond this, advocates contend that the assessment should reflect some real-life, complex problem. Such assessments are necessarily time-consuming and, therefore, expensive. Thus, multiple-choice tests are more frequently chosen, especially for large scale accountability purposes. In instructional settings, authentic, direct assessments are usually preferable.

11. WHAT ARE PERFORMANCE TESTS?

Performance tests are closely related to authentic assessment. We can say that all authentic assessment involves performance, but there might be some trivial performances that do not qualify as authentic assessment. For instance, assessing student's writing involves a performance: they must write. But what are they writing about? If the assignment is trivial or banal, authentic assessment is not happening.

Authenticity can also be subverted by scoring. For instance, when students do write, someone has to score their essays. If the writing is part of a statewide testing program, the expositions will not be scored by local teachers, but some organization specialized in the scoring of writing samples. A statewide assessment generates many essays and as a consequence, the scorers have to score the essays very fast, as fast as one every ten seconds. Such speed precludes thoughtful attention to the essay. Essays are judged according to formula and genuine creativity might well be punished. Thus instruction as well as assessment is subverted.

12. WHAT ARE PORTFOLIOS?

Portfolios are one variety of performance assessment, usually collections of various kinds of displayed productions. Some districts also use mathematics portfolios which can be collections of problem-solving activities and examples of how the students solved the mathematics problem. In science portfolios, the results of experiments or other investigations can be collected. Writing portfolios are the most common, however, and are considered analogous to the portfolios of artists, collections that show a range of writing exhibitions from expositions, to narrative to poetry. School-to-work programs frequently involve portfolios to demonstrate real work accomplished.

13. WHAT IS A "HIGH STAKES" TEST?

A high stakes test is one that results in some kind of punishment for those who score low or some kind of reward for those who score high, or, occasionally, both. For many years, the most common high stakes tests were the SAT and ACT. Students who scored high had a better chance of admission to selective colleges. This is still true, although after the baby boom passed through, many colleges switched from selecting students to recruiting them in order to maintain or increase the number of their programs and faculty. IQ tests were sometimes also high stakes, resulting in children being placed in gifted and talented programs on the one hand, or low track or special education programs on the other.

The problem with high stakes tests is that they cause people to pay too much attention to increasing scores, to the detriment of a more comprehensive education. When a lot is riding on the outcome, teachers will teach test-taking skills, too closely align their teaching with the test, and even cheat occasionally in order to look good, to make their principal or district look good, or to keep their jobs.

14. WHAT IS AN IQ TEST?

An IQ, or Intelligence Quotient, test measures certain thinking skills that are mostly school-related, but not school specific. They were developed initially in France to determine which children could not benefit from regular school programs. When transported to this country, IQ tests were seen as measuring "g" or a general mental factor that determined much of a person's thinking abilities. The "g" factor, in turn, was seen as an entity controlled by a single gene. The role of genetics in determining intelligence is still very much in debate, as seen in the controversy over the book *The Bell Curve*.[2]

The genetic theory of intelligence was countered by those who believe experience and environment are more important, giving rise since the 1920's to the "nature-nurture" debate. The either-or characterization of this debate is now recognized as naïve, but the size and importance of the roles played by genes vs. the environment is still argued. Some contend that genes account for as little as 20% of intelligence and others hold out for an 80% determination.

From the outset, not everyone subscribed to the "g" factor theory. Another popular theory argues that intelligence is composed of a number of specific abilities. In 1983, Howard Gardner put forth a theory of seven separate-but-equal intelligences that has become popular among educators (he has since added two additional intelligences).

IQ tests consist of specific skills, such as repeating a number of random digits, or using blocks to copy a design shown by the tester. The Stanford-Binet, one of the most popular IQ tests,

uses 15 subtests grouped into four categories: verbal reasoning, quantitative reasoning, abstract/visual reasoning and short-term memory.

IQ tests have been criticized for being culturally biased. This is a highly charged area, much too complex for resolution here. Suffice to say that children from more affluent families are more likely to have life experiences that contribute to higher performance on IQ tests. For example, one study found that middle class parents talked to their children four times as much as parents living in poverty. They also get earlier exposure to books.

Tests like the Stanford-Binet or the Wechsler IQ tests must be administered individually by highly trained administrators. Some group-administered IQ tests have been developed, but these typically are not called IQ tests. These are usually given in school in conjunction with achievement tests. The scores on these tests are used to "predict" the student's academic achievement test scores. IQ tests are generally thought of as one kind of "ability" test in contrast to achievement tests. This is not conceptually sound as noted in the next section.

---

[2]*Herrnstein, R. J., and C. Murray. 1994.* The Bell Curve: Intelligence and Class Structure in American Life. *New York: The Free Press.*

15. WHAT IS THE DIFFERENCE BETWEEN AN ABILITY OR APTITUDE TEST AND AN ACHIEVEMENT TEST?

Principally, achievement tests are more directly connected to what is taught in school than are ability tests. Most people think these are different "types" of tests, and the distinction has caused a great deal of mischief. Most people think of ability in terms of "potential." Students who score lower on an achievement test than an ability test are often labeled "underachiever," that is, not living up to their ability. "Overachiever" labels go to people whose achievement test scores are higher than their ability test scores.

However, a single test can never measure "potential." All it can measure is what the students know and can do at the single point in time when they take the test. When looked at closely, the distinctions between achievement and ability become conceptually fuzzy; tests with the different names don't necessarily measure different things. About all we can validly say is that the knowledge and skills that are tested on achievement tests look like the kinds of things that are taught in schools, while the skills tested on ability tests don't seem school-based and rely less on specific knowledge. Some ability tests have analogy questions. Analogical thinking is important to success in school, and probably life too, but teachers rarely explicitly teach children to figure analogies.

Some ability tests also present perceptual items. These often take the form of a series of four or five geometrical figures. The student's task is to select which of another group of figures would be the next one in the series. Students who are good at this sort of item often develop good skills at videogames or perceptual games such as chess. Students who score high on these "nonverbal" tests and low on the verbal and quantitative parts of ability tests have a difficult time in school. These children are perceptually oriented, but schools are all about symbols: numbers and letters.

Some have argued that ability tests predict future achievement and that achievement tests summarize past achievement. This is merely a convention describing common uses. Achievement tests can be used to predict future achievement. When we make such predictions, we are correlating the test scores with something in the future, like

college success. Any test can be used to make such predictions.

In fact, the correlations involved can be calculated for any two variables. We could use height or weight or density of eyebrows to predict future achievement—all we need to do is plug the different heights, weights or number of hairs of people and their college grades into the equation. Whether such predictions would yield meaningful, statistically significant results or not is another question. We might find that eyebrow density did not predict anything, in which case we would have to stop using it as a predictor. And even if it predicted grades, it's not clear that admissions policies should then be changed to take that into account. Typically, achievement tests given in high school predict college grades as well as the most-often-used "aptitude" or "ability" test, the SAT.

16. WHAT ARE THE ITBS, ITED, TAP, STANFORD-9, METRO, CTBS AND TERRA NOVA?

Except for the ITED, these are all popular commercially-produced, norm-referenced achievement tests: the Iowa Tests of Basic Skills, the Iowa Tests of Educational Development; Tests of Achievement and Proficiency; the 9th version of the Stanford Achievement Tests; the Metropolitan Achievement Tests; the Comprehensive Tests of Basic Skills; and a new version of the Comprehensive Tests of Basic Skills with a fancy name, Terra Nova.

A complete "battery," as they are often called, offers tests of reading, mathematics, language arts, vocabulary, science and social studies. The latter two tests are not used as often as the first three because of wide program variation in science and social studies curricula among schools. Unless the science and social studies curricula have been specifically aligned with the tests, the tests might not reflect what is being taught at a particular grade.

The ITED, for grades 9-12, is not used in many places because it is considerably more difficult than the others. It contains long reading passages, requires students to solve multi-step mathematics problems, and to analyze simulated science experiments. Most states and districts substitute the easier TAP.

17. WHAT IS A MINIMUM COMPETENCY TEST?

As originally conceived, a minimum competency test was an assurance that high school seniors were leaving school "minimally competent." In the 1970s, as now, people worried that students were being "socially promoted" on the basis of "seat time" and leaving school without having a minimal level of skill. It was soon seen, though, that the minimal level could not be specified through technical means. There was always some arbitrariness in establishing what skills would be tested and what the cut score would be.

Minimum competency tests became very popular, at one point existing in some form in 35 states. They have been replaced more recently by what is generally known as "the standards movement" which calls for "high standards" and "high expectations" "a challenging curriculum for all students"—something more than minimum. The cut scores were usually set so that sufficient numbers of students failed initially to satisfy those who had called for the tests in the first place, but so that by graduation time virtually everyone had passed. One court decision coming out of the minimum competency test era that might come around again held that in order for a state to withhold diplomas on the basis of a test, it had

to prove that the children had actually been provided opportunities to learn the material on the test (*Debra P v. Turlington*, 1981).

18. WHAT ARE ADVANCED PLACEMENT (AP) TESTS?

Advanced Placement Tests are taken by high school students to gain college credit. Since their inception in 1900, the College Board has attempted to "drive" instruction with assessments. Advanced Placement (AP) tests are the culmination of an effort to provide high school students with high quality instruction in areas of school study built around a particular curriculum and leading to tests based on that curriculum. Currently, although only about one half of U. S. high schools offer AP courses, over a million students take AP tests each year, a ten-fold increase over a twenty-year period.

The major incentive for taking the tests is college credit. Trained scorers mark AP tests on a five-point scale, and many colleges grant credit for a score of three or better. Since the tests cost much less than college courses, students who pass them get an accelerated start in college and save money at the same time. It is not necessary to take an AP course before taking an AP exam. Many high schools offer "advanced" or "accelerated" or "honors" courses which accomplish much the same thing without adhering strictly to the AP syllabus.

A secondary incentive for taking AP tests is college admission. Admissions officers have favored students who take AP examinations, particularly those who take more than one. An incentive for parents is money. The AP courses are provided free through the public schools and the tests cost around $75, considerably less than tuition for the same course and test in college.

19. WHAT IS THE INTERNATIONAL BACCALAUREATE?

The International Baccalaureate (IB) is a rigorous program of study that originated in Switzerland. IB examinations are sometimes compared with the Advanced Placement, but there is a difference. Students can take an AP test without having taken an AP course. To be eligible for an IB examination, students must be enrolled in a school that has been accredited through the fairly rigorous IB accreditation process and be taking the course for which they wish to be examined. In the IB system, the exam will count for 75% of the course grade. While the number of IB examinations given in this country has tripled in the last decade, it is still tiny in comparison to AP, with some 14,000 exams being taken annually.

20. WHAT IS THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS?

The National Assessment of Educational Progress (NAEP) began as a nationwide study of what students and young adults know and can do in the areas of reading, mathematics and science. Since its early days in the late 1960's, NAEP has added history, geography, writing and, most recently, art and civics to its assessments. The original idea behind NAEP was simply to establish what a sample of people know and don't know. Its creators viewed it very much like a health survey that might determine the incidence of various diseases. Without knowing the frequency of, say, tuberculosis, it would be difficult to know how much of an effort would be needed to eradicate it.

In 1982, the Educational Testing Service won the federal contract for administering the NAEP (the money is part of the budget of the U. S. Department of Education) and dubbed it "The Nation's Report Card." This can only partly be true because NAEP is aligned with no particular curriculum. Students learning mathematics in a "Connected Math" program might well have different NAEP scores from those learning in a Saxon Math district. Therefore, one cannot specify that one curriculum is "better" than the other.

When proposed, many people and organizations feared NAEP would lead to a standard national curriculum and federal control of education. As a consequence, NAEP was housed in a state-supported policy agency, the Denver-based Education Commission of the States, and forbidden to report data in any aggregation smaller than "region." In 1988, new federal legislation permitted state-level reporting and about 40 states now participate in NAEP state-by-state assessments.

21. WHAT IS THE NATIONAL ASSESSMENT GOVERNING BOARD?

In the 1980's a National Assessment Governing Board (NAGB, pronounced, NAG bee) was formed to provide policy guidelines for conduct of NAEP. NAGB undertook to change NAEP from a "what is" assessment to a "what should be" program. That is, NAGB took NAEP from being descriptive to being prescriptive.

To do this, NAGB established "proficiency levels" for each of the tests calling performances either "basic," "proficient," or "advanced" (it is possible to score "below basic," but this is not really a level like the others). These proficiency levels have been beset with criticism from studies conducted by the General Accounting Office, the Center for Research in Evaluation, Standards, and Student Testing, and various eminent psychometricians around the country. In the spring of 1999, the National Research Council declared that the proficiency levels were "fundamentally flawed" and should be replaced.

The proficiency levels do not provide a perspective on student performance that is corroborated by other indicators. For instance, in the most recent NAEP mathematics and science assessments, few fourth graders attained "proficient" and virtually no fourth graders garnered "advanced." Yet these same fourth graders scored above average in mathematics when compared to students in 26 nations, and third in the world in science. In addition, NAEP math and science scores have risen since 1977, the first year for which long-term trend data were collected.[3]

[3]*See* Do You Know the GOOD NEWS About American Education? *Washington, D.C.: Center on Education Policy and American Youth Policy Forum, 2000, pp. 12-15.*

22. WHAT IS THE THIRD INTERNATIONAL MATHEMATICS AND SCIENCE STUDY (TIMSS)?

TIMSS is the third attempt by educators to compare achievement in mathematics and science across nations. International comparisons have become a popular barometer of how schools around the world are doing. The Third International Mathematics and Science Study (TIMSS) is, at this writing, the largest, most recent and best-controlled study of its sort.

It has its problems, though. For one thing, the reliability of the tests is not impressive, something that has been largely overlooked. TIMSS

administered tests to students in 26 countries at grade 4, 41 countries at grade 8, and 16 to 21 countries, depending on the test, at the Final Year of Secondary School. It is called Final Year because, in many instances, it does not correspond to 12th grade in the United States.

TIMSS has generated a popular, but false, cliché: the longer American students stay in school, the farther behind their foreign peers they fall. The cliché derives from the fact that American students score very high in both math and science at the 4th grade level, average at the 8th grade level and nearly last in the Final Year study. The slippage from 4th grade to 8th grade is probably real, but the further decline between 8th and 12th probably is not.

One of the findings from the curriculum study segment of TIMSS was that American educators consider the middle school years the culmination of elementary school while most other industrialized nations view it as the start of high school and more intense academic study. The consequence is that 7th and 8th grades in numerous other countries feature the study of algebra and geometry, while only about 15% of American 8th graders receive instruction in algebra. The rest receive a review of earlier topics. In part, this review is necessary because of another curriculum finding from TIMSS: American textbooks are about three times as thick as those in other nations. Teachers in other countries teach fewer topics and spend a longer time on each. American teachers try to teach everything in the texts. This makes coverage often brief and superficial.

The Final Year results indicating poor performance by American students are fatally flawed. Only five nations met the criteria established by the study itself for valid data. Moreover, the educational systems of most other nations are no longer comparable to that of the United States after the 8th grade (or, in many instances, to each other). In other nations, students enter focused programs; some receive intensive study in math and science, others enter technical or vocational programs, others receive instruction in arts and the humanities. The length of these programs varies but the students in other nations averaged more than a year older than American students and some were as old as American *college* seniors.

There are also cultural differences among the countries that produce large differences in test scores. In most other nations, students are students, not both students and workers. But 55% of American students in the study indicated that they worked more than 21 hours a week. Research on the relationship between working and school performance finds that working up to 20 hours a week is associated with improved performance but, beyond that, working has a detrimental impact on schooling: students don't get enough sleep, slack off on homework and skip meals, especially breakfast.

American students who did not work a lot had scores at the international average, just where they were in the 8th grade. Those who worked 21-35 hours a week (28%) were well below average and those who worked more than 35 hours a week (27%) fell off the chart. When one parses out subgroups of American students who most closely resemble their foreign peers on other dimensions, they, too, have average scores as they did in eighth grade.

23. WHAT IS "HOW IN THE WORLD DO STUDENTS READ?"

This is the name of a 1992 book summarizing an international reading study conducted by the same organization that produced TIMSS. It is virtually unknown. American 10-year-olds and 14-year-olds were outscored by students in only one nation, Finland. There were 27 countries participating at the younger age, 31 at the older.

American students have consistently done well in international comparisons of reading. This is likely due to the concerted effort that elementary teachers make to teach reading and the lesser amount of time spent on math and science.

24. WHAT IS THE COLLEGE BOARD?

The College Board began in 1900 as a small collection of northeastern colleges and universities. It was for many years called the College Entrance Examination Board and its initial purpose was to bring coherence to the curricula of high schools. The colleges had found that students whose transcripts looked alike often had vastly difference experiences in terms of the sophistication and rigor of the courses taken. The Board thought it could eliminate the confusion by developing examinations in various topics. From these examinations, the high schools could determine what it was that the colleges valued and change their curricula accordingly.

Impressed by the testing procedures developed by the military during World War I, the Board decided to develop a single test to predict success in college. In 1926, it introduced the Scholastic Aptitude Test, almost always referred to by its initials, SAT. Most of the Board's activities are geared to support some aspect of the 3,300 institutions that constitute the Board's membership. It still views its major function as easing the transition from high school to college. Chartered as a nonprofit corporation, in the fall of 1999 the Board announced its first for-profit venture, a web site that will offer low-cost tutoring for the SAT and AP courses and financial aid information. The College Board might in the future provide AP courses online.

25. WHAT IS THE EDUCATIONAL TESTING SERVICE (ETS)?

ETS is a large testing and research-about-testing organization headquartered in Lawrence Township, New Jersey, near Princeton. It was spun off from the College Board in 1947. Its best known products are the Scholastic Assessment Test (nee Scholastic Aptitude Test) and, since 1982, the National Assessment of Educational Progress (NAEP). It also develops and administers law and medical school admissions tests and tests for use in business and industry.

26. WHAT IS THE SAT?

The College Board developed the SAT in 1926. Until 1994, the letters stood for "Scholastic Aptitude Test." The first SAT contained both multiple-choice and essay questions. When the onset of World War II prevented the administration of the essay portion, the Board decided to use only the multiple-choice section for all future administrations.

When ETS changed the name of the SAT in 1994 to Scholastic Assessment Test, it also started referring to the test as a "reasoning" test, but little was changed except an increased emphasis on "critical reading" and the deletion of the antonyms section.

There are 138 items on the new SAT and 180 minutes of testing time, so not a lot of "deep" reasoning is possible on each question. ETS

converts the raw scores into scale scores such that the mean is 500 and the standard deviation is 100, producing a scale that runs from 200 to 800.

Annually, about 1,200,000 seniors currently take the SAT. When juniors and sophomores are added in, ETS administers about 2,000,000 SATs a year.

27. WHAT IS THE PSAT?

The PSAT is the "Preliminary Scholastic Assessment Test." It is a shortened version of the SAT containing old SAT questions. It is sometimes taken by 10th graders for practice. It is also the sole criterion that qualifies students for National Merit scholarships. The latter use is problematical because boys score better on the PSAT and SAT than girls. As a consequence, boys win up to two-thirds of the scholarships. ETS added a writing test to the PSAT and, because girls do better than boys on this test, the differential awarding of scholarships has been cut by about 50%.

28. WHAT IS THE NATIONAL MERIT SCHOLARSHIP CORPORATION?

The National Merit Scholarship Corporation is an independent non-profit organization in Evanston, Illinois that administers two scholarship programs, the National Merit Scholarship Program and the National Achievement Scholarship Program. The Corporation uses the PSAT to qualify students.

Each year, about 35,000 students with the highest PSAT scores receive "Letters of Commendation," while another 15,000 are designated as semi-finalists. They are asked to fill out scholarship applications and eventually about 6,500 scholarships are awarded.

29. WHAT IS THE ACT?

These letters denote both a set of college admissions tests and the organization that makes them, the American College Testing Program located in Iowa City, Iowa. (The "P" is most often dropped.) Whereas the developers of the SAT wanted to identify academically gifted students and to bring them to Eastern seaboard universities, the ACT's developers were more interested in providing both academic and counseling information for *all* students who would be attending state schools, especially the land-grant colleges of the Midwest. About 900,000 seniors currently take the ACT battery. Most colleges now accept either the SAT or the ACT for admissions purposes.

30. WHAT IS FAIRTEST?

"FairTest" is the name most often used for what is formally the National Center for Fair and Open Testing in Cambridge, Massachusetts. FairTest began largely as an anti- ETS organization with its attention focused on the SAT. Since it's founding, it has widened its scope to be concerned with matters of gender and ethnic equity and with issues surrounding the "standards movement."

31. WHAT IS A STANDARD?

The word admits of many definitions: It can be a banner or something that records a magnitude like a platinum rod that sets the standards for length. Or it can be something ordinary or familiar like a standard grade of meat or standard equipment on a car. In the realm of education, though, standard is usually used in reference to a "degree or level of requirement, excellence or attainment" (a definition in the *American Heritage Dictionary"*).

The "standards movement" is not a formal organization or effort, but has grown out of a concern both that American students are not learning enough and that what they are learning is not of sufficiently high quality or rigor. Whether or not this is true is a matter of considerable debate.

32. WHAT IS A CONTENT STANDARD? WHAT IS A PERFORMANCE STANDARD?

Content standards specify **what**, performance standards, **how** much. Since the National Council of Teachers of Mathematics published their curriculum standards in 1989, most standards have been content standards, setting out what standards writers thought students should know or, at least, be exposed to. The tests that have been constructed around these content standards, with their accompanying cut scores for passing, can be considered performance standards. The NAEP proficiency levels discussed earlier were attempts to set performance standards on the various NAEP assessments.

33. WHAT IS ALIGNMENT?

Alignment refers to the degree to which the curriculum is aligned with a test and vice versa. Bringing the test into alignment with a curriculum is important. Otherwise the test will test things that are not taught. On the other hand, aligning a curriculum with a test has its pitfalls because the test covers only a small part of any curriculum. Alignment might well serve to narrow the curriculum.

In the evaluation of educational programs, it is important to have the test aligned with the program's objectives. Without alignment, an effective program might not look like one. In one evaluation of the remedial program "Success For All," for example, the goals of the test (the Comprehensive Tests of Basic Skills in this case) did not completely match the objectives of the instructional program. This might have attenuated the apparent program impact.

34. WHAT IS CREDENTIALING?

Credentialing is the use of tests to award or deny credentials or licenses for specific professions. A number of states use tests to credential or certify that teachers know enough to enter the classroom. The use of tests for this purpose has been hotly debated over the years. Some argue that much of teaching involves a set of skills unrelated to specific content knowledge and that these skills cannot be measured by paper and pencil tests. Others contend that all teachers need some minimal level of knowledge independent of whatever teaching skills they possess. There are also credentialing tests for lawyers, doctors, CPAs, and many other professionals. These tests are also developed by one of the private test publishing firms, usually working in coordination with the professional organization that oversees the profession, such as the American Medical Association, American Bar Association, etc.

PART III
SOME ISSUES IN TESTING

1. WHY IS TEACHING TO THE TEST A PROBLEM IN EDUCATIONAL SETTINGS, BUT NOT IN ATHLETIC SETTINGS?

About 75 years ago, an educator observed that tennis coaches "taught to the test." That is, they instructed their students in precisely those things they would need to be successful in their sport: how to serve, how to lob, how to volley, how to come to the net. This is teaching to the test, and it is widely accepted practice. Indeed, we would think any coach insane to do otherwise.

Why, then, is teaching to the test a problem in education? The answer is that the coaching of tennis or football incorporates all aspects of the sport and the coaching for a specific test in education usually does not. Football coaching might go awry if the opposing team installed new plays but, in this case, the "test" also becomes a teaching tool: the players will learn something in coping with the opposition's new plays.

The curriculum of, say, mathematics can be thought of as a large circle incorporating the entire field. The test is a series of smaller circles that sample parts of the large one. As long as teachers are working on the whole domain, the test is a valid representation of what is happening, just as a vein of ore represents the larger deposit. But if one concentrates only on the part of the domain covered by the test, education suffers. Theoretically, tests could cover an entire domain, but they would take many hours and many dollars to administer.

The achievement tests commonly used in schools usually have only 25-40 items to cover a subject. In some performance assessments, we approach a system like sports. We can teach aspects of writing and then have students write and observe how well they have learned those aspects. As we saw in the section on performance tests, though, this practice can be degraded if the student writing samples are graded rapidly using a formula that concentrates on a few elements and ignores or even punishes creativity.

2. WHO DEVELOPS TESTS?

Virtually all tests in this country are developed by for-profit publishing houses such as CTB McGraw-Hill, Riverside, or Harcourt Educational Measurement, or by private non-profit testing firms such as the American College Testing Program or Educational Testing Service. A few firms specialize: Measurement Incorporated in North Carolina scores writing samples; National Computer Systems in Iowa specializes in mass scoring of answer sheets; National Evaluation Systems in Massachusetts specializes in teacher tests; and Advanced Systems in New Hampshire specializes in custom-developing tests.

In recent years, more and more test development has taken place at the state level. At the initiative of a governor, legislature or state board of education, a testing program has been designed specifically for a particular state. Thus, there is the Texas Assessment of Academic Skills, the Virginia Standards of Learning tests, the Massachusetts Comprehensive Assessment System, and others.

In some cases, like Virginia's, the tests have been derived from a particular curriculum framework. These tests are initially developed by the private testing firms according to specifications from the states. The tests are then reviewed by teachers, supervisors, and university professors in the

various states. Some states, such as Virginia and North Carolina, have contracted with university researchers to determine if the tests meet technical requirements concerning reliability and validity.

3. WHAT AGENCIES OVERSEE THE PROPER USE OF TESTS?

There is virtually no regulation of the testing industry. The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education jointly developed and adopted *Standards for Test Use*, but little attention is paid to these standards except by persons conducting research using tests. When the tests have been misused, as recently in Chicago and California, neither the test publishers nor any of the three organizations named above have raised public objections to the violations.

In both Chicago and California, test scores alone are being used to determine whether or not children are promoted or retained in grade. This violates at least two standards: that a test alone should not be used to make decisions about people, and that a test designed for one purpose should not be arbitrarily applied to another purpose. The tests used in Chicago and California are norm-referenced achievement tests, which were not designed for promotion or retention decisions, nor technically are they up to the task. They are precise enough to say that a child is above or below average, but not precise enough to say that a particular child should spend another year in the same grade.

Various educators have called for some kind of "watchdog agency" to monitor testmakers. George Madaus of Boston College would like to see an "FDA for testing." Larry Cuban of Stanford University also argues that the advent of "high stakes" testing increases the need for an oversight agency since the true meaning of test numbers is easily distorted.

4. WHY DO CORRELATION COEFFICIENTS CAUSE SO MUCH MISCHIEF?

The correlation coefficient is the source of much misunderstanding because human brains appear to be wired to infer causation from mere correlation. However, given only a correlation coefficient statistic, we *cannot* infer causality. The two variables might be causally linked, or they might both be affected by a third variable, or the correlation might just happen by some artifact. There is, for example, a correlation between SAT scores and freshman college grades, but we cannot say that the SAT *caused* the college grades.

5. WHY IS THERE NO MEANINGFUL NATIONAL AVERAGE FOR THE SAT OR ACT?

Around the end of August each year, the College Board and ACT release the latest results for the SAT and ACT "national average." Much has been made about these numbers ever since a 1977 report analyzed the causes of what was at the time a 14-year decline in the average SAT score.

The "national average" is not meaningful for a number of reasons. First, the students who take the test are a self-selecting group and a larger and larger proportion of all seniors has taken the test each year. Thirty years ago, about 30% of the entire senior class took the SAT; today the figure is around 43%.

The growing percentage of students taking the SAT and ACT represents an ever deeper dig into the talent pool. In addition, the demographic

characteristics of who takes the SAT have been changing, especially since the 1960's. The SAT was standardized (see WHAT IS STANDARDIZED ABOUT A STANDARDIZED TEST?) on a small, elite group of white students mostly living in the Northeast and planning to attend Ivy League and Seven Sisters colleges and universities.

Beginning in the 1960's, however, as colleges opened up to women and minorities, more of these two groups have taken the SAT. In addition, more and more students from low-income families and students with non-stellar high school grade point averages have aspired to college and taken the test. Under these circumstances, it is small wonder that the SAT average fell. For a variety of arguable reasons, women and minorities (except for Asian students on the math section) do not score as well on the SAT as men.

6. WHY DID THE SAT AVERAGE SCORE DECLINE?

The demographics of who has been taking the tests have been changing over time and all of the changes are associated with lower test scores — more women, more minorities, more students from low-income families, more students with low grade point averages. Indeed, one study found that from 1975 to 1990, the SAT average score would have risen if just one variable — students' high school class rankings — stayed the same. But more and more students in the bottom 40% of the high school ranks took the SAT.

When the College Board assembled a panel in 1976 to study the falling SAT average score, the panel concluded that a host of factors caused the decline. Indeed, one of the background papers for the panel simply listed the various hypotheses that had been advanced to explain the decline. There were 74 of them!

The distinguished panel called the period of the decline a "decade of distraction." During this period the country had been rocked by the assassinations of John F. Kennedy, Jr., Robert F. Kennedy Jr., Martin Luther King, Jr., and Malcolm X. It had endured an unpopular war and protests against it. It had suffered through Watergate. Virtually all urban areas had experienced serious rioting. During the decade, newspapers had almost an "outrage-of-the-day" to show: police beating demonstrators at the 1968 Democratic National Convention, a young woman crying over the body of a friend at Kent State University, etc. Recreational drugs had become popular and television ubiquitous. Little wonder that people were paying less attention to parsing sentences and factoring equations. Other indicators of achievement in this period fell along with the SAT.

7. WHY WAS THE SAT "RECENTERED?"

The College Board took this action in 1996 to make a score of 500 once again reflect the average score of people applying to college. As noted in the section, WHY IS THERE NO MEANINGFUL NATIONAL AVERAGE FOR THE SAT?, the standard-setting group in 1941 was an elite. Specifically, they were 10,654 students living in the Northeast. Ninety-eight percent were white, 61% were male, and 41% had attended private, college-preparatory high schools. This hardly represented the body of students taking the test in 1996, the year of the recentering. In that year, over 1,000,000 students huddled in angst on Saturday mornings to take the SAT. Twenty-nine percent of them were minorities, 52% were women, and 83%

of them had attended public schools. The test-taking pool had broadened substantially and become quite democratized. However, the scaled score of 500 had been assigned to the average verbal and mathematics score of the standard-setting elite (this transforming of test scores from raw scores – i.e., number correct — to some kind of scale is quite straightforward and occurs in virtually all tests. IQ tests are scaled scores, as are scores from the National Assessment of Educational Progress and the Third International Mathematics and Science Study; see WHAT IS A SCALED SCORE?).

In 1941, a scaled score of 500 represented an average score of those planning to attend college, at least in the Northeast. By 1996, it represented no one's average score. Students receiving a 464, say, in 1996 might believe they were "below average" because, after all, 500 was "average." But it was average only for that initial standardizing group in 1941. So, in 1996, the College Board decided to make a 500 once again represent the average score of everyone who took the SAT.

The Board's action was not without controversy because it appeared that scores rose for no good reason or, at least, for no reason relating to how the students were actually performing. "The greatest dose of educational Prozac in history," was one wag's comment. People also worried that trend data would be lost. However, ETS provides scales which translate back and forth between the old and new scales. People can follow trends with whichever scale they prefer. The recentering accomplished the Board's purpose: to make 500 again represent the true average score of all SAT test takers.

8. DO THE SAT AND ACT "WORK?"

The answer depends in part on one's perspective and in part on how one defines "work." The function of both tests is to predict freshman college grades. Both tests do this but the predictions are hardly perfect. The typical correlation between test scores and freshman grades is about +0.45. This means that the test accounts for about 20% of what goes into the grades. Other factors account for about 80% (see WHAT IS A CORRELATION COFFICIENT?).

This 0.45 correlation shrinks at highly selective colleges because SAT scores are more tightly bunched at such schools. The more people look alike, the less successful we can make differential predictions about their success. Suppose, for instance, you wished to predict the effect of body weight on success as a defensive lineman. If everyone who showed up to play weighed 275 pounds, you could make no predictions because everyone has the same "score," 275. As scores become more and more differentiated, better predictions become possible.

At non-selective colleges — the overwhelming majority of colleges in the country — grades combined with rank-in-class predict freshman success at least as well as test scores. At selective colleges, tests usually predict better than grades because grades are even more tightly bunched than test scores.

9. DO COLLEGES OVER RELY ON THE SAT?

Probably not. The popular culture believes that the tests work and that low SAT scores doom a student's chance of admission to a selective college. *USA Today* recently carried a cartoon showing a mother reading to her child in bed. The caption had the mother saying "And the little pig with the higher math and verbal lived happily ever after. The other two were swallowed by the wolf." In *None of the Above*, David Owen declared that "People who forget their shoe sizes remember what they got on the SAT."

In fact, colleges use many factors to make admissions decisions and glean information from things like portfolios, videotapes and personal histories. One of the myths surrounding college admissions is that all applicants are in competition with all other applicants. In fact, selective colleges admit by categories. They want "brains" to be sure, but they also want "the All-American Kid" and "legacies" (children of alumni). They make adjustments for "special talent." This includes not only athletes, but many in the fine and performing arts who tend not to do well on paper and pencil tests. "Social conscience" has also been a category since the 1960's, but it is on the decline as the courts have ruled against at least some affirmative action programs. Finally, deans of admission prefer "paying guests"—those who can pony up the $20,000+ a year costs without college financial assistance.

As evidence that colleges do not use only SAT scores, consider the entering freshman class at Brown University, one of the nation's most selective colleges. In 1998, Brown could have filled *two* freshman classes using only students scoring between 750 and 800 on the SAT verbal. In fact, they admitted students with scores ranging from 350 to 800. Only one-third of applicants who scored between 750 and 800 were admitted. Looking at the number of admitted students who were ranked high in their class, Brown appeared to be more interested in rank-in-class than test scores.

There is also evidence that the selective colleges don't need the SAT or ACT. Some years ago Bates and Bowdoin Colleges made the SAT optional for admissions, but still required it for placement and counseling. Students who submitted SAT scores with their applications scored about 150 total points higher than those who didn't. But they did not have higher grade point averages. The university noticed that it became more diverse geographically and ethnically and by intended-major. The faculty was happier with the character of the SAT-optional classes.

Finally, one person who often addresses admissions officers states that he always asks for a show of hands of those who would continue to use the SAT if the colleges, not the students had to pay for it. He says he has yet to see one arm rise.

WHY "ASSESSMENT LITERACY"?

It is to be hoped that a reader having perused the previous pages will come away better informed but not overwhelmed. Testing is a much more complex undertaking than usually presented in the media. Indeed, the media do not probe announced test results but accept them quite uncritically. When the NAEP Civics results were released in November 1999, only *Education Week* noted that the NAEP proficiency levels are flawed. Everyone else played the story as statistically correct.

And the tests keep on coming. As this primer neared completion these items made news:

- The National Test proposed by the Clinton Administration had been developed and was ready for field testing, awaiting only funding from Congress. The funds were not forthcoming.
- Arizona released the results of its first-ever state testing program and 89% of the students failed. (Interestingly enough, the state superintendent of public instruction took the tests and barely passed.)
- Evaluations of charter schools in Michigan and Ohio concluded that it was too early to tell if charters improved test scores.
- John Stossel hosted a segment on ABC's 20/20 lauding Catholic schools for getting higher test scores than public schools—and at much lower costs (like most such reports, this one failed to take into account low salaries, the cost of buildings, and subsidies provided by the church).
- A court case against a Chicago teacher who published some test items in a newspaper moved towards trial.
- Republican presidential candidate, Governor George W. Bush made Texas test score gains a feature of his campaign. Critics immediately questioned the validity of these alleged gains.
- California completed its development of a test-based Academic Performance Index for evaluating schools.
- A new Michigan social studies test flunked most students.
- The National Association for the Advancement of Colored People announced sponsorship of SAT-ACT prep courses for minority students.
- The superintendent of schools in Kansas City, Missouri announced a program to get scores up in order to restore state accreditation to the school district.
- Asked why the annual whale watching trip had been canceled, the superintendent of schools in East Palo Alto, California replied: "Students are not tested on whale watching, so they are not going whale watching."

There might come a time for education when tests and test scores recede from such prominence, but that time is not now. In view of the pervasiveness of tests and assessments, "assessment literacy" seems like a must.