Introduction

Achievement Testing in U.S. Schools

Achievement tests have been a prominent feature of American elementary and secondary schooling for generations, and a characteristic of educational and occupational decision making for centuries. Achievement tests have a single, simple purpose: to measure what a student has learned.

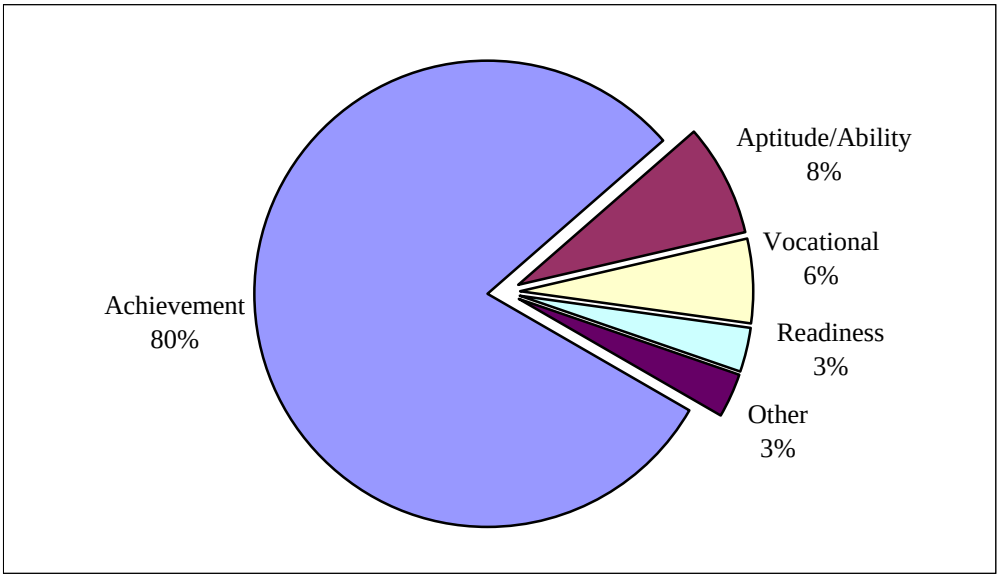
Today, achievement tests administered to U.S. students span a broad range of designs, but share a common purpose. At one end of the continuum we find tests such as the straightforward, classroom-specific, teacher-constructed, end-of-week, spelling test that seeks to determine whether Mrs. Garcia's second-grade students have grasped when *I* comes before *E*. At the other end are tests such as the National Assessment of Educational Progress (NAEP)--large-scale, corporately-constructed tests of exquisitely complex design that are administered to nationally-representative samples of students; these tests attempt to assess the country's overall educational health. Near the middle of the continuum are state-level competency tests used by many states as "gatekeepers" for grade-to-grade promotion or graduation.

Between the extremes are the tests that parents and policy makers have traditionally been most concerned with, and that have frequently been used (or mandated) as markers in efforts to improve the education of American children. These tests, such as the *Iowa Tests of Basic Skills* (*ITBS*), the *California Achievement Test* (*CAT*), and others, are long-lived measures of achievement. More recent incarnations of these tests have less familiar names, such as *Terra Nova.* It is these tests that are most often the grist of parent-teacher conferences, the stuff of parental bragging rights, and the topic of heated policy deliberations at all levels of the American education and political systems.

All of these tests seek to measure student *achievement*, which makes them different by design from ability tests, aptitude tests, and other tests used in the American education system. As shown in Figure 1, achievement testing accounts for the majority of standardized testing in American schools. This booklet focuses on achievement testing, in particular on standardized, norm-referenced tests such as the *ITBS* with which parents and policy makers are most familiar. (The distinction between achievement tests and other kinds of tests is explored in greater detail later in the report.)

The American public has a long-held affection for achievement testing. Results of opinion surveys from 30 years ago and those conducted today reveal broad and durable support for *even more* achievement testing. Contrary to assertions that tests are foisted upon the public by self-serving politicians, the evidence is clear that consumers of U.S. education favor testing. Parents believe that testing promotes accountability and improvement in the education system, and that tests should be relied upon to help make important decisions about students, such as grade-to-grade promotion or high school graduation. Students acknowledge that the presence of high-stakes tests motivates their work in school. Even among school personnel--teachers, principals, district and state level administrators--testing is

Figure 1 Types of Standardized Tests in American Schools



From USGAO, 1993, p. 21

acknowledged to be a useful mechanism with net positive benefits (Phelps, 1998).

Long before any talk of voluntary national tests, standardized tests such as the *ITBS* and *CAT* provided those interested in the educational well-being of American students with information regarding specific education outcomes. However, two caveats are warranted regarding the information provided by standardized tests. First, as we shall see, although current standardized tests are *capable* of providing information of exceptionally high quality because of advances in psychometric theory, in practice standardized tests have often been abused and misused. As a result, the quality of the information they provide has frequently been corrupted.

Second, in recent years it has become fashionable for critics to proclaim that these tests either 1) do not measure *all* educational outcomes, or 2) do not measure *the most important* outcomes, such as a pupil's potential for making a positive contribution to our democratic republic.

Such criticisms have been ignored by most parents and policy makers who understand that to demand that these tests measure *everything* valuable is to impose a burden that is impossible for any test to shoulder. Historically, standardized achievement tests have performed a narrowly-defined purpose well: they efficiently provide accurate information about students' skills in areas such as reading comprehension, mathematical computation, locating and using resource materials, and placing correct punctuation in a sentence. The first criticism is as easily dismissed as calls to abolish barometers because they are incapable of telling temperature or humidity.

The second criticism--that these tests don't measure the most important outcomes--has also been refuted. The relationship of standardized achievement tests to the essential goals of education--say, becoming a responsible, productive citizen--relies to a great extent on the principle of "necessary but not sufficient." It is easy to see that acquiring proficiency in reading, mathematics, writing, and so on, is *necessary* to accomplishing that goal. Admittedly, other student characteristics--such as personal responsibility and creative thinking or problem solving--are also necessary for the goal to be achieved. However, an extra measure of personal

responsibility or creative thinking cannot compensate for deficits in a student's knowledge of language or mathematics, or in the student's ability to organize and communicate his or her ideas. Students *may* become productive and responsible contributors to society if they master fundamental academic skills; they will almost certainly be unable to do so without them. Thus, certain fundamental academic proficiencies are judged to be necessary, although not sufficient, for attaining the ultimate goals of schooling. That current standardized tests do not address the ultimate goals of schooling is not so much a criticism as a form of wishful thinking: no tests are available to gauge attainment of the ultimate aims of education. Until they are--if ever--American schools must continue to assess students' acquisition of the tools that predict their eventual success.

We are left, then, with three ineluctable facts about testing and American education at the dawn of the 21st century: 1) standardized achievement tests are a widely-used mechanism by which parents and policy makers have defined and continue to report the standing and progress of students in K-12 schools; 2) although incapable of providing information on ultimate educational outcomes, standardized achievement tests can yield highly accurate, dependable information about a finite but vital constellation of knowledge and skills; and 3) abuses of standardized tests by those who deploy them can distort the information they provide and misinform students, parents, and policy makers regarding educational health.

This booklet makes possible an understanding of the purpose, construction, and results of standardized achievement tests—an understanding that is essential to both consumers and producers of education. Parents demand, and these tests provide, relatively easy-to-comprehend snapshots of children's standing in and progress through the education system. Teachers must understand the information provided by such tests in order to adapt students' education programs to their individual needs and communicate with parents about pupil strengths and weaknesses. Educational administrators must master the essentials of testing in order to evaluate programmatic strengths and weaknesses and provide informed, data-based decision making.

For better or worse, it seems that many contemporary suggestions for reforming education in the United States include some element of testing. It obviously behooves those concerned with education reform to be more knowledgeable about this element. Because policy makers—from local boards of education to legislators and lobbyists—frequently invoke testing as a means of tracking education progress and promoting accountability, they too must possess a thorough understanding of the appropriate uses and misuses of standardized tests.

This booklet provides these audiences with information relevant to their needs. The first section presents an overview of the current market for standardized achievement tests. Also provided are key definitions, distinctions between ability and achievement testing, comparisons of norm-referenced, criterion-referenced, and standards-referenced testing, and examples of appropriate interpretations of norm-, criterion-, and standards-referenced test scores.

The next section provides information on the most frequently-used standardized achievement tests, such as their format, content coverage, cost, sample items, availability of non-English language versions, and amenability to customization. Tests covered include: the *California Achievement Test,* the *Stanford Achievement Test,* the *Metropolitan Achievement Test,* the *Iowa Tests of Basic Skills,* the *Comprehensive Test of Basic Skills,* and *Terra Nova.*

The focus of the following section is the uses and misuses of tests. This section describes how test information can be used by

various audiences, including policy makers, educators, and parents; and summarizes cautionary information from relevant professional guidelines and standards.

The final section analyzes current issues and controversies in large-scale achievement testing and speculates about future trends and areas of concern. "References and Resources" appears at the end and provides information that readers can use to contact test publishers and obtain test reviews and other information related to testing in U.S. schools.

The Basics

Although the purpose of achievement tests--measuring a student's knowledge, skills, and abilities--seems simple enough, the picture becomes more complex when tests are used to gauge the learning of groups of students. For example, determining whether an individual fourth-grader can multiply fractions correctly can be accomplished using a teacher-made achievement test, or simply through observation of the student's daily classroom work. However, discovering how his or her performance stacks up against the math skills of students in other classrooms across the state or nation or determining whether fourth-graders know enough math to progress to fifth grade poses a greater challenge.

This section supplies background information on large-scale student achievement testing, which includes state-mandated competency tests and standardized, norm-referenced achievement tests. First, some data on the scope of testing in the United States are presented. Then a few necessary definitions are provided. Finally, key distinctions among different types of tests are explained.

The Marketplace for Standardized Testing

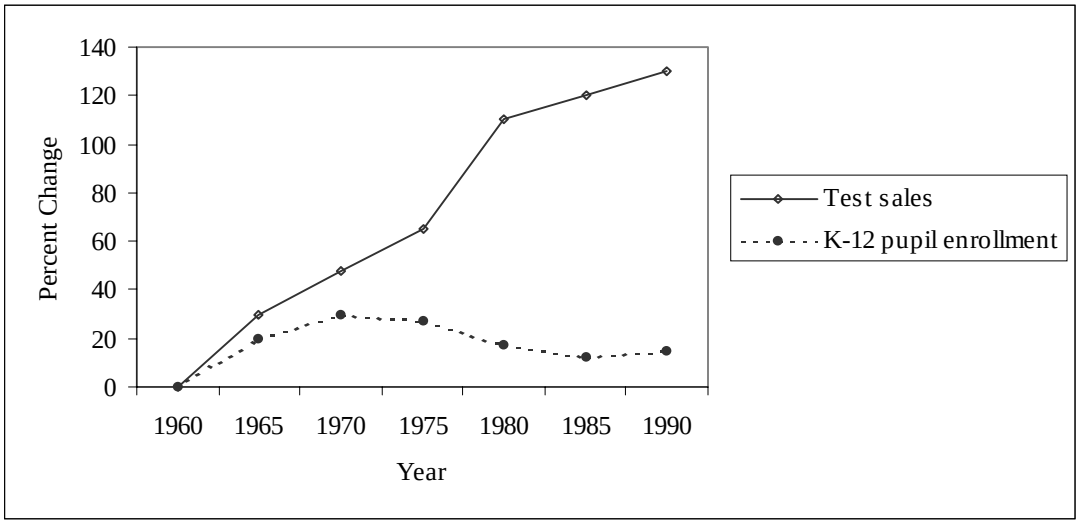
American elementary and secondary education witnessed a great expansion in achievement testing during the 1970s and 1980s with the introduction of state-mandated student competency tests and national tests (such as NAEP). During the 1990s, the increase in testing has slowed somewhat, with increases in some types of testing offset by declines in other areas. NAEP testing and high school graduation testing requirements have expanded, while the use of intelligence testing, standardized testing in the early elementary grades, and required college admissions testing have decreased.

Table 1
Annual Test Administrations

Type of Test	Minimum Estimate	Maximum Estimate
State-mandated testing	33,000,000	71,500,000
School district testing	85,621,429	271,626,602
Special-needs student testing (e.g., learning disabled, gifted, bilingual, etc.)	11,520,000	30,600,000
College admissions testing	12,034,318	21,759,548
TOTAL	143,175,747	395,486,150

From Haney, Madaus, & Lyons, 1993, p. 61

Figure 2 Growth in Test Sales and Enrollment, 1960-1990



From Office of Technology Assessment, 1992, p. 4.

Precise estimates of the number of tests administered annually to students in grades K-12 are difficult to come by. An annual survey of achievement testing is conducted by the Council of Chief State School Officers. The most recent results indicate that 48 of the 50 states have student achievement testing programs. Only Nebraska and Iowa do not have any state-mandated testing, although Nebraska is considering legislation to require it and Iowa has a long history of nearly all districts using the home-grown Iowa Testing Program on a voluntary basis. Twenty-two states use commercially-produced, norm-referenced tests alone or in combination with criterion-referenced measures. The remaining 26 states use commercially-produced or locally-developed criterion-referenced tests (Roeber, Bond, & Connealy, 1998).

One estimate of the number of achievement tests administered each year was reported by Haney, Madaus, and Lyons (1993) who provided low- and high-end estimates for various types of school testing (see Table 1). They estimate the number of tests administered to be between 140-400 million per year. These figures represent from three to eight standardized tests administered annually to each of the approximately 50 million students enrolled in K-12 public and private schools. Unfortunately, these estimates counted each subtest or portion of a test as a test in itself, and therefore overstate--by up to ten times--the amount of testing in U.S. schools.

Other estimates of the amount of testing are considerably lower and more plausible. One publisher estimated that 30 to 40 million tests are administered per year. The General Accounting Office (GAO) estimated that between 30 million and 127 million tests are administered annually at an overall cost (including teacher time) of $516 million for the 1990-1991 school year (USGAO, 1993, p. 3). The likeliest estimate seems to be in the range of about one to two standardized tests per year per student, representing a total of 50 to 100 million tests administered annually. A recent review of standardized testing during the 1980s and 1990s concluded that state-wide and district-wide testing accounts for less than two days per year per student (Phelps, 1997).

The growth of testing is also visible in the rise of revenues from test sales. Figure 2, based on a report by the U.S. Congress's Office of Technology Assessment, shows the changes in test sales and enrollment in grades K-12 for the period 1960-1990, during which revenues from sales of commercially

published standardized tests increased from approximately $35 million to about $95 million (in 1982 dollars).

Despite the growth in testing, the actual dollar amounts and time spent on testing remain small. Annual sales of commercially-produced standardized tests amount to only about $2 per student tested. Assuming total expenditures on American elementary and secondary education of $350 billion in the year 2000, the annual figure of $105 million represents only .003% of total spending. Of the 38 states reporting total spending on testing in the 1998 CCSSO report, the median budget for state achievement testing programs was $2.8 million, with Texas spending the most ($23,600,000) and Wyoming and Alaska tied for lowest spending ($100,000). The 1993 GAO report concluded that "U.S. students do not seem to be overtested," and that "the average student spent only 7 hours annually on system wide testing" (USGAO, 1993, pp. 2-3).

Definitions

Understanding standardized achievement testing in American schools requires familiarity with a few key concepts. Some of the terms defined below have both a common usage and also technical definitions. The common usages can differ substantially from the way testing specialists use the terms, which frequently confuses discussions. Where appropriate, these distinctions will be highlighted.

Basic Concepts

Test - any structured, purposeful sample of knowledge, skill, or ability. Ideally, those interested in student achievement would like to know everything about what the student can do but, because of cost or time considerations, must settle for a subset of observations. For example, to gauge whether a student knows the multiplication tables from 0 to 12, the student could be asked every multiplication fact, beginning with 0x0, 0x1, 0x2 and so on, and ending with 12x12--a total of 169 questions. However, it is more practical to give students a test, consisting of a random assortment of, say, 20 questions. From the student's performance on the 20-question test, it is possible to infer, with reasonable accuracy, how the student would have done on the entire set of 169 math facts.

It is important to note that the term test is used regardless of the format of the testing. A test could consist of multiple-choice questions, true/false questions, a performance task such as making a table and chairs in wood shop, an oral examination, an essay examination, a physical demonstration, etc.--or a combination of these.

Item - a test question. If the test question is written in the multiple-choice format, the item has a *stem* (the part that introduces the question) and several *options*, usually labeled A, B, C, D, and E or similarly. If the test question asks the student to engage in a performance or demonstrate a skill such as writing an essay, the item is called a *prompt*--the written or other material introducing the task.

Item format - the style in which the item has been constructed. Item formats have been classified as either *select-response* in which the student chooses the correct answer from alternatives provided (e.g., multiple-choice, matching, true/false) or *constructed-response* in which an answer must be supplied by the student (e.g., essay, short-answer, speech, project, etc.).

Authentic - a description of item formats that attempt to present questions to students in contexts that are valuable in themselves, or in contexts that simulate problems as they might be encountered in "real life" situations. For example, a mathematics problem that asked a student about the time it would take two trains to meet, starting a specified number of miles apart and traveling toward each other at differing speeds, would not be considered "authentic" because that particular problem does not naturally occur in real life. On the other hand, simple addition or subtraction problems couched in terms of balancing a checkbook would be considered authentic because the task of balancing a checkbook is routinely performed in everyday life.[1] Occasionally, the term "authentic assessment" is used interchangeably with "alternative assessment" as a way of contrasting these formats with objectively-scored items (e.g., multiple-choice).

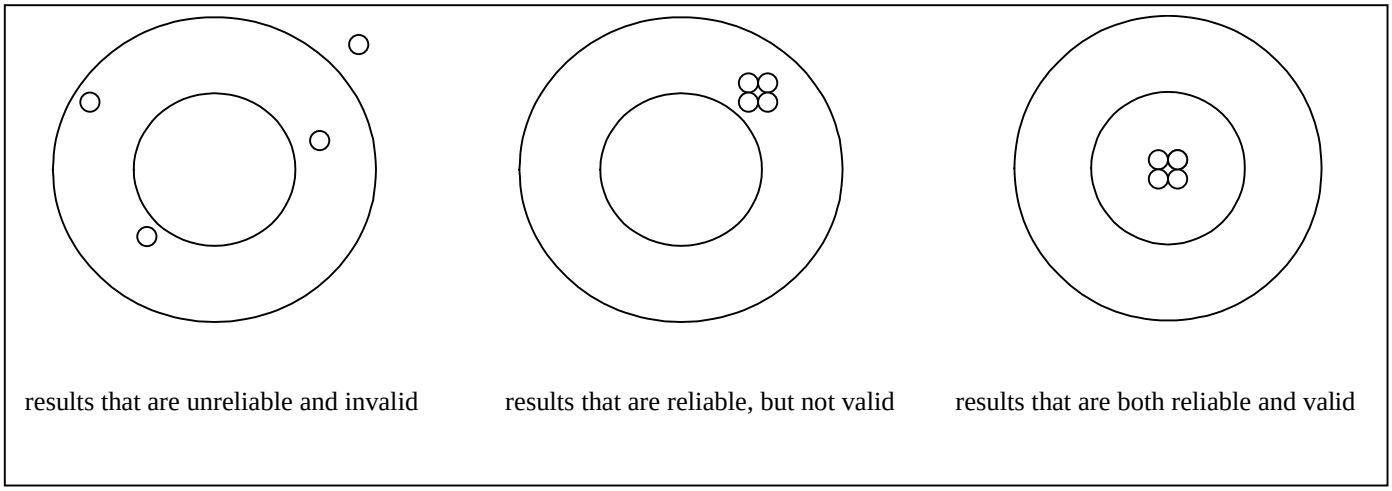
Item difficulty - an index of how easy or difficult a test item is (also referred to as a *p-value*) which can range from 0.0 to 1.0. The index is derived by dividing the number of students who answered an item correctly by the total number of students who attempted the item. Two notes about item difficulty are in order. First, the item difficulty index is counterintuitive: an item that all or nearly all students answer correctly has a *high* difficulty index (i.e., near 1.0); an item that all or nearly all student get wrong has a *low* difficulty index (i.e., near 0.0). Second, no item has an intrinsic difficulty level; the difficulty of any item is determined strictly by students' performance on the item.

Item discrimination - an index of how well the item differentiates between students with high overall mastery and those with low overall mastery. The index is usually calculated as the correlation between students' scores on a particular item or task and their overall scores on the test. It can range from -1.0 to +1.0. Positive values (i.e., closer to 1.0) are usually preferred; they indicate that students who performed well on the item tended to be those who performed well on the total test. Negative values (i.e., less than 0.0), which reveal that students who performed poorest on an individual item tended to perform well on the total test, usually indicate a flaw in the construction of the test item.

Reliability - the test score characteristic of being dependable. Because a test consists only of a sample of questions or tasks and because both students and those who score tests are susceptible to various unpredictabilities in performance (called *random errors*), no test score can be considered to be a perfectly dependable snapshot of a student's performance. Various methods can be used to quantify the degree of confidence that can be placed in students' scores on a test. All of the methods result in a number, called a *reliability coefficient*, that can take on any value from zero (0.0) to one (1.0). A reliability coefficient of 0.0 would denote test scores that are completely undependable--no more useful for making decisions about students than flipping a coin or guessing. A reliability coefficient of 1.0 indicates perfect dependability--the complete absence of those pesky errors mentioned above--and the potential for the test scores to be used with great confidence for decision making. Values between 0.0 and 1.0 indicate relative poorer (nearer to 0.0) or greater (nearer to 1.0) dependability. Obviously, all tests aspire to yield reliability coefficients as close to 1.0 as possible. Reputable publishers of standardized tests explicitly state a test's reliability coefficient and the methods they used to determine it.

Validity - the degree to which the conclusions yielded by a test are meaningful, accurate, and useful. Validity is the extent to which a test provides information about student performance or yields inferences about student ability that are "on target." This characteristic is different from reliability, which merely refers to consistency. Figure 3 illustrates three potential results of target shooting. As shown in the second panel of the figure, a marksman can produce consistent, but completely "off-target" results. Similarly, although a test may yield highly consistent scores--

Figure 3
Relationships Between Reliability and Validity



i.e., be very reliable--the test may or may not provide accurate information about student performance.

The degree to which a test can be said to yield valid results is often *not* expressed in statistical terms as reliability coefficients, but in logical and empirical terms based on evidence. For example, the use of test scores to make decisions about placing students into an appropriate mathematics course in high school would be considered more valid if: 1) the test were drawn from typical high school mathematics content, or reviewed and approved by high school math teachers (what is sometimes called *content validity evidence*); 2) scores on the test were shown to be related to success in the courses in which students were placed (*predictive validity*); 3) students who perform well on the test tend to be those with the highest previous math achievement (*construct validity*), and so on. The greater the weight of validity evidence that is presented, the more confidence test users can have that they are making accurate (i.e., correct) decisions about students based on their test performance. Publishers of standardized achievement tests rely most heavily on content validity evidence.

Battery - a collection of tests. The California Test of Basic Skills is one of several large-scale testing programs that offer what is called a complete battery. Taken together, its individual tests in language, mathematics, study skills, and so on, form an achievement battery. (In some cases, the individual tests comprising a test battery are called *subtests*.) Some publishers offer both a complete battery and a trimmed-down version of the complete battery, called a *survey edition*, that is intended for use when available testing time is minimal.

Test form - a version of a test that can be used interchangeably with other versions. Because a test may be needed on more than one occasion, several versions of the test are usually developed. For example, the SAT is administered several times each year. Each time the items are different, although there is no advantage to the student regardless of when the test is taken because the versions--or forms--are statistically equated to make scores on all versions comparable.

Standardized - any test that is developed, administered, and scored under controlled conditions. Standardized tests are frequently

developed under controlled conditions because of the need to create interchangeable forms. They are administered under controlled conditions (including scripts for proctors, strict time limits, assigned seating, etc.) so that differences in students' scores can be more confidently attributed to real differences in the students, as opposed to differences in the conditions under which they took the test. Finally, they are scored under controlled conditions--often involving computer scoring or highly trained human raters--to eliminate differences in scoring due to human error or subjective judgment.

The public labors under several profound misconceptions about the meaning of standardized. First, the term standardized is unrelated to the format of the test. A standardized test may include any format--multiple-choice, true/false, performance tasks, oral, essay, etc.--although most current standardized achievement tests include some multiple-choice items.

More important, *most standardized tests do not contain "standards" as that term is commonly understood.* Some standardized tests--specifically those known as *criterion-referenced* (see below)--are designed to provide information about student performance relative to standards. However, the most common standardized tests--called *norm-referenced*--do not contain normative "standards" at all and are not intended to prescribe levels of acceptable performance.

Evaluation - ascribing value or worth to a score or performance. Saying that a student correctly answered 18 of 20 multiplication items is simply measuring the student's performance. Going beyond simple reporting to say that 18 of 20 correct should be judged to be Proficient or awarding a grade of B+ represents evaluation of the student's performance.

Assessment - gathering and synthesizing numerous sources of information for the purpose of describing or making decisions about a student. The term assessment has been borrowed from fields such as counseling psychology, in which a client may be given a variety of tests, the results of which must be synthesized--that is, analyzed and interpreted--by a single professional or team of experts. In education, this process has been used frequently in the context of special education, in which an Individualized Education Program (IEP) is planned for a student based on a variety of sources of information about the student. In education, the term assessment is increasingly used simply as a synonym for test.[2]

Key Distinctions

High stakes vs. **Low stakes**

The terms high stakes and low stakes were coined to represent the severity of the consequences associated with performance on a test. For example, if a test is used to determine whether a high school student can graduate, the consequences of passing or failing it are obviously serious, and the test would be called high stakes. On the other hand, weekly classroom tests may count toward a student's semester grades, but serious consequences or decisions about a student do not ordinarily follow from his or her performance on a single quiz. Some testing programs (e.g., NAEP) do not even report scores for individual students. These tests would be termed "low stakes."

Achievement vs. **Ability**

Achievement tests are designed to measure *attainment* of knowledge, skill or ability. They answer the questions: "What does the student know and what can he or she do?" Examples of achievement tests include weekly spelling tests in elementary school, chemistry lab tests in high school, and driver's license tests. Standardized achievement tests include the *Iowa Tests of Basic*

Skills, the *California Achievement Test*, and others.

Ability tests are designed to measure *potential* for achievement. They answer the question, "What is the student capable of?" Standardized ability tests include the *Otis-Lennon School Abilities Test* and the *Cognitive Abilities Test*, and others. Such tests are frequently used in conjunction with or in addition to intelligence tests such as the *Weschler Intelligence Scale for Children--Revised* (WISC-R) to identify students for placement in certain programs (e.g., special education, gifted education). Ability tests are also sometimes used in conjunction with standardized achievement tests to derive ability/achievement comparisons which describe the extent to which a student is "underachieving" or "overachieving" in school, given his or her measured potential.

Norm-referenced, Criterion-referenced, and **Standards-referenced tests**

These three types of tests differ primarily in their purposes. It is useful to think of the purpose of a test in the same way one might think of the purpose of a research study: each study embodies a research question that it attempts to answer and employs methods best suited to answering that question.

Norm-referenced tests

Norm-referenced tests (NRTs) are constructed to cover content that is considered fairly universal at each grade level. However, the purpose of an NRT is to describe relative *rank* among students at a particular grade level. NRTs provide information about how a student's performance compares with a reference group of students, called the *norm group*. The norm group is a sample of students that is intended to be a miniature portrait of all U.S. school children--representative in terms of ages, grade levels, sex, ethnicity, public and private school settings, community size, and so on.[3] The norm group takes the NRT and percentages of students at each score level on the NRT are calculated. These values are called the *norms*. For example, suppose that the NRT language test contains a total of 40 items. Further, suppose that, in the norm group, answering 19 items represents performance that is better than half (i.e., 50%) of the students in the norm group. A score of 19 is therefore established as the 50th percentile on the NRT language test. Subsequently, in actual use of the NRT, any student who answers 19 of 40 questions correctly would be reported as being at the 50th percentile. The same process is used to establish the percentile ranks on any other tests within the battery, as well as for the total or composite test score.

Three additional points about *norms*. First, it is possible to create comparison groups that are not nationally-representative. Norms can be calculated based only on the performance of urban school students, or non-public school students, and so on, which in certain circumstances may be more informative. For example, a Catholic school system might perform at the 90th percentile compared to national norms, but at the 68th percentile using Catholic school norms. The latter information may be more useful if educators in the Catholic system want to compare themselves to schools with a similar purpose, curriculum, pedagogy, etc., rather than to all schools nationally.

Second, because classroom instruction, familiarity with test forms, and the characteristics of American students evolve over time, NRT publishers must update test norms on a regular basis to ensure that comparisons based on the norm group remain accurate. Test users must evaluate the results of any NRT in light of the recency of the norms upon which current scores are based. This need for current norms was highlighted by the "Lake Wobegon" report issued by Cannell (1988) who found that, by using outdated norms (among other things), all states were able to claim that their students' performance on NRTs was above average. (Cannell also illustrated some of many ways in which test scores are misused. Additional information on these abuses is found in the section entitled “Uses and Misuses”)

Third, the meaning of "norm" as in "norm-referenced test" is distinctly different from the use of the term in common parlance. According to *Webster's New Collegiate Dictionary*, a norm is defined as "an authoritative standard." In everyday usage, norm carries a prescriptive ("normative") connotation. However, in the context of NRTs, the term norm is narrowly used to mean "average" without any prescriptive connotation. The distinction is clearly stated in a leading textbook on testing: "*Norms are not standards*. Norm information tells us how people actually perform, *not how they should perform*" (Mehrens & Lehmann, 1991, p. 229, emphasis in original).

This distinction illustrates a key point about NRTs: A student's performance at, say, the 50th percentile *does not necessarily indicate anything about the knowledge or skills a student has mastered, nor whether scoring at the reported percentile represents acceptable progress, nor whether instruction has been of sufficient quality, nor whether the content is sufficiently challenging or the outcomes measured desirable.* Although some such information may be teased from the data, the primary purpose and construction techniques of NRTs are focused on answering the question "Where does this student stand compared to others at his or her grade level?"

Notice also that concepts such as performing "at grade level" are murky in the context of NRTs. In this realm, *performing at grade level means only that a student is performing about as well as the average performance of the norm group; no evaluation is made regarding whether the norm group as a whole is performing superbly or terribly*. A student performing "at grade level" on an NRT could be well-prepared for global competition or woefully lacking in even the most rudimentary areas; NRTs simply aren't designed to tell us which is the case.

Criterion-referenced tests

The purpose of criterion-referenced tests (CRTs) is to gauge whether a student knows or can do specific things. Student competency tests developed and used by individual states for purposes of determining grade-to-grade promotion or high school graduation are examples of CRTs (e.g., the MEAP tests in Michigan, the HSPT-11 in New Jersey, the TAAS in Texas, and so on.) CRTs are based on content or objectives judged to be important in the particular state.

The criteria for success on a CRT are established in a judgmental fashion; experts in a given subject or grade level determine a passing score, i.e., the level of performance that will be deemed acceptable. In addition to being used as gatekeepers to determine whether students have mastered specified knowledge or skills, CRTs are frequently (and appropriately) used as diagnostic measures because of their ability to permit clear inferences about an individual student's strengths and weaknesses.

The simplest illustration of a CRT is the road portion of a driver's license test. In the parallel parking portion, the candidate for a license must meet certain *criteria:* for example, park the car within a marked area, in four minutes or less, without knocking over more than one orange pylon. A person's performance on the test--manifested in whether or not the person gets a driver's license--does not depend on how well other candidates perform. In theory, all candidates could pass or all could fail. There is no distinction between one candidate who parks the vehicle perfectly in the middle of the space, in only two minutes, with no pylons knocked over, and the candidate who parks awkwardly within the space, in just under four minutes, and knocks one pylon over. Both candidates have met the criteria.

Scores on CRTs are expressed in different ways from NRTs. Because gauging whether the criterion for success has been met is the primary objective of CRTs, performance on CRTs is most often reported as simply passing or failing. Of course, it is also possible to distill some norm-referenced information from CRTs. For example, driving students could be ranked in terms of how many pylons were knocked over, the time it took them to park, and so on.

A key point about CRTs is this: The single most accurate inference about a student who meets the criterion on a CRT is that the student has performed up to the expectations of those who

established the criterion. The student's performance reflects to some degree on instructional quality, parental support, and so on, but *it does not necessarily indicate anything about whether the student is better or worse than average, nor whether the criteria represent noteworthy expectations given the student's age or grade level, nor whether the content is challenging or the outcomes measured desirable.* Again, although some of this information can be teased from the data, CRTs are designed to answer the research question "Can the student demonstrate knowledge or skill to a specified level?"

Standards-referenced tests

Standards-referenced tests (SRTs) are similar to CRTs in that both attempt to describe the knowledge, skill, or abilities that students possess. Whereas CRTs express standards in terms of quantity and category (e.g., a percentage correct and passing/failing), SRTs link students' scores to concrete statements about what performance at the various levels means.

Typically, SRTs are constructed to match *content standards.* Content standards are developed by curriculum specialists and represent "academic statements of what students should know and be able to do in specific subjects or across several subjects" (CCSSO, 1998, p. 28). An SRT consists of a set of items or tasks constructed to measure these knowledge and skills. *Performance standards* are then established, based on judgment, which describe "how well students need to be able to perform on a set of content standards in order to meet pre-defined specified levels of expected performance" (p. 29). For example, three levels of performance could be described, such as Beginning, Proficient, and Expert, each of which is linked to specified content performance. A student classified as Beginning would have demonstrated mastery of certain knowledge and skills; a student labeled as Proficient would have demonstrated a different (greater) set of capabilities, and so on. The process of determining the correspondence among students' performances on the test, their standing with respect to the content standards, and their classification is called *mapping*. An example of an SRT is the National Assessment of Educational Progress (NAEP), which reports students' performance as Basic, Proficient, and Advanced.

SRTs permit classifications of students according to content standards. However, a student's performance and corresponding classification (e.g., "Proficient") *do not necessarily indicate anything about whether the student is better or worse than average* (in fact, "average" performance may be a lower or higher level), *nor whether the criteria represent noteworthy expectations given the student's age or grade level, nor whether the content standards associated with the performance are particularly challenging.* And, because performance standards--like the criteria of CRTs--are established in a subjective manner, classifications such as Proficient or Expert are inextricably linked to the conceptions of competence held by those who establish them. If those who set the standards have high expectations for performance, a classification such as "Proficient" might mean magnificent accomplishment; if the standard-setters have low expectations, the same classification could represent mediocrity.

Drawing It All Together

It is frequently the case that when a term is used ubiquitously and in diverse contexts, it loses much of whatever meaning it may have originally carried. One need only think of terms such as *world-class, quality,* or *excellence* as examples of concepts that mean different things to different people and are shop-worn to the extent that they no longer carry strong connotations or commonly-held meanings.

This problem also affects the term *standards*. First, standards may represent levels of mastery, expectations of performance, or desired content, depending on the context and the type of testing at issue. Of course, many tests of educational achievement are also termed *standardized*, although that description is unrelated to content or performance *standards*, and the most common type of standardized tests--norm-referenced tests--do not refer to content or performance standards at all.

All those concerned about American education and its potential for improvement--parents, educators, policy makers, and others--must be critical consumers of test information. Each kind of test conveys certain information, while leaving other issues largely unaddressed. For example, should parents and taxpayers applaud the performance of their local school district if its students score, on average, at the 70th percentile on a norm-referenced test? Knowing that local students outperformed 70% of their peers may be encouraging, but only if the content the students are tested on is useful, challenging, and comparable to that mastered by other students preparing to enter a global economy. If average performance in a norm group is abysmal compared to desired content mastery or international performance, then achievement at the 70th percentile may not be laudable at all. Accordingly, appropriate aims of education reform might be wrongly specified if "above average" performance is labeled "success."

Conversely, it may be desirable for a student to master all the objectives on a criterion-referenced or standards-referenced test, but only if those objectives are truly rigorous and only in comparison to the accomplishments of the student's peers. One story illustrating this point concerns the mother who proudly announced to a group of friends that her son had just been potty-trained. The group received the information warmly until they learned that the son was 17 years old. In this case, knowledge of the criterion-referenced outcome (i.e., sphincter-control) is only interpretable in the context of norm-referenced information (i.e., data on what children at a given age or grade level are, on average, able to do; that is, what *normal* performance is).

In the end, all standard-setting is judgmental, requiring consensus about which aims of education are valued and what content is worthy of pursuit, as well as decisions about what level or levels of performance should be expected. These expectations, in turn, are influenced by notions of where students currently stand, aspirations for future levels of performance, cognitive and developmental constraints, and information from other relevant sources, such as international comparisons.

Decisions about standards and criteria are, ultimately, policy decisions, not scientific or technical ones. It is entirely possible that policy makers could establish *standards-referenced tests* with classifications such as Developing, Mastering, and Exemplary to define truly internationally-relevant and demanding performance levels. On the other hand, these levels may only represent performance at, say, the 10th, 15th, and 20th percentiles of an international norm group. Because no single approach currently provides a complete picture of student achievement, those responsible for mandating, conducting, or interpreting the results of testing programs must demand as much standards-based and norm-referenced information as possible.

Comparisons of Major Tests

Norm-referenced tests (NRTs) have been widely used to assess how U.S. students compare with each other. Such information has been valuable to school districts and parents. For example, in a certain school district, average student performance at the 80th percentile by fourth graders on the appropriate NRT would be praiseworthy. Although any particular student might be well above or below that level of performance, district leaders could (correctly) conclude that performance in the district as a whole was substantially superior to that of most other districts,[4] real estate agents would make hay with the findings, and the media might use this information to make (dubious) comparisons of instructional quality across local districts. Classroom teachers, largely underprepared to interpret or use NRT information,[5] would struggle to incorporate the information into their plans for addressing individual students' strengths and weaknesses.

However, as described in the previous section, interpretations of pupil achievement and performance are necessarily linked to the content and standards associated with the particular test. Also, public demands for accountability and legislative responses tied to testing have created the need for tests that serve many masters and purposes. Responding to pressures to address these diverse concerns, commercial test publishers have attempted to develop products that attempt to serve multiple purposes. Concurrently, many states have developed their own pupil achievement testing programs to replace or supplement traditional norm-referenced testing. These programs, like the MEAP in Michigan, TAAS in Texas, and HSPT-11 in New Jersey, are exclusively criterion-referenced; that is, they do not attempt to provide the comparative information that NRTs do, but only to determine whether a student performs at or above a level of performance judged to be adequate or minimal.

To remain competitive, commercial publishers of norm-referenced tests have retained their traditional goal of providing comparative information, but have also begun marketing tests that are keyed to the content standards promulgated by professional organizations such as the National Council of Teachers of Mathematics (NCTM, 1989),[6] and that are capable of providing diagnostic information about students' areas of strength and weakness along the lines of criterion-referenced tests. These diverse aims blur traditional terminology and conceptions of NRTs, CRTs, and SRTs. As new tests evolve, it is increasingly important for parents, educators, and policy makers to understand what the major commercially-available tests offer.[7]

This section provides a comparison of the major norm-referenced achievement tests: the *Comprehensive Test of Basic Skills*, the *California Achievement Test*, and *Terra Nova*, all published by CTB/McGraw-Hill; the *Stanford Achievement Test* and the *Metropolitan Achievement Test*, both published by Harcourt-Brace Educational Measurement;[8] and the *Iowa Tests of Basic Skills*, published by Riverside Publishing. Several aspects of these tests will be compared, including: stated purpose and development methods; content coverage; technical quality; and special characteristics.[9]

Together, these tests substantially define large-scale, norm-referenced achievement testing in the United States. Nearly 60% of the state-mandated achievement tests used across the country are commercially published, with the achievement tests of these three major publishers accounting for 43% of

all system-wide tests (USGAO, 1993, p. 3).

Purposes and Development

Nearly all major standardized tests share similar purposes and methods of development. The purposes of the tests are, primarily, to yield accurate comparisons or rankings of students and, secondarily, to provide information about student achievement vis-à-vis specific educational objectives. Both of these purposes presuppose the existence of certain fundamental or common educational goals that cut across school districts, states, and institutional types (e.g., public and private). This common foundation permits norm-referenced comparisons—rankings only have meaning when the basis for ranking is known—and criterion-referenced conclusions, which must be based on content knowledge and skills.

To identify these fundamental objectives, development of the major tests generally follows the same procedures. An example of the procedures is found in a manual for *Terra Nova*, which describes four sources of information for developing the content framework of that test:

- meeting with teachers, administrators, and education specialists across the country to define the content of major curriculum areas;
- reviewing curriculum guides from states, districts, and dioceses;
- examining the National Educational Goals and national content standards; and
- analyzing the content of widely used textbooks and basal series (CTB/McGraw-Hill, 1997a, p. 6).

In order to permit test users to evaluate how well this task was carried out, test publishers usually provide appendices or separate manuals detailing the names and affiliations of curriculum and content area specialists, textbook series, and state and district curriculum frameworks that were consulted in the course of preparing the tests.

Characteristics of Major Achievement Batteries

Because the development processes followed by major test publishers are nearly identical (and reflect the fairly homogeneous curricula, textbooks, and objectives of education across the U.S.), the content coverage of the various standardized achievement tests looks remarkably uniform. Nearly all of the basic batteries consist of the same content area tests. Figures 4 and 5 illustrate this by showing the content coverage of the five major batteries[10] at the early elementary and late elementary school levels. Similar development procedures also translate into similar (and uniformly strong) technical quality on characteristics such as reliability and validity.

Each publisher's major batteries are also very strongly related to each other. For example, the major products of Harcourt-Brace Educational Measurement (HBEM) in current use include the *Metropolitan Achievement Tests, Seventh Edition* (MAT-7), and the *Stanford Achievement Test, Ninth Edition* (SAT-9). The MAT-7 bears a copyright date of 1993, while the SAT-9 was published in 1996. As in this case, it is common for a publisher to have two or more series that assess many of the same content areas but bear different copyright dates. This is due to the lengthy and complex process of test development, review, validation, and norming, which can take up to 10 years. By having two (or more) products developed on staggered cycles, a publisher can always offer at least one battery that is reasonably current. Beginning with the batteries published by HBEM, the following sections describe each publisher's major products, covering the norm-referenced achievement tests in widest use in U.S. schools.

Figure 4
Content Coverage of Major Tests -- Early Elementary

	Tests				
	MAT-7 Level P1	SAT-9 Level P1	CAT-5 Level 11	CTBS/Terra Nova[1] Level 12	ITBS[1] Level 7
Grade Levels	1.5-2.5	1.5-2.5	1.6-2.2	2.0-3.2	1.7-2.6
Content Areas					
Reading					
	Vocabulary Comprehension Word Recognition	Vocabulary Comprehension Word Study Skills	Vocabulary Comprehension Word Analysis	Vocabulary Reading Word Analysis	Vocabulary Reading Word Analysis
Mathematics					
	Concepts & Problem Solving Procedures	Problem Solving Procedures	Concepts & Application Computation	Mathematics Math Computation	Math Concepts Math Problems Math Computation
Language					
	Language	Language[2] Spelling Listening	Mechanics Expression	Mechanics Spelling	Language Listening
Other					
	Science[3] Social Studies[3]	Environment[3]	Science[3] Social Studies[3]	Science Social Studies	Science Social Studies Sources of Information

[1] Elements listed in column 4 are common to CTBS and Terra Nova
Information for the CTBS/Terra Nova Level 12 is for the Complete Battery Plus, Form A. Information for the ITBS is for the Complete Battery.

[2] Alternate form SA is available to replace Form S Language, Spelling, and Study Skills.

[3] This test is not part of the basic battery; it is an optional part of the complete battery.

Figure 5
Content Coverage of Major Tests --Late Elementary

	Tests				
	MAT-7 Level I4	SAT-9 Level A2	CAT-5 Level 18	CTBS/Terra Nova[1] Level 18	ITBS[1] Level 14
Grade Level(s)	8.5-9.5	8.5-9.9	7.6-9.2	7.6-9.2	8.0
Content Areas					
Reading					
	Vocabulary Comprehension	Vocabulary Comprehension	Vocabulary Comprehension	Vocabulary Reading	Vocabulary Reading Comprehension
Mathematics					
	Concepts & Problem Solving Procedures	Problem Solving Procedures	Concepts & Application Computation	Mathematics Math Computation	Math Concepts & Estimation Math Problems Solving & Data Interpretation Math Computation
Language					
	Prewriting Composing Editing	Language[2] Spelling Listening	Mechanics Spelling Expression	Mechanics Spelling	Usage and Expression[3] Spelling Capitalization Punctuation
Other					
	Science[4] Social Studies[4] Thinking Skills[5] Research Skills[5]	Science[4] Social Science[4] Study Skills	Science[4] Social Studies[4] Study Skills	Science Social Studies	Science Social Studies Reference Materials Maps and Diagrams

[1] Elements listed in column 4 are common to CTBS and Terra Nova
Information for the CTBS/Terra Nova Level 18 is for the Complete Battery Plus, Form A. Information for the ITBS is for the Complete Battery.

[2] Alternate form SA is available to replace Form S Language, Spelling, and Study Skills.

[3] The ITBS Complete Battery in Language consists of either the four subtests listed or an integrated assessment of writing skills.

[4] This test is not part of the basic battery; it is an optional part of the complete battery.

[5] This test does not consist of distinct items; a student's score on this test is derived by scoring a collection of items that already exist in other subtests (e.g., reading, language, etc.).

The Metropolitan Achievement Tests, Seventh Edition (MAT-7)

The MAT-7 series, published by HBEM, consists of 14 different levels designed for use from the youngest kindergarten student (Level PP) to the oldest high school student (Level S4). The MAT-7 can be used in conjunction with the *Otis-Lennon School Ability Test, Sixth Edition*--an aptitude test published by HBEM--in order to obtain a separate measure of a student's scholastic aptitude or an indication of over- or underachievement.

The 14 levels of the MAT-7 can be viewed in three groupings. The first consists of Levels PP and PR, which span kindergarten through first grade. These are available only as a "basic battery" and consist of tests in Reading, Mathematics, and Language. Administration time is less than two hours.

The second group (a total of 8 levels) focuses on the elementary school level, and spans approximately grades 1 through 9. The testing time for each of these is approximately four hours; the administration manual provides a schedule for giving the test in portions over several sittings. Ascending through the grade levels, these levels carry the designations P1, P2, E1, E2, and I1 through I4 (the letters corresponding to "primary," "elementary," and "intermediate," respectively). The basic battery for each of these levels includes tests in Reading, Mathematics, and Language. Alternatively, a school district might choose the "complete battery," which adds Science and Social Studies. For 6 of the levels, the language test is subdivided into three subtests: Prewriting, Composing, and Editing. For all 8 of the levels, the mathematics test is divided into two subtests of Concepts and Problem Solving, and Procedures; the Reading test is divided into three subtests covering Word Recognition, Vocabulary, and Comprehension.

As is common on other publishers' tests, the items in subtests such as Concepts and Problem Solving can count toward a student's score in more than one way; for example, a question on problem solving might count toward the student's Concepts and Problem Solving score, toward the student's overall Mathematics score, and also toward the complete battery or composite score. However, the MAT-7 also uses multiple scoring of items to create subtests of Research Skills and Thinking Skills. For example, the 31 items shown as making up the Research Skills subtests of Level E1 do not appear on the MAT-7 as a separate "Research Skills" segment. Instead, items that assess research skills in the Mathematics, Language, and other tests are combined to form the Research Skills subtest.

The third grouping of MAT-7 tests is geared to the high school grades (Levels S1 to S4). These maintain the three reading subtests, three language subtests, research skills, thinking skills, and science and social studies options for the elementary and intermediate forms; however, the mathematics portions do not retain the Concepts and Problem Solving and Procedures subtests. They provide only an overall math score.

Reviews of all the major norm-referenced tests appear in a publication titled *Mental Measurements Yearbook*. These reviews--generally two independent reviews by specialists in educational testing--provide potential users of tests such as the MAT-7 with information about reliability, validity, norms, and other technical information; analysis of how a test compares to its competitors; and summary recommendations. Reviews of the MAT-7 appear in the *Twelfth Mental Measurements Yearbook* (Conoley & Impara, Eds., 1995, pp. 601-610). Overall, reviews of the MAT-7 describe a battery that possesses adequate reliability and validity for its purpose, though the two reviews are more cautious in their recommendations than are reviews of the other major batteries.

The Stanford Achievement Test, Ninth Edition (SAT-9)

The SAT-9 series (sometimes referred to as the "Stanford 9") is also published by HBEM and consists of 13 levels, one fewer than the MAT-7. The SAT-9 is also designed for use from kindergarten (Level S1) to high school (Level T3). To be precise, the first two levels of the *Stanford* series bear the title *Stanford Early School Achievement Test*; the elementary and junior high school levels are called the *Stanford Achievement Tests*; and the high school levels are titled the *Stanford Tests of Academic Skills*.

The SAT-9 is a close cousin of the MAT-7, though somewhat broader in coverage. The basic battery consists of Reading, Mathematics, Language, Spelling, Study Skills, and Listening tests. Again, Science and Social Science are optional extras that make up the complete battery. The broader coverage comes at the cost of additional testing time, which ranges from 2¼ hours in kindergarten to nearly 5½ hours at the upper elementary level.

The content of the SAT-9 is said to emphasize thinking skills to a degree not found in previous versions of the battery. The test manual asserts that "all of the items in Stanford 9 assess either Basic Understanding or Thinking Skills, with more items than ever before assessing the higher order skills" (HBEM, 1996, p. 8). In contrast to the MAT-7, which assesses research and thinking skills by multiply-scoring items in other subtests such as reading and mathematics, the SAT-9 has a distinct subtest for study skills (levels I1 to T3).

The content of the SAT-9 is also said to be aligned with the National Assessment of Educational Progress (NAEP) in reading comprehension, and with the National Council of Teachers of Mathematics (NCTM) *Standards* in math problem solving and procedures. In recognition of lingering debates about the teaching of literacy, two versions of the SAT-9 language test are available; one version (Form S) offers a more traditional mechanics-and-expression approach, while the other (Form SA) offers an alternative approach that emphasizes the writing processes of prewriting, composing, and editing.

Like the MAT-7, the SAT-9 can be used in conjunction with the *Otis-Lennon School Ability Test* to obtain achievement/ability comparisons. The SAT-9 also permits users to customize the battery by using an abbreviated version, an expanded version with open-ended questions in reading, math, science and social studies, and the opportunity to supplement either version with locally-developed items to enhance the test's usefulness in providing locally-relevant norm- or criterion-referenced information. An option to score the tests locally is also permitted. Although the test is generally scored to produce norm-referenced information, the SAT-9 can also be scored to provide standards-referenced information, with the descriptors *partial mastery, solid academic performance*, and *superior performance* used to denote increasing levels of content mastery.

Reviews of the SAT-9 can be found in the *Thirteenth Mental Measurements Yearbook* (Impara & Plake, Eds., 1998, pp. 921-930). Reviewers judge the SAT-9 to meet relevant professional standards for reliability and validity. Because it incorporates both constructed-response and multiple-choice items, the issues of reliability and validity are not as straightforward as they would be if only one format had been used. Inclusion of both formats frequently has the beneficial effect of enhancing validity; the skill or ability being tested is more adequately covered by a variety of assessment methods. On the other hand, reliability is often lower for constructed-response formats.[11] In the case of the SAT-9, reliability coefficients for open-ended assessments in Reading, Mathematics, Science, and Social Studies are considerably

lower than for their multiple-choice counterparts.

The California Achievement Tests, Fifth Edition (CAT-5)

The CAT-5 series is one of two major test series published by CTB/McGraw-Hill. The CAT-5 consists of 21 levels spanning kindergarten (Level K) to high school (Level 21/22). The CAT-5 is available in three versions of varying length: the Survey version is the shortest and is designed to minimize testing time; the Basic Skills version is comparable to other publishers' basic batteries; the Complete Battery contains all basic and optional subtests.

In addition to the basic elements of the CAT-5 complete battery, a number of optional components or services are available. Performance assessment modules can be included to provide "integrated outcome scores that cut across several content areas" (CTB/McGraw Hill, p. 2). An optional writing assessment can be included; the design permits users to select the writing prompts that seem most congruent with the district's writing program.

Estimates of over- and underachievement (called "Anticipated Achievement" in the CAT-5) can be obtained if the publisher's aptitude measure, the *Test of Cognitive Skills,* is administered together with the CAT-5. Like the other major tests, the CAT-5 can be scored by the publisher or locally. A unique scoring option permits users to obtain predicted performance levels for individual students on the SAT, ACT, and NAEP.

Reliability and validity data for the CAT-5 are similar to those of the other major batteries. Primary validity evidence for all NRTs is content validity; that is, the validity of the test scores and interpretations is most strongly based upon the extent to which the test reflects appropriate content for the ages, grade levels, and subject areas tested. All major NRTs begin the test development process with a review of current curriculum materials, textbooks, teaching practices, and so on, yielding--for the CAT-5 as well as the other batteries--strong evidence of content validity.

As with the other batteries, reliability coefficients are high (i.e., generally in the range of .80 to .90) and follow predictable patterns. For example, the reliability of the Complete Battery is higher than those of the individual content area tests (e.g., Reading). Reliability coefficients for longer tests tend to be higher than for shorter tests; the reliability of whole content area tests (e.g., Reading) tends to be higher than their subtests (e.g., Word Analysis). Reliability of the Complete Battery version is usually superior to the reliability of the Survey version.

The Comprehensive Tests of Basic Skills (CTBS)/Terra Nova

The other major test series published by CTB/McGraw-Hill is a combination of a traditional norm-referenced battery and alternative assessments. Strictly speaking, the series is called *Terra Nova*. However, *Terra Nova* can refer to various combinations of up to five components: 1) a traditional *CTBS* component; 2) a Multiple Assessments component that comprises both the traditional, multiple-choice portions of the *CTBS* and constructed-response items (e.g., in which the student provides an ending for a story, speculates about the reasons for a character's actions, constructs a bar graph, explains reasoning or shows work for a math problem, etc.); 3) a Performance Assessment component which increases the extent of constructed-response items in communication arts, mathematics, science, and social studies; 4) a Writing Assessment; and 5) a Custom Component that permits inclusion of items to assess educational objectives peculiar to a specific state's or district's curriculum.

When *Terra Nova* is mentioned, perhaps the most common configuration that comes to

mind is the Multiple Assessments component, which includes both traditional and alternative item types within the same test booklet. (See the example in Figure 6). The example also illustrates the recent trend of integrating assessments across content areas and incorporating assessments that tap both cognitive and affective dimensions.

As with other major batteries, *Terra Nova* was developed with attention to "thinking skills." Item development was organized around six cognitive skills: gathering information, organizing information, analyzing information, generating ideas, synthesizing elements, and evaluating outcomes. In addition, the teachers' guide for *Terra Nova* states that "each content area reflects the intent and processes described in the Secretary's Commission on Achieving Necessary Skills (SCANS) competencies" (CTB/McGraw-Hill, 1997b, p. 12).

As with other batteries, a variety of objective-based and norm-based reporting alternatives are available for *Terra Nova*, including the potential for users to generate local norms and obtain customized reports. Performance standards were also developed as another reporting option. The performance levels are intended to provide an overall description of a student's proficiency. Five proficiency levels are used for *Terra Nova* reports. In ascending order, these are: Starting Out/Step One, Progressing, Nearing Proficiency, Proficient, and Advanced. Finally, some components of the *Terra Nova* can be obtained in Spanish language versions.

For adequate technical information on the *Terra Nova*, potential users should consult the appropriate technical manuals available on request from the publisher. *Mental Measurements Yearbook* does not contain a review of this product because its development and documentation were not completed in time. However, because revised versions of standardized achievement batteries maintain strong likenesses to previous versions in their lineage, published reviews of the *Comprehensive Tests of Basic Skills, 4th edition* (in Kramer & Conoley, Eds., 1992, pp. 213-220) may be useful to potential users of *Terra Nova*. These reviews show that, like other major batteries, the CTBS-4 has generally strong content validity and reliability in the .90s for the complete battery and in the .80s for individual tests. However, the reviews also point to an area of weakness in the CTBS, that of its norms, which one reviewer called "fuzzy" (p. 217).

Reviews of the CTBS-4 point out another feature common to most standardized achievement batteries: reliability values tend to increase with the level of the test. That is, more reliable scores are seen as children progress through the grade levels. This caution against putting too much confidence in scores for students at the lowest levels (e.g., kindergarten, first grade) applies to all the major batteries.

The Iowa Tests of Basic Skills (ITBS)

The ITBS series is one of three major test series published by Riverside Publishing Company. The ITBS series consists of 10 levels for use from kindergarten (Levels 5 and 6) to ninth grade (Level 14). The other two achievement test series, the *Tests of Achievement and Proficiency (TAP)* and the *Iowa Tests of Educational Development (ITED)* extend the measurement of achievement from ninth through twelfth grade (Levels 15 to 18). If a district decides to implement a testing program using Riverside products, the most common configuration involves use of ITBS for elementary school children and the ITED in high school. The balance of this section applies to the ITBS.

The ITBS is available in three versions of varying length: shorter Core and Survey versions and a Complete Battery version.

Figure 6
Sample from Terra Nova Showing Integration of Multiple-Choice and Open-Ended Formats

The Moon

Perhaps you have gazed at the moon and wondered why it looks different at different times. This article will help explain why the moon seems to change shape. Read the article. Then do Numbers 1 through 4.

Throughout the ages, the moon, our closest neighbor in space, has excited curiosity. Have you ever heard of the dark side of the moon? It is the side that never faces the earth. We are always looking at the same side of the moon! And what do we really see when the moon shines? Moonlight? Actually, the moon has no light of its own. It is like a mirror, reflecting the sun's light. Perhaps the most curious thing about the moon is that even though the side we see is always lighted by the sun, it appears to change its shape. Sometimes we see a full moon, sometimes we see a half moon, and other times we see just a sliver of a moon.

The moon seems to change shape because we see different amounts of the moon's lighted side as it revolves around Earth. These apparent changes are called phases. In the first phase, called the new moon, we see no moon at all. In the nights following, the moon seems to grow from a sliver of light to a crescent moon. After a week, the moon has moved far enough in its circle around Earth for us to see half of its lighted side. This phase is called the half-moon phase. About one week after the half-moon phase, the entire side of the moon facing Earth is lighted by the sun. This is the full-moon phase. As the moon continues on its journey, it appears to grow smaller again, shrinking to a sliver and then disappearing altogether to become, once again, a new moon.

1. The words *full, half,* and *crescent* describe *phases* of the moon. Find the word that means about the same as *phases*.

A names
B lights
C colors
D stages

(Questions 2-7 omitted for this illustration.)

8. "The Path on the Sea" and the article about the moon's phases are examples of how two writers can choose different ways to write about the moon. The categories in the chart below will help you identify some of these differences. Write the missing information in the appropriate boxes.

Categories	"The Path on the Sea"	Article about moon's phases
author's point of view		third person
author's purpose	to describe how the moonlight on the sea looks to her	
author's approach		factual and educational
language	figurative	literal
organization	a new line for each image	

9. The author of the poem and the author of the article chose different ways to write about the moon. Which did you enjoy reading more, the poem or the article? Support your answer by choosing one of the elements from the chart that identifies what you liked about the poem or the article. Give an example from the text that illustrates that element.

From CTB/McGraw-Hill, 1997b, pp. 38, 42.

Like *Terra Nova*, the ITBS series has integrated constructed-response and performance-type items in a variety of ways. In Language Arts, the ITBS can be ordered in a traditional, four-part configuration (spelling, capitalization, punctuation, and usage and expression), or as a single, integrated writing skills test. Supplemental constructed-response items are available for the Reading, Language, and Mathematics subtests of both the survey and complete batteries. The *Integrated Performance Assessment Series (IPAS)* which offers performance measurement in Integrated Language Arts, Mathematics, Social Studies, and Science can be administered in conjunction with, or independent from the ITBS. A nationally-normed performance-style writing test, the *Iowa Writing Assessment* is another available performance assessment module.

In addition to national norms, students' performance on the ITBS can be compared with norms for Catholic/private schools, large city schools, and high and low socioeconomic groups. International norms are also available, as are national performance standards based on categories similar to those used by the National Assessment of Educational Progress (i.e., Basic, Proficient, Advanced). Predicted achievement scores and ability/achievement comparisons can be obtained if the ITBS is administered with the *Cognitive Abilities Test (CogAT)*, and numerous scoring options are available, including Windows-based software for local scoring and reporting. Braille and large-print editions of the ITBS are available.

The newest version of the ITBS derives from a long lineage of tests which began in 1935 as the *Iowa Every Pupil Test of Basic Skills*. Reviews of the ITBS in *Mental Measurements Yearbook* describe a long history of stable measurement of school abilities and basic skills. Reliability is a distinguishing characteristic of the ITBS; it has consistently high dependability of scores, especially at the upper levels. Overall, reviews of the ITBS are perhaps the most positive of the major batteries. For example, one reviewer concludes that the ITBS "is one of the oldest and best in the business. It is a set of standardized tests of basic skills that is supported by exemplary research and documentation" (Brookhart, 1998, p. 542). A reviewer of the high school extension of the ITBS, the ITED, concludes that "The Iowa Tests of Educational Development continue a long tradition of excellence. They represent valid tests of the general verbal and numerical abilities that high school students need in adult life" (Subkoviak, 1998, p. 552).

The ITBS is also somewhat more cautious than the other major batteries in its inclusion of new formats and alignment with emerging pedagogical and curricular trends. This caution is deliberate, as the content coverage and approach of the ITBS are designed to strike a balance between what currently occurs in classrooms and what major professional organizations recommend ought to occur.

Other Differences among the Major Batteries

Overall, Figures 4 and 5 portray highly similar content coverage for the major achievement batteries. However, the differences are substantial enough that attempts to link or *equate* scores from one major battery to another have proven to be problematic. The director of the Iowa Testing Programs, Robert Brennan, has stated the essential problem succinctly: "the various tests measure things that are too different from one to the other" (National Tests, 1998, p. 5).

These differences are easily discernible to the non-technical observer when the content of the tests is examined at a finer level, called the *cluster* and *subskill* levels. Clusters are small groupings of test items, consisting of several subskills. Clusters represent a narrow focus within a particular test; subskills focus

Figure 7
Comparison of Social Studies Cluster Classifications (Early Elementary)

CAT-5 (Level 11)		ITBS (Level 7)		CTBS/Terra Nova (Level 12)	
Cluster	subskills	Cluster	subskills	Cluster	subskills
Geographical Concepts		Geography		Geographic Perspectives	
	location, place, and regions		physical features		location, place, and region
	human/environment interaction		people and environment		human-environment interaction
					process/investigation
Economic Dimensions		Economics		Economic Perspectives	
	basic concepts and terms		work and workers		production, distribution, and consumption
	economic roles and communities		supply and demand		science/technology/society
	resources and technology		material needs and wants		process/investigation
	interdependence		impact of technology		
Historical Perspectives		History		Historical and Cultural Perspectives	
	time cognition and chronology		people who shaped history		time, continuity, and change
	significant events and people		change: chronology		
	change and continuity				
Government and Citizenship		Political Science		Civics and Government Perspectives	
	basic concepts and terms		citizen responsibilities		basic concepts
	governmental structure		rules		American ideals and citizenship
	laws and rules		laws		process/investigation
Sociological Patterns		Sociology/Anthropology			
	basic concepts and terms		social interactions		
	responsible behavior		human needs and wants		
	individual and group roles		psychology: learning		

even more specifically on a single skill or ability. Figure 7 shows the fine-grained content classifications at the cluster level in the area of Social Studies for three of the major batteries, the CAT-5, the CTBS/Terra Nova, and the ITBS. The figure shows how the subskills that comprise the content area clusters on a test such as social studies can vary from battery to battery. Again, with this information, users can better evaluate the match of any particular test to a local curriculum or philosophical beliefs about the structure of a discipline.

In addition to measuring somewhat different things, the various batteries differ slightly in terms of such technical characteristics as difficulty level and reliability. These differences are difficult to quantify because direct comparisons of the batteries are not usually performed. It would be nice to know, for example, how performance at, say, the 80th percentile on one battery compared to the percentile rank on the others. The ideal method of comparison--administering two or more complete batteries to the same group of students--is not feasible in terms of cost, sample size requirements, and additional student testing time.

However, the differences can be identified in other ways. For example, a simple comparison of the item difficulty indices (i.e., p-values) for the major batteries reveals that the ITBS is comprised of slightly more difficult items and the CTBS of slightly easier items, with the other major batteries falling in between. State-mandated achievement tests for home-schooled children provide another source of evidence. Because many states' regulations require performance at a specified percentile for continuation of home schooling, home schoolers are often drawn to the standardized test that is likely to yield the highest percentile score for their child, i.e. the "easiest" test. Experience suggests that the ITBS yields lower percentile ranks than the Stanford tests, which yield lower ranks than the CTBS.[12]

In summary, the major batteries all meet minimum requirements for confident use to differentiate between individual and group performance on fundamental school-related knowledge and skills. Much additional information on these tests is available from their respective publishers and potential test users are urged to gather all available information in order to make thorough comparisons. Overall, the similarities in these batteries far exceed their differences in both technical quality and content coverage. However, the differences across the tests--particularly in content coverage--are significant enough to make cross-battery comparisons tenuous at best.

Uses and Misuses

Everyone has taken a test. Surely that makes everyone capable of using and interpreting test results. Or so it would seem. People who would never dream of writing a book, making a speech, or crafting legislation on the perplexing problems of AIDS, economic recession, or reading instruction (to name just a few) are not reticent at all about making pronouncements on testing. This fact has led to increased use of tests generally, increased use of tests as policy instruments specifically, and attempts to use tests to answer questions about American education that they are ill-suited to address. However, standardized tests can provide valuable and accurate information if users recognize their strengths and limitations.

Disadvantages and Limitations of Standardized Tests

At least four significant drawbacks of standardized tests are recognized by most psychometricians and many educators.[13] These have to do with the tests' format; the use of their results; their effects on teaching and learning; and invalidity caused by misuse. Outright cheating in the administration of tests also continues to be a problem faced wherever high-stakes testing occurs.

Format

Standardized achievement tests have frequently relied on multiple-choice and other formats that can easily be coded onto machine-scannable sheets for rapid scoring. As recently as 1993, the General Accounting Office reported that 71% of all standardized achievement tests included only multiple-choice items (USGAO, 1993, p. 3). Although that percentage has probably decreased over the last five years, multiple-choice and other so-called *select-type* formats dominate the large-scale testing market. These "objective" formats are known to be most amenable to testing lower-order thinking skills (e.g., knowledge and comprehension). Although poorly constructed select-type items can promote simple recognition of a correct response (as opposed to generation of correct or unique responses), well-written select-type items (such as those in the major NRTs and most state-mandated CRTs) can tap higher-order cognitive skills.

Even when well-constructed, however, select-type items are limited in terms of the educational objectives, subject areas, and outcomes they can be crafted to address. A multiple-choice item can be used to test a student's ability to *identify* correct chronological sequencing in a story, but it can not be made to assess whether the student can *produce* a story that is interesting. For this reason, standardized achievement tests have increasingly been incorporating extended writing samples and other *constructed-response* formats, which permit the testing of a broader array of learning objectives.

Results

Standardized tests typically yield large amounts of quantitative information. The information is also expressed using concepts such as percentile ranks, normal curve equivalent scores, and so on. Yet many educators are uncomfortable or unaccustomed to dealing with quantitative information and, as mentioned previously, are generally

Figure 8
Curriculum Narrowing Effects

Examples of Irish Primary Certificate Examination Compositions, 1946-1948.

A Bicycle Ride (1946)

I awakened early, jumped out of bed and had a quick breakfast. My friend, Mary Quant, was coming to our house at nine o'clock as we were going for a long bicycle ride together.

It was a lovely morning. White fleecy clouds floated in the clear blue sky and the sun was shining. As we cycled over Castlemore bridge we could hear the babble of the clear stream beneath us. Away to our right we could see the brilliant flowers in Mrs. Casey's garden. Early summer roses grew all over the pergola which stood in the middle of the garden.

A Day in the Bog (1947)

I awakened early and jumped out of bed. I wanted to be ready at nine o'clock when my friend, Sadie, was coming to our house. Daddy said he would take us with him to the bog if the day was good.

It was a lovely morning. The sun was shining and white fleecy clouds floated in the clear blue sky. As we were going over Castlemore bridge in the horse and cart we could hear the babble of the clear stream beneath us. Away to our right we could see the brilliant flowers in Mrs. Casey's garden. Early summer roses grew all over the pergola which stood in the middle of the garden.

A Bus Tour (1948)

I awakened early and sprang out of bed. I wanted to be ready in good time four our bus tour from the school. My friend, Nora Greene, was going to call for me at half-past eight as the tour was starting at nine.

It was a lovely morning. The sun was shining and white fleecy clouds floated in the clear blue sky. As we were going over Castlemore bridge in the horse and cart we could hear the babble of the clear stream beneath us. From the bus window we could see Mrs. Casey's garden. Early summer roses grew all over the pergola which stood in the middle of the garden.

From Madaus, 1988, p. 94

ill-prepared academically with respect to even the most fundamental concepts of testing and grading. Consequently, test results are routinely ignored or used only in the crudest fashion, such as for simple rankings of students or comparisons among schools, as opposed to being used for individualizing teaching, instructional improvement, etc. E. F. Lindquist, developer of the Iowa Testing Programs and inventor of high-speed optical scanning technology, presaged current concerns about the use of test results:

> Tests seem to me to have gone farther away from higher and higher precision and more accuracy in measurement. There seems to be less of an effort to provide a really faithful, dependable picture of the abilities and aptitudes of the individual child, and more concern with group achievement along the lines that are of interest to school administrators... (quoted in Kohn, 1975, p. 20-21).

In the quarter century since Lindquist's remarks, some test publishers have struggled to make both norm-referenced information and diagnostic, criterion-referenced information available to users of their tests, as well as to make reports of test results more "user friendly." However, because the purposes and construction of NRTs and CRTs are substantially different (see discussion pp. 11-13), the marriage has not been uniformly successful. Additionally, only modest advances have been accomplished in terms of making score reports from standardized tests more useful to and interpretable by their intended audiences.

Teaching and Learning

Perhaps the greatest concern about standardized achievement testing relates to its effects on teaching and learning. The power of testing to influence what happens in classrooms was acknowledged by Popham (1980) who called the power of mandated tests to influence teaching *measurement driven instruction* (MDI). The same phenomenon was later termed--less flatteringly--*psychometric imperialism* by Madaus (1988, p. 84) who documented numerous untoward effects of externally-mandated testing on teaching and learning. The essence of the MDI principle is that, when an achievement test is mandated in a high-stakes environment, teachers will work to ensure that their students perform well on it. Abundant research on the phenomenon of MDI has documented that teachers' efforts may go beyond the desired effects of emphasizing certain educational objectives to narrowing the curriculum to focus almost exclusively on a limited set of knowledge or skills.

This effect is not only a consequence of testing formats, to be sure, nor are the curriculum-narrowing effects recent. Figure 8 illustrates the curriculum-narrowing effects of a mandated writing assessment used in Ireland in the 1940s. The sample compositions produced by students who were asked to write a brief narrative story show that teachers can teach to any type of test, not just multiple-choice tests, in a way that causes their students to approach learning in a rote fashion.

Misuse of Tests

Although not an inherent characteristic of the tests themselves, the abuse of standardized tests by educators has become a national scandal. The first wave broke in the late 1980s with the publication of Cannell's report on how public educators cheat on standardized achievement tests--better known as the "Lake Wobegon Report" (1988). Cannell discovered that, by using old versions of NRTs with outdated norms, each of the 50 states was able to claim that its students were above the national average--a logical impossibility. Further explication by Cannell (1989) provided additional reasons for the phenomenon, including lax test security, and inappropriate test preparation practices. Subsequent investigation of the Lake Wobegon phenomenon by testing specialists confirmed the sham and prompted test publishers to update norms more frequently.

Nonetheless, abuse of NRTs continues. In school districts with gifted education programs, for example, where a student's participation is tied to his or her performance on an ability test, the number of eligible students can be manipulated simply by selecting an older or more recently-normed ability test. Qualification for special education placement can be manipulated in the same way. When a school district wants to appear to be improving, an "easier" NRT can be administered than the one given the previous year; an increased percentile rank for the district can be demonstrated even in the absence of any true change in students' learning. Also, use of the same form of a test for several years can result in familiarity with the test content among both teachers and pupils, causing inflated scores. The same score-raising effect can be produced by encouraging low-achieving students to be absent on test-taking day, by removing their answer sheets from the batch of sheets to be scored, by excluding limited-English proficiency (LEP) students and special education students from testing, and so on--all practices that destroy the credibility and comparability of aggregated test results.

Inappropriate methods of preparing students to take tests can also make the results virtually meaningless. These can range from practices of questionable integrity to downright cheating. Table 2 provides a summary of the responses of various educator groups who were asked their perceptions

Table 2
Educators' Views of Questionable Test Administration Practices

	Percent of respondents considering the practice to be appropriate					
	Midwest		California			
	Teachers	Administrators	Teachers	Principals	Supts.	Board Mbrs.
Behavior						
Student practice with previous test form	34	47	57	25	60	68
Student practice with current test form	14	17	36	6	17	21

From Popham (1991)

about inappropriate testing practices. The table shows that even those practices that would clearly be proscribed by test publishers--such as having students practice on a current, secure form of the test--are viewed favorably by fairly large percentages of educators.

Outright Cheating

The problem of outright cheating on the part of teachers and administrators began to surface in the 1980s and has become even more widespread since that time. In one study, 3rd, 5th, and 6th grade teachers in two large school districts were asked how much cheating they believed was practiced by teachers in their schools (Shepard & Doughtery, 1991). Their responses, shown in Table 3, reveal at least the perception that inappropriate practices are not uncommon.

Cheating scandals have also begun to receive national attention. In 1985, an investigation of cheating in the Chicago Public Schools was undertaken because some schools showed unusual patterns of score increases and unnecessarily large orders of blank answer sheets for the *Iowa Tests of Basic Skills*. After finding a high percentage of erasures and other anomalies on one administration of the ITBS among 7th and 8th graders, a second form of the test was administered to students at a group of "suspect" schools and a group of control schools under more secure conditions. It was found that, even accounting for the reduced level of motivation students would have had on the retesting, "clearly the suspect schools did much worse on the retest than the comparison schools" and that, compared to original suspicions of the amount of cheating, "it's possible that we may have under-estimated the extent of cheating at some schools" (Perlman, 1985, pp. 4-5).

In more recent Chicago-area cases, the principal of Cherokee Elementary School in Lake Forest, Illinois, was suspended and then demoted as a result of an investigation into extraordinarily high scores by her students on the *Stanford Achievement Test*. A hearing officer found that the principal had distributed materials covering content that would be on the *Stanford* and actual copies of the test itself. In addition to giving them the forbidden materials, she had encouraged

teachers to cheat, instructing them to erase and correct answers that students had written in their test booklets before they were sent in for scoring, and to erase all answers on a math test that had not been completed by a student so that it would be invalid and not counted in the school's average (Cheating Scandal, 1992). In 1996, the Chicago school district initiated termination proceedings in the case of a curriculum coordinator who purportedly copied forms of the *Iowa Tests of Basic Skills* and distributed them to teachers at an elementary school. The teachers are believed to have led students in practicing on the same version of the test that was to be used in the district (Lawton, 1996).

In New York, the superintendent of the Barker School District resigned over his role in a cheating scandal. The superintendent and a principal were alleged to have told teachers at an elementary school to correct wrong responses on the answer sheets of third graders who had taken the New York Pupil Evaluation Program tests.

Several attempts have been made to identify inappropriate test administration practices and to quantify the extent of cheating by teachers. In one study, Kher-Durlabhji and Lacina-Gifford (1992) asked 74 teachers-in-training to indicate how appropriate they believed certain behaviors to be. Only 1.4% thought that changing answers on a student's answer sheet was appropriate, and only 2.7% said that allowing more time than allotted for a test was acceptable. However, 8.1% thought that practicing on actual test items was okay, 23.4% believed rephrasing or rewording test questions to be acceptable, and 37.6% judged practice on an alternate test form to be appropriate.

Sadly, the beliefs of pre-service teachers probably translate into actual classroom practices. Third-, sixth-, eighth-, and tenth-grade teachers in North Carolina were asked to report how frequently they had witnessed certain "test irregularities." Overall, 35% of

Table 3
Prevalence of Inappropriate Test Administration Practices

Question: To what extent do you believe these are practiced by teachers in your school?

	Percent of respondents				
Behavior	Never	Rarely	Often	Frequently	No Idea
1. providing hints on correct answers	28.5	20.8	16.9	5.8	28.0
2. giving students more time than test directions permit	38.0	19.7	15.2	4.4	22.7
3. reading questions to students that they are supposed to read themselves	38.8	22.2	11.9	2.2	24.9
4. answering questions about test content	43.2	20.5	8.9	2.8	24.7
5. changing answers on a student's answer sheet	58.4	7.8	5.5	0.6	27.7
6. rephrasing questions during testing	36.3	20.8	16.1	1.9	24.9
7. not administering the test to students students who would have trouble with it	50.7	15.8	7.5	5.8	20.2
8. encouraging students who would have trouble on the test to be absent on test day	60.1	10.8	5.5	1.9	21.6
9. practicing items from the test itself	54.6	12.5	8.0	3.3	21.6
10. giving students answers to test questions	56.8	11.6	6.4	1.9	23.3
11. giving practice on highly similar passages as those in the test	24.9	15.8	10.5	19.7	19.1

From Shepard and Doughtery (1985)

the teachers said they had observed cheating, either engaging in inappropriate practices themselves or being aware of unethical actions of others. The behaviors included giving extra time on timed tests, changing students' answers, suggesting answers to students, and directly teaching sections of a test. The teachers reported that their colleagues engaged in the behaviors from two to ten times more frequently than they had personally. The cheating was both flagrant and subtle. More flagrant examples included students being given dictionaries and thesauruses by teachers for use on a state-mandated writing test; one teacher said that she checked students' answer sheets "to be sure that her students answered as they had been taught" (Gay, 1990, p. 99).

Perhaps the mother of all cheating scandals occurred in 1996 in Fairfield, Connecticut, and involved one of that district's most respected schools, Stratfield Elementary. The Fairfield school district comprises nearly 7000 students and is widely considered to represent educational excellence. Stratfield Elementary was itself twice singled out (in 1987 and 1994) to receive "Blue Ribbon" awards for excellence from the U.S. Department of Education; in 1993 it was honored by the magazine *Redbook* as one of the best elementary schools in the country.[14] So good, in fact, was the performance of Stratfield's students, that between 1990 and 1992, composite ITBS scores for the school's 3rd and 5th graders never fell below the 98th percentile. Perhaps too good.

When the third- and fifth-grade students' answer sheets for the January 1996 administration were sent in for scoring, an analysis turned up an extremely high rate of erasures. Not just erasures, but highly unusual patterns of erasures. The analysis showed that 89% of the erasures were from a wrong response to a correct one. And the rate of erasures at Stratfield was up to five times greater than at other schools in the same district.

Because of the highly unusual patterns, Stratfield students were retested in March 1996, using an alternate form of the ITBS. The district widened its investigation into the matter and instituted a public relations campaign to control damage to its image in what was dubbed "Erasergate."[15] The retesting resulted in substantial discrepancies; scores were significantly lower on the March 1996 administration. One observer familiar with the investigation concluded that the probability of tampering with the students' answer sheets was "95% certain" but that officials would "never find the smoking eraser" (Lindsay, 1996, p. 29).

Advantages of Standardized Achievement Tests

Recognizing undesirable effects of standardized tests does not negate their advantages. The finest hammer is ill-suited to drive screws. The same hammer can even be used illegally, to break into a house, vandalize property, or commit murder. Intentional misuse of a tool does not reflect on the advantages it holds for accomplishing its intended purpose. For standardized achievement tests, these advantages include efficiency, usefulness, technical characteristics, and content coverage.

Efficiency

Among available options, standardized achievement tests currently yield the greatest amount of information about student performance for the resources invested. They provide more information about students, at less cost, and with less student testing time, than other available alternatives, such as portfolios or performance assessments. These facts pertain not just to traditional tests using multiple-choice formats, but also to tests that include alternative formats, extended writing samples, and other constructed-responses. A

second aspect of efficiency involves the turn-around time for score reporting. Typically, results from standardized achievement tests are available to policy makers, school districts, teachers, and parents within a few weeks of test administration.

Ease of Use

Although many teachers lack the training and experience to make optimal use of test information for improving classroom instruction, test publishers and scoring contractors have developed reporting options that synthesize and summarize data on student performance in ways that are easily used by various audiences. These include summary reports for policy makers, student performance reports for use by school personnel, and narrative reports designed for parents. To some extent, these reports ameliorate the problem of lack of expertise in interpreting performance reports.

A second aspect of usefulness is familiarity. Standard ways of reporting performance on standardized tests have become recognizable indicators for policy makers, educators, and parents. For example, the simple percentages of students who correctly answered questions in subjects as diverse as mathematics, geography, history or literature are easily presented to a variety of audiences and are not easily misinterpreted. Similarly, an individual student or school district's performance at, say, the 75th percentile is readily interpreted by most people as meaning that the student or district performed better than 75% of other students or districts in the comparison group. Although alternative reporting systems are being proposed, they are still poorly developed and just as apt to lead to confusion as to provide useful information.

Standardized achievement tests also permit comparisons of individuals and groups in ways that alternative measures are not yet capable of doing. Portfolio assessments, for example, are good for gathering and representing the unique accomplishments of individual students, and may be well-suited to the needs of the student and classroom teacher. However, portfolios are less useful for *system* uses such as reporting aggregated performance, or for accountability and monitoring uses. This characteristic is, of course, related to their purpose and design: current standardized achievement tests were constructed for ease of comparing students and systems; portfolios seek primarily to portray individual accomplishments.

Reliability and Validity

Current standardized tests provide exceptionally *reliable* and *valid* information about student achievement. Standardized achievement batteries have reliability coefficients of .95 or greater on the scale from 0.0 to 1.0 (see previous sections). This accomplishment may be due in part to the fact that the technology of standardized testing has benefited from decades of development, compared with some newer alternatives. One investigation of large-scale alternative testing in Vermont was unusually frank, noting that the state's testing program

> … has been largely unsuccessful so far in meeting its goal of providing high-quality data about student performance. The writing assessment is still hobbled by unreliable scoring, and the mathematics assessment has yet to demonstrate that it has addressed the vexing problems of validity that confront...unstandardized assessments embedded in instruction. (Koretz, Stecher, Klein, & McCaffrey, 1994, p. 11)

The fact that the objectives tested by standardized tests are derived from widely-used textbook series, state- and district-level curriculum guides, and professional association guidelines enhances the content validity of the tests. Additionally, extensive

Figure 9
Breadth of Coverage

Iowa Tests of Basic Skills, Level 10

- Mathematics
  - Math Concepts and Estimation
    - Numeration and Operations
    - Geometry
    - Measurement
    - Fractions/Decimals/Percents
    - Probability and Statistics
    - Equations and Inequalities
    - Rounding
    - Order of Magnitude
    - Compensation
  - Math Problems and Data Interpretation
    - Single-step addition or subtraction
    - Single-step multiplication or division
    - Multiple-step problems with whole numbers or currency
    - Problem solving strategies
    - Reading amounts from graphs
    - Determining differences and finding ratios
    - Identifying trends or underlying relationships; drawing conclusions
  - Math Computation
    - Adding whole numbers with and without renaming
    - Subtracting whole numbers with and without renaming
    - Multiplying whole numbers with and without renaming
    - Division (basic facts, computation with and without remainder)

From Hoover, et al., 1996a, pp. 47-55

review by content experts and studies demonstrating the tests' predictive ability give weight to the argument that they measure some of the outcomes deemed important in U.S. schools.

Comprehensiveness

A final advantage of standardized achievement tests is related to the first one (i.e., efficiency): they assess a broad range of educational objectives. As noted in the previous section, commercially-published batteries typically measure pupil achievement in numerous content areas, including language arts, mathematics, study skills, and science. Because of their efficiency, considerable information for sub-areas can also be obtained. For example, the mathematics portion of the *Iowa Tests of Basic Skills*[16] consists of the three sub-areas and numerous further subdivisions illustrated in Figure 9. (Of course, the tradeoff for attaining substantial breadth of content coverage is reduced ability to measure student achievement in all these areas in similar depth.)

Promoting the Appropriate Use of Standardized Tests

Because standardized test results are often used in political contexts, their advantages can be overstated or under-valued, and their limitations can be accentuated or ignored, depending on the aims of the user. When used improperly, standardized tests and test results can fail to yield the accurate, dependable information and conclusions that they were designed to produce. To address the need for clear guidance about their proper use and interpretation, and to educate various audiences about their potential misuse, a number of professional organizations have developed guidelines that articulate sound testing practices. The following describe the current guidelines promulgated by three major groups: testing specialists, teachers, and administrators.

Guidelines Developed by Testing Specialists

Standards for Educational and Psychological Testing

Since 1954, the American Psychological Association (APA) has participated in the publication of guidelines for test development and use. Other organizations, namely the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME), have joined with the APA to cosponsor revisions of the guidelines. The current version is called the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1985); a revision is anticipated by the year 2000. According to the current version

> The purpose of publishing the *Standards* is to provide criteria for the evaluation of tests, testing practices, and the effects of test use. Although the evaluation of the appropriateness of a test or application should depend heavily on professional judgment, the *Standards* can provide a frame of reference to assure that relevant issues are addressed (p. 2).

Despite these modest purposes, the *Standards* have proven to be a sophisticated and highly influential compilation of technical and procedural guidance; they have been extensively relied upon in test development and reporting, and in litigation concerning tests. The *Standards* are organized around key principles of measurement, with section titles such as "Validity," "Reliability and Errors of Measurement," and "Scaling, Norming, Score Comparability, and Equating." Among the guidelines described in this section, the *Standards* are considered to be the most authoritative statement regarding appropriate test development and use.

Code of Fair Testing Practices in Education

In addition to the organizations that sponsor the *Standards for Educational and Psychological Testing*, three other groups joined them to produce the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988). These groups are the American Association for Counseling and Development, the Association for Measurement and Evaluation in Counseling and Development, and the American Speech-Language-Hearing Association. The *Code* is "meant to apply broadly to the use of tests in education" but is "directed primarily at professionally developed tests such as those sold by commercial test publishers or used in formally administered testing programs" (p. 1). The *Code* presents separate guidelines for test users and test developers, and covers designing and selecting tests, interpreting scores, promoting fairness, and informing test takers. A sample of the guidelines most relevant to readers of this booklet--those for

Figure 10
Guidelines from the *Code of Fair Testing Practices in Education*

Guidelines for Developing/Selecting Appropriate Tests

Test Users Should:

1. first define the purpose of testing and the population to be tested. Then, select a test for that purpose and that population based on a thorough review of the available information.

2. investigate potentially useful sources of information, in addition to test scores, to corroborate the information provided by tests.

3. read the materials provided by test developers and avoid using tests for which unclear or incomplete information is provided.

4. become familiar with how and when the test was developed and tried out.

5. read independent evaluations of a test and of possible alternative measures. Look for evidence required to support the claims of test developers.

6. examine specimen sets, disclosed tests or samples of questions, directions, answer sheets, manuals, and score reports before selecting a test.

7. ascertain whether the test content and norms group(s) or comparison group(s) are appropriate for the intended test takers.

8. select and use only those tests for which the skills needed to administer the test and interpret scores correctly are available.

Guidelines for Interpreting Scores

Test Users Should:

1. obtain information about the scale used for reporting scores, the characteristics of any norms or comparison group(s), and the limitations of the scores.

2. interpret scores taking into account any major differences between the norms or comparison groups and the actual test takers. Also take into account any differences in test administration practices or familiarity with the specific questions in the test.

3. avoid using tests for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use.

4. explain how any passing scores were set and gather evidence to support the appropriateness of the scores.

5. obtain evidence to help show that the test is meeting its intended purpose(s).

From Joint Committee on Testing Practices, 1988

designing and selecting tests and interpreting scores--is provided in Figure 10.[17]

Code of Professional Responsibilities in Educational Measurement

One group that participated in the 1985 *Standards for Educational and Psychological Measurement* has developed a separate set of standards to guide the conduct of members who engage in any type of educational assessment. The National Council on Measurement in Education has published the *Code of Professional Responsibilities in Educational Measurement* with the intent that its guidance should apply "to any type of assessment that occurs as part of the educational process, including formal and informal, traditional and alternative techniques for gathering information used in making educational decisions at all levels" (NCME, 1995, p. 2). The guidelines in the *Code* are addressed primarily to those who make and use tests, and include sections on developing assessments, marketing assessments, selecting assessments, administering assessments, and so on.

Guidelines Developed by Education Associations

Two sets of guidelines have been prepared primarily under the leadership of education associations; one contains guidelines for teachers while the other offers guidance for administrators.

Standards for Teacher Competence in Educational Assessment of Students

The *Standards for Teacher Competence in Educational Assessment of Students* were jointly developed by the American Federation of Teachers, the National Education

Figure 11
Guidelines from the *Standards for Teacher Competence in Educational Assessment of Students*

Teachers should be skilled in:

1. choosing assessment methods appropriate for instructional decisions.

2. developing assessment methods appropriate for instructional decisions.

3. administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.

4. using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.

5. developing valid pupil grading procedures which use pupil assessments.

6. communicating assessment results to students, parents, other lay audiences, and other educators.

7. recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

From AFT/NCME/NEA, 1990.

Figure 12
Guidelines from *Competency Standards in Student Assessment for Educational Administrators*

A. Competencies associated with **assisting teachers**:

1. Have a working level of competence in the *Standards for Teacher Competence in Educational Assessment of Students*.

2. Know the appropriate and useful mechanics of constructing various assessments.

B. Competencies associated with **providing leadership** in developing and implementing assessment policies:

1. Understand and be able to apply basic measurement principles to assessments conducted in school settings.

2. Understand the purposes (e.g., description, diagnosis, placement) of different kinds of assessment (e.g., achievement, aptitude, attitude) and the appropriate assessment strategies to obtain assessment data needed for the intended purpose.

3. Understand the need for clear and consistent building- and district-level policies on student assessment.

C. Competencies needed in using assessments in **making decisions** and communicating assessment results:

1. Understand and express technical assessment concepts and terminology to others in nontechnical but correct ways.

2. Understand and follow ethical and technical guidelines for assessment.

3. Reconcile conflicting assessment results appropriately.

4. Recognize the importance, appropriateness, and complexity of interpreting assessment results in light of students' linguistic and cultural backgrounds and other out-of-school factors and in light of making accommodations for individual differences, including disabilities, to help ensure the validity of assessment results for all students.

5. Ensure that assessment and information technology are employed appropriately to conduct student assessment.

6. Use available technology appropriately to integrate assessment results and other student data to facilitate students' learning, instruction, and performance.

7. Judge the quality of an assessment strategy or program used for decision making within their jurisdiction.

From AASA/NAESP/NASSP/NCME, 1997 (emphasis in original)

Association, and the National Council on Measurement in Education (AFT/ NCME/NEA,1990). They consist of the seven standards shown in Figure 11. According to the *Standards* document, they are "intended to guide the preservice and inservice preparation of educators, the accreditation of preparation programs, and the future certification of all educators" (p. 1).

Competency Standards in Student Assessment for Educational Administrators

Four organizations collaborated to produce the *Competency Standards in Student Assessment for Educational Administrators*. These were the American Association of School Administrators (AASA), the National Association of Elementary School Principals (NAESP), the National Association of Secondary School Principals (NASSP) and the National Council on Measurement in Education (NCME). Like the assessment standards for teachers, the administrator standards were intended to guide preservice and inservice education and to influence personnel certification and institutional accreditation. The document comprises 12 standards, summarized in Figure 12, and organized around the themes of assisting teachers, providing leadership, and making decisions using assessment information.

Summary

Three conclusions about the uses and abuses of standardized tests can be drawn from this and preceding sections. First, when used appropriately, norm-referenced achievement tests can provide valuable information about student performance. Second, the recognized benefits and limitations of these tests can be viewed as two sides of the same coin. For example, the benefit of efficiency has recently been faulted for promoting lower-order thinking skills and artificial contexts at the expense of more creative problem solving in authentic situations. Finally, adherence to proper use and avoidance of abuse of tests requires that those who use them be educated and conscientious. Professional guidelines exist to promote sound testing practices. The extent to which these guidelines are followed is the extent to which consumers of test information are accurately informed.

Issues and Trends

Many forces affect testing in U.S. schools. In this final section, seven issues are addressed that will concern parents, educators, and policy makers in the coming years.

Educator Preparation

American education suffers from a serious deficiency in the training of school personnel to administer, interpret, and use information generated by tests (Gullickson, 1986; Hills, 1991; Ward, 1980). Despite the existence of standards for competence in assessment (described in the previous section), the vast majority of states still do not require explicit training in assessment as a condition for teacher certification; administrators are often even less well prepared in basic assessment than are teachers. One testing expert observed that generally we are "a nation of assessment illiterates" (Stiggins, 1991, p. 535).

Parents need basic, easily accessible information about their children's progress. Often, the conduit for that information is the classroom teacher and local administrators. Thus, it is essential that teachers be prepared to interpret test performance in ways that parents can use, and that administrators provide critical oversight and security when high-stakes tests are administered.

Teachers also need to understand certain fundamental principles of testing--such as reliability and validity--so that they are prepared to address concerns such as the limitations in generalizability of a student's performance, test administration and security, and the inherent imprecision of test scores. Along with principles that apply to large-scale achievement testing, teachers need to be better prepared to design, administer, and interpret high-quality classroom assessments of their own.

Administrators and policy makers require test data as part of the information they use to monitor and suggest improvements in education systems.[18] Obviously, they need to know good data from bad; they need the ability to analyze how the data at hand bear on important education problems; and they need skill in translating evidence about student performance into sound proposals for change.

Part of adequate preparation in assessment is awareness of ethical issues. As we saw in the previous section, educators' self-interests can motivate inappropriate testing practices, obfuscation of results, and patently fraudulent activities. At an individual level, education is needed to increase awareness of the inappropriateness of these activities and the need for test security. At the state and national levels, continuing education on ethical testing practices, clarifications of regulatory guidelines, and aggressive prosecution of those who engage in unethical test-related activities may be necessary to address these problems.

Developing Test Formats

Although multiple-choice continues to be the format of choice for the majority of large-scale achievement testing, many innovative formats have recently been introduced, and many test developers are reintroducing traditional formats such as extended writing

samples. The primary challenge associated with alternative assessment techniques is refining the formats to be more efficient in terms of student testing time, more accurate in scoring, and more valid in terms of providing a representative picture of student achievement. Testing specialists continue to work on these problems and on potential solutions (e.g., computer scoring of essays[19]) that may have significant effects on the cost, accuracy, and dependability of test information.

One popular alternative is portfolio assessment. Portfolios are collections of work samples, anecdotal records, and other information about the student. Many states currently require or are planning to implement portfolios as a method of representing students' accomplishments and preparation to enter the world of work. On a smaller scale, many teachers have begun to replace traditional tests and grades with student portfolios. In both contexts, portfolios often represent a departure from valued features of traditional assessments. For example, although they do display actual student performances, portfolios frequently contain smaller, less representative samples of a student's work. They can also yield less accurate and dependable information due to subjectivity and variability in the scoring of portfolios by human evaluators. Work samples may vary from student to student, reducing the ability to make comparisons between individuals or to a common standard. And, because portfolios frequently contain products that result from group effort, the unique contributions of individual students may not be discernible.

Two key questions must be answered in the area of portfolio assessment. The first question is a technical matter, the second a policy concern. On the technical side, it is still unclear whether portfolios will be able to provide sufficiently accurate and dependable information for making decisions about individual students. Advances in the quality of portfolio assessment systems and their *instructional* utility notwithstanding, their eventual place as student *evaluation* tools remains an open question. Second, it is essential to recognize the longstanding demands of parents and policy makers for aggregate data that can be useful for making comparisons. It is unclear what role, if any, portfolios can play in the system-monitoring functions that large-scale achievement testing programs typically provide.

Amount of Testing Time

With nearly every state mandating some form of achievement testing, many districts instituting their own testing requirements, and the federal government involving itself in testing through the National Assessment of Educational Progress, Title 1, and the proposed Voluntary National Tests to name a few, the amount of time devoted to testing continues to expand. It is not surprising that there is interest in gauging the extent to which some of the information generated by the constellation of tests that students take may be redundant, unnecessary, or simply not an effective use of limited education dollars and hours. The so-called *linking,* or at least coordinating, of local, state, and national tests represents the hope that an efficient alternative to currently fragmented efforts will be found.

Integrity of Test Data

The information generated by standardized achievement tests is only as good as the quality-control procedures employed in gathering it. As mandated testing spread in the 1980s, researchers began investigating the unintended consequences of those tests. The director of testing for Austin (Texas) Independent School District, Glynn Ligon, candidly observed:

> Teachers cheat when they administer standardized tests to students. Not all teachers, not even very many of them; but enough to make cheating a major concern to all of us who use test data for decision making. (Ligon, 1985, p. 1)

As shown in the previous section, cheating is a serious problem affecting achievement testing programs. In the wake of well-publicized cheating scandals, many states are attempting to control the problems of prior access to secure test materials, "teaching to the test," failures to follow test administration guidelines, and outright cheating. Some of these problems can be easily addressed, while others are quite difficult to affect.

For example, many large-scale achievement testing programs have begun to offer better training for those responsible for secure test materials. Shipping test materials to school sites at the latest possible time has helped reduce the temptation (and opportunity) to access materials for inappropriate purposes. Shrink-wrapping of materials, providing only the exact number of tests required, and specifying accounting procedures for test materials have also helped reduce test security problems.

Still, problems persist. So long as standardized test results reflect on the instructional and administrative quality of schools, they will be viewed as "high stakes" for teachers and principals, and there will be incentives for cheating. To the extent that tests become less prominent as accountability tools, test security concerns are also likely to diminish.

A challenge for the future is deciding how best to obtain accurate data on educational performance *and* address accountability concerns. Current accountability systems might be described as "internal audits" because they are implemented by participants within the system. This means that those whose performance is, in part, the object of accountability systems are also charged with gathering the data. Such systems present inevitable conflicts of interest and are naturally susceptible to corruption in ways discussed previously. Substantial developmental work remains to be done regarding what an "external audit" system might look like, and regarding the feasibility and likely consequences of such a system.

Methods of Reporting Educational Achievement

The familiar *percentile rank* and *grade equivalent* scores of traditional norm-referenced standardized tests have reasonably straightforward interpretations and have been shown to provide useful information to parents, educators, and policy makers. Yet concerns persist about the potential for misinterpretation of these and other ways of reporting on student achievement. For example, grade-equivalent (GE) scores, reported on many standardized tests to show relative achievement, are commonly misinterpreted as providing information about students' functional level. A fourth-grade student with a GE score in Reading of 9.7 has certainly outperformed most of the norm group on the Reading test, but the GE score of 9.7 does not mean that he or she is capable of reading or functioning at the 9th grade level.

In the past few years, initial attempts have been made to develop new reporting methods, with research being conducted by organizations (e.g., the National Assessment Governing Board for public reporting on NAEP) as well as individual researchers (e.g., Ronald Hambleton at the University of Massachusetts). The development of new ways to report on student learning may represent the area of greatest growth in test development over the next decade.

Validation of Test Uses

Who is responsible for assuring that test results are used properly? This question is at the heart of another current debate. In a desire to prevent tests from being used as accountability tools, proposed revisions to the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1985) currently include a mandate that evidence be provided by test users whenever a claim is made about an intended use or consequence of testing. This new standard would apply to standardized achievement tests, such as the *Iowa Tests of Basic Skills* and could profoundly affect the use of standardized achievement tests.

For example, if legislators mandated the *ITBS* in order to promote more attention to basic skills in a state's classrooms, or to "raise educational standards," the legislature (or state department of education) would need to provide evidence that its use did, in fact, result in increased attention to basic skills and raise standards. In practice, this specific standard would be nearly impossible to satisfy in the context of the global rationales for testing often proffered by policy makers. The practical effect could be the proscribing of many current testing programs aimed at increasing educational accountability.

Standard Setting

The problem of how best to establish performance standards continues to plague achievement testing. The practice of using norms as standards--despite being universally condemned by testing specialists--continues in many testing contexts. The use of norms as standards occurs when relative performance is mandated as a standard for admission to a program, selection for an honor, etc. One example is the use in many states of percentile standards for accountability, in which parents must submit evidence that a student attained a certain score (i.e., performance at the 30th percentile) on a nationally-normed standardized test in order to remain eligible for home schooling.

A number of alternatives exist. Standard-setting has received increased attention in the last several years; numerous technical advances have been made and new methods are being investigated. However, the technical issues related to standard setting are likely not as germane to parents, educators, and policy makers as the conceptual issues.

Recently, much has been written about *standards* generally, and concepts such as *performance standards*, *content standards*, and *opportunity to learn standards*, in particular. Setting *performance standards* refers to the process of deciding how much a student should know or how well a student should perform on a test. The process is complicated because of the judgment that is necessarily involved. Measurement specialist Robert Linn has identified performance standards as reflecting four potential uses: "(1) exhortation, (2) exemplification of goals, (3) accountability for educators, and (4) student certification" (1994, p. 3). Depending on the goals of those who set them, standards can be developed that simply validate the status quo, represent aspirations for the future, provide concrete indications of the knowledge and skills a student possesses, promote accountability, or some combination of these.

For standards to have real impact on educational reform, however, two characteristics are essential. First, they must represent attainment of challenging knowledge and skills that are relevant to joining the work-force or continuing education. That is, standards must necessarily be criterion-referenced or content-referenced to some extent in order that performance on any test can be used to gauge students' preparation for competition in a college classroom or workplace. Frameworks such as those used by the National Assessment of Educational Progress represent attempts, through a classification system including the levels "Basic," "Proficient," and "Advanced,"

to imbue performance levels with specific content-mastery relevance.

Second, to be meaningful to parents, teachers, and others, and to allow us to make real judgments, some indication of the context of performance is also necessary, including knowledge of how students in other school districts, states, and countries perform on the same (or parallel) sets of items or tasks. For example, knowing that 10% of U.S. students would be classified at (or below) the Basic level in reading, with 70% Proficient and 20% Advanced might be--given information about the knowledge and skills represented by each level--highly informative and encouraging. On the other hand, if international comparisons revealed that the percentages in other advanced countries were 10, 20, and 70, respectively, for the Basic, Proficient, and Advanced levels of performance, such news would be cause for alarm. Only by gathering and monitoring both kinds of information can policy makers and the public be assured that whatever standards are invoked actually represent the levels of accomplishment that are truly desirable for American school children.

References and Resources

The first of two following sections provides references for works cited in this document. In the second section, contact information is provided for the major standardized test publishers in elementary and secondary education and related resources.

References

American Association of School Administrators, National Association of Elementary School Principals, National Association of Secondary School Principals, National Council on Measurement in Education. (1997). *Competency standards in student assessment for educational administrators*. [cooperative publication by sponsoring organizations]

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Federation of Teachers, National Council on Measurement in Education, National Education Association. (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: Authors.

Balow, I. H., Farr, R. C., & Hogan, T. P. (1993). *Directions for administering, Metropolitan Achievement Tests, seventh edition; Complete/Basic battery*. San Antonio, TX: Psychological Corporation.

Brookhart, S. M. (1998). Review of *Iowa Tests of Basic Skills*. In J. C. Impara & B. S. Plake (Eds.), *The thirteenth mental measurements yearbook* (pp 539-542). Lincoln, NE: University of Nebraska, Buros Institute of Mental Measurements.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above average. *Educational Measurement: Issues and Practice, 7*(2), 5-9.

Cannell, J. J. (1989). *The "Lake Wobegon" report: How public educators cheat on standardized achievement tests*. Albuquerque, NM: Friends for Education.

Cheating scandal jars a suburb of high achievers. (1992, January 1). *New York Times*, p. 32.

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. E. (1995/1996). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment, 3*, 159-179.

Conoley, J. C., & Impara, J. C. (Eds.) (1995). *The twelfth mental measurements yearbook.* Lincoln, NE: University of Nebraska, Buros Institute of Mental Measurements.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Holt, Rinehart, & Winston.

CTB/McGraw-Hill. (nd). *A look inside the CAT/5.* Monterey, CA: Author.

CTB/McGraw-Hill. (1997a). *The only one: Terra Nova*. Monterey, CA: Author.

CTB/McGraw-Hill. (1997b). *Teacher's guide to Terra Nova.* Monterey, CA: Author.

Finn, Jr., C. E. (1993, January 20). What if those math standards are wrong? *Education Week*, p. 36.

Gay, G. H. (1990). Standardized tests: Irregularities in administering of tests affect test results. *Journal of Instructional Psychology, 17*(2), 93-103.

Gullickson, A. R. (1986). Teacher education and teacher-perceived needs in educational measurement. *Journal of Educational Measurement, 23*, 347-354.

Haney, W. M., Madaus, G. F., & Lyons, R. (1993). *The fractured marketplace for standardized testing*. Boston, MA: Kluwer.

Harcourt-Brace Educational Measurement. (1996). *Directions for administering, Stanford Achievement Test, ninth edition; Complete/Basic battery multiple choice.* San Antonio, TX: Author.

Hills, J. R. (1991). Apathy concerning testing and grading. *Phi Delta Kappan, 72*, 540-545.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., Dunbar, S. B., Oberley, K. R., Bray, G. B., Lewis, J. C., & Qualls, A. L. (1996a). *Iowa tests of basic skills, content classifications with item norms, complete/core/survey batteries, levels 5-14, form M*. Itasca, IL: Riverside.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., Dunbar, S. B., Oberley, K. R., Bray, G. B., Lewis, J. C., & Qualls, A. L. (1996b). *ITBS interpretive guide for teachers and counselors levels 5-14, form M, complete and survey*. Itasca, IL: Riverside.

Impara, J. C., & Plake, B. S. (Eds.) (1998). *The thirteenth mental measurements yearbook.* Lincoln, NE: University of Nebraska, Buros Institute of Mental Measurements.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington, DC: American Psychological Association.

Kher-Durlabhji, N., & Lacina-Gifford, L. J. (1992, April). Quest for test success: Preservice teachers' views of high stakes tests. Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, TN.

Kohn, S. D. (1975). The numbers game: How the testing industry operates. *Principal, 54*(6), 9-23.

Koretz, D., D. McCaffrey, S. Klein, R. Bell, and B. Stecher. (1993). Reliability of scores from the 1992 Vermont portfolio assessment program (CSE Technical Report 355). Los Angeles, CA: RAND Corporation.

Kramer, J. J., & Conoley, J. C. (Eds.) (1992). *The eleventh mental measurements yearbook.* Lincoln, NE: University of Nebraska, Buros Institute of Mental Measurements.

Lawton, M. (1996, November 13). Alleged tampering underscores pitfalls of testing. *Education Week*, p. 5.

Ligon, G. (1985, March), Opportunity knocked out: Reducing cheating by teachers on student tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Lindsay, D. (1996, October 2). Whodunit? Officials find thousands of erasures on standardized tests and suspect tampering. *Education Week*, pp. 25-29.

Linn, R. L. (1994, October). The likely impact of performance standards as a function of uses: From rhetoric to impact. Paper presented at the NAGB-NCES Conference on Standard Setting, Washington, DC.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the National Society for the Study of Education* (pp. 83-121). Chicago, IL: University of Chicago Press.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology, fourth edition*. Fort Worth, TX: Holt, Rinehart and Winston.

Michaels, H. R., & Ferrara, S. (forthcoming). Using data to drive state policy and local reform. In G. J. Cizek (Ed.), *Handbook of educational policy*. San Diego, CA: Academic.

National Council on Measurement in Education, Ad Hoc Committee on the Development of a Code of Ethics. (1995). *Code of professional responsibilities in educational measurement*. Washington, DC: Author.

National Tests: Will We Go There? (1998, Spring). *University of Iowa College of Education Perspectives*, p. 5.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Lawrence Erlbaum.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*(7), 561-565.

Perlman, C. L. (1985, April). Results of a citywide testing program audit in Chicago. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Phelps, R. P. (1998). The demand for standardized student testing. *Educational Measurement: Issues and Practice, 17*(3), 5-23.

Phelps, R. P. (1997). The extent and character of system-wide student testing in the United States. *Educational Assessment, 4*(2), 89-121.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan, 68*(9), 679-682.

Popham, W. J. (1991, April). Defensible/indefensible instructional preparation for high-stakes achievement tests. Presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Roeber, E., Bond, L., & Connealy, S. (1998). *Annual survey of state student assessment programs.* Washington, DC: Council of Chief State School Officers.

Shepard, L. A., & Doughtery, K. C. (1991). Effects of high-stakes testing on instruction. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(7), 534-539.

Subkoviak, M. (1998). Review of *Iowa Tests of Educational Development*. In J. C. Impara, & B. S. Plake (Eds.), *The thirteenth mental measurements yearbook* (pp. 550-552). Lincoln, NE: University of Nebraska, Buros Institute of Mental Measurements.

U. S. Congress, Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (Report No. OTA-SET-519). Washington, DC: U.S. Government Printing Office.

U. S. General Accounting Office. (1993). *Student testing: Current expenditures, with cost estimates for a national examination.* (Report No. GAO-PEMD-93-8). Washington, DC: U.S. Government Printing Office.

Ward, J. G. (1980). Teachers and testing: A survey of knowledge and attitudes. In L. Rudner (Ed.), *Testing in our schools* (pp. 15-24). Washington, DC: National Institute of Education.

Resources

Buros Institute of Mental Measurements Publisher of the two most comprehensive and authoritative resources on published tests, the *Mental Measurements Yearbook* and *Tests in Print*. Both of these resources contain objective reviews of tests by measurement specialists, and provide information on technical characteristics, costs, administration time, ordering information, etc. Both books are nearly universally available in college and university libraries; more information about these resources can be found at the Institute's web site. The web site also contains links to a test locator data base (ERIC) and an on-line test review locator.

Contact Information: Buros Institute of Mental Measurements
Barbara S. Plake, Director
135 Bancroft Hall
University of Nebraska
Lincoln, NE 68588-0348

Internet: www.unl.edu/buros/index.html
Telephone: 402-472-6203
Email: bplake@unlinfo.unl.edu

CTB McGraw-Hill Publisher of the *California Achievement Test (CAT)*, the *Comprehensive Tests of Basic Skills (CTBS)*, and *Terra Nova*. A limited amount of information on these products is available at the company's web site.

Contact Information: CTB/McGraw-Hill
20 Ryan Ranch Rd.
Monterey, CA 93940

Internet: www.ctb.com/index.htm
Telephone: 800-538-9547
Email: ctbwebmaster@ctb.com

Harcourt Brace Educational Measurement Publisher of the *Stanford Achievement Test* and the *Metropolitan Achievement Test*. The company's web site has extensive product information on its ability, achievement, and attitude tests, including information on the latest versions of its *Stanford* (SAT-9) and *Metropolitan* (MAT-7) tests; follow the Trophy Case link. Links are also provided to the *Code of Fair Testing Practices in Education* and to a "Checklist for Reviewing Standardized Achievement Tests" by following the Library link.

Contact Information: Harcourt Brace Educational Measurement
555 Academic Court
San Antonio, TX 78204-2498

Internet: www.hbem.com/index.htm
Telephone: 800-211-8378
Email: customer_service@hbtpc.com

Riverside Publishing Company Publisher of the *Iowa Tests of Basic Skills (ITBS)* and the *Iowa Tests of Educational Development (ITED)*. The company's web site has ample information on test development, products, and scoring services found by following the Products & Services link. Regardless of which of the major achievement tests a user is considering, ordering a copy of the test content and administration manuals for the *ITBS* or *ITED* is a good idea; Riverside's references such as the *ITBS Manual for School Administrators* provide accurate, complete, and in-depth information about achievement, ability, and aptitude assessment that rivals some textbooks in the field of testing.

Contact Information: Riverside Publishing Company
425 Spring Lake Dr.
Itasca, IL 60143-2079

Internet: www.riverpub.com/
Telephone: 800-767-8420
Email: rpcwebmaster@hmco.com

Notes

[1] In truth, it is incorrect to classify items dichotomously as *authentic* or *unauthentic*. Strictly speaking, all testing would be classified as unauthentic and contrived, because achievement testing essentially ceases (in one form) when a student enters the world of work. It is more accurate to speak of "more authentic" and "less authentic" in relationship to the degree to which items present tasks or problems that are likely to be encountered in real life. Unfortunately, teachers cannot be certain which tasks will be germane--i.e., "authentic"--with respect to any given student. In the end, it seems wisest to prepare students in ways that maximize the potential applicability of knowledge to variegated practical situations, perhaps by stressing the generality of the knowledge as opposed to its specific application to any "real-life" context.

[2] A concrete example of current confusion about the terms *test* and *assessment* is apparent in a recent name change for the SAT, which went from the *Scholastic Aptitude Test* to the (redundant) *Scholastic Assessment Test*.

[3] A variation of this process, called *rolling norms* allows the norms to be continually updated, usually on an annual basis.

[4] Of course the district would recognize, as was noted in the previous section, that average performance at the 80th percentile only represents where the district stands with respect to others, not necessarily that the content covered by the test was rigorous, appropriate, relevant to students' success, or indicative of globally competitive performance.

[5] A number of writers have documented the inadequate preparation of teachers in assessment of any kind. See, for example, Cizek, Fitzgerald, and Rachor (1995/1996), Gullickson (1986), and Ward (1980). Hills (1991) has also documented a corresponding apathy regarding testing.

[6] The linkage of test content to standards promulgated by professional organizations such as the NCTM is not as straightforward as it might seem. Two obvious problems--technical and conceptual--have not yet been addressed satisfactorily. First, a student's performance on a *standards-referenced test* is rarely perfectly mapped to the student's unique state of knowledge and understanding vis-à-vis the content standards. In the field of psychometrics, this concern is being addressed in the study of *cognitively diagnostic testing* (see Nichols, Chipman, & Brennan, 1995), although progress in this area has been limited and no cognitively diagnostic tests are currently available for widespread use. Second, the incorporation of content standards begs the question of whether the standards developed by organizations such as the NCTM are worthy in the first place, in the sense of bearing a relationship to the way students learn, the way teachers teach, standard curricula in place in the U.S., and so on (see Finn, 1993).

[7] The preceding description of the evolutionary changes in content, standards, and testing is presented as if the changes are broadly and uncontroversially accepted. This is not necessarily the case. The following example of "The Evolution of the Math Problem" (source unknown) illustrates, with humor, serious concerns about the ways in which related changes in teaching and learning can affect achievement and assessment.

> 1950 (math): A logger sells a truck load of lumber for $100. His cost of production is 4/5 of this price. What is his profit?

1960 (traditional math): A logger sells a truck load of lumber for $100. His cost of production is 4/5 of this price; in other words, $80. What is his profit?

1970 (new math): A logger exchanges as set L of lumber for a set M of money. The cardinality of set M is 100, and each element is worth $1. Make one hundred dots representing the elements of the set M. The set C of the cost of production contains 20 fewer points than set M. Represent set C as a subset M, and answer the following question: What is the cardinality of the set P of profits?

1980 (developmentally appropriate math): A logger sells a truckload of wood for $100. His cost of production is $80, and his profit is $20. Your assignment: underline the number 20.

1990 (NCTM Standards math): The government protects public lands by taxing big lumber companies at a rate of $10,000 annually per acre under production. Lumber production has increased at 20% per year, beginning with 50 million acres in 1990. Take your calculator to someone who can help you determine how much revenue this policy generated from 1990 to 1993. Or draw a graph.

1995 (Outcomes Based Education math): By cutting down virgin forest, a logger makes $20 in profit. What do you think of her way of making a living? (2 points)
Bonus: How do you think the forest birds and squirrels feel about the destruction of their environment? (10 points)

[8] Harcourt-Brace Educational Measurement is a testing research, development, and publication division of Harcourt-Brace & Company. This division consists largely of the former Psychological Corporation.

[9] In the course of preparing to review these tests, I requested information from the three companies that publish the five tests. Each company produces what are called *specimen sets*, which are ordinarily only made available to qualified school personnel and are provided with strict security guidelines. These sets usually contain--or *should* contain--samples of the tests themselves, and enough information about content, technical characteristics, cautions, and scoring and reporting options to permit a potential user to make an informed evaluation and decision regarding the suitability of the product for local use.

I requested the specimen sets in the same way as a typical school district testing coordinator might do so. I telephoned regional representatives from each company and requested the same materials--specimen sets for two grade levels and technical overview materials for the respective tests. I was surprised at the variability in the materials I received.

Both Riverside (publishers of the *Iowa Tests of Basic Skills* and *Iowa Tests of Educational Development*) and CTB/McGraw-Hill (publishers of the *California Achievement Tests* and *Terra Nova*) provided a wealth of material, including sample tests, detailed information on test development, content outlines, and technical characteristics. The materials from Riverside Publishing were exceptionally well-prepared, accurate, and complete.

[10] The tests listed for each battery were taken from two current forms for each battery. To provide a broad picture of content coverage, forms were chosen to represent an early elementary level (designed for grades 1-2) and a later elementary level (grades 7-8).

[11] Two different types of reliability tend to be lower for constructed-response formats than for select-response formats. First, because scoring is usually automated for selected-response formats (e.g., multiple-choice) via optically scanning equipment, there is almost never a difference between repeated scorings of the same set of responses. For constructed-response items, which must be scored by human readers using scoring rubrics, variation in repeated scoring is much more likely. Second, the internal consistency of

selected-response formats is generally higher because the items tend to be more unidimensional, whereas constructed-response formats--by design--often assess multidimensional skills.

[12] According to some experts associated with the major batteries, the difference between the "easier" and "harder" batteries may be as much as five percentile ranks depending on the tests and the location in the ability distribution.

[13] It has become fashionable to begin a list of limitations of standardized tests in education with the historical missteps of early psychometricians, such as phrenology, eugenics, and bias in intelligence testing. For the interested reader, there are numerous sources for this information but, for purposes of this booklet, that information is omitted. Indeed, it is questionable whether such information is even relevant to modern educational achievement testing. For example, although I only occasionally skim the journals of the medical profession, I have concluded that medical specialists do not feel compelled to recall, in painful detail, the use of leeches in order to motivate the advances they proffer.

[14] The material presented in this and following paragraphs regarding the Fairfield, Connecticut schools is drawn from information on the scandal published in Lindsay (1996).

[15] Apparently, the damage was severe. Lindsay (1996) reports that the district hired a person named Thomas Failla, the same media consultant who managed public relations for Union Carbide Company in the aftermath of the Bhopal, India chemical spill that killed over 2000 people in 1984.

[16] This information is for the *ITBS* complete battery, Level 10 (see Hoover, et al., 1996a).

[17] Although guidelines in these categories have been published for test developers and test users, only the guidelines for test users are presented in the figure.

[18] Of course, a beginning step toward these goals is to recognize the importance of using data to inform educational decision making--a step that may not necessarily be assumed to have been taken. For more information on this essential perspective, see Michaels and Ferrara (forthcoming).

[19] One interesting development that may have application to standardized achievement testing is that of computer scoring of extended writing samples. Page and Petersen (1995), who coordinate "Project Essay Grade," have developed software capable of reproducing the actual average scores given by six human raters better than smaller sets (e.g., 1 to 3) of human raters could do. Because the smaller number of raters is typically used in large-scale testing programs, such software might represent a significant advantage.