Chapter 1

Issues and Measurement Practices in the Schools

We must start where men who would improve their society have always known they must begin—with an educational system restudied, reinforced, and revitalized.

—Lyndon B. Johnson
January 12, 1965

Issues and Themes

Testing is an integral part of the life story of all American children. The American public supports verifying the quality of public education through accountability. This press for educational accountability has increased both the number and importance of educational assessments in the schools. An assessment is a multidimensional method for collecting data that usually includes testing. Educational testing is normally conducted to measure the status of the child on one dimension, such as arithmetic, whereas an assessment is designed to provide multiple sources of data. A battery of tests is often the core component of the assessment. Batteries are a collection of tests that are designed to measure different parts of the curriculum—e.g., mathematics, reading, and science.

High-stakes assessments are educational measures that have significant negative consequences for failure. The No Child Left Behind Act (NCLB; P.L. 107-110, 2002) is the latest major stimulant for the use of high-stakes assessments in public education. During the past 20 years, federal court rulings have supported both the use of these high-stakes tests and the sanctions that are applied when children and schools fail to measure up. One outcome of the No Child Left Behind Act mandates is a major boon for the publishers of educational tests. Over a billion dollars are now being spent by the states to develop, administer, and score these new mandated tests. The whole face of testing is changing rapidly with the introduction of technology into the assessment process. With all of these changes, the education profession has scrambled to provide statements of ethical principles and practitioner guidelines for testing programs.

Learning Objectives

By reading and studying this chapter you should acquire the competency to

- Describe the relationship between accountability, assessment, and testing
- Explain the major reasons why children are assessed
- Describe the three assumptions made by test developers
- Discuss the relationship of formative and summative testing in the classroom
- Contrast norm-based and criterion-referenced tests
- Describe the case law related to the selection of students for admissions into advanced high school programs
- Explain how case law has affected the college admission process
- Explain how case law has affected the practice of educational testing and assessment
- Present and discuss the core ethical canons related to educational testing

Accountability, Assessment, and Testing

Accountability refers to the linkage and balance between the outcome of an enterprise and the efforts and resources used to achieve that outcome. In education, **assessment** provides an accounting of how much children learn

in school and what resources are expended on achieving those learning outcomes. The need for accountability grew in this country as the cost of education grew. In the 1960s, the cost of public education was not just the largest part of each states' budget; it was actually equal to the cost of running every other state agency *combined*.

Educational accountability requires that all students be assessed to quantify what they have learned and what skills they have developed. Commercially published tests of student achievement have been around since the 1920s. The publishers of these tests have formed a cozy relationship with the schools that bought into their use for school and student **evaluation**. By the 1980s, virtually every school system could make the illogical boast that their school district was above average. A West Virginia physician (John J. Cannell, M.D.) was first to ask the question: If the average is in the middle of the data, how it is possible for everyone to be above average (Phelps, 2005)?

The 19th century saw the first efforts in the United States to determine the success of schools in meeting local and statewide goals for the education of children. The first attempt to assess a large American school system involved a test administered to the children of Boston in 1845. That testing program was organized by Massachusetts's new state school superintendent, Horace Mann (Crocker, 2003).[1] With this effort to assess educational outcomes, the students of Boston were tested for their understanding of the facts that they were learning in the new "**common schools**" of Boston. In 1864, the Board of Regents of New York initiated a statewide "preliminary test" for junior high–aged school students. This measure was used as a basis for allocating state funds to the various school systems.[2] This was supplanted by a mandated set of examinations known as the New York Regents Examinations in 1878, which assessed all high school students (New York State Education Department, 1987).[3]

The assessments of student educational progress can be accomplished using several different methods. In this text there is a chapter describing alternative approaches to assessment (Chapter 9) and another describing the use of essay tests (Chapter 7). However, the paper-and-pencil test with **multiple choice questions** is the dominate method of constructing all the state-mandated tests designed for the assessment of achievement. The advantages to this approach to assessment include the fact that these "**objective tests**" are less expensive to score and involve less effort by local school personnel to develop and administer than do alternative assessment approaches. Also, data from these tests are readily quantifiable and familiar to the policy makers and the general public. The downside of this approach to assessment is that the multiple choice questions used to build these measures tend to

stress rote learning, favor students from middle-class backgrounds, and are based on a technology born under a cloud of tacit racism (see Chapter 2).

In summary, accountability in education is an inevitable requirement that is associated with spending large amounts of public tax funds. Educational accountability requires that an assessment be made of the outcome of the educational process. One method of assessment involves using paper-and-pencil tests. Typically these tests are primarily composed with multiple choice format questions. When there are several tests that make up the assessment, the term **test battery** is appropriate to describe the **measurement**.

WHY DO WE TEST?

At one level we test to protect the health and physical status of the child. This type of testing starts at the moment of birth. Typically, it is the attending obstetrician who welcomes the newborn into the world, and it is this physician who is first to formally assess the child. That medical status examination, the **Apgar**, is an observational rating scale that is administered one minute after the birth of the child and again four minutes later (see Table 1.10). From this starting point, the growing child will be assessed and measured by a pediatrician on a regular basis.

By the age of four, children enrolled in **Head Start programs** are tested to determine how well those centers are doing their jobs. This is one of the accountability functions of testing. This assessment process is then repeated in the public schools under a federal mandate beginning in third grade and extending into the high school years.

In 1965 President Lyndon Johnson's legislative package included a new act directed toward providing federal assistance to the nation's public schools. This act, the **Elementary and Secondary Education Act** (**ESEA**, 89-10, 1965) provided money to improve the educational achievement of children living in poverty. One part of this act created a method to measure the impact of the improvements being instituted throughout the schools of the country. That method resulted in the development of what is known as the Nation's Report Card (**National Assessment of Educational Progress** [**NAEP**]). Today this measure provides a picture of how well each state is doing in educating its children.

Also, we test to measure how much progress each child is making toward developing proficiency in core areas of learning. The driving issue here is also one of accountability in education. This type of testing was begun during the 1990s and became mandatory in 2002 under terms of the revised ESEA, now

Table 1.1 A Proposal of a New Method of Evaluation of the Newborn Infant

A score is given for each sign at one minute and five minutes after the birth. If there are problems with the baby, an additional score is given at 10 minutes. A score of 7–10 is considered normal, while 4–7 might require some resuscitative measures. A baby with Apgars of 3 and below requires immediate resuscitation.

	Sign	*0 Points*	*1 Point*	*2 Points*
A	Appearance (Skin Color)	Blue-gray, Pale All Over	Normal, Except for Extremities	Normal Over Entire Body
P	Pulse	Absent	Below 100 bpm	Above 100 bpm
G	Grimace (Reflex Irritability)	No Response	Grimace	Sneeze, Cough, Pulls Away
A	Activity (Muscle Tone)	Absent	Arms and Legs Flexed	Active Movement
R	Respiration	Absent	Slow, Irregular	Good, Crying

SOURCE: From "A proposal of a new method of evaluation of the newborn infant," by V. Apgar, 1953, *Current Researches in Anesthesia & Analgesia, 32*, p. 261–267. Reprinted with permission from Eric Apgar.

known as the **No Child Left Behind** (**NCLB**) Act. Today, all 50 states have both specified what all children are expected to learn at each grade level in the core subjects (reading, mathematics, and science) and have developed tests to measure achievement in those areas. Naturally, our schools have revamped their curriculums in an effort to stress these core subjects. This has had a negative impact on the variety of subjects and disciplines taught in the schools (Jennings & Rentner, 2006).

The curriculum most public school children are exposed to has been skewed away from the arts and humanities and toward the content covered by the **mandated assessments** (Dillon, 2006; Manzo, March 2005). This has done great damage to those curriculum areas that are not tested. As a result, the curriculum followed in most schools now de-emphasizes areas such as the arts, social studies, modern language, and physical education while providing extra doses of basic skills-development drill and practice in reading, mathematics, and science.

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

Case in Point (1a)

A group of music educators in Florida spent three years and $90,000 developing a music test that can be used as a part of a statewide assessment (Gupta, 2005). This test is presented on a CD and is answered by students on machine-scorable multiple choice forms. This test is scheduled to be a part of the statewide assessment program in 2008. By going on the offensive and forcing their subject into the state testing system, music educators have saved a place for themselves in the school curriculum. Many other subjects such as art and physical education may find that they have less space in the curriculum as they are not part of the statewide tests.

To verify the old adage that no bad idea stays dead forever, a number of states are using these student data from the NCLB Act to award teachers with pay raises. **Merit pay** is incredibly difficult to institute fairly. There are so many differences between schools and the children who populate them that a system based on student **achievement test** scores is inherently flawed. The pressure for merit pay is an extension of the **accountability** focus of policy makers who look at schools in much the same way as they look at corporations. The dismal results from almost two centuries of trying this in Great Britain have been ignored in the United States (Wilms & Chapleau, 1999).

Those same state-mandated tests perform a second accountability function: They provide the data needed to determine if various groups of students within each school are making progress toward the goal of being proficient in the core subjects. This function of the mandated tests is linked to the NCLB requirement that all identified groups of children are proficient in the core subjects by the year 2014. Each school must show that all groups—including children receiving **special education**, Native American children, non-Hispanic Black children, Hispanic children, children with one or more disabling conditions, Anglo-White children, and children from impoverished homes—are making Adequate Yearly Progress (AYP) toward the goal of universal proficiency by 2014. The various states have each identified specific standards for student learning and have established proficiency targets for the children of each grade level. Those targets or **benchmarks** are unique to each state but must be approved by the U.S. Department of Education (U.S. Department of Education, 2007). The argument can be made that the policy of evaluating all schools exclusively on arbitrarily established standards for learning, and fixed levels of achievement, do not do justice to

the diversity of communities, students, and schools. The central requirement of the NCLB legislation (viz., that all children achieve at a proficient level by 2014) is one that will result in nearly all schools failing to meet the mandate (Linn, 2007b).[4]

School Strategies

Schools have taken steps to reduce the number of children scoring below the level of proficient on these mandated assessment tests. One step involves the introduction of a published achievement test during first and second grades. This achievement testing provides the teachers with data that can identify weakness in the curriculum and spot those children who may be at risk of failure on the mandated assessment in third grade. Other steps schools can take may involve after-school and summer remediation programs or revising the school's curriculum to emphasize the elements measured by the assessment program. Yet another strategy involves the addition of supplementary instructional staff to tutor and provide individualized assistance to children identified as being at risk for failure.

Case in Point (1b)

Perhaps the best strategy that a school can take to improve test scores is to involve the teachers in reviewing the curriculum and the tasks required on the mandated test. One component of such a review relates to the learning standards that are being measured on the assessment and verifying that the school's curriculum provides all students with instruction in those areas (Kristoback & Wright, 2001). The second component is one that is frequently overlooked. This involves assuring that all the test's modalities are familiar to all students. For example, a school that teaches spelling by having children memorize spelling lists, and then tests its students by having them write down the dictation of their teacher, may do badly on a standardized spelling test. Standardized tests measure spelling achievement by having children mark all the words in a printed passage that are not spelled correctly. If children never saw this method for testing spelling achievement, they will not score well no matter how good they may be at spelling. Their low scores will not indicate what they know, only their lack of familiarity with that particular modality of testing.

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

Testing is also done in the classroom by classroom teachers. For the most part the measures used by teachers are created by teachers or borrowed and modified by the teachers from the publisher of classroom textbooks. These teacher-made tests may be designed to check on student understanding and used in "real time" to inform the ongoing classroom instruction (see Chapter 6). In this way, testing occurs in a formative environment. A *formative test* is one used to check on the efficacy of the teaching-learning process. It can identify problems in understanding and guide the teacher in reteaching difficult topics and assisting students in achieving the instructional objective.

Alternatively, teacher-made tests may be designed to provide an "end of the instructional unit" summary of what each child has learned. This is known as the summative function of tests. **Summative testing** provides data needed to make objective judgments about the child, including assigning report card **grades**. To the extent possible, teachers should always work toward assigning grades that do not appear to be either arbitrary or capricious. For this reason, test data built from measures of the curriculum that were actually taught should be at the core of the report card grades.

Testing is also done to determine the special needs that some children may have for additional learning support and individualized education. This type of testing is part of the process of making an **entitlement decision.** An entitlement decision can provide extra assistance to a child who has fallen significantly behind his or her peers in terms of classroom achievement. The regulations of the various states provide that all children are entitled to a thorough and efficient education. An assessment is one approach that can provide the necessary data to document the need for assistance.

Finally, testing is also done to determine which children are selected to receive advanced or specialized educational opportunities. This type of testing can be as prosaic as deciding if a young child is ready to attend kindergarten or should wait for a year.

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

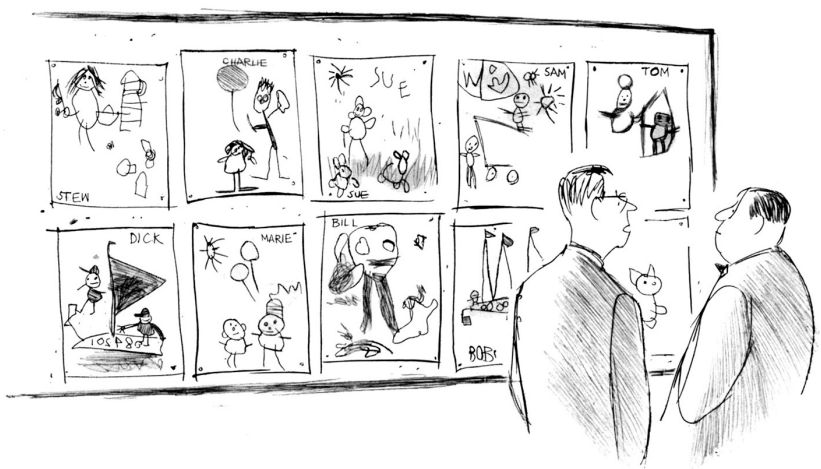
Case in Point (1c)

Eighty years ago, the first tests for infants and young children were devised by Arnold L. Gesell at Yale University. It was Gesell who introduced the word ***readiness*** into the lexicon of educators. His research demonstrated that children have points in their development when they are mentally or physically ready to acquire a new skill or learn a new concept. Instruction before that time of readiness is futile, but once the child has reached readiness, then learning proceeds rapidly (Gesell & Thompson, 1929). Until recently, this construct was

widely employed by public schools to decide which children were ready for admission into kindergarten and first grade and which should wait a year. In 2000, the use of these assessments for readiness screening was deemed unacceptable by the National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE, 2000). School readiness tests are still widely used and are even mandated in several states (Stephens, 2006).

Today, parents are more likely to hold their children back from enrolling in kindergarten (Datar, 2003; Gootman, 2006; Russell & LaCoste-Caputo, 2006). This voluntary delay in starting public education is being done more frequently in states where grade promotion from third grade is contingent on a test score.[5] Parents who do this want their children to be a year older and more mature than their peers entering school (Brock, 2006). This practice is so widespread that it has its own sobriquet: "**academic redshirting**."

About half of the states require that specialized educational programs be made available to the brightest and/or most **gifted** students in the schools. Admission into these programs typically involves cognitive tests covering dimensions such as mental ability and **creativity**. This type of cognitive testing



"Which is yours?"

Figure 1.1 "Which Is Yours?"

SOURCE: The New Yorker Collection, 1962. James Stevenson, from cartoonbank.com. All rights reserved.

is also linked to college admission and certain scholarship award programs. Parents of gifted children are generally not satisfied with the NCLB testing program and the resulting curriculum modifications in the public schools. These parents see the present curriculum, with its heavy emphasis on drilling basic skills in reading, mathematics, and basic science, as not meeting the needs of their children (DeLacy, 2004; Reed, 2004; Tierney, 2004). These parents are secure in the belief that their children will pass any state assessment and want them to have an educational experience rich in complex thinking tasks, and they want an educational program that is supplemented with advanced coursework in the sciences and arts. Research has supported this parental concern. A study using **computer adaptive testing** in Idaho demonstrated that as educational resources are focused on children who have the greatest educational needs, advanced students experience minimal achievement growth. This reflects that observation that the new curriculum over-teaches concepts that the advanced students already understand and know (Clark, 2005).

To make matters worse, there is a clear and direct relationship between measures of **cognitive ability** and the outcome on statewide assessments (Burson & Wright, 2003). In other words, those children who are the most gifted in academic ability do well on the high-stakes tests while those who have less cognitive ability are less likely to score at the level of proficient.

This raises a question of what the state-mandated high-stakes tests actually measure. These assessments are supposed to be a measure of how well students have achieved the state's approved learning standards. A general criticism of the tests is that they lack cognitive richness and are not designed to elicit complex thinking (Lane, 2004). Research has documented that many of the questions included on statewide assessments are written in a way that requires less complex thinking than the state's own standards may require (Webb, 2002, 2005).

Because there is no federal mandate requiring the states to address the differential educational needs of academically gifted children, many of those programs for the academically talented or gifted have been truncated or even eliminated to provide the resources needed to help the academically less able reach the goals of the No Child Left Behind Act (Berger, 2007). This has not gone unnoticed by the parents of gifted children, who have become vocal critics of the No Child Left Behind Act (Clark, 2005; Cloud/Thornburg, 2004; DeLacy, 2004; Goode, 2002).[6]

TYPES AND VARIETIES OF TESTS

There are four sources of achievement tests administered in the public schools today: classroom tests and quizzes, published achievement tests, and

the high-stakes tests required by the state education departments. In addition to these there are a myriad of other tests that are administered as part of guidance activities, by reading specialists, by **school psychologists**, and by speech and language therapists. A primary focus of this book is on achievement testing in the schools.

The four sources of achievement tests can be further organized into two groups. One group consists of tests and quizzes made by the classroom teacher. Teacher-made tests are described in some detail in Section III, Chapters 6, 7, 8, and 9. The second group consists of the published tests and assessments that are developed by private contractors and public agencies.

Published Tests

One type of school-based achievement test is published by the large conglomerated education publishing houses. Each year hundreds of millions of published tests are taken by students in the United States. This represents billions of dollars in revenue for the test publishers each year. These corporations publish numerous products and offer a range of consulting services in addition to providing educational measurements. Included among the major publishers are Pearson Assessments (www.pearsonassessments.com), Riverside Publishing (www.riverpub.com/products/index.html), CTB McGraw-Hill (www.ctb.com), and Harcourt Assessment (https://harcourtassessment.com).

A second source of published tests is the various state departments of education. Under the provisions of the No Child Left Behind Act of 2002, all states must assess their public and charter school children starting in the third grade. This assessment must involve a test based on the state's approved learning standards. Because there are consequences for schools that do not meet the Adequate Yearly Progress (AYP) mandate of the NCLB Act, these tests are referred to as **high-stakes tests**. Not only can schools and educators be excoriated over poor student performance, but in eight states grade promotion for the children is contingent on achieving good scores on these measures.

Case in Point (1d)

In addition to parents, school administrators have also begun to encourage the **grade retention** of primary grade children who are at risk for not passing the statewide assessment test in third grade. A clear example of this is the State of North Carolina, which initiated a required test, the North Carolina End of

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

Grades Test. Since requiring that test, the number of students retained in kindergarten and first and second grades has increased twofold. Grade repetition is a major expense for the schools. It now costs North Carolina approximately $140 million a year to educate these extra children (those who were not promoted) in the primary grades.

In 22 of the 50 states there is also a mandated test for high school graduation (Olson, August 2006). In 2007, approximately 65% of all high school seniors were required to pass a state test to qualify for their diploma. For these students, the term *high-stakes* is especially poignant. With a few exceptions, like Oregon, most state education departments do not actually write or score their own tests. These tasks are outsourced to private contractors (for more detail, see Chapter 11).

The final group of published tests used in the schools is provided by the **College Board** and the **American College Tests**, the **ACT**. There is a description of these two competing admission testing programs in Chapter 12.

Scoring Criteria

One way to classify tests is by the way they are scored. A published test may be scored using an absolute standard or criterion. These tests are referred to as being **criterion referenced** and are used to demonstrate whether the student has reached a required level or standard of achievement. For example, the 50 statewide mandated assessment tests each have a required level of success that children must reach to be graded as proficient. Other examples of criterion-based tests include licensing tests such as the PRAXIS published by ETS and the Class III Pilots Written Examination administered by the Federal Aviation Administration. In each of these cases there is a required passing score expressed in terms of the number of questions answered correctly that the test taker must obtain to pass or be proficient.

A second approach to scoring published tests is to employ a norm-reference group. A **norm-referenced test** is scored by comparing the **raw scores** from a current test taker with the scores achieved in the past by a sample of subjects used to establish the expected scoring pattern for the test. This group used to set the expected pattern of scores is referred to as the "norm group." Thus, each test taker is assigned a score that has been compared to

a standard established by previous test subjects. The comparison group may have been established once in the past when the test was originally published. This model is followed by the popular Terra Nova achievement test published by CTB/McGraw-Hill.

Many states have it both ways. They report the scores of individual students in terms of a criterion based on **cut scores** and also report norm-based scores to the schools. These norm-based scores can be aggregated and the average score calculated for each school. These data are then reported in the local press and published as part of a required **school report card**. This school report card is designed to make it easy for parents to see which schools in a community are doing better than others. Unfortunately, private and parochial schools are not required to use the state assessment and rarely make a public report of the results of their own testing programs.

Administration

Published tests can be administered to groups of students in classrooms or even given to hundreds at a time in large halls. Group-administered educational measurements can involve a single dimension test, such as a test of mental ability (e.g., Otis Lennon School Ability Test [OLSAT] from Harcourt Assessment), or, they can be multifaceted, covering a range of different curriculum areas (e.g., Iowa Tests of Basic Skills [ITBS] from Riverside Publishing). Tests such as the ITBS are referred to as **test batteries**. In this context, the word *battery* refers to the fact that there are two or more parts to the test, much like a musical batterie (homophone), which can be used to describe the different drums in the percussion section of a band, or an artillery battery, which describes two or more cannons able to fire projectiles together.

Tests are also administered one-on-one. These are usually diagnostic measures designed to identify specific areas in which the student is experiencing learning problems. These tests may also measure mental ability with individually administered instruments, such as the battery of mental ability tests of the Wechsler Intelligence Test for Children, third edition (WISC III), from Harcourt Assessment.

Individually administered **diagnostic tests** can be customized for the child by the psychologist who is doing the testing. An example of this is **curriculum-based assessment** (**CBA**), which involves a brief series of problems or tasks taken from the curriculum material that the child is studying. The repeated measurement over time using these brief tests, known as **curriculum probes**, provides a picture of the progress a child is making toward learning the subject area or required skill.

Information Sources

With the hundreds of different tests available to use in the schools, there is a need to find unbiased sources of information about these measures. Not all educational tests are created equal, and many commercially published assessments are rubbish. Others may be well designed but are still inappropriate for a particular school to use. The first step in selecting the optimal published test for a school's use involves identifying the goals for testing. It is an axiom of measurement, that the test must match the purpose and goals defined by the educators of the school. Once that has been decided, it is then appropriate to review all of the possible measures that could meet the identified goals for testing.

The largest collection of tests is the one maintained by the **Educational Testing Service**. That collection extends from the present time and goes back over a hundred years. All in all, there are over 25,000 published tests maintained in the ETS collection. Online test descriptions can be reviewed on the ETS Web page (http://ericae.net/testcol.htm#ETSTF then open "ETS Test Collection Page ETS"). Independent reviews and evaluations of tests are available from the Buros Institute of Mental Measurements on the campus of the University of Nebraska, Lincoln. A total of over 4,000 tests are described and reviewed in this collection. These reviews and descriptions can be found on the Buros Web page (http://buros.unl.edu/buros/jsp/search.jsp). In 2007, the fee for this service is $15 per test review.

ASSUMPTIONS MADE BY TEST DEVELOPERS

With the plethora of assessments and tests used to measure our children, few educators ever consider the basic assumptions that underlie this endeavor. The first assumption is empirical. This assumption is one implying the belief that by carefully observing small aspects of a child's behavior it is possible to make an informed conclusion about what he or she has learned or is able to do.

The second assumption is one linked to the adequacy of the coverage provided by the test. It is never possible to exhaustively assess every bit of knowledge and every skill a child has learned through instruction. Therefore, all educational tests are only a small sample of what the child knows or can do. This leads to questions of how adequate is the curriculum coverage provided by the test. Does the test evaluate only a select portion of the content being evaluated? Or does the test measure the full domain? To assure complete coverage, the test should include an array of items selected randomly from the whole domain of possible content areas designated for assessment.

This is a goal rarely achieved by professional publishers and almost never by classroom teachers. A test should be viewed as a sample of behaviors. As a sample, a test is not a perfect representation of the full domain of knowledge or skills being evaluated.

Another point is that all tests and assessments represent the performance of the child at one moment in time. The condition of the child changes on a regular basis, and therefore we should anticipate that all test and assessment scores are going to vary. These two points combine to imply that the outcome of all tests includes a component of error and that all test scores have a degree of instability. This point should always be emphasized, as our society relies on the scores from tests and other assessments to make critical personnel and **placement decisions**.

TRENDS IN TESTING

The use of tests and assessments in public education will increase in both the near and long-term future. The format of educational tests will also evolve as testing becomes more closely integrated into ongoing instruction. This instructional integration will be facilitated by the application of modern online technologies.

Early Use of Computers in Testing

Over the past 50 years, schools have made a number of attempts to integrate computers into instruction. These efforts date from the first interactive learning systems of the late 1960s. Systems such as the first generation of PLATO (Programmed Logic for Automatic Teaching Operations) and TICCIT (Time-Shared Interactive Computer Controlled Information Television) were linked to university-based mainframe computers. These learning systems required a great deal of expertise at both the university computer center and in the public school to operate. These interactive systems also required expensive hardware and were very expensive to maintain. Naturally, without federal funding this first generation of interactive instructional systems was soon considered déclassé. In part, this was also owing to the natural limitations of the technology of the era. By 1972, it was only possible for 1,000 students to log on and use the PLATO system at the same time (McNeil, 2004). Today, most schools have scores of computers with broadband Internet connections. Interactive tests and tutorials were a part of these early mainframe-supported learning systems. These first online

instructional programs were the true precursors of what is happening in educational measurement today.

New Technology

The educational world was changed forever when Steve Jobs and Steve Wozniak began to market the Apple I in 1976. Through the next three decades, one generation of desktop computers followed another. In the schools these computers were usually kept under lock and key in a "computer laboratory." The computer applications typically involved word processing and the use of instructional software that was in a game format (Wright & Lesisko, 2007).

Today there is a new technology, and the children entering school are already sophisticated in its use. Schools are now scrambling to catch up with the technological skills of the offspring of the X generation. Schools are now employing online testing and assessment, including online diagnostic evaluations that are matched with individualized tutorials. In a number of school districts it is now possible for parents to communicate with teachers online and read up-to-the-moment evaluations of their child's progress.

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

Case in Point (1e)

In 2003, Tennessee parents became able to access the full file of school assessment data records for their children and could read the predicted likelihood of their children passing the state-mandated state assessment. These parents could even access and read a projected score for their children on the SAT and ACT examination programs.

Online Report Cards

After 100 years, the era of the quarterly report card is almost over. The online system is also eliminating the old paper grade book. Teachers can now post grades from school or home and parents can read those grades when they get home from work each day. This allows parents to see problems as they develop and take action as needed. Parent conferences will no longer

hold surprises for the parents, who will be well versed as to the daily progress of their children. This form of parent–school linkage is already in large-scale use in Arizona and will soon impact all schools (Ryman, 2005).

Computerized Grading

The evaluation of student writing is now also being done by computers online. The state of California has contracted with the Educational Testing Service to provide an online essay examination as a part of that state's required high school graduation test. In 2004, 17 states administered mandated state assessment tests to students over the Internet. That number is growing and will soon include all schools in this country. But it is not only the children who are being assessed online; schools have started to use online talent tests to prescreen prospective teachers (Keller, 2004, May).

It is possible to tour this new system online and have a practice session with it by visiting the California Education Department's Web site: www.cde.ca.gov/ta/tg/sr/resources.asp.

Qualitative Assessments

The second new direction in the future of testing will see the reemergence of a more qualitative form of assessment. Standardized tests are designed so that the student is required to get the one correct **answer**. The new trend is toward a more open format of assessment. The introduction of written essays and open-ended mathematics problems on standardized assessments is a step in this direction. Perhaps a clearer picture of the future is the new statewide testing program in Nebraska. That state has attempted to meet part of the NCLB mandates by allowing school systems to employ a **portfolio** assessment system from the third through the eighth grades. This portfolio assessment serves as an optional part of the mandated statewide testing program. The Nebraska program, the School-based, Teacher-led Assessment and Reporting System (**STARS**), has been shown to provide a good measure of achievement in mathematics; however, the reading assessment has not proven to be as reliable as the typical standardized assessment test (Brookhart, 2005).

Qualitative assessment such as that used in Nebraska requires a holistic, open-ended scoring system. **Holistic scoring** implies that the evaluation is of the whole of the child's work considered in its totality and not conducted as

a simple summation of the quality of the various parts. In a real sense, this process is much like selecting a Most Valuable Player (MVP) in sports. The electors do not grade individual components of each player's performance but make one overall judgment of quality. This type of assessment provides classroom teachers with information that is far more useful in planning for instruction with individual children. This reflects the fact that evaluation data gathered in this informal way can be integrated into the instructional program immediately and the results of instruction can be assessed in real time as the teaching is occurring.

LAW AND TESTING

For over 40 years the federal government has established laws mandating various testing programs. State legislatures have also been involved in establishing large-scale testing programs by writing mandated tests and test policy into state codes. Yet, there is another level where testing programs and the legal system interact. That interface occurs through the courts and what is referred to as "case law."

Legislation

The use of test data, and other school records, was first addressed by the federal government in the **Buckley Amendment**, or the Family Educational Rights and Privacy Act (20 USC S. 1232g, 1974). This law details the parent's right to inspect and offer corrections to the educational records maintained by the school. It also assures that information from those records is not released without the signed permission of the parents. The exceptions to this release rule provide for ongoing educational research, accreditation reports, school-to-school transfers, and for the local planning of the educational program for the child.

The last quarter of the 20th century saw a number of reports and reviews that were highly critical of American education. These included the **First and Second International Math Study** (**FIMS & SIMS**) and the **Third International Math and Science Study** (**TIMSS**; U.S. Department of Education, National Center for Educational Statistics, 2004), as well as the widely read report, *A Nation at Risk: The Imperative for Educational Reform* (National Commission on Excellence in Education,1983). The media coverage and political fallout from these and other reviews of American

education created an added impetus for statewide assessment programs. They also served to provide the background arguments for the **Improving America's Schools Act of 1994**. Much of the testing-based reform movement was started in 1994 under the program initiated under President Clinton known as the Improving America's Schools Act (IASA; P.L. 103-382, 1994). This act was the first to require that each state develop learning standards and monitor schools by using tests of those standards. Prior to that time, several states had taken the lead in implementing statewide testing programs. The states at the forefront of this movement in the 1970s and 1980s were Florida, New Jersey, New York, Pennsylvania, and Texas. In states like New Jersey and Texas, the business community was the force behind instituting testing-based educational reforms.

Miracle of Texas

One example of this process can be seen in the home state of President George W. Bush. Texas has been a leading state in the educational assessment movement since 1983. That year, Governor Mark White asked fellow Texan and business leader H. Ross Perot to chair a select committee of business leaders to identify ways to improve education in Texas. The Perot Commission reported ideas for a number of major revisions to public education, which were quickly passed into law and were well funded by the Texas legislature in 1984. Part of these reforms was a call for standardized achievement tests. This resulted in the publication of the Texas Educational Assessment of Minimal Skills (TEAMS) in 1985 (Haney, 2000).

This was replaced with the Texas Assessment of Academic Skills (TAAS) in 1990. The TAAS was used to document what became known as the "**miracle of Texas**" and served as the template for much of the No Child Left Behind Act of 2002 (Haney, 2000). The miracle of Texas was the name given to what was perceived to be a significant improvement in the number of children who scored at the proficient level on the TAAS. The Texas Education Department required many of the same things that are now a part of the NCLB Act. These include public reports of school scores, penalties for schools where children do poorly, and a required graduation test.

Recently there has been a reexamination of these outcomes, and much skepticism has been expressed in the educational research literature (Haney, 2000; Kellow & Willson, 2001). Criticism of this assessment, and of the reported success of the Texas model, led to the development of a new assessment, the Texas Assessment of Knowledge and Skills (TAKS), in 2003.

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

Case in Point (1f)

The largest city in Texas is Houston. The superintendent of Houston's schools in the 1990s was Rod Paige (later U.S. Secretary of Education 2001–2005). He reported that the assessment test scores in his district had soared and that dropouts were an issue of the past. All of this miraculous good news was attributed to the "tough love" of the new Texas reforms. When Governor Bush became president, those reforms became the core of the federal legislation known as No Child Left Behind. An investigation by the TV news show *60 Minutes* (Fager, 2004) demonstrated how students who were being given their exit interview on deciding to drop out of school were asked about their future plans, including their education plans. If the students said that they planned to "get a diploma someday" through the GED, they were not counted as being a dropout. Thus, virtually no student was classified as a school dropout.

The average school test scores were also found to be manipulated. Students who were at risk for failure were retained in a lower grade before they had to face the test. Once they repeated that grade once or twice, they were double promoted over the grade level where the test was required. The result was that Houston's high schools had bulging enrollments in grade 9 and anemic enrollments in grade 10, the grade where the high-stakes test was given.

One reason for this dubious policy was that Dr. Rod Paige only gave his high school principals one year contracts with reappointment being contingent on the school's test scores and dropout rate.

CASE LAW

Special Education

Case law has focused on three central issues related to testing. In general, the courts have followed a practical approach to determining the **appropriateness** (**validity**) of a test, focusing on the consequences of the measurement and the content that was included in the test (Sireci & Parker, 2007). The first of these cases involved the use of tests and assessments in placing children into special education programs. The central issue in these cases had been related to whether there was some form of measurement bias against minorities on the standardized measures. Two old cases demonstrate this issue. In a series of state and federal court cases between 1972

and 1984 referred to as *Larry P. v. Riles*, the schools of the City of San Francisco were prevented from using standard **intelligence tests** to place African American children in special education classes. This was based on the observation that all racial sub-groups within the population do not have the same profile of test scores. Also, it was argued that the professional community was not sensitive to group differences when making placement decisions. The later decisions (1984) upheld the lower court ruling and eventually enjoined all school systems in California from using any of 20 different measures of mental ability when making placement decisions for African American children.

A few years later, and a half a continent away, the case of *Parents in Action on Special Education (PASE) v. Hannon* resulted in the opposite outcome for special education testing. Here the use of standard intelligence measures was found to be without bias and was permitted for all placement decisions. The difference between the two outcomes is that in the latter case the judge read and evaluated each item on the measures to determine if he could see any obvious bias. Also, these intelligence measures were not used as the sole criteria for a special education placement but were a part of a larger, multidimensional assessment of the child.

High-Stakes Tests

Another area where case law has had an impact on testing is with high-stakes tests in the public schools. Florida was among the first states to require students to pass a graduation test before they could be awarded a high school diploma. This was challenged in 1978 when 10 African American students from Hillsborough County, Florida, who failed their competency tests, sued the state for being denied a high school diploma. The plaintiffs argued that the disproportional number of minority students who had been denied a diploma was a violation of the 14th Amendment of the U.S. Constitution (see Table 1.2).

The resolution of the case, known as *Debra P. v. Turlington*, happened in 1981. The courts ruled in favor of the state of Florida. This ruling came after the court noted the fact that Florida had aligned the test items with the curricula taught in the schools, and that all students had several opportunities to learn what was required to pass the assessment and earn a diploma. Florida awarded those students who failed the assessment a "Certificate of Completion," which allowed them to enroll in adult education through which they could work toward their diplomas.

The issue of denying a diploma to students with disabilities who cannot pass a state-mandated graduation test was resolved shortly after the Florida decision. The parents of a child with disabilities in Illinois who was denied a diploma after failing a graduation test sued and lost in the case of *Brookhart v. Illinois State Board of Education* (1983). In this case the courts expressed the opinion that a school district's desire to "ensure the value of a high school diploma" is admirable, and that the courts should avoid interfering in educational policy unless a constitutional or statutory right of the child has been clearly violated.

Table 1.2 Pass Rates on Exit Exams

Percentages of Students Passing State Exit Exams on the First Attempt						
	English					
States	*Reading*	*ELA*	*Writing*	*Math*	*Science*	*Social Studies*
Alabama	88%	86%		83%	82%	
Alaska	66%		47%	44%		
Arizona	67%		68%	31%		
California		64%		44%		
Florida	58%			72%		
Georgia		94%	92%	91%	68%	80%
Indiana		68%		65%		
Louisiana		78%		65%		
Massachusetts	82%			75%		
Minnesota		80%	91%	75%		
New Mexico	92%	82%	95%	82%	80%	79%
South Carolina	85%		86%	81%		
Tennessee				76%	95%	
Virginia	82%	82%	84%			

Percentages of Students Passing on the First Attempt by Subgroups for Three States						
Student Groups	*Indiana Math*	*Indiana English/LA*	*Minnesota Math*	*Minnesota English/LA*	*Massachusetts Math*	*Massachusetts English/LA*
All students	65%	68%	75%	80%	75%	82%
Asian	79%	72%	62%	61%	84%	79%
Black	31%	38%	33%	46%	46%	59%
Hispanic	46%	49%	43%	52%	41%	51%
White	70%	73%	80%	86%	82%	87%
Free/ reduced lunch	42%	45%	51%	59%	Not Available	Not Available
Students with disabilities	24%	19%	33%	40%	39%	46%
English language learners	33%	28%	32%	30%	42%	39%

SOURCE: From Center on Education Policy; based on information collected from state departments of education. Copyright 2002 by Center on Education Policy. Reprinted with permission.

Advanced Program Admission

Another area in which case law is guiding the use of educational tests is in the admission of students into advanced programs. The question is one of the selection of children for gifted programs. An example is that of the three selective secondary schools for the gifted in Boston. In Boston, the use of a test score to place students into these programs for the academically talented resulted in racial and ethnic disparities in enrollments. To correct the imbalance, the school admission policy was modified to include racial set-asides. Differential admissions programs for minority groups at these schools have come under court review, and the three schools have all been ordered to

stop racial set-asides in their admissions (*Wessmann v. Gittens*, 1998). In 2005, it was noted that the end of the set-aside policy in Boston has resulted in a skewing of the enrollment in programs for the gifted toward White and Asian students and away from African American students (Sacchetti, 2005).

Higher Education

Most of the case law issues regarding admissions have taken place with colleges and graduate schools. Most higher education institutions have the admissions goal of creating a diverse population of undergraduate students. This tenet is established to assure that different communities and ethnicities are part of the culture of the institution. The assumption is that living in such a culture becomes an important part of the learning experiences of undergraduate students. It should also be noted that not all higher education institutions seek diversity in all dimensions of the student body. For example, almost all the students attending Bob Jones University (evangelical Protestant) and Ave Marie University (Roman Catholic) share a common campuswide doctrine and religious faith. Alverno College only accepts women, while Wabash College is an all-male institution. Historically there are a number of colleges that were originally established to educate African American students. Most of these institutions of higher education are still primarily attended by African American students.

The easiest solution for the admissions office to promote student diversity is an open-door approach. The problem occurs for those institutions that maintain a selective admissions policy while still seeking student diversity. This can be a daunting challenge.

Early Case Law on Admissions

Over the past 50 years the legal battles over enrollment in America's schools and colleges have gone through a 180-degree turn. In the 1950s and 1960s the fight led by the attorney general and the Supreme Court under Chief Justice Earl Warren was to desegregate unwilling school systems and colleges. Most of these were located in the South and in the antebellum border states. The landmark Supreme Court decision was *Brown v. Board of Education* (347 U.S. 483 [1954]). This decision did away with the concept of "separate but equal" in all matters of public accommodation, including education.[7] During the 1960s, Attorney General Robert Kennedy brought desegregation lawsuits against hundreds of school systems and the state departments of education in a dozen states.

A major shift in the zeitgeist occurred during the late 1960s, and by the 1970s many school systems and most colleges were working to remove all vestige of segregation from their student populations. During that era, colleges and professional schools began to take positive (affirmative) steps to increase minority enrollment.[8]

Score Gap

The persistence of significant differences between the average scores of different racial groups on all high-stakes and admissions tests is a vexing and longstanding problem for educators (see Table 1.3). In an earlier era this differential would have been explained as being a function of inherited differences in ability. Today, a number of observers are willing to explain the difference in

Table 1.3 Tables From ACT Scores by Gender and Ethnicity

National Average ACT Composite Score by Gender, 1994–2006													
	1994	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*
Males	20.9	21.0	21.0	21.1	21.2	21.1	21.2	21.1	20.9	21.0	21.0	21.1	21.2
Females	20.7	20.7	20.8	20.6	20.9	20.9	20.9	20.9	20.7	20.8	20.9	20.8	21.0

National Average ACT Composite Scores by Race/Ethnicity, 5-Year Trends					
	2002	*2003*	*2004*	*2005*	*2006*
All Students	20.8	20.8	20.9	20.9	21.1
African American/Black	16.8	16.9	17.1	17.0	17.1
American Indian/Alaskan Native	18.6	18.7	18.8	18.7	18.8
Caucasian American/White	21.7	21.7	21.8	21.9	22.0
Hispanic	18.4	18.5	18.5	18.6	18.6
Asian American/Pacific Islander	21.6	21.8	21.9	22.1	22.3
Other/No Response	20.3	20.6	20.9	20.9	21.1

SOURCE: From www.act.org. Copyright © 2007. Reprinted with permission from ACT, Inc.

achievement between groups in terms of group attitudes and motivation (Thernstrom & Thernstrom, 2003) (see Table 1.4).

Confounding this problem of differential average scores on achievement tests is the fact that there is a direct, monotonic relationship between the income level of families and the admissions test scores of children in those families. Likewise, there is a direct relationship between the quality of the high school that students attend and the admissions test scores students earn. In part, this reflects the availability of a competitive academic environment in the high schools of high-scoring students (Bridgeman & Wendler,

Table 1.4 Tables From SAT Scores by Gender and Ethnicity

2006 COLLEGE-BOUND SENIORS TEST SCORES: SAT **Approximately 1.47 million test takers, of whom 53% were female**				
	Verbal	*Math*	*Writing*	*Total*
Gender				
Female	502	502	502	1506
Male	505	536	491	1532
Ethnicity				
American Indian or Alaskan Native	487	494	474	1455
Asian, Asian American, or Pacific Islanders	510	578	512	1600
African American or Black	434	429	428	1291
Mexican or Mexican American	454	465	452	1371
Puerto Rican	459	456	448	1363
Other Hispanic or Latino	458	463	450	1371
White	527	536	519	1582
Other	494	513	493	1500
No Response (5%)	487	506	482	1475

SOURCE: *2006 College-Bound Seniors*. Copyright © 2006 the College Board, www.collegeboard.com. Reproduced with permission.

2004). The complexity and rigor of high school curriculums followed by students have been shown to predict how well those students do in undergraduate college (Glickman & Babyak, 2006).

Parental Factors

Middle-class families are likely to pay the required tuition for their children to attend a test-preparation course and to hire academic and test-preparation tutors (Chiles, 1997). It is clear that family income and parent education level have a lot to do with student achievement (Joireman & Abbott, 2004) (see Table 1.5).

Table 1.5 Average SAT II Reasoning Scores by Family Income Group

Family Income	*Reading*	*Math*	*Writing*	*Total*
Less than $10,000/year	429	457	427	1313
$10,000–$20,000/year	445	465	440	1350
$20,000–$30,000/year	462	474	454	1390
$30,000–$40,000/year	478	488	470	1436
$40,000–$50,000/year	493	501	483	1477
$50,000–$60,000/year	500	509	490	1499
$60,000–$70,000/year	505	515	496	1516
$70,000–$80,000/year	511	521	502	1534
$80,000–$100,000/year	523	534	514	1571
More than $100,000/year	549	564	543	1656
No Response (35%)	Scores not reported			
ALL TEST TAKERS	503	518	497	1518

SOURCE: *2006 College-Bound Seniors*. Copyright © 2006 the College Board, www.collegeboard.com. Reproduced with permission.

These factors are linked to parental behaviors such as the enforcement of homework time and the setting of limits on television watching. Home factors are central to the academic success of all students. In a recent study of the students in several middle schools in Pennsylvania, it was found that 15% of the **variance** in success on the mandated statewide assessment was accounted for by a vector of **variables** that measure home life. These variables included parent education level, parent expectations for the child's education, hours spent reading for pleasure, the number of books and magazines at home, family **mobility**, frequency of absenteeism by the child, and the number of hours the child spent watching TV (Holbrook & Wright, 2004). Howard Gardner once quipped that we can predict with surprising accuracy whether a youngster will eventually graduate from college by only knowing his/her zip code (personal communication, Howard Gardner, April 2007).

The **score gap** on admissions tests has been examined by a number of authors, including Claude Steele. He presents evidence that there may be at least two extraneous factors in the SAT scores of African American and Latino children. One source is produced by the "**stereotype threat**" created by the test situation (Steele, 1997, 1999). Steele has demonstrated that "stereotype threat" occurs when minority students are placed in a high-stakes test situation. His research has shown that in those situations minority students feel the added stress of the stereotypical expectations held for them. The extra pressure caused by the fear of proving the stereotype correct correlates with lower performance.

The second problematic area is related to the linguistic aspects of the test items (Freedle & Kostin, 1997). Roy Freedle (2002) has identified differences in word utilization patterns between White and Black adolescents. This has the potential of producing differential test item performance (Dorans & Zeller, 2004). The unexpected and unexplained finding is that African American test takers perform better on the hardest items and less well on the easiest items. This is the opposite of what occurs with a population of middle-class White test takers. If the harder items were to be given extra weight on the test, the score gap would be reduced by a third. Also, Freedle points out the **correction for guessing** that ETS applies to all the SAT II scores disproportionately lowers the scores of minority students who make more errors on the easier items of the test.

Standard formula for the correction for guessing on a test composed of multiple choice questions that each have four answer options,

$$\text{total corrected raw score} = \frac{\textit{total correct} - \textit{total wrong}}{4}$$

There is also reason to believe that tests involving essay writing may also present a problem. Essay tasks that provide background information about the topic for the writing assignment tend to be easier to write than essay tasks that provide no structure or guidelines. The performance of African American adolescents is best when the essay format is the hardest (i.e., no structure).

The move toward the use of **computer adaptive testing** may prove to lead to an even wider score gap between White and Black students (see Chapter 2). This reflects the way computer adaptive systems are designed to first attempt to determine an approximate ability level for the test taker (Cheng & Chang, 2007). This is done by presenting several items of a mid-level of difficulty at the start of the test. Based on the student's performance on those items, the test taker is then presented with items assumed to match his/her ability. If Freedle's model is correct, it is possible that African American students could never be presented with the more difficult level of test items and thereby have test scores based on easier questions. The result will be low scores that are assigned to those answering easier questions.

Affirmative Action in Admissions

It was a natural step for highly selective institutions to adopt an **affirmative action** model for their admissions processes. One reason for this decision is that the pool of available minority students who have high SAT scores is very small. Research sponsored by ETS has shown that only 3.3% of African American test takers scored over 1300 on the SAT I, while 9% of Hispanic students and 39% of Anglo-White students scored at that level (Bridgeman & Wendler, 2004).

Case in Point (1g)

Admission test scores are one of only two primary predictors of college potential. The other is high school **class rank**, which is derived from the high school **grade point average** (**GPA**) of students. The importance of tests like the SAT II in the admission process is increasing because high schools have begun not reporting either a GPA or class rank (Finder, 2006). This reflects an effort by a number of suburban high schools to remove unhealthy levels of letter-grade stress on students. The movement began with private schools that were trying to provide a way for the good students who attend their highly competitive schools to be admitted into the most elite colleges (Zweigenhaft, 1993).

Those college admission officers who evaluated the entering class of 2009 reported that half or more of the applicants' transcripts did not report the class rank or the high school GPA. The result was that SAT II and ACT scores were more important in the admissions decision than ever.

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

Affirmative action plans made it possible for both selective public schools[9] as well as institutions of higher education to set minority enrollment targets and actively work to achieve them. Such strategies took the form of simple quotas or in some cases the addition of bonus points to the admissions file of minority students.

Recently reported research into the admissions process at highly selective institutions demonstrates that several classes of students were given preference in admission.[10] The advantage given to African Americans is the equivalent of 230 extra points on the 1600-point (SAT I) scale being added to their test score. The advantage to Hispanic Americans has been worth 185 extra points (Espenshade & Chung, 2005). The advantage to Asian students has been negative, equaling a loss equivalent to 50 points.

Naturally, when higher scoring students of one racial group were rejected for admissions and lower scoring students from a minority were accepted, tension ensued. It had to be anticipated that these affirmative admissions systems would be criticized and come under legal challenge.[11]

Court Challenges

When an otherwise well-prepared non-minority student receives a letter of rejection from a college or professional school, the personal pain can be excruciating (Kinzie, 2007). The rejected student may well blame the admissions process or others who are perceived as receiving special treatment.

The first challenge to affirmative action in admissions was a case involving admission into the Medical School of the University of California at Davis. In this case, a White male applicant was rejected for admission in 1974 and sued the regents of the university on the basis of Title VI of the Civil Rights Act of 1964 and the **equal protection** clause of the 14th Amendment of the Constitution of the United States (***University of California Regents v. Bakke***, 438 U. S. 265 [1978]). The university had established a special admissions program for the economically disadvantaged and members of targeted minority groups, including African Americans, Chicanos (Mexican Americans), Asians, and Native Americans. This special admissions process used a set-aside program that guaranteed that there would be 16 seats available for the targeted groups. This approach to admissions was rejected, but a separate opinion written by Justice Powell, and affirmed by the Court, did recognize student diversity as a compelling state interest.

The second major blow to affirmative action in admissions came from a case in Texas. In that case, ***Hopwood et al. v. The State of Texas*** (1994), the issue before the courts involved admission into the Law School of the

University of Texas, Austin. Among the many Anglo-White students who were rejected for admission into the Law School were four who sued. The Law School had previously created a second parallel admissions process reserved for African American and Mexican American applicants. These targeted minority groups were subjected to a less selective admissions test score requirement than were Asian and Anglo-White students, including the four who brought the lawsuit. Based on their scores on the Law School Admission Test (LSAT) and undergraduate grades (GPA), the four who sued all would have been admitted if they were members of either of the targeted minority groups. At the first level, the state courts agreed with the university and its admissions system, rejecting the Hopwood petition. On appeal, the U.S. Fifth Circuit Court overturned that decision and effectively ended that affirmative admissions practice (*Hopwood et al v. University of Texas*, 861 F. Supp. 551, 578–579 [WD. Tex. 1994] 5th Circuit [1996]).

The final defining cases for affirmative action in admissions involved two of the colleges of the University of Michigan in 2003. The first, ***Gratz et al. v. Bollinger et al.***, involved admission of two undergraduates into the College of Literature, Science, and the Arts. The university used a point system to assist in the admission decision process. Admission was granted to all students who reached a total of 100 points using a system that included a **weighted combination** of test scores, GPA, and **Advanced Placement** (**AP**) course completion. All targeted minority group members (African American, Hispanics, and Native Americans) were automatically awarded a bonus of 20 points on their admissions files. The Supreme Court held for the plaintiffs and rejected the admissions model employed for undergraduates at the University of Michigan (*Gratz v. Bollinger* [02-516]. U.S. [2003]). The court rejected the argument that Michigan's admissions office needed a simplified point system because the number of applications was too high to give individual attention to each case.

A second case, ***Grutter v. Bollinger et al.***, at the University of Michigan has provided guidance for all admissions systems that strive to be both highly selective and also enroll a significant number of minorities. The university's law school employs a **holistic approach to admissions**, which includes LSAT scores and undergraduate GPA. It also looks at "soft variables" including enthusiasm, recommendations, the quality of the applicant's essay, life experiences, and the difficulty of the undergraduate course of study. The stated goal of this admissions system is to select motivated and able students with the best potential to contribute to the practice of law in Michigan.

Barbara Grutter had excellent grades and a high LSAT score, but she was not admitted. She argued that the admissions process was biased in favor of minority students. She won her case in the district court but had it reversed by the U.S. Sixth Circuit Court. On appeal, that decision was upheld by the

U.S. Supreme Court (*Grutter v. Bollinger et al.,* 02-241. U.S. [2003] 288 F.3d 732, affirmed). The opinion of the Supreme Court was that the admissions system of the law school met the requirements as specified in the opinion of Justice Powell from the Bakke case of 1978.[12] The justices recognized that there is a compelling reason for an agency of the state (University of Michigan) to consider the race of potential students. Thus, it is possible to include race as a variable in admissions decisions when those decisions are made following a holistic approach to student selection.[13]

The compromise implied by the Bollinger cases not withstanding, in 2007 the regents of the University of Wisconsin voted to approve a policy requiring that the race of applicants be considered as one salient factor in the admissions process. In addition, a special scholarship program available only to students who are members of one of three different ethnic minority groups was continued for undergraduate students of the university (Schmidt, 2007). When UCLA initiated a holistic system in 2006 for use in the admission of undergraduate students, that institution was able to report an increase in the proportion of African American students of approximately 40% (Schmidt, 2007).

EDUCATIONAL ETHICS AND TESTING

Professional associations hold themselves to be keepers of the best traditions and practices of their fields and assume that the general public will have confidence in their work and respect for their members. To assure this continuing regard of the public, the various professional associations publish guidelines for the ethical behavior and practice of their members. In the field of educational testing, the primary associations with an interest in the issues of educational measurement have published a single document on ethics. This document is the combined effort of the American Counseling Association (ACA), the American Educational Research Association (AERA), the American Psychological Association (APA), the American Speech–Language–Hearing Association (ASHA), the National Association of School Psychologists (NASP), the National Association of Test Directors (NATD), and the National Counsel on Measurement in Education (NCME).

The following are four principles that are drawn from that document (Joint Committee on Testing Practices, 2005).[14]

1. The first of these principles for the ethical practice of testing involves communication with those taking the test. The purpose of the test and the areas that are to be measured should be fully understood by the test taker prior to the time of the test. The use of scores from the

test should be explained and the test takers should be told how long their results will be kept on file. This communication includes providing test takers with practice on similar materials to familiarize them with the mechanics of the measure. Also, the test administrator should be aware of the need for special **accommodations** that test takers may require prior to the time of the administration of the test.

2. The second area involves confidentiality. It is necessary for the test administrator to put into place procedures that ensure that the scores from individual students are never disclosed to people not having professional need for those data. The students' parents are included in the group who should have access to the test score data.

   At another level, confidentiality involves the test itself. It is critically important that test materials be stored in a secure location and never released for review by interested others.

3. The third point is that the interpretation of the test scores should be carried out in a way consistent with the guidelines provided by the test developer and publisher. This also implies that the person interpreting the scores should be trained in the process and knowledgeable of the test and its scores. Parents and students should be informed of the scores and their interpretation in a developmentally appropriate way, using understandable language. Educators who discuss score reports with parents and students should avoid educational jargon and provide clear descriptions to the parents and students. This includes the process used by the various agencies in setting cut scores and minimal standards for success. If there is a scoring error it should be corrected immediately and that correction carried through on all of the student's records.

4. Finally, a single score on a test should never be used to determine the placement of a student. Interpretations should always be made in conjunction with other sources of information.

A last point involves the development and selection of tests. A test or assessment should never be used for a purpose for which it was not designed and has not been standardized. The test should provide a manual documenting that the measure is valid and reliable for the tasks it is designed to accomplish. Also, the measure should provide evidence that there is no consistent gender bias or ethnic or racial group bias represented in the scores. The test should provide users with detailed directions for the test administration as well as for those who score the test.

Summary

The people of this country are investing enormous amounts of their resources in the endeavor of public education. Therefore, it is not a surprise that the public has expressed a need for accountability in education. To have accountability there must be regular assessments of educational outcomes. The most common vehicle for those assessments is through the use of batteries of paper-and-pencil format achievement tests. This accountability system has been ratcheted up under the provisions of the No Child Left Behind Act of 2002. That legislation was designed to close gaps in the average levels of achievement between groups of students. This act also furthered the use of high-stakes tests in the assessment process and required that schools make adequate yearly progress toward the goal of having all students achieving at a proficient level. These assessments are established by each state and are based on specific standards that each state has established. This legal position established by case law does not conform with the ethical principles established by the learned and professional societies that have a vested interest in educational testing programs.

Educational technology is changing the format and nature of educational assessments. Increasingly, assessment tests are administered online. The same technology is improving the communication between teachers and parents and making it increasingly possible for parents to be partners in their children's education.

Legislation and case law have provided guidance in the use of educational tests and assessments. Federal legislation ushered in a new era for children with disabilities with the passage of laws including the Individuals with Disabilities Educational Act (1997) and the Individuals with Disabilities Educational Improvement Act (2004). Also, the rights of parents to control the flow of testing data were established in the Family Educational Rights and Privacy Act of 1974. The federal courts have also shaped testing policy by approving the use of high-stakes assessments in making graduation and even grade promotion decisions. These high-stakes uses of tests and assessments are permitted even if the sanctions that are imposed are felt disproportionally by one or more of the minority groups of students.

Discussion Questions

1. How does the British experience over the 18th and 19th centuries with testing school students and awarding merit pay to teachers compare with the current reforms in American education?

2. Should parents be able to hold their children out of public education for an extra year ("academic redshirting") to give them a developmental advantage over their peers? Explain your position.

3. What arguments can be made for and against retaining children in third grade who do not achieve a score of proficient (pass) on the state-mandated high-stakes tests? What do you believe should be done with low-performing children on the state tests?

4. Should the 50 states be permitted to replace the use of standardized tests composed of multiple choice questions with a more open-ended form of assessment (i.e., portfolio assessments)? Explain your position on this issue.

5. Should the business community have the ability to initiate educational reforms in the public schools? What advantages and disadvantages are there to having business leaders and entrepreneurs set goals and create priorities for educational reform?

6. What specific procedures should a state department of education take to standardize how its many school districts count the number of students who have dropped out of school?

7. In the selection of students for special programs, and in college admission, should educators take specific steps to ensure a proportional ethnic/racial mix in the student body? If you feel that diversity is an appropriate goal to work toward, how can this be achieved within current case law as adjudicated?

8. Check with your local school system and ask to see a copy of its written guidelines for the maintenance and distribution of student test data. After reviewing that policy, determine if it seems to meet the requirements under the Buckley Amendment.

Student Study Site

Educational Assessment on the Web

Log on to the Web-based student study site at www.sagepub.com/wrightstudy for additional Web sources and study resources.

NOTES

1. The first sophisticated educational testing system emerged in Great Britain. The British system reflecting a new nationalism started in 1710, just three years after the Crowns of England and Scotland were united into one United Kingdom. That testing program assessed students against national educational standards in reading, writing, and arithmetic. By the 1860s, teachers throughout Great Britain were even paid on the basis of student test scores, a system that lasted for another 30 years (Troen & Boles, 2005). The result was the truncation of the curriculum, and instruction focused on only those three core areas included on the test, the famous three R's. It also brought about the demoralization of the teaching profession and widespread cheating and corruption.

2. A number of terms are used throughout this book and have both vernacular and technical meanings. One is *evaluation*, a process of judging or making a decision based on observations and/or measurements; another is *measurement*, a method or device used to assign a numerical value to a characteristic of people or objects; *assessment* is a broad term encompassing all the methods employed to gain information about a person or object, including tests and other measures; and finally **test**, a word describing on organized task or series of tasks employed to represent and demonstrate knowledge, a skill, or a trait of an individual.

3. Many of the Regents examinations from the past 50 years can be seen at www.nysl.nysed.gov/regentsexams.htm.

4. In 2006 at total of 22,873 public and charter schools failed to make Adequate Yearly Progress (AYP; Packer, 2006).

5. The states that require a score of proficient on the state's assessment to be promoted to the next grade include Delaware, Florida, Georgia, Louisiana, North Carolina, Texas, and Wisconsin.

6. Bowing to parent pressures, several states reconsidered their lack of educationally challenging programs for gifted students (Samuels, 2007).

7. This did away with the notorious case law set by the Supreme Court in 1896 known as *Plessy v. Ferguson,* which had permitted states and local governmental units to provide "separate but equal" facilities, schools, and programs for identified racial groups. This court decision legalized the policy of American apartheid.

8. A subsequent exception was in Virginia, where the Virginia Military Institute (V.M.I.), a public college, fought against the admission of women students (*United States v. Virginia et al.,* and *Virginia et al. v. the United States*, 94-1941 & 94-2107, § 64 U.S.L.W. 4638 [1996]). In August of 1997, following the decision of the U. S. Supreme Court (7–1), V.M.I. enrolled 30 freshman women.

9. There are a number of public schools that are designated for the most talented and/or mentally gifted students in the country. Many of these "admission by testing" and highly specialized schools are located in urban centers. They include

Lowell High School, San Francisco; Boston Latin High School; Eastern Sierra Academy, Bridgeport, California; Philadelphia's Central High School; Dallas School for the Talented and Gifted; Bronx High School for the Sciences; Buffalo's (NY) Honors High School; Bloomfield Hills, Michigan's, International Academy; and Alexander W. Dreyfoos School of the Arts, West Palm Beach.

10. There is no evidence that a degree from a highly selective or elite college is related to the level of success experienced by the graduate later in life. More critical are the factors of motivation, ability, and personal desire (Easterbrook, 2004).

11. The administration of President George W. Bush was at the forefront in the fight against any consideration of race or ethnicity in admissions decisions. Critics of the administration have raised the issue of the provisions of the No Child Left Behind Act, which require performance assessments that are tabulated and reported by race when those data are never permitted for use in the admission process (Gershberg & Hamilton, 2007).

12. Allan Bakke went to the Medical School at the University of California, Davis, and eventually became a respected anesthesiologist in Minnesota.

13. Following the Grutter decision, in 2006 the American Bar Association began requiring all law schools in the United States to demonstrate concrete steps toward developing student bodies, faculties, and staffs that are racially and ethnically diverse. In 2006, Jennifer Gratz, a plaintiff in the *Gratz v. Bollinger* case, led a statewide drive in Michigan to pass Proposition 2 outlawing all forms of affirmative action in that state. The measure passed on November 7, 2006. As written, the new law prevents any consideration of ethnicity in admission to any publically supported program, school, college, or professional and/or graduate school. This law stands in opposition to the admission policies required for accreditation of various professional programs (e.g., law).

14. The ethical principles for classroom-level testing by teachers are similar to those for large-scale assessment tests. Those issues are presented in Chapters 6 and 7, which focus on the use of high-stakes measures and other standardized measurements in the schools.