National Association for
College Admission Counseling

Guiding the way to higher education

College Admission Testing

February 2007

Dr. Rebecca Zwick
Professor at the Gevirtz Graduate School of Education
University of California, Santa Barbara

This report was commissioned by the National Association for College Admission Counseling as part of an ongoing effort to inform the association and the public about the state of college counseling in America's high schools.

The views and opinions expressed in this report are solely those of the author and not necessarily those of NACAC.

Introduction by NACAC

Over the past 10 years, standardized admission tests have become an increasingly important factor in undergraduate admission, as a burgeoning number of applications have initiated a more methodical approach to admission at an increasing number and variety of undergraduate institutions. At the same time, reforms in elementary and secondary education at both the state and federal level have vaulted standardized tests to previously unrivaled heights as a tool to measure educational outcomes. However, long-standing concerns with standardized tests have persisted, and the role of the ACT and SAT in determining who gains entry into the nation's colleges and universities continues to be a hotly debated topic. A growing number of postsecondary institutions have adopted "test-optional" admission policies, and recent scoring errors on the SAT have ignited stakeholder anxieties about the role of standardized tests in the decision to admit students to college.

In its 1999 *Myths and Tradeoffs* report, the National Research Council cast the debate over standardized tests in larger terms, suggesting that the use of standardized tests in admission raises important questions about the social goals that underlie the admission of students into institutions of higher education. In the American system of higher education, institutions exercise great autonomy in determining admission standards and in making admission decisions. Standardized tests are only one of the tools—albeit a frequently used and therefore important tool—at their disposal in making these decisions. Ultimately, each college is uniquely situated to resolve the debate over the fairness or usefulness of standardized tests for admission to its campus. Admission officers must therefore exercise due diligence in understanding how to properly interpret test scores. Colleges and universities must continue to conduct research that determines how or whether test scores, as well as other admission criteria, predict student performance at their institutions.

To help admission officers, counselors and other stakeholders better understand the role of standardized tests in undergraduate admission, NACAC commissioned a white paper by Dr. Rebecca Zwick, professor of education at the University of California, Santa Barbara. This paper is intended to provide a concise summary of the history of standardized tests, the role of testing in undergraduate admission and current research on the tests' effectiveness in providing meaningful data to admission offices about applicant qualifications for postsecondary study.

About the Author

Dr. Rebecca Zwick specializes in educational measurement and statistics, test validity and testing policy. She has been a professor at the Gevirtz Graduate School of Education at the University of California, Santa Barbara since 1996. She received her doctorate in Quantitative Methods in Education at the University of California, Berkeley and her M.S. in Statistics from Rutgers University. After a postdoctoral year at the L. L. Thurstone Psychometric Laboratory at the University of North Carolina at Chapel Hill, she was a statistical researcher at Educational Testing Service (ETS) in Princeton for 12 years. While

at ETS, she served on the technical staff of the National Assessment of Educational Progress (NAEP), including one year as director of data analysis. More recently, she served on NAEP's Technical Advisory Committee on Standard-Setting, and she currently serves on NAEP's Design and Analysis Committee. She is also a member of the College Board's Psychometric Panel for the SAT and PSAT, and is advisory editor for *Journal of Educational Measurement* and *Educational Measurement: Issues and Practice*. She is the author of more than 50 journal articles and books in educational measurement and statistics, and is the principal investigator for a three-year National Science Foundation project, Instructional Tools in Educational Measurement and Statistics (ITEMS) for School Personnel.

Key Questions

One important purpose of this white paper is to generate discussion about the role of standardized tests in undergraduate admission. Data from *NACAC's State of College Admission* report suggests that standardized tests have increased in importance over the past decade, coinciding with the record number of students and applications flowing through the admission system.

As this paper suggests, there are a number of key questions about standardized tests as admission factors that remain unresolved. As we introduce this paper into the national conversation about college admission, we urge readers to ponder these difficult questions.

- What does a test score tell an admission office about an applicant to college?
- What influence do test scores have in predicting whether students will succeed in college?
- How much research about the predictive validity of standardized tests is conducted independent of the agencies that sponsor the tests?
- Do institutions clearly articulate the reasons why standardized tests are included as requirements for admission?
- How do students and parents view standardized tests?
- How would the admission process differ if tests were not available?

College Admissions Testing[i]

By Rebecca Zwick

A Brief History of Admission Testing

Although standardized testing was used by the Chinese Imperial Civil Service at least two thousand years ago, university admission tests did not make their debut until centuries later. Admission testing apparently began in Europe, but there is some disagreement about the time and location of the first such test. According to Webber (1989, p. 37), most historians agree that "[t]esting for admissions to universities…did not begin in Europe until the eighteenth century." A report by the Congressional Office of Technology Assessment, however, alludes to a thirteenth-century Sorbonne entrance examination (Office of Technology Assessment, 1992), and a College Board publication suggests that university admission testing began in sixteenth-century Spain (Stewart, January 1998). Most accounts agree that admission testing had been instituted in Germany and England by the mid-1800s. In most countries, the use of tests to get out of universities preceded the use of tests to get in. In the early nineteenth century, it was still the case that anyone who could afford a university education could get into prestigious institutions, such as Oxford and Cambridge.

Standardized admission testing was first established in the U.S. in the early twentieth century. At that time, college applicants were faced with a bewildering array of entrance examinations that differed widely across schools. In an attempt to impose order on this chaos, the leaders of 12 top northeastern universities formed a new organization, the College Entrance Examination Board, in 1900. The College Board created a set of examinations that were administered by the member institutions and then shipped back to the Board for hand scoring. Initially, the Board developed essay tests in nine subject areas, including English, history, Greek and Latin; it later developed a new exam that contained mostly multiple-choice questions—the Scholastic Aptitude Test. This precursor to today's SAT was first administered in 1926 to about 8,000 candidates. The first SAT consisted of questions similar to those included in the Army Alpha tests, which had been developed by a team of psychologists for selecting and assigning military recruits in World War I. These Army Alpha Tests, in turn, were directly descended from IQ tests, which had made their first U.S. appearance in the early 1900s.

In World War II, as in World War I, tests played a role in screening individuals for military service and assigning them to jobs. World War II also fueled an expansion in the use of standardized testing by creating an urgent need for well-trained individuals who could be recruited into the military; this led to an increased emphasis on college study in the U.S. The passage of the GI Bill in 1944 sent thousands of returning veterans to college as well, boosting the popularity of the efficient multiple-choice SAT.

Another development that was to have a major impact on the testing enterprise was the machine scoring of tests. Beginning in 1939, scoring the SAT, a task that had once

required many hours of training and tedious clerical work, was done automatically. This change effectively transformed testing from an academic venture to a bona fide industry, setting the stage for the establishment of Educational Testing Service (ETS). ETS was founded in Princeton, New Jersey, in 1947 through the merger of the testing activities of three companies: The College Entrance Examination Board, the Carnegie Foundation for the Advancement of Teaching and the American Council on Education. (All three continue to exist as separate organizations.)

In 1959, ETS gained a competitor in the college admissions test market. The American College Testing Program was founded in Iowa City "with no equipment and not even one full-time employee," according to the organization's own description.[ii] (Today, the company is ACT, Inc., and the test is simply the "ACT." Like SAT, ACT is no longer considered an acronym.) ACT, Inc. was founded by E. F. Lindquist, a University of Iowa statistician and a man of many talents. Lindquist was the director of the Iowa Testing Programs, which instituted the first major statewide testing effort for high school students. Remarkably, he was also the inventor, with Phillip Rulon of Harvard, of the "Iowa scoring machine," the first device to use electronic scanning techniques (rather than simply a mechanical approach) to score test answer sheets. The founding of ACT, Inc. was, in fact, closely tied to the development of this scoring machine, "a marvel of blinking panels … that could … emit a record of achievement from the brief encounter of small black marks on paper and the photocells in a reading head" (Peterson, 1983, pp. 111, 114). In 1953, Lindquist formed the not-for-profit Measurement Research Corporation, which was to continue the development of test processing systems and offer services to other testing programs. ACT, Inc., in turn, was a spin-off of the MRC and the Iowa Testing Programs (Peterson, 1983, p. 164).

The Early Promoters of Admissions Tests in the U.S.

Some early proponents of standardized testing attempted to use test results to bolster their racist beliefs. This is particularly true of Carl Brigham, an early College Board advisor who has been called "the father of the SAT." Although he later modified his views, Brigham concluded, based on an analysis of test results from World War I Army recruits, that immigrants were less intelligent than native-born Americans, and that Americans of Nordic heritage were superior in intelligence to those of Alpine or Mediterranean heritage. He also warned that American intelligence was expected to deteriorate rapidly if action was not taken to halt "the importation of the negro" (Brigham, 1923, p. xxi).

Many opponents of standardized testing believe that today's SAT remains tainted by the views of Brigham and his ilk. Yet some early champions of college admissions tests were, in fact, staunch supporters of equal opportunity. The prime example is James Bryant Conant, the Harvard president who in the late 1930s promoted the idea that the leading US testing agencies be merged into a single centralized company, and who eventually served as the first chairman of the ETS Board of Trustees. In a series of *Atlantic Monthly* articles in the early 1940s, Conant cautioned against the development of a caste system

in America and argued for a fluid society where people's roles would be determined by their merit (Lemann, August 1995). Conant continued to be an advocate of educational reform throughout his career, deploring the existence of segregated schools and the "evil influence" of racial prejudice (Conant, 1964).

The sharp distinction between the bigotry of Carl Brigham's early writings and the egalitarian stance of James B. Conant is reflected in the dual perceptions of the role of admissions testing which exist today: To some, admissions tests are harsh and capricious gatekeepers that bar the road to advancement; to others, they are gateways to opportunity.

Admissions Tests in Use in the United States

The SAT and ACT are the two primary college admission tests used in the U.S.; these are discussed in detail below.[iii]

The SAT

The SAT testing program is sponsored by the College Board; the tests are administered by ETS under a contract with the Board. The SAT Reasoning Test is claimed to measure "developed" critical thinking and reasoning skills needed for success in college.[iv] Until recently, the SAT provided math and verbal scores; as described below, it now provides scores in math, critical reading and writing. Three hours and 45 minutes are allotted for students to complete the SAT Reasoning Test. When they register for the SAT, students can choose to complete the SAT Questionnaire, which asks about demographic background, course preparation, interests and plans. This information (with the exception of some items that are deemed confidential) is then passed on to the colleges to which students send their scores.

In addition to the SAT Reasoning Test, the current SAT program also includes the SAT Subject Tests, which assess the candidates' knowledge in particular areas. Twenty SAT Subject Tests are available, in literature, U.S. and world history, math, biology, chemistry, physics, and foreign languages. (When the new SAT Writing Test was unveiled in 2005, the SAT Subject Test in writing was eliminated.)

The SAT has changed substantially since it was first administered in 1926, complete with instructions indicating that pencil is preferable to fountain pen for responding to the test. The 1926 test, for example, included a set of deductive reasoning items, as well as a set of items requiring test-takers to translate sentences to and from an artificial language whose vocabulary and rules were provided. A history of the changes in the content of the SAT is provided by Lawrence, Rigol, Van Essen, and Jackson (2004).

Two key events in the history of the SAT occurred in the mid-1990s. In 1994, major changes in content and procedures were implemented. Math items that required students to compute the answer rather than merely select it from several alternatives were introduced. Also, an earlier prohibition on the use of calculators was lifted. In addition,

antonym items were eliminated from the verbal section, reading comprehension items were made more complex, and sentence completion items were added. (An early plan to include an essay section in the SAT was dropped, but a writing test that included an essay component was incorporated in the SAT Subject Tests.) Beginning in 1995, scores for the mathematics and verbal sections were reported on scales that had been "recentered" so that a score of 500 would represent an average score for each section, as in the original SAT. The SAT Subject Tests were also rescaled at this time. Post-recentering scores are not comparable to pre-recentering scores without adjustment; equivalence scales have been created to facilitate such comparisons.

Before the recentering, SAT scores were reported on a scale established in 1941. At that time, it was decided that, on each section (verbal and math), scores would range from 200 to 800, with an average score of 500—the midpoint. Or, to put it another way, the average score of the 1941 test-takers was arbitrarily labeled "500." (The standard deviation was set to 100.) Then, through the process of test equating, subsequent versions of the test were linked to the 1941 version. But over the years, the scores "drifted" downward and lost their intended meaning. (See Turnbull, 1985, for a discussion of the reasons for the SAT score decline.) A score of 500 was no longer the average on either section, and the math and verbal averages were no longer the same. By 1993, the math average was 478, while the verbal average was 424. The recentering was, in essence, an adjustment procedure that assigned a label of 500 on the "new" SAT scale to the average score obtained by a special sample of about one million 1990 high school seniors (see Dorans, 2002). This adjustment made it possible, once again, to interpret individual test scores relative to a mean of 500. The recentering did not change the percentile rank of an individual's score. If a student scored at the 65th percentile (better than 65 percent of students) on the old scale, he would be at the 65th percentile on the recentered scale, but scores assigned to him would be different on the two scales.

In 2005, the SAT changed once again, reflecting modifications agreed upon following a nationwide controversy about the SAT that came to a head in 2001, with a speech by Richard C. Atkinson, then the president of the University of California. Atkinson recommended the elimination of the SAT Reasoning Test as a criterion for admission to the University and advocated an immediate switch to college admissions tests that were tied closely to the high school curriculum. In 2002, after months of discussion with UC representatives, the College Board Trustees approved several significant changes to the SAT. The new SAT, which made its debut in March 2005, substitutes short reading items for the verbal analogy items that were formerly part of the verbal section, incorporates more advanced math content, eliminates "quantitative comparison" items, and adds a writing section. All the critical reading questions and most of the math questions are multiple-choice. Each SAT also includes some math questions that require "student-produced" answers—there are no response choices. The newly added writing section includes both multiple-choice questions and an essay. (Essay and multiple-choice subscores for writing are provided, along with an overall writing score.)

Field trials of the new SAT were conducted in 2003, based on more than 45,000 students at 680 high schools. According to the College Board, changes to the verbal and math sections "will not affect the difficulty or reliability of the test" and will not exacerbate score disparities among ethnic or gender groups. Furthermore, the Board stated that math and critical reading score scales on the new SAT can be considered equivalent to the previously existing math and verbal score scales, so that "longitudinal data will be maintained." Based on a smaller study of the new test, the College Board also expected the addition of the new writing section to enhance the predictive validity of the SAT ("The new SAT 2005," 2004). Results of a validation study of a prototype version of the SAT writing test are given in the section, "Prediction of college grades."

The ACT

In 1959, when the ACT program began, the SAT was already well-established. Why start a new college admissions testing program? In Iowa testing circles, the SAT was considered to be geared toward the elite institutions of the east, and its developers were viewed as sluggish and resistant to change. From the beginning, the ACT was somewhat different from the SAT in terms of underlying philosophy: While the SAT consisted only of verbal and mathematical sections, the ACT was more closely tied to instructional objectives. The original version of the ACT had four sections—English, mathematics, social studies reading, and natural sciences reading. It is no coincidence that these subject areas were also included in the Iowa Tests of Educational Development, which had been used to assess Iowa high schoolers since 1942. In fact, because of scheduling constraints, the first form of the ACT assessment was constructed from the same pool of test items that was being used to assemble new forms of the ITED. In its early years, the ACT was administered primarily in midwestern states, but it is now used nationwide.

The content of the modern-day ACT is based on an analysis of the material that is taught in grades seven through 12. The test specifications and items are developed from information obtained from regular surveys of secondary school teachers and curriculum experts which ask about the major themes being taught in the ACT subject areas. All questions in these subject areas are multiple-choice. Slightly more than four hours is allotted for students to complete the ACT (excluding the writing test).

In 1989, major changes in the test content were implemented and the current four subject areas were introduced: English, mathematics, reading, and science reasoning (now renamed "science"). At the same time, the scoring of the test was changed. Scores on this "enhanced ACT" cannot be compared to scores on the original ACT without adjustment. Students receive a score in each subject area, as well as a composite score. Seven subscores are also reported—two in English, three in mathematics and two in reading.

In 2002, after the College Board announced that a writing component would be added to the SAT, ACT, Inc. announced that it would add a writing test to the ACT. Unlike the SAT writing section, however, the ACT writing test, first administered in 2005, is optional.

Students who elect to take it along with the ACT receive two additional scores: a writing test score and a combined English/writing score.

As well as being more strongly linked to instructional goals than the SAT, the ACT also places a greater emphasis on facilitating course placement and academic planning. In keeping with this goal, the ACT registration booklet includes a questionnaire on high school courses and grades, educational and career aspirations, extracurricular activities, and educational needs, as well as a career interest inventory (UNIACT). Information from these questionnaires is then passed on to the colleges to which students send their scores.

How Tests Are Used in Undergraduate Admissions

The sorting process that ultimately leads to college admission starts with the applicants themselves, who typically consider a combination of academic and nonacademic factors in deciding where to apply. For candidates who pick one of the "open-door" colleges, tests play no role in the admissions process: All that is required is to complete an application and, in some cases, show proof of high school graduation. Eight percent of the 957 four-year institutions that responded to a survey conducted in 2000 by ACT, Inc., the Association for Institutional Research, the College Board, Educational Testing Service, and the National Association for College Admission Counseling (referred to hereafter as "the joint survey"), fell into the open-door category; 80 percent of the 663 two-year institutions were open-door (see Breland, Maxey, Gernand, Cumming, & Trapani, 2002, p. 15). But even for applicants who prefer not to attend open-admission schools, chances of getting admitted to some institution are still quite good, since about 71 percent of four-year institutions admit at least 70 percent of their applicants (see Breland et al., 2002, p. 23).

Of course, the degree to which standardized test scores and other academic criteria are regarded as useful in a school's admissions decisions depends entirely on the goal of the institution's admission policies, and, more broadly, on its educational mission. As Harvard public policy professor Robert Klitgaard pointed out in his thought-provoking 1985 book, *Choosing Elites*, the "first question to ask about selective admissions is why it should be selective at all" (p. 51). Klitgaard notes that we as a society have mixed feelings about selectivity. On one hand, we think it "has unpleasant connotations of elitism, unfairness, snobbishness, and uniformity." On the other hand, we "laud excellence, recognize its scarcity and utility, and endorse admissions on the basis of merit ..." (p. 51). Another argument for selectivity in college admissions is that it encourages high schools to provide a quality education. But most institutions are selective for a more immediate reason: They consider it desirable to admit candidates who are likely to be able to do the academic work required of them. Standardized admissions tests, along with other criteria, are considered in an attempt to identify these candidates.

Just how widespread is the use of standardized tests in undergraduate admissions? According to the joint survey, the percentage of four-year colleges requiring either the SAT

or ACT held steady at slightly over 90 percent between 1979 and 2000 (Breland et al., 2002)[v] The number of students taking either of these tests increased from about half of those graduating from high school in 1979 to about two-thirds of the 1998 graduates (Breland, 1998, pp. 3, 7). Although this has not always been true, the ACT and SAT are now used interchangeably by the majority of institutions. According to ACT, Inc., the ACT is taken by more than half of graduating seniors in 25 states. In Illinois and Colorado, all juniors in public schools have taken the ACT since the 2001–02 school year. Of all US students who graduated from high school in 2005, the College Board reports that 1,475,623 students took the SAT, and ACT, Inc. reports that 1,186,251 students took the ACT. (Some students take both tests.)[vi]

To allow at least rough comparisons between the ACT and SAT, tables of "concordance" between ACT and SAT scores can be created (e.g., see Dorans, Lyu, Pommerich, & Houston, 1997). The linkage, however, is only approximate. Because the content of the tests is not identical, the association between SAT and ACT scores may not be the same for all types of test takers. In particular, the relationship between the scores is likely to depend on whether the test takers have been exposed to the curricular content in the ACT.[vii]

How heavily are test scores weighted in undergraduate admissions decisions? Two major sources of information, the joint survey and the more recent National Association for College Admission Counseling (NACAC) Admission Trends Survey (see Hawkins & Lautz, 2005)[viii] indicate that test scores are the second-most important factor, after high school grades. Four-year institutions responding to the joint survey rated high school grade-point average (GPA) or class rank as the most important factor in admissions, as was the case in similar surveys conducted in 1979, 1985, and 1992. Admission test scores had the second-highest average rating, and showed a slight increase in average importance between 1979 and 2000 (Breland et al, 2002, p. 67). The third-most important factor in all four surveys was "pattern of [high school] course work."

About 70 percent of four-year institutions reported that test scores were "routinely considered in reaching an overall judgment regarding admissibility"; another six percent of these schools said they used scores only when other credentials were weak (Breland et al., 2002, p. 61). Roughly 40 percent of four-year schools reported that they had minimum test score requirements for admission; 57 percent had minimum requirements for high school GPA (see Breland et al., 2002, p. 59). Scores on achievement tests such as the SAT Subject Tests "were not viewed as highly important in admissions decisions in any of the four surveys between 1979 and 2000" (Breland et al., 2002, p. xi).

According to the *State of College Admission* (Hawkins & Lautz, 2005), a report based on the NACAC Admission Trends Survey of colleges, to which 661 colleges and universities responded in 2004, grades in college preparatory courses and admission test scores were the two factors most likely to be accorded "considerable importance" in admission

decisions, with 80 percent giving this response for college prep course grades and 60 percent doing so for admission test scores. In general, the importance assigned to test scores increased with college size, with 55 percent of institutions with less than 3,000 students attributing considerable importance to tests and 92 percent of institutions with 20,000 or more students doing so (p. 42). Other factors given considerable importance by at least one-quarter of the institutions were "grades in all courses," class rank, and "essay or writing sample" (p. 39). Annual NACAC survey results from previous years (1993–2004) showed that the percentage of schools assigning considerable importance to grades in college prep courses remained quite constant, but that the percentage of institutions reporting that considerable importance was assigned to admission test scores increased fairly steadily from 46 percent in 1993 to the 2004 value of 60 percent (p. 39).

What do college admissions tests measure?

Despite the entrenchment of standardized admissions tests, questions have persisted about their precise function: Are these tests intended to measure specific academic achievements, or to assess intellectual aptitude? Testing experts have not been particularly helpful in clarifying the niche that admissions tests are intended to fill, and disputes on this point have been prominent in recent debates on the fairness of the SAT. From the perspective of most testing professionals, achievement tests and aptitude tests can be viewed as endpoints of a continuum, with exams that focus on specific course material lying closer to the "achievement test" pole, while those that are less reliant on mastery of particular content falling near the "aptitude test" end.

The ACT is based on an analysis of the material that is taught in grades seven through 12 in each of four areas of "educational development"—English, math, reading and science and is, therefore, closer to the "achievement" end of the continuum. By contrast, the SAT has not been linked to particular high school courses; instead, it has been claimed to measure "developed verbal and mathematical reasoning abilities" that are relevant to success in college. According to the College Board, the new SAT that emerged in 2005 remains "a test of developed reasoning," but is "more closely tied to what students learn in the high school classroom than ever before. The college success skills measured by the exam have been identified through research and discussions with college faculty, high school teachers, and subject area experts across the country." (College Entrance Examination Board, 2004, p. 3). Another notable feature of the new SAT is the absence of analogy items, which, in the eyes of critics, exemplified the SAT's poor linkage to classroom learning (e.g., see Atkinson, 2001). The new SAT, then, is somewhat closer to the achievement end of the continuum than its predecessor. (The SAT Subject Tests are, of course, based on specific curricular content.)

Reinforcing the view that aptitude and achievement are not clearly distinguishable is the high correlation between "aptitude" and "achievement" measures. For example, although the SAT Reasoning Test and the ACT were created using different frameworks, the (pre-2005) SAT total score (verbal score plus math score) and the ACT composite

score are highly correlated—.92 in a recent large-scale study (Dorans, 1999). Similarities in the functioning of the SAT Reasoning Test and SAT Subject Tests are discussed by Crouse and Trusheim (1988), Kobrin, Camara, and Milewski (2004), Bridgeman, Burton, and Cline (2004), Geiser & Studley (2004), and Zwick (2004).

Nevertheless, because the designation of admission tests as aptitude or achievement tests does have implications for the perceived fairness of these tests and the educational practices that result from their administration, the controversy is unlikely to die. In a recent reemergence of this debate, Richard C. Atkinson, then president of the University of California, announced in 2001 that he opposed the use of the SAT Reasoning Test as a university admissions criterion, arguing that it is viewed as being "akin to an IQ test" and hence as unfair, and that it promotes undesirable instructional practices, such as the implementation of analogies drills in the classroom. He recommended that standardized tests be developed which would be directly tied to college preparatory courses, and added that he hoped to eventually move away from quantitative admission formulas in order to "help all students, especially low-income and minority students, determine their own educational destinies" (Atkinson, 2001).

A different view is presented by Lohman (2004), who suggests that "aptitude tests that go beyond prior achievement have an important role to play in admission decisions, especially for minority students." He presents evidence that scores on "well-constructed measures of developed reasoning abilities" show smaller disparities among ethnic groups than scores on good achievement tests, and argues that tests of reasoning ability can help admission officers to identify students who do not do well on curriculum tests but can succeed academically if they try hard. According to Lohman, the "problem with the [2001] version of the SAT I may not be that it is an aptitude test, but that it is not enough of an aptitude test" (2004, p. 50).

Despite the fact that these debates are unresolved, SAT and ACT scores continue to be a key factor in admission decisions at most institutions.

The Predictive Validity of College Admissions Tests

Although uncertainty may exist about precisely what they measure, admission test sponsors are unambiguous about the claim that these tests are useful for predicting first-year college grades. Predictive accuracy alone is, of course, insufficient evidence of test validity. According to the eminent validity theorist Samuel Messick, validity "is an overall evaluative judgment … of the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1988, p. 33; italics in original). A comprehensive evaluation of a test's validity, then, must involve a consideration of the test's design, development, content, administration, and use. Typically, however, the validity of admission tests as a selection tool for higher education institutions is judged largely by the degree to which test scores can predict first-year college grade-point average (FGPA).

Why is first-year GPA the most common criterion for college success in these studies rather than, say, cumulative GPA at graduation? One reason is that, the later the grade data are acquired, the more students will have dropped out or transferred to other schools. Also, freshman courses tend to be more similar across fields of study than are courses taken later in college, and the resultant GPAs are, therefore, more comparable. As described below, however, some validity studies have used other measures of college success, such as cumulative GPA at graduation or degree completion.

Prediction of College Grades

How can the predictive value of a standardized test—say, the SAT—be measured? Conducting a predictive validity study requires that the FGPAs for the cohort of interest be available so that the predicted FGPAs (estimated using test scores and high school grades) can be compared to the FGPAs actually earned by the admitted students. Predictive validity studies are usually conducted within a single institution, although results may later be averaged across institutions.

Linear regression analysis is typically applied to estimate an equation for predicting FGPA using high school GPA, SAT math score, and SAT verbal score. The resulting multiple correlation provides an index of the effectiveness of the prediction equation. One way to think about a multiple correlation in this context is to note that it is equal to the simple correlation between the observed and predicted FGPAs for a particular analysis. But squared correlations are often considered to be easier to interpret: Here, a squared (simple or multiple) correlation can be interpreted as the proportion of variability in FGPA that is "explained by" or "associated with" the predictor(s) in the regression equation. For example, if the multiple correlation is .3 for an equation in which FGPA is predicted using high school GPA, SAT math score, and SAT verbal score, we could say that nine percent of the variability in FGPA ($.3^2$, multiplied by 100 to convert it to a percentage) is associated with this set of three predictors (implying that 91 percent is associated with other factors). The initial regression analysis can then be repeated using high school GPA alone as a predictor. Comparing the predictive effectiveness of the two equations gives an estimate of the "value added" by using SAT scores, sometimes called incremental validity.

For example, analyses of this kind were performed as part of a comprehensive study of the utility of the SAT as a predictor of college grades, conducted by Ramist, Lewis and McCamley-Jenkins (1994). The research was based on 1985 data from a total of about 45,000 students from 45 colleges. All analyses were performed separately within each school and then averaged. The regression analysis using only high school GPA as a predictor yielded a moderately high correlation of .39. (Using only the SAT produced a correlation of .36.) When high school GPA, SAT math and SAT verbal scores were used in combination, the correlation rose to .48, yielding an "SAT increment" of .09 (.48 minus .39). (Correlations given in the present article are not "corrected," except where noted. Corrected correlations are often larger by as much as .2, as described in a later section.) These findings parallel the results of many other test validity analyses in two basic ways.

First, prior grades alone were more effective in predicting subsequent grades than were admission test scores alone. Second, adding test scores to prior grades improved the prediction.

How important is the improvement that is achieved by using the SAT to predict FGPA? In their book, *The Case Against the SAT*, Crouse and Trusheim (1988) argued that the typical SAT increment is so small as to make the SAT useless. Essentially, their claim was that SAT scores are largely redundant with high school grades. By contrast, a recent study by Bridgeman, Pollack and Burton (2004), based on data from 41 institutions, showed that even for students with similar high school grades and course backgrounds, SAT scores contributed substantially to the prediction of college "success," defined as the attainment of a college GPA above a particular criterion level. (Cutpoints of 2.5 and 3.5 were considered, as were GPAs at the end of freshman and senior year.) From an institutional perspective, even a small improvement in prediction accuracy is often perceived as worthwhile, especially by large schools that do not have the opportunity to interview candidates or review applications in elaborate detail. In fact, institutions often view admissions tests as extremely cost-effective: Students themselves pay to take the tests, and schools are required to spend only a minimal amount to collect and process the scores.

Another reason that admission tests can be valuable is that using only high school grades, without test scores, to predict freshman grades tends to produce predictions that are systematically off target for some ethnic groups, a problem that can occur despite the sizeable correlation between high school and college grades. Including test scores in combination with high school grades to predict college performance often reduces these systematic distortions, as explained in subsequent sections.

What does current research say about the predictive validity of college admission tests? An examination of large-scale studies, (focusing on multi-institution studies and reviews published in 1985 or later) reveals some consistent patterns. The multiple correlation of ACT score (all four section scores considered together) or SAT score (verbal and math scores considered together) with FGPA is about .4, on average (ACT, 1997; Camara & Echternacht, 2000; The College Board and ETS, 1998; Noble, 1991; Ramist et al., 1994; Rigol, 1997; Willingham, 1998). This correlation—the validity coefficient—is usually slightly lower than the correlation between high school GPA and FGPA. Considering ACT or SAT scores as predictors along with high school grades yields correlations with FGPA that average about .5.[ix]

Validity of Writing Tests Used in College Admission

Because the ACT and SAT writing tests are quite new, insufficient time has elapsed for the college grade data that are needed for validity studies to have been obtained from test-takers. In the case of the SAT writing test, some research results are available on a prototype version, which was administered in the summer and fall of 2003.[x] The College Board commissioned the American Institutes for Research to conduct a study of the

predictive and incremental validity of this prototype version of the writing test at 13 colleges and universities (Norris, Oppler, Kuang, Day, & Adams, 2005, p. ii). In the context of this study, predictive validity referred to the degree to which the scores on the writing test could predict first-year college GPA, as well as first-year GPA in English composition courses. Incremental validity referred to the degree to which including writing scores improved the prediction of GPA, over and above the prediction accuracy that could be achieved using SAT math and verbal scores and self-reported high school GPA. (Incremental validity was investigated only for the case in which overall first-year GPA was used as a criterion.) The sample size was 1,248 for the analyses using overall first-year GPA and 891 for the analyses based on English composition GPA. All validity analyses were conducted within institution; a weighted average of the results was then obtained (using weights based on the sample size within each institution).

Norris et al. reported (p. 16) both uncorrected correlations and correlations corrected for range restriction (see the section, "Effects of range restriction and criterion unreliability on validity coefficients"). The corrected correlations are given in parentheses below. The correlation between the total SAT writing test score and first-year GPA was .32 (.46); the correlation obtained using only the multiple-choice writing score was nearly as large. By contrast, the score on the essay portion of the SAT writing test had a correlation with first-year GPA of only .16 (.20). The finding of a modest validity coefficient for the total and multiple-choice scores and a small coefficient for the essay scores parallels the findings obtained in studies of the SAT Subject test in writing; (e.g., see Breland, Kubota, & Bonner, 1999). Norris et al. found a similar pattern of results for the prediction of English composition GPA (p. 16): The total SAT writing test score had a correlation of .24 (.32) with the GPA, the multiple-choice writing score had a nearly equivalent correlation, and the essay score had a considerably lower correlation of .14 (.18). Incremental validity studies showed an improvement of .01 or .02 (depending on the statistical adjustments applied) in the size of the multiple correlation when SAT writing was added as a predictor to an equation that already included SAT verbal and math scores and self-reported high school GPA (p. 18).

Occasionally, test scores are found to do a better job than high school grades in predicting college GPA, as in a 1990 study at Dartmouth ("SAT's better freshman predictor than grades," January 16, 1991). In addition, Ramist et al. (1994) found that SAT scores tended to be more effective than high school GPA in predicting grades in individual college courses and that, among African-American students, SAT scores were slightly more effective than high school grades in predicting FGPA.

The validity of the SAT Subject Tests (then called the College Board Achievement Tests) was investigated by Ramist, Lewis and McCamley-Jenkins (2001) using data from 1982 and 1985. They found that validity coefficients for the individual Subject Tests ranged from .17 for Spanish and German to .58 for Chemistry and Mathematics II (p. 36). (These coefficients were corrected for restriction of range and for the upward bias of the sample

multiple correlation.) Ramist et al. also found that the average of a student's SAT Subject Test scores tended to be a slightly better predictor of FGPA than was the SAT Reasoning Test. The correlation between the SAT Subject Test average and FGPA was about the same as the correlation between high school GPA and FGPA. Adding the SAT Subject Test scores to a prediction equation that included SAT Reasoning Test score and high school GPA further boosted the multiple correlation by a small amount (p. 35). More recently, several studies based on University of California data have found the SAT Subject Tests, particularly the writing test, to be more predictive of FGPA than the SAT Reasoning Test in most instances (Geiser & Studley, 2004; Kobrin, Camara, & Milewski, 2004; Zwick, Brown, & Sklar, 2004).

College Board studies have provided fairly strong evidence that admission tests can be useful in predicting grades beyond the first year of college. A recent report summarized the results of 19 studies (all appearing since 1985) of the association between students' SAT scores and their cumulative grade-point averages upon completing college (Burton & Ramist, 2001). These studies were based on results from 227 institutions and over 64,000 students. SAT verbal score and SAT math score each had correlations averaging about .4 with the final college GPA, as did high school achievement (grades or class rank). These correlations are at least as large as those typically reported for first-year college GPA.

In another research project sponsored by the College Board, researchers conducted a meta-analysis of more than 1,700 previous studies of the predictive value of the SAT. They concluded that the SAT was useful for predicting grades obtained both early and late in college, as well as other success criteria. The corrected correlations between SAT scores and grade-point averages earned in the second, third, and fourth years of college ranged from roughly .35 to .45 (Hezlett, Kuncel, Vey, Ahart, Ones, Campbell, & Camara, 2001).

Prediction of Graduation

Isn't it more important to predict who will complete a degree than to forecast first-year grades? This very reasonable question often arises in discussions of the utility of admissions tests. Willingham (1974, p. 275) neatly summarized the pros and cons of graduation as a measure of academic success. (Although his article was about graduate education, his remarks are equally applicable at the college level.) He noted that while graduation "is probably the single most defensible criterion of success … one must wait a long time for this criterion. Another difficulty is the fact that whether or not a student graduates may frequently depend upon extraneous influences. [Also,] this criterion places a premium on academic persistence and probably does not differentiate very well the most promising scholars and professionals."

How well can admission test scores predict who will graduate? Burton and Ramist (2001) reviewed the research conducted in the last twenty years about the association between

SAT scores and college graduation, and concluded that "there is a solid academic component to graduation that is measured by [SAT scores and high school record]" (p. 17). Among the studies they considered were those of Astin, Tsui, and Avalos (1996) and Manski and Wise (1983). In a study of more than 75,000 freshmen at 365 institutions, Astin et al. found that even among the 9,000 students with high school GPAs of A or A+, the SAT was a valuable predictor: The graduation rate was 28 percent for those with total SAT scores of less than 700 and rose steadily as SAT score increased, reaching 80 percent for students with scores over 1300. In their analysis of data from the National Longitudinal Study of the High School Class of 1972, Manski and Wise (1983, p. 15) also found a strong relationship between SAT score and persistence in college, even among students with similar class ranks in high school. On the other hand, several analyses of large national databases have yielded only moderate correlations between SAT scores and graduation (about .3, slightly greater than the typical correlation between high school grades and college graduation), and some analyses conducted within single institutions found correlations of only .1 or .2 between test scores and graduation (see Burton & Ramist, 2001, pp. 16-19).

Some additional relevant studies, not included in the Burton and Ramist review, are those of Adelman (1999), Astin and Osequera (2002), Hezlett et al. (2001), and Zwick and Sklar (2005). Adelman found a strong relationship between a "mini-SAT" (a one-hour test with items drawn from old SATs) and the likelihood of completing college. Results from a national sample of more than 7,000 college students showed that only seven percent of those who scored in the bottom fifth on this test completed a bachelor's degree, compared to 67 percent of those who scored in the top fifth (Adelman, 1999, p. 15). A College Board guide for admission officers, drawing on an analysis by Astin and Osequera (2002) based on nearly 57,000 students at 262 institutions, displays a chart that shows a rise in degree attainment rates as SAT scores and high school grades increase (College Board, 2005, p. 11). Hezlett, et al., (2001) showed, based on 11 earlier analyses, that SAT scores were somewhat useful for predicting both persistence through the first year of college and completion of the bachelor's degree. My co-author Jeffrey C. Sklar and I used survival analysis, a statistical technique for modeling the time until the occurrence of an event, to study college graduation patterns over seven years for the sophomore cohort of the High School and Beyond Survey conducted by the National Center for Education Statistics. We considered four groups of students: Hispanic students who said their first language was Spanish, and Hispanic, Black, and White students who said their first language was English. Total SAT score was a statistically significant predictor of college graduation, after taking high school GPA into account, in the Hispanic/English and White/English groups, but not in the remaining two groups (Zwick & Sklar, 2005).

One pattern that becomes evident in the results of this body of research is that single-institution studies tend to find smaller correlations between test scores and degree completion than studies based on large national data bases. In a large study that includes many colleges, there will be a much larger range of test scores and graduation rates than in a

single school. Multi-institution analyses of graduation are usually based on the combined data from all the schools (unlike multi-institution GPA prediction studies, which usually involve analyses that have been conducted within institutions and then averaged). To some extent, then, the apparent association between test scores and graduation reflects the fact that some schools have both higher test scores and higher graduation rates than others. (This phenomenon was noted by Willingham, 1985, p. 105.)

In summary, while admission test scores may be of some use for predicting graduation, their predictive value *within a particular school* is likely to be quite small. Whether students remain in college is determined in part by nonacademic factors involving finances, mental and physical health, and family responsibilities. Therefore, it is unlikely that any measure of academic performance can do a very accurate job of predicting who gets a degree.

Prediction of Persistence in College

Researchers at ACT, Inc. have studied the degree to which ACT scores predict persistence in college, although graduation itself was not considered. One study (Noble, Maxey, Ritchie, & Habley, 2005) examined factors related to retention in college for the nearly 800,000 ACT-tested students who graduated from high school in 2003 and enrolled in college in fall 2003. Students were considered "retained" if they returned to the same college in fall 2004. The authors found that students who exceeded "college readiness benchmark" scores on the ACT (Allen & Sconing, 2005) were more likely to persist than those who did not. In particular, students who met or exceeded benchmarks in all four ACT subject areas (English, math, reading, and science) had a retention rate of 84 percent, compared to a rate of 70 percent for students who did not meet any of the four benchmarks (Noble et al., Table 5). The benchmark criteria were particularly predictive of retention for African-American students (Noble et al., Figures 7–8).

Another study of retention (Perkhounkova, Noble, & McLaughlin, 2006) was conducted by ACT, Inc. in collaboration with DePaul University, a large private Midwestern institution. The study produced some results that are counterintuitive. Results of a logistic regression analysis based on more than 5,000 students showed that ACT composite score had a small but statistically significant negative effect on second-year retention for freshmen (including both new and transfer students) and for nonfreshman transfer students (Perkhounkova et al., 2006, p. 6). That is, the higher the ACT score, the less likely it was that the student would remain at the university (with all other factors, including high school GPA, held constant). According to the authors, the explanation for this finding may be that "very capable students at DePaul tend to transfer to other institutions after the first year" (p. 8). This explanation is not entirely satisfactory, however, since high school GPA was positively related to retention.[xi]

Effects of Range Restriction and Criterion Unreliability on Validity Coefficients

A factor that complicates the interpretation of validity coefficients is selection, or restriction

of range: A very basic limitation of test validity research is that only a portion of the college applicants is available for analysis. For example, in an ACT validity study conducted by a particular college, students whose ACT scores were too low to allow admission will not, of course, have freshman GPAs. Some high-scoring applicants who were, in fact, admitted may also be unavailable for analysis because they chose another school. Because of this restriction of range (of ACT scores, and, as a result, of other predictors and of FGPAs as well), validity coefficients tend to be smaller for the admitted students than they would be for the entire population of applicants. As a result, the apparent association between test scores and FGPA is smaller than it would be if the entire group of applicants could be considered. A simple correlation or regression approach to the analysis of test validity will, therefore, produce a more pessimistic picture than is warranted, given that the intended goal of a validity study is to estimate the usefulness of tests in selecting students from the overall applicant pool.

To compensate for the effects of selection, statistical corrections are sometimes applied in an attempt to estimate how big the association would have been if the range had not been restricted (see Gulliksen, 1987, pp. 165-166). These adjustments, which are based on the disparity between the sample covariance matrix of the predictors for the admitted students and the covariance matrix for the applicant pool (or some other reference population) are only approximate since they require the unrealistically simple assumption that selection takes the form of truncation of the predictor distributions (as well as other statistical assumptions). In actuality, the determination of which applicants are accepted at a particular college is quite complex, involving a combination of self-selection and institutional selection.

An additional drawback of traditional validity studies is that they ignore the inaccuracies and inconsistencies of GPA as a criterion of academic performance. As in the case of range restriction, statistical corrections can be applied to validity coefficients to adjust for the unreliability of grades. In evaluating the results of a validity analysis, it is important to determine whether validity coefficients have been adjusted for range restriction or criterion unreliability, since the effect of these corrections can be substantial. For example, Ramist et al. (1994) found the uncorrected multiple correlation of verbal and math SAT scores with FGPA to be .36; with adjustments for restriction of range and criterion unreliability, the correlation rose to .57, a sizable increase.

Several researchers, using approaches of varying sophistication, have attempted to improve the precision of grade-point averages by taking into account the fact that some college courses are harder than others and that some fields of study have more stringent grading practices than others (see Johnson, 1997; Stricker, Rock, Burton, Muraki, & Jirele, 1994, and Willingham, Pollack, & Lewis, 2000 for reviews.) Adjusting GPAs for course difficulty—not a trivial task—usually leads to a slight increase (up to .1) in the correlation between test scores and GPAs. Even a relatively simple refinement—achieved by using only specific academic courses as the basis for computing GPA—has been

found to make GPAs more comparable across students and more highly correlated with test scores.[xii]

Another subtle aspect of validity analysis is the problem of "underprediction" and "overprediction." Suppose a college uses data from all freshmen to obtain a regression equation for use in predicting FGPA for future applicants. Even if the validity coefficient is high, it is still possible that the equation will lead to predicted GPAs that are systematically too high or too low for certain student groups. This phenomenon is discussed further below.

Research on the Admission Test Performance of People of Color, Women, and Other Special Populations

College admission test results often reveal substantial average score differences among ethnic, gender, and socioeconomic groups. In the popular press, these differences are often regarded as sufficient evidence that these tests are biased. From a psychometric perspective, however, a test's fairness is inextricably tied to its validity. According to Cole and Moss (1989), test bias occurs (i.e., fairness is violated) "when a test score has meanings or implications for a relevant, definable subgroup of test takers that are different from the meanings or implications" for other test takers. "… [B]ias is differential validity of a given interpretation of a test score…" (p. 205).

Psychometric assessment of the fairness of college admission tests typically comprises two broad types of investigations. One type consists of analyses of differential prediction of a criterion, usually first-year grade-point average (FGPA). Two distinct questions are typically addressed in differential prediction studies: First, are the test scores equally effective as predictors of FGPA for all groups? This question is investigated by obtaining separate prediction equations for each group and then comparing correlation or regression coefficients across groups. Second, if we obtain a single prediction equation for students as a whole, how accurate are the predictions, on average, for particular student groups? In a least squares regression analysis, the sum of the prediction errors across all observations must be zero, but this need not be true within each group of interest. It is useful, therefore, to determine whether the equation produces predicted GPAs that are systematically too high or too low for some groups. Analyses of this kind often include key predictors in addition to test scores. For example, high school GPA would typically be included in the regression equations. Analyses of differential prediction allow a determination of whether a test is biased according to Cleary's definition, which says that a test is biased against a particular subgroup of test-takers "if the criterion score [in this case, GPA] predicted from the common regression line is consistently too high or too low for members of the subgroup" (Cleary, 1968, p. 115; see also Humphreys, 1952). Most researchers today would use the term "prediction bias" rather than "test bias," to reflect the fact that prediction errors may stem from causes external to the test.

Another common component of fairness assessment is an examination of item content. Before an item is approved for inclusion in an admissions test, it undergoes a "sensitivity

review" to make sure its content is not disturbing to certain student groups or offensive in some other way. Later, after the test is administered, a differential item functioning (DIF) screening is performed to determine whether equally skilled members of different groups (e.g., men and women) have statistically different rates of correct response on some items.[xiii] The ultimate purpose of the DIF analysis is to identify test items with content that may be problematic for some student group. For example, the item may include content that is irrelevant to the construct being assessed, and is more familiar to some student groups than others, such as sports content in a mathematics test. Items found to have DIF are either discarded immediately or "flagged" for further study. Typically, a flagged item is reviewed by panelists from a variety of backgrounds who are expert in the subject-matter of the test. If the question is considered legitimate despite the differential functioning, it remains on the test. If the item is determined to be biased, it is modified or eliminated.

The sections below summarize the findings on differential prediction and differential item functioning on college admissions tests and also address fairness issues related to test coaching.

Differential Performance and Differential Prediction Findings for Ethnic Groups

White and Asian-American test-takers typically receive higher scores on standardized tests than African-American, Hispanic and Native American test-takers; some differences show up as early as preschool (Nettles & Nettles, 1999, p. 2). A wide array of reasons has been offered for these score gaps, including socioeconomic, instructional, cultural, linguistic, and biological factors, as well as test bias. Given the disturbing patterns of performance differences, it is particularly important to determine how well admission tests work as a measuring device for people of color.

Two recurrent findings in SAT validity studies involving ethnic groups are that correlations of test scores with FGPA tend to be somewhat smaller for Black and Hispanic students than for White students (see Young, 2004, p. 291), and that, counter to intuition, use of a common regression equation to predict FGPA using SAT scores and high school grades produces overpredictions (predicted grades higher than actual grades) for these groups. Based on 11 studies, Young (2004, pp. 293–294) found the average overprediction for African-American students to be .11 (on a 0-4 grade-point scale); based on eight studies, the average overprediction for Hispanic students was .08. The lower correlations and the tendency toward overprediction also occur when high school GPA only is used to predict FGPA. In fact, although high school GPA is usually more highly correlated with FGPA than is SAT, overprediction tends to be worse if only high school GPA is included in the prediction equation (e.g., Ramist et al., 1994; Zwick & Schlemer, 2004; Zwick & Sklar, 2005). The overprediction of college achievement for Black and Hispanic students has also been found in research on the ACT (Noble, 2004, Sawyer, 1985). Educational researchers have long been aware of the overprediction phenomenon (e.g., Cleary, 1968; Linn, 1983). With the publication of *The Shape of the River* (Bowen & Bok, 1998) and T*he*

Black-White Test Score Gap (Jencks & Phillips, 1998), overprediction has become more widely recognized. A brief overview of the most prevalent hypotheses about the reasons for the overprediction findings appears here. (See also Zwick, 2002, pp. 117–124.)

One conjecture is that minority and White students are likely to differ in ways that are not fully captured by either their test scores or their high school grades. For example, a Black student and a White student with the same high school GPAs and admissions test scores may nevertheless differ in terms of the quality of their early schooling, which could influence their academic preparation. A related technical explanation is that overprediction occurs because neither SAT scores nor high school grades are perfectly reliable measures of academic abilities. They can be influenced by factors other than the students' academic knowledge and skills; that is, they are susceptible to measurement error. The effect of this imprecision is simplest to understand in the case of one predictor. Suppose a test score is being used to predict subsequent GPA. Under typical classical test theory and regression assumptions, the unreliability of the score can be shown to produce a regression line that is less steep than the line that would theoretically be obtained with a predictor that was free of measurement error, implying that groups with lower test scores will tend to be overpredicted while those with higher scores will tend to be underpredicted (see Snedecor & Cochran, 1967, pp. 164–166). But two findings argue against this factor as an all-purpose explanation: First, women often score lower than men on admission tests, yet their college grades tend to be underpredicted. Second, since the reliability of college admission tests is typically very high, the impact of measurement error on prediction accuracy is likely to be small.

Another hypothesis about overprediction is that minority students do not fulfill their academic potential in college, which is assumed to be accurately captured by the tests. This "underperformance" could occur because of outright racism, an inhospitable campus environment, or life difficulties, such as inadequate financial resources. It has also been conjectured that anxieties, low aspirations, or negative attitudes may interfere with the academic success of minority students (e.g., Bowen & Bok, 1998, p. 84).

The "stereotype threat" theory of Steele and Aronson (1998) has been offered as another possible explanation for overprediction (e.g., Bowen & Bok, 1998, p. 81). Stereotype threat—"the fear of doing something that would inadvertently confirm [a negative] stereotype"—is assumed to produce stress, which ultimately causes students to perform more poorly (Steele, 1999, p. 4). Standardized testing situations are considered to be particularly evocative of stereotype threat (Steele & Aronson, 1998, pp. 425–426). However, if stereotype threat depressed admissions test performance, but didn't affect subsequent academic work, it would be expected to lead to underprediction because the affected students would perform better in college than their (depressed) test scores would indicate. Therefore, stereotype threat does not seem like a plausible explanation for overprediction of minority students' test scores.

It seems evident that unmeasured differences between White students and Black and Hispanic students with the same test scores and previous grades play a role in the recurrent finding of overprediction. It also seems likely that a greater incidence among minority students of life difficulties and financial problems in college contributes to the phenomenon.

How does test content contribute to ethnic group differences in test performance?
Differential Item Functioning Findings

What kinds of items have shown evidence of DIF in ethnic group analyses? It has been a recurrent finding that African-American, Hispanic, and Asian-American test-takers do not perform as well as a matched group of Whites on verbal analogy items (Bleistein & Wright, 1987; Rogers & Kulick, 1987; Schmitt, 1987). (As of 2005, verbal analogy items no longer appear in the SAT.) The same is true for test questions containing homographs––words that have two (or more) completely different meanings, such as "light," which can mean "not heavy" or "not dark" (Schmitt & Dorans, 1988; O'Neill & McPeek, 1993). Schmitt and her colleagues also found that items containing similar words with common roots in English and Spanish—true cognates—favor Hispanic test-takers if the Spanish version is used more frequently than its English cognate (O'Neill & McPeek, 1993, p. 266). There is also some evidence that Hispanic test-takers are disadvantaged by false cognates—similar words that have different meanings in the two languages.

DIF findings on math items are quite mixed. Some studies have found that minorities perform better than a matched group of Whites on "pure math" items—those involving algebraic manipulations in the absence of any context—and do worse on word problems. One speculation about the reason for this is that pure math items tend to resemble textbook problems, which may be the focus of instruction at schools with fewer resources (O'Neill & McPeek, 1993, p. 270). Some research has also found that Black test-takers do not perform as well as a matched group of Whites on test questions that include graphs, charts and diagrams, although the reasons for this remain unclear.

Other studies have shown that questions on topics of "minority interest" show evidence of DIF in favor of people of color. For example, one study of the SAT found results of this kind on a reading comprehension passage about a Black mathematician and on passages about civil rights and poverty (Rogers & Kulick, 1987, p. 7; see also O'Neill & McPeek, 1993, p. 262–263). These findings suggest that DIF can be caused by differences in test-takers' interest in the content of the test items. Differences across groups in training and course background can also result in DIF.

To what degree do problem test items contribute to overall test score differences among ethnic groups? Burton and Burton (1993) examined ethnic group performance differences on the SAT before and after DIF screening began at ETS in 1989. Essentially, there was no change over time in the score disparities among ethnic groups. For one thing, the number of test questions found to have problems was fairly small. Also, on some items

that were eliminated, ethnic minorities had an advantage. Even in the absence of evidence that it affects overall scores, however, DIF screening is important as a precaution against the inclusion of unreasonable test content and as a source of information that can contribute to the construction of better tests in the future.

Differential Performance and Differential Prediction Findings for Men and Women

On average, men score better than women on the SAT math and verbal sections, the ACT math test, the ACT science test and the ACT composite. Women tend to score better then men on the ACT English and reading tests. (At one time, women scored better than men on the verbal SAT but this changed beginning in 1972.) National results are not yet available from the new ACT and SAT writing tests. However, in 2002–2003, Breland, Kubota, Nickerson, Trapani, and Walker (2004) conducted a study of student performance on two essay items that used a new prompt type intended for the new SAT (which had not yet been released) and on two items of an older prompt type that was already in use on the SAT Subject Test in Writing. The study, based on approximately 2,500 11th graders sampled from 49 schools, showed that females outperformed males by .2 to .3 standard deviation units on all four of the essay prompts. (Gender differences on the new prompt type were similar to those on the old prompt type.)

What accounts for the fact that men perform better than women on most sections of college admission tests? It has been suggested that this score gap occurs in part because women and men differentially self-select to take these tests, producing sex differences in average economic and academic background. More young women than men take both the ACT and SAT, for example, and, according to a College Board research summary, a "much higher proportion of females than males taking the SAT come from families with lower levels of income and parental education" (The College Board, 1998, p. 2).[xiv] The College Board summary (1998, p. 2) also noted that "important differences still persist in the proportion of males and females completing advanced courses in math, science, and computer programming" in high school, but more recent analyses show that disparities in course background have decreased substantially over the last 10 years (The College Board, 2000; see also Coley, 2001). Countless other reasons have been offered to explain the gender gap in test scores, such as test bias, biological differences, diverging interests and aspirations, and societal influences, including stereotype threat (Steele, 1997, p. 619).

Two other recurrent findings involving gender differences in admission test results are that test validity evidence tends to be stronger for women, but that, paradoxically, women's college grades tend to be underpredicted by test scores. Each of these findings is addressed in turn.

In SAT research, it is typical to find that validity coefficients are higher for women, although the reasons are not clear. One frequent speculation is that men are more likely to skip classes and homework assignments, making their college grades less predictable. In

more selective colleges, where both men and women are presumably more dedicated to their academic work, validities for men and women tend to be more similar to each other (see Young, 2004 for a review; see also Bridgeman, McCamley-Jenkins, & Ervin 2000, Table 4, page 5; Ramist et al., 1994, Table 18, p. 25).

The second persistent finding is that, when a common regression equation, based on men and women, is used to predict college grades using SAT scores and high school GPA, women's predicted grades tend to be lower than the FGPAs they actually earned, and the reverse is true for men (e.g., Ramist et al, 1994, p. 15). Bridgeman et al. (2000), who studied SAT validity at 23 (mostly selective) colleges found that it was primarily White women who were underpredicted (by .09, nearly a tenth of a grade point), while the college grades of men from all ethnic groups tended to be overpredicted. Based on consideration of 17 studies, Young (2004) found an average underprediction of .06 (on a 0-4 scale) for women. Although this is smaller than the average overprediction found for African-American and Hispanic test-takers (see above), Young noted that, because women constitute the majority of college students in the U.S., "the net impact of the differential prediction by sex has a much greater overall effect than the overprediction problem for minority students" (p. 296).

The SAT, the ACT, and most of the SAT Subject Tests have been found to underpredict women's college grades (see Willingham & Cole, 1997, Leonard & Jiang, 1999, and Young, 2004 for reviews; see also Ramist et al., 2001). What accounts for this phenomenon? Some research suggests that women are less likely than men to take college courses that are stringently graded (Ramist et al.,1994) or to select demanding majors (Pennock-Román, 1994). Underprediction seems to be less evident at more selective institutions (e.g., Ramist et al, 1994, p. 27), perhaps because nearly all students at these institutions take difficult courses, reducing the differences between men and women in the grading stringency of their coursework. Another explanation is that women are better prepared and more studious than their male counterparts (Stricker, Rock, & Burton, 1993; see also Dwyer & Johnson, 1997; Willingham, Pollack, & Lewis, 2000). According to this conjecture, women actually do perform better in college than men with equivalent academic preparation, and this is appropriately reflected in their college grades. It has also been found that underprediction of women's grades is reduced when writing test scores play a substantial role in the prediction equation (e.g., Ramist et al., 1994; Leonard & Jiang, 1999). This finding suggests that the underprediction of women's college grades may be reduced as a result of the addition of a writing component to the SAT and ACT, though new validity studies will be needed to determine if this is the case.

How does test content contribute to sex differences in test performance? Differential Item Functioning Findings

Awareness that certain test content may be disadvantageous to women dates back as least as far as 1923, when Carl Brigham, a rather unlikely forefather of today's test critics, noted with regard to the Army Alpha exams, "As a rule women object to the information

test more than men because the test samples rather heavily the fields of sport, mechanical interests, etc. The chances are that this test would penalize women rather heavily..." (Brigham, 1923, p. 30). Today, testing companies take great pains to avoid test questions about topics that might be relatively unfamiliar to women or people of color. As noted earlier, test items go through "sensitivity review" to eliminate content thought to be racist, sexist, or potentially offensive in some other way. Among the types of test questions that should generally be avoided, according to an ETS document, are those that involve violence or harm, sports knowledge, or military topics (Educational Testing Service, 1999). Test items that unnecessarily include these subjects are thought to give men an unfair advantage.

Certain DIF findings have emerged fairly consistently from comparisons of men and women. Women tend not to do as well as a matched group of men on verbal SAT items (predating the strictures above) about scientific topics or about stereotypically male interests, like sports or military activities (e.g., Bridgeman & Schmitt, 1997, p. 194). On the other hand, women tend to perform better than their male counterparts on questions about human relationships or questions about the arts (O'Neill & McPeek, 1993, p. 262).

It seems likely that these particular performance disparities stem from differences in interests and pastimes, and perhaps high school course work. But for the vast majority of items that show evidence of DIF, the reasons are murky at best. In their review of DIF findings, O'Neill and McPeek (1993) noted that on several ETS tests and on the ACT, women perform better on algebra questions than men with equivalent quantitative scores; men do better on geometry and mathematical problem-solving. Also, analyses of the SAT and ACT have shown that women do better on "pure mathematics" problems (23/2 + 23/3 + 23/6 = ?), and men tend to perform better on word problems framed in terms of an actual situation (O'Neill & McPeek, 1993, p. 269).

As in the case of ethnic group score differences, Burton and Burton (1993) found that there was essentially no change in the SAT score gap between men and women after DIF screening began at ETS in 1989. A fairly small number of items were found to have problems, and some of those eliminated were items on which women had an advantage.

Performance of People with Disabilities on Admission Tests

A standardized test is meant to be administered under uniform conditions and time constraints, but fairness dictates that test scores should not be affected by any limitations of the test-taker which are not relevant to the skills being assessed. In this spirit, various types of special accommodations are made available to admission test candidates with disabilities. Test-takers with visual impairments, for example, are offered Braille, cassette, or large-type versions of the test, or are provided with assistants who read the test aloud. Other special arrangements that are typically available include scribes for individuals with physical impairments that make writing impossible and sign language interpreters who can relay spoken instructions to deaf test-takers. Extended time is also permitted

for candidates with disabilities. The rationale for offering these accommodations is that "the standard procedures … impede [these] test-takers from performing up to their ability" (Mandinach, Cahalan, & Camara, 2001, p. 5). Ideally, scores on the accommodated admissions tests should be comparable to scores obtained from nondisabled test-takers under standard conditions—they should measure the same cognitive abilities and should be of equivalent difficulty and precision.

The provision of special testing arrangements gives rise to a vast array of questions. What should "count" as a disability in an admissions testing situation? How can we determine whether the difficulty of an accommodated test for a candidate with a disability is equal to the difficulty of the standard test for a nondisabled test-taker? Should scores that are sent to schools be "flagged" if they have been obtained under nonstandard conditions? Do admissions test scores predict grades as well for people with disabilities as for other test-takers?

At the recommendation of a National Academy of Sciences panel, a research program focusing on SAT and GRE candidates with disabilities was undertaken during the 1980s, under the sponsorship of ETS, the College Board, and the Graduate Record Examinations Board (Willingham, Ragosta, Bennett, Braun, Rock, & Powers, 1988). The researchers concluded that in general, the scores of SAT-takers who received accommodations were roughly comparable to scores obtained by nondisabled test-takers under standard conditions. The one major exception, described below, involved test-takers who were granted extended time. Prediction of subsequent grades was found to be somewhat less accurate for candidates with disabilities, whether test scores or previous grades were used as predictors. The researchers speculated that one reason may be the wide range in the quality of educational programs and grading standards for these students. Individuals with disabilities may also be more likely than other students to experience difficulties that affect their academic performance, such as inadequate funds or support services.

In general, students who received extended time were found to be more likely to finish the test than candidates at standard test administrations. Willingham et al. (1988, p. 156) stated that, for SAT-takers claiming to have learning disabilities, "the data most clearly suggested that providing longer amounts of time may raise scores beyond the level appropriate to compensate for the disability." In particular, these students' subsequent college grades were lower than their test scores predicted, and the greater the extended time, the greater the discrepancy. By contrast, the college performance of these students was consistent with their high school grades, suggesting that their SAT scores were inflated by excessively liberal time limits. Similar conclusions have been obtained in more recent studies of the SAT (Cahalan, Mandinach, & Camara, 2002) and ACT (Ziomek & Andrews, 1996). Developing fair policies for candidates with learning disabilities has always been particularly troublesome for testing companies because even the definition of "learning disability" is unclear and subject to manipulation (e.g., see California State Auditor, 2000; Leatherman, December 4, 2000; Weiss, 2000).

A longstanding controversy about testing accommodations for people with disabilities is whether score reports should contain a "flag" indicating that the test was given under nonstandard conditions. Proponents of flagging (who include most college admission officers and high school guidance counselors, according to a recent survey [Mandinach, 2000]) say that information about testing conditions is needed to interpret test scores correctly. Test users, such as universities, are misled when this information is withheld, they contend, possibly to the test-taker's disadvantage. Advocates of flagging say that it can also help to discourage dishonest "game-players" from requesting undeserved extra time, and can thus increase the fairness of the test to those who play by the rules. Those who argue against flagging, however, say that it stigmatizes test-takers with disabilities and constitutes both a privacy violation and a form of discrimination that is prohibited by law.

The *Standards for Educational and Psychological Testing* offer a reasonable guideline for determining when flagging is appropriate. "[I]mportant information about test score meaning should not be withheld from test users who interpret and act on the test scores," the Standards say, "and ... irrelevant information should not be provided. When there is sufficient evidence of score comparability across regular and modified administrations, there is no need for any sort of flagging" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 105). The one accommodation for which comparability evidence is clearly lacking is the provision of extended time to candidates claiming learning disabilities. From this perspective, flagging the scores from these administrations seems appropriate.

The flagging debate, however, has been more heavily influenced by legal than by psychometric considerations. In response to a federal lawsuit filed by a Graduate Management Admission Test candidate with a disability, ETS announced in 2001 that it would discontinue flagging scores of test-takers who received extra time on some tests (Foster, 2001). The decision did not affect the SAT because it is owned by the College Board, which was not a defendant in the suit. However, in 2002, the College Board decided to discontinue flagging for the SAT as well, and a corresponding decision regarding the ACT soon followed ("ACT to stop flagging scores of disabled students who need extra time on test," 2002).

The Effectiveness and Fairness Implications of Admissions Test Coaching

The effectiveness and ethics of commercial test preparation for admission tests, particularly the SAT, have long been the subject of controversy. During the last 15 years, several well-designed research studies have produced consistent results about the magnitude of score improvement that results from SAT coaching. Becker (1990), Powers and Rock (1999), and Briggs (2001; 2004) all concluded that the average gain from SAT coaching is between 6 and 8 points on the verbal section and between 14 and 18 points on the math section. Coaching studies on tests other than the SAT are quite scarce. Research suggests that coaching produces small benefits on the ACT (Briggs, 2001; Scholes &

McCoy, 1998). (See Zwick, 2002, Chapter 7, and Kaplan, 2005, for additional reviews of coaching research.) Although many testing companies long maintained the position that test preparation programs were largely ineffective, the sponsors of all major admissions tests now produce test preparation materials, seemingly a tacit acknowledgment that preparation can be beneficial.

Currently, the coaching debate tends to focus on the question of whether coaching, because it is likely to be most accessible to those who have already benefited from a lifetime of educational advantages, presents an impediment to test fairness for poor and minority test-takers. Powers and Rock (1999) and Briggs (2001) found that coached SAT-takers came from more affluent families than uncoached candidates; they were also more motivated and were more likely to be Asian-American.

Although average coaching effects are apparently quite small, it is legitimate to question the fairness of a system in which some test-takers can afford coaching and others cannot. It is clear that coaching programs are here to stay, and that it is impractical, if not impossible, to create admissions tests that are not susceptible to coaching. Minimizing the impact of coaching on test fairness, then, requires that the availability of free and low-cost test preparation be increased.

A Crucial Issue for the Future: The Performance of Students with Limited English Skills on College Admission Tests

An increasing proportion of US college applicants are immigrants or children of immigrants. In response to a question about the first language they learned, 22 percent of college-bound high school seniors in 2005 responded, “English and another language” or “another language” (College Entrance Examination Board, 2005). The appropriateness of standard college admission criteria for non-native speakers of English is therefore an issue of great practical importance.

Recent findings on the performance of students with limited English skills on college admissions tests (mainly the SAT) have been quite mixed, making this a difficult issue to resolve. Whereas prediction of college performance has sometimes been found to be less effective for language minorities than for native speakers of English, the reverse has been true in other cases (see Zwick & Schlemer, 2004; Zwick & Sklar, 2005). In addition, the use of SAT scores to predict FGPA has sometimes led to underprediction (Ramist et al., 1994) and sometimes to overprediction (Zwick & Schlemer, 2004; Zwick & Sklar, 2005. The inconsistency of results is probably due in part to the fact that language minority groups have been defined in varying ways (see Zwick & Sklar, 2005).

Of particular importance is the impact of the new SAT and ACT writing tests on non-native English speakers. The addition of a writing section to the SAT has long been a source of controversy because of the anticipated effect on recent immigrants and minorities. In 1988, a member of a blue-ribbon commission that was assembled to consider an

overhaul of the SAT strenuously objected to the addition of a writing requirement because of the possible adverse impact on non-native English speakers (Pitsch, 1990). These concerns played a role in the College Board's decision against adding a writing component to the main part of the SAT. Instead, a separate and optional test—the SAT II Writing Test—was instituted, and further research was promised.

Now that the SAT and ACT both have writing components (though the ACT's is optional), it will be necessary to monitor the impact on language minorities. The effect could be positive if otherwise promising language minority students who score poorly on the writing section are admitted, but routed to writing improvement courses. It could be detrimental, of course, if it leads to the exclusion of talented students without allowing them the opportunity to improve their writing skills. This is a key area for future work.

What Do Admissions Personnel Need to Know about Admissions Tests

As NACAC's Statement of Counselor Competencies states, it is essential that college admissions counselors understand "the proper administration and uses of standardized tests and be able to interpret test scores and test-related data to students, parents, educators, institutions, agencies, and the public" (NACAC, 2000, p. 11). In order to develop a thorough understanding of test use and interpretation, counselors need to have a command of the fundamentals of educational measurement and statistics. However, like K–12 school personnel, who must also interpret test scores and explain them to multiple audiences, admissions counselors may not have had the opportunity to acquire training in the area of academic assessment and score interpretation.

One possible avenue for professional development for admission counselors is the use of Web-based instruction in educational measurement and statistics. Web-based instructional modules on test score interpretation have been developed for K–12 teachers and administrators as part of the Instructional Tools in Educational Measurement and Statistics (ITEMS) project (Zwick, Sklar, & Wakefield, 2006) under a grant from the National Science Foundation. These 25-minute modules are intended to help prepare K–12 personnel to use test results effectively and explain them to students, parents, the school board, and the press. The modules can be viewed at a time and place of the user's own choosing.

The ITEMS modules, two of which have been developed so far, use animated vignettes to explain issues of test score interpretation in a non-technical fashion. The modules include no equations; instead they use graphical displays and realistic examples of test results to present statistical and psychometric concepts. The first module, "What's the Score?" focuses on test score distributions and their properties (mean, median, mode, range, standard deviation), types of test scores (raw scores, percentiles, grade-equivalents), and norm-referenced and criterion-referenced score interpretation. The second, "What Test Scores Do and Don't Tell Us," explains why test scores are not perfectly precise measures of student achievement, focusing on the effect of measurement error on

individual student test scores and on the effect of sample size on the precision of average scores for groups of students.[xv] Navigation features and rewind capabilities ensure that users can revisit content that is still unclear after the first viewing. Each module has an accompanying "handbook" (available online and in paper form), which provides supplementary material and references.

Of course, admission personnel who have not had adequate instruction in educational measurement and statistics can also obtain the information they need by attending workshops and university courses, and by studying measurement textbooks and other resources on their own. It is also important that admissions counselors stay up-to-date on research findings about admissions tests, which is best achieved by subscribing to key journals in the field and by consulting the ACT, College Board, and ETS Web sites, which often feature reports of admissions testing research long before they appear in the journals.

Finally, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), mentioned in an earlier section, should be on the bookshelf of every admission officer. This important (though not always accessible) volume includes explanations of fundamental testing concepts, such as validity, reliability, measurement error, score scales, norms, and test fairness, as well as widely accepted professional guidelines for test use and interpretation. In general, the Standards cautions against over-reliance on test scores, noting that "[t]he improper use of tests…can cause considerable harm to test-takers and other parties affected by test-based decisions" (p. 1).

On a more positive note, the *Standards* also states that, "[a]lthough not all tests are well-developed nor are all testing practices wise and beneficial, there is extensive evidence documenting the effectiveness of well-constructed tests for uses supported by validity evidence. The proper use of tests can result in wiser decisions about individuals and programs than would be the case without their use…" (p. 1).

How, then, can we protect against the improper use and interpretation of tests and promote their wise and beneficial use? One essential step is to provide high-quality training to all test users, including college admission counselors.

The Future of College Admissions Testing

What kind of material will the college admission tests of the future include, how will they be administered, and what will the public think of them?

A longstanding public debate exists between those who believe that college admission tests should be based on classroom learning and those who argue that the tests should be less dependent on learned material and should instead assess "aptitude." Most testing professionals view achievement tests and aptitude tests as endpoints of a continuum,

rather than as qualitatively different types of assessment, and in fact, "aptitude" and "achievement" measures tend to rank students in very similar ways. Nevertheless, the designation of admission tests as assessments of aptitude or achievement does influence public attitudes toward these exams. Despite the evidence that students do not have equal access to high-quality schools, there is a widespread perception today that classroom-based tests are more fair because, in theory, all students have an opportunity to learn the required content. The college admission tests of the future, therefore, are likely to be more focused on material that is taught in the classroom and on skills that clearly resemble those needed for college study.

The proponents of aptitude tests and the advocates of classroom-based exams typically agree on a significant point: An assessment of writing ability should be included in admission tests. Writing assessments can incorporate cognitively complex tasks, and they also have obvious practical applications. Although a valid concern exists about the effect of writing tests on the admissions prospects for students who are not native English speakers, the requirement of a writing assessment as part of the college admission process seems eminently sensible because writing undeniably plays a key role in college-level work. Ideally, using a writing assessment as an admission criterion can help to identify students who need to improve their writing skills before college, allowing an opportunity for remediation. Writing is here to stay as a part of the admission process, and its role is likely to expand.

How about test administration? Within the next 15 years, the ACT and SAT will almost certainly be administered by computer. (Because of public concerns about fairness, among other reasons, they are unlikely to be adaptive tests—those that use a test-taker's responses to previous questions to determine which questions should be administered next.) Computerized testing has many advantages, including convenient scheduling and immediate reporting of provisional scores (at least on machine-scoreable sections). It also facilitates the use of more innovative test questions, better graphical displays, and improved accommodations for test-takers with disabilities. Now that computerized tests are well-established in the higher education admission arena, ACT, Inc. and the College Board are in a position to take advantage of the accumulation of research and practical experience, as well as the establishment of computerized-testing centers. Furthermore, the greater computer familiarity of each successive cohort of high school students should help to allay concerns about the fairness of administering a college admission test via computer.

Will changes in the content and mode of administration reduce the degree of controversy surrounding college admissions tests? In a word, no. These tests are a central component of a mechanism for allocating scarce resources—places at the nation's most prestigious universities. Simply because of that role, college admissions tests will always be a focus of public debate.

References

ACT to stop flagging scores of disabled students who need extra time on test. (2002, August 9). *The Chronicle of Higher Education*, p. A36.

Critics of SAT and ACT hail decline in colleges that use them. (1997, August 8). *The Chronicle of Higher Education*, p. A41.

SAT's better freshman predictor than grades. (1991, January 16). T*he Chronicle of Higher Education*, A35.

The new SAT 2005. (2004). Retrieved December 22, 2004 from http://www.collegeboard.com.

ACT, Inc. (1997). *ACT Assessment Technical Manual*. Iowa City, IA: Author.

ACT, Inc. (1999). *ACT Assessment User Handbook 1999*. Iowa City, IA: Author.

Adelman, C. (1999). *Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment*. Washington, DC: U. S. Department of Education.

Allen, J., & Sconing, J. (2005, August). *Using ACT Assessment scores to set benchmarks for college readiness*. (ACT Research Report 2005-3) Iowa City, IA: ACT, Inc.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Astin, A. W., & Osequera, L. (2002). *Degree attainment rates at American colleges and universities*. Los Angeles: Higher Education Research Institute, Inc.

Astin, A., Tsui, A., & Avalos, J. (1996). *Degree attainment rates at American colleges and universities: Effects of race, gender, and institutional type*. Los Angeles: University of California, Los Angeles, Higher Education Research Institute.

Atkinson, R. (2001, February 18). *Standardized tests and access to American universities*. The 2001 Robert H. Atwell Distinguished Lecture, delivered at the 83rd annual meeting of the American Council on Education, Washington, D.C.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research*, 60, 373-417.

Bleistein, C. A., & Wright, D. J. (1987). Assessment of unexpected differential item difficulty for Asian-American examinees on the Scholastic Aptitude Test. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test* (ETS Research Memorandum No. 87-1). Princeton, NJ: Educational Testing Service.

Bowen, W. G., & Bok, D. (1998). *The shape of the river: Long-term consequences of considering race in college and university admissions*. Princeton, NJ: Princeton University Press.

Breland, H. M. (1998, December). National trends in the use of test scores in college admissions. Paper presented at the National Academy of Sciences Workshop on the Role of Tests in Higher Education Admissions, Washington, DC.

Breland, H. M., Kubota, M. Y., & Bonner, M. W. (1999). *The performance assessment study in writing: Analysis of the SAT II: Writing Subject Test* (College Board Report 99-4). New York: College Entrance Examination Board.

Breland, H., Kubota, M., Nickerson, K., Trapani, C., & Walker, M. (2004). *New SAT writing prompt study: Analyses of group impact and reliability*. (College Board Report 2004-1). New York: College Entrance Examination Board.

Breland, H., Maxey, J., Gernand, R., Cumming, T., & Trapani, C. (March 2002). *Trends in use college admission 2000: A report of a survey of undergraduate admissions policies, practices, and procedures*. (Sponsored by ACT, Inc., Association for Institutional Research, The College Board, Educational Testing Service, and the National Association for College Admission Counseling.) Retrieved October 6, 2003 from http://www.airweb.org.

Brennan, R. L. (1999, July). *A perspective on educational testing: The Iowa testing programs and the legacy of E. F. Lindquist.* Paper Presented to the National Institute for Testing and Evaluation, Jerusalem.

Bridgeman, B., Burton, N., & Cline, F. (2004). *Replacing reasoning tests with achievement tests in university admissions: Does it make a difference?* In R. Zwick (ed.), Rethinking the SAT: The Future of Standardized Testing in University Admissions, pp. 277-288. New York: RoutledgeFalmer.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Prediction of freshman grade-point average from the revised and recentered SAT I: Reasoning Test* (College Board Report 2000-1). New York: College Entrance Examination Board.

Bridgeman, B., Pollack, J., & Burton, N. (2004). *Understanding what SAT Reasoning Test scores add to high school grades: A straightforward approach*. (College Board Research Report No. 2004-4). New York: College Entrance Examination Board.

Bridgeman, B., & Schmitt, A. (1997). *Fairness issues in test development and administration*. In W. W. Willingham & N. Cole, Gender and fair assessment (pp. 185-226). Mahwah, NJ: Lawrence Erlbaum Associates.

Briggs, D. (2001). The effect of admissions test preparation: Evidence from NELS: 88. *Chance*, 14 (1), 10-18.

Briggs, D. C. (2004). Evaluating SAT Coaching: Gains, effects and self-selection. In Zwick, R. (ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions*, pp. 217-233. New York: RoutledgeFalmer.

Brigham, C. C. (1923). *A study of American intelligence*. Princeton, NJ: Princeton University Press.

Burton, E., & Burton, N. W. (1993). The effect of item screening on test scores and test characteristics. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 321-336). Hillsdale, NJ: Lawrence Erlbaum Associates.

Burton, N. W., & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980* (Research Report 2001-2). New York: College Entrance Examination Board.

Cahalan, C., Mandinach, E., & Camara, W. (2002). *Predictive validity of SAT I: Reasoning test for test takers with learning disabilities and extended time accommodations.* (College Board Research Report RR 2002-05). New York: College Entrance Examination Board.

California State Auditor (2000, November). Standardized tests: Although some students may receive extra time on standardized tests that is not deserved, others may not be getting the assistance they need (Summary of Report 2000-108). Sacramento, CA: Bureau of State Audits.

Camara, W. J., & Echternacht, G. (2000, July). *The SAT and high school grades: Utility in predicting success in college* (College Board Research Note RN-10). New York: College Entrance Examination Board.

Cleary, T. A. (1968). Test bias: Prediction of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5 (2), 115-124.

Cole, N. S., & Moss, P.A. (1989). Bias in test use. In Linn, R. L. (ed)., *Educational Measurement* (Third Edition). pp. 201-219. New York: American Council on Education/Macmillan.

Coley, R. J. (2001). *Differences in the gender gap: Comparisons across racial/ethnic groups in education and work* (ETS Policy Information Report). Princeton, NJ: Educational Testing Service.

The College Board (1998, February). *SAT and gender differences* (College Board Research Summary RS-04). New York: The College Board, Office of Research and Development.

The College Board (2005). *The new SAT: A guide for admission officers*. New York: Author.

The College Board and Educational Testing Service (1998). *Admission staff handbook for the SAT program 1998-1999*. Princeton, NJ: Authors.

College Entrance Examination Board (2000). *College-bound seniors 2000*. Retrieved September 1, 2000 from http://www.collegeboard.org.

College Entrance Examination Board (2004). *The SAT Program Handbook 2004-2005*. New York: Author.

College Entrance Examination Board (2005). *2005 College-bound seniors: Total group profile report*. Retrieved November 11, 2005, from http://www.collegeboard.com.

Conant, J. B. (1964). *Shaping educational policy*. New York: McGraw-Hill.

Crouse, J., & Trusheim, D. (1988). The case against the SAT. Chicago: University of Chicago Press.

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (College Board Report 99-1). New York: College Entrance Examination Board.

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: How and why. *Journal of Educational Measurement*, 39, 59-84.

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University*, 73 (2), 24-32.

Dwyer, C. A., & Johnson, L. M. (1997). Grades, accomplishments, and correlates. In W. W. Willingham & N. Cole, *Gender and Fair Assessment* (127-156). Mahwah, NJ: Lawrence Erlbaum Associates.

Educational Testing Service (1999). *Overview: ETS fairness review*. Author. Retrieved August 30, 1999 from http://www.ets.org.

Foster, A. L. (2001, February 23). ETS agrees to alter policy on reporting tests taken under modified conditions. *The Chronicle of Higher Education*, p. A49.

Geiser, S., & Studley, R. (2004). UC and the SAT: Predictive validity and differential impact of the SAT and SAT II at the University of California. In Zwick, R. (ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions*, pp. 125-153. New York: RoutledgeFalmer.

Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Hawkins, D. A., & Lautz, J. (2005, March). *State of college admission*. Alexandria, Virginia: National Association for College Admission Counseling. Retrieved January 27, 2006 from http://www.nacacnet.org.

Hezlett, S. A., Kuncel, N. R., Vey, M., Ahart, A. M., Ones, D. S., Campbell, J. P., & Camara, W. (2001, April). *The effectiveness of the SAT in predicting success early and late in college: A meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Seattle, WA.

Hoover, H. D., & Han, L. (1995, April). *The effect of differential selection on gender differences in college admission test scores*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Humphreys, L. G. (1952). Individual differences. *Annual Review of Psychology*, 3, 131-150.

Jencks, C., & Phillips, M. (Eds.), *The black-white test score gap*. Washington, DC: Brookings Institution Press.

Johnson, V. E. (1997). An alternative to traditional GPA for evaluating student performance. *Statistical Science*, 12 (4), 251-278.

Kaplan, J. (2005). The effectiveness of SAT coaching on math SAT scores. *Chance*, 18, 25-34.

Klitgaard, R. E. (1985). *Choosing elites*. New York: Basic Books.

Kobrin, J. L., Camara, W. J., & Milewski, G. B. (2004). The utility of the SAT I and SAT II for admissions decisions in California and the nation. In Zwick, R. (ed.), *Rethinking*

the SAT: The Future of Standardized Testing in University Admissions, pp. 251-276. New York: RoutledgeFalmer.

Lawrence, I., Rigol, G., Van Essen, T., & Jackson, C. (2004). A historical perspective on the content of the SAT. In Zwick, R. (ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions*, pp. 57-74. New York: RoutledgeFalmer.

Leatherman, C. (2000, December 4). California study finds racial disparities in granting of extra time on SAT. *The Chronicle of Higher Education*, N5.

Lemann, N. (1995, August). The structure of success in America. *The Atlantic Monthly*, pp. 41-60.

Leonard, D., & Jiang, J. (1999). Gender bias and the college prediction of the SATs: A cry of despair. *Research in Higher Education*, 40 (4), 375-408.

Lewis, C., & Willingham, W. W. (1995). *The effects of sample restriction on gender differences* (ETS Research Report 95-13). Princeton, NJ: Educational Testing Service.

Linn, R. L. (1983). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord* (pp. 27-40). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lohman, D. F. (2004). Aptitude for college: The importance of reasoning tests for minority admissions. In R. Zwick (ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions*, pp. 41-55. New York: RoutledgeFalmer.

Mandinach, E. B. (2000, April). Flagging: Policies, perceptions, and practices. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Mandinach, E. B. & Cahalan, C., & Camara, W. J. (2001, April). *The impact of flagging on the admissions process: Policies, practices, and implications*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Manski, C. F., & Wise, S. A. (1983). *College choice in America*. Cambridge, MA: Harvard University Press.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequence of measurement (pg no). In H. Wainer & H. I Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

National Association for College Admission Counseling (2000). Statement on Counselor Competencies. Retrieved January 27, 2006 from http://www.nacacnet.org.

Nettles, A. L., & Nettles, M. T. (1999). Introduction: Issuing the challenge. In A. L. Nettles & M. T. Nettles (Eds.), *Measuring up: Challenges minorities face in educational assessment* (pp. 1-11). Boston: Kluwer Academic.

Noble, J. P. (1991). *Predicting college grades from ACT assessment scores and high school course work and grade information* (ACT Research Report 91-3). Iowa City, IA: American College Testing Program.

Noble, J. (2004). The effects of using ACT composite scores and high school averages on college admissions decisions for ethnic groups. In R. Zwick (ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions*, pp. 303-319. New York: RoutledgeFalmer.

Noble, J., Maxey, J., Ritchie, J., & Habley, W. (2005, October). *Enhancing college student retention: Identification and intervention.* Paper presented at the National Symposium on Student Retention, Dallas.

Norris, D., Oppler, S., Kuang, D., Day, R., & Adams, K. (2005). *The College Board SAT writing validation study: An assessment of the predictive and incremental validity*. Washington, DC: American Institutes for Research.

Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions*. Washington, DC: U.S. Government Printing Office.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pennock-Román, M. (1994). *College major and gender differences in the prediction of college grades* (College Board Report 94-2). New York: College Entrance Examination Board.

Perkhounkova, Y., Noble, J. P., & McLaughlin, G. (2006, Spring). Factors related to persistence of freshman, freshman transfers, and nonfreshman transfer students. *AIR Professional File*, No. 99. The Association for Institutional Research.

Peterson, J. J. (1983). *The Iowa testing programs*. Iowa City, IA: Iowa University Press.

Pitsch, M. (1990, October 10). College Board trustees postpone vote on S.A.T. revision.

Education Week. Retrieved May 22, 2002 from www.edweek.com.

Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement*, 36 (2), 93-118.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Report 93-1). New York: College Entrance Examination Board.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (2001). *Using Achievement Tests/SAT II Subject Tests to demonstrate achievement and predict college grades: Sex, language, ethnic, and parental education groups* (Research Report No. 2001-5). New York: College Entrance Examination Board.

Rigol, G. W. (1997, June). *Common sense about SAT score differences and test validity* (College Board Research Notes RN-01). New York: College Entrance Examination Board.

Rogers, H. J., & Kulick, E. (1987). An investigation of unexpected differences in item performance between Blacks and Whites taking the SAT. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test* (ETS Research Memorandum No. 87-1). Princeton, NJ: Educational Testing Service.

Sawyer, R. L. (1985). *Using demographic information in predicting college freshman grades.* (ACT Research Report No. 87) Iowa City: ACT, Inc.

Schmitt, A. P. (1987). Unexpected differential item performance of Hispanic examinees. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test* (ETS Research Memorandum No. 87-1). Princeton, NJ: Educational Testing Service.

Schmitt, A. P. & Dorans, N. J. (1988). *Differential item functioning for minority examinees on the SAT* (ETS Research Report 88-32). Princeton, NJ: Educational Testing Service.

Scholes, R. J., & McCoy, T. R. (1998, April). The effects of type, length, and content of test preparation activities on ACT assessment scores. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods* (6th ed.). Ames, Iowa: The Iowa State University Press.

Steele, C. M. (1997). A threat in thin air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52 (6), 613-629.

Steele, C. M. (1999, August). Thin ice: "Stereotype threat" and Black college students. *The Atlantic Monthly*. Retrieved September 19, 1999 from http://www.theatlantic.com

Steele, C. M., & Aronson, J. (1998). Stereotype threat and the test performance of academically successful African Americans. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 401-427). Washington, DC: Brookings Institution Press.

Stewart, D. M. (1998, January 25). *Why Hispanic students need to take the SAT*. The College Board. Retrieved April 4, 1999 from http://www.collegeboard.org.

Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in predictions of college grades from Scholastic Aptitude Test scores. *Journal of Educational Psychology*, 85 (4), 710-718.

Stricker, L. J., Rock, D. A., Burton, N. W., Muraki, E., & Jirele, T. J. (1994). Adjusting college grade point average criteria for variations in grading standards: A comparison of methods. *Journal of Applied Psychology*, 79 (2), 178-183.

Turnbull, W. W. (1985). *Student change, program change: Why SAT scores kept falling* (College Board Report 85-2). Princeton, NJ: Educational Testing Service.

Webber, C. (1989). The mandarin mentality: University admissions testing in Europe and Asia. In B. R. Gifford (Ed.), *Test policy and the politics of opportunity allocation: The workplace and the law* (pp. 33-57). Boston: Kluwer Academic.

Weiss, K. R. (2000, January 9). New test-taking skill: Working the system. *The Los Angeles Times*. Retrieved December 20, 2000 from http://www.latimes.com

Willingham, W. W. (1974). Predicting success in graduate education. *Science*, 183, 273-278.

Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability*. New York: College Entrance Examination Board.

Willingham, W. W. (1998, December). Validity in college selection: Context and evidence. Paper presented at the National Academy of Sciences Workshop on the Role of Tests in Higher Education Admissions, Washington, DC.

Willingham, W. W., & Cole, N. (Eds.) (1997). *Gender and Fair Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Willingham, W. W., Pollack, J. M., & Lewis, C. (2000). *Grades and test scores: Accounting for observed differences* (ETS Research Report 00-15). Princeton, NJ: Educational Testing Service.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (1988). *Testing handicapped people*. Boston: Allyn and Bacon, Inc.

Young, J. W. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In Zwick, R. (Ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions*. pp. 289-301. New York: RoutledgeFalmer.

Ziomek, R. L., & Andrews, K. M. (1996). *Predicting the college grade point averages of special-tested students from their ACT assessment scores and high school grades.* (ACT Research Report 96-7). Iowa City, IA: ACT, Inc.

Zwick, R. (2002). *Fair Game? The Use of Standardized Admissions Tests in Higher Education*. New York: RoutledgeFalmer.

Zwick, R. (2004). Is the SAT a "wealth test?" The link between educational achievement and socioeconomic status. In R. Zwick (ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions*, pp. 203-216. New York: RoutledgeFalmer.

Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.), pp. 647-679 Westport, CT: American Council on Education/Praeger.

Zwick, R., Brown, T., & Sklar, J. C. (2004, July). *California and the SAT: A reanalysis of University of California admissions data*. Center for Studies in Higher Education, UC Berkeley, Research and Occasional Papers Series. Retrieved August 10, 2006 from http://cshe.berkeley.edu/publications/rops.htm.

Zwick, R., & Schlemer, L. (2004). SAT validity for linguistic minorities at the University of California, Santa Barbara. *Educational Measurement: Issues and Practice*, 25, 6-16.

Zwick, R., & Sklar, J. C. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language. *American Educational Research Journal*, 42, 439-464.

Zwick, R., Sklar, J. C., & Wakefield, G. (2006, August 8). I*nstructional Tools in Educational Measurement and Statistics (ITEMS) for School Personnel: Evaluation of Two Web-based Training Modules*. Presented at the annual meeting of the American Statistical Association, Seattle.

Notes

[i]Portions of this chapter are adapted from *Fair Game: The Use of Standardized Admis-*

sions Tests in Higher Education, RoutledgeFalmer, 2002. Copyright © 2002 by Rebecca Zwick. See also Zwick (2006). I am grateful to James Maxey, Senior Research Scientist, Special Research Initiatives, ACT, Inc. and to Amy Schmidt, former Executive Director of Higher Education Research, The College Board, for their assistance.

[ii] Historical material on the ACT is based on ACT, Inc. (1999), Brennan, (1999), Haney, Madaus, & Lyons (1993), and Peterson (1983).

[iii] In addition to the tests discussed here, the PSAT/NMSQT (formerly the Preliminary Scholastic Aptitude Test) serves as a practice SAT and is used in awarding National Merit Scholarships, the PLAN assessment (formerly the P-ACT+) is a "pre-ACT" test typically administered to high school sophomores, and the Test of English as a Foreign Language is required of foreign students who attend US colleges or graduate schools.

[iv] Originally, "SAT" stood for "Scholastic Aptitude Test," which was later changed to "Scholastic Assessment Test." The test was then renamed the SAT I: Reasoning Test, and is now the SAT Reasoning Test. "SAT" is no longer considered to be an acronym, but the actual name of the test. The SAT Subject Tests were formerly called the College Board Achievement Tests and later, the SAT II: Subject Tests.

[v] The watchdog organization FairTest claimed in 1997 that this percentage had recently declined ("Critics of SAT and ACT hail decline in colleges that use them," Chronicle of Higher Education, August 8, 1997). As of January 2006, FairTest included over 730 schools on its "College Admissions Test Score Optional List," available at www.fairtest.org. The Website indicates that the list includes institutions that "deemphasize the use of standardized tests by making admissions decisions about substantial numbers of applicants who recently graduated from U.S. high schools without using the SAT I or ACT."

[vi] The 2005 figures were confirmed by Amy Schmidt, The College Board, July 26, 2006, and by James Maxey, ACT, Inc., August 1, 2006. The additional ACT information appears in an unpublished ACT report supplied by James Maxey, October 21, 2005.

[vii] According to ACT, Inc., an updated concordance table between the ACT and SAT Reasoning Test, reflecting recent changes in the tests, will probably not be available until at least 2007 (personal communication, James Maxey, ACT, Inc., October 14, 2005).

[viii] Because respondents to these surveys do not constitute random samples of the corresponding populations of institutions, survey results should be interpreted with caution.

[ix] It is often observed that much of the published research on the validity of admissions tests comes from the testing companies themselves. (For a major recent exception, see Geiser & Studley, 2004) Institutional research offices, particularly those at large universities, often conduct their own validity studies, but these are rarely published. Unless there

is a methodological innovation, or a surprising finding, they are not ordinarily of interest outside the institution. Also, some institutions prefer not to make their research findings publicly available.

[x] No information is expected to be available on the validity of the ACT writing test until 2007 (personal communication, James Maxey, ACT, Inc., October 14, 2005).

[xi] The signs of the logistic regression coefficients are consistent with the signs of the biserial correlations between the predictors and a dichotomous persistence variable (Perkhounkova et al., p. 5) and therefore do not appear to result primarily from multi-collinearity. Julie Noble (personal communication, February 3, 2006) suggests that the reason ACT had a negative coefficient, while high school GPA had a positive coefficient is that GPA incorporates many "noncognitive characteristics … , some of which could be related to second-year retention."

[xii] A more troublesome question about the validity of GPAs, which is rarely investigated, is whether grades reflect biases against particular groups of students. In general, grades are subject to far less scrutiny than tests in this regard.

[xiii] In the DIF methods used by most testing companies, the measure of overall skill is the total score on the test or test section that is under investigation. Although this approach introduces a certain circularity into the process, the test scores are typically the only relevant measures of skill that are available.

[xiv] Also, average score differences between males and females who have been "selected" on the basis of academic criteria are likely to be larger than the differences that exist before the selection criteria are applied. See Hoover & Han, 1995; Lewis & Willingham, 1995. Similar phenomena may apply in the case of ethnic group differences.

[xv] "What's the Score?" and "What Test Scores Do and Don't Tell Us" are freely available at http://items.education.ucsb.edu/. DVD and CD copies are also available on request. Each module has been formally evaluated through administration to more than 100 school personnel and teacher education students. The modules were shown to be effective in increasing assessment literacy, particularly among teacher education students (Zwick, Sklar, & Wakefield, 2006). A third module is now near completion.