High-Stakes Testing in Higher Education and Employment

Appraising the Evidence for Validity and Fairness

Paul R. Sackett, Matthew J. Borneman, and Brian S. Connelly
University of Minnesota, Twin Cities Campus

The authors review criticisms commonly leveled against cognitively loaded tests used for employment and higher education admissions decisions, with a focus on large-scale databases and meta-analytic evidence. They conclude that (a) tests of developed abilities are generally valid for their intended uses in predicting a wide variety of aspects of short-term and long-term academic and job performance, (b) validity is not an artifact of socioeconomic status, (c) coaching is not a major determinant of test performance, (d) tests do not generally exhibit bias by underpredicting the performance of minority group members, and (e) test-taking motivational mechanisms are not major determinants of test performance in these high-stakes settings.

Keywords: employment testing, admissions testing, selection, validity

As young adults complete high school in the United States, they typically pursue one of three options: continue their education, enter the civilian work force, or join the military. In all three settings, there is a long history of using standardized tests of developed cognitive abilities for selection decisions. In these domains, the tests themselves often are very similar. For example, Frey and Detterman (2004) reported a correlation of .82 between scores on the SAT, widely used for college admissions, and a composite score on the Armed Services Vocational Aptitude Battery. Given the similarities in tests, test-taking populations, and questions that commonly arise about test use in these domains, in this article we examine both educational admissions and personnel selection.

Testing is one aspect of the field of psychology with which virtually the entire public comes into contact. As members of the broader community, psychologists, regardless of their area of specialization, are likely to have contact with family members, friends, and neighbors who are asked to take tests as part of the educational admissions or occupational entry processes. In their role as psychologists, they are likely to be asked to comment about a range of testing issues. We believe that there is much myth and hearsay regarding standardized tests and that it is important for psychologists to be aware of the central findings of the testing literature. Thus, in this article, we summarize key findings about a number of criticisms commonly made about testing. In many cases, these claims are contrary to findings considered well established within the testing research community, and they are generally expressed in contexts outside of the scientific literature, such as the popular press. In other cases, they reflect issues still being investigated and debated within the testing community. We attempt to differentiate claims for which there is general agreement within the testing community from claims that are as yet unresolved. For these as yet unresolved claims, we document specific instances of debate and summarize relevant research in the area. We hope both types of information will be helpful to psychologists in responding to questions about test use. We focus on the following set of assertions commonly made about testing:

Assertion 1: Tests predict badly, if at all. Correlations with commonly used criteria (such as first-year college grades) are small, typically in the .25–.35 range. The squared correlation gives the percentage of criterion variance accounted for by the test; thus a correlation of .30 indicates that the test accounts for less than 10% of the variance in the criterion. Given the small amount of variance accounted for by the tests, tests should not play a significant role in high-stakes decisions.

Assertion 2: Tests do not measure all important determinants of all important criteria. Even if tests do predict to the modest degree outlined above, they predict performance in the short term only (e.g., first year grades, performance in job training) and do not predict criteria in the long term, such as earnings, or job/academic success. There are additional determinants of these criteria other than those measured by tests of developed abilities, such as diligence, persistence, energy, and drive.

Assertion 3: Even if tests have some predictive value, they are

Paul R. Sackett, Matthew J. Borneman, and Brian S. Connelly, Department of Psychology, University of Minnesota, Twin Cities Campus.

Full disclosure of interests: Paul R. Sackett has research grants from and serves on the SAT Psychometric Panel of the College Board and serves on the Research Advisory Committee of Psychological Services, Inc. Matthew J. Borneman is an employee of PreVisor, an employment testing firm. Brian S. Connelly has a student research grant from the College Board.

Correspondence concerning this article should be addressed to Paul R. Sackett, Department of Psychology, University of Minnesota, Twin Cities Campus, Elliott Hall, 75 East River Road, Minneapolis, MN 55455. E-mail: psackett@umn.edu

Vol. 63, No. 4, 215–227 DOI: 10.1037/0003-066X.63.4.215

Paul R. Sackett

valuable only for screening out individuals with low scores. Above a certain threshold, higher scores do not matter; thus, it is inappropriate for colleges or employers to use test scores to differentiate among those above this threshold.

Assertion 4: Tests serve merely as a proxy for wealth and privilege; they reflect socioeconomic status (SES) rather than developed abilities. Any predictive power that tests appear to have disappears once one controls for SES.

Assertion 5: Tests are readily coached; those with knowledge of this fact and the financial resources to pay for coaching programs can substantially increase their scores.

Assertion 6: Tests are biased against members of racial and ethnic minority groups, and sometimes against women as well, as evidenced by the common finding of substantially lower mean scores for minority groups.

Assertion 7: While minority group members obtain lower mean scores on tests, they perform just as well as majority group members once admitted or hired.

Assertion 8: Motivational mechanisms, such as stereotype threat, explain majority–minority group mean differences.

We respond to each of these assertions, presenting what we view as the most compelling data on each. We focus on meta-analytic syntheses and/or large, nationally representative samples whenever these are available. Individual small-sample studies are prone to features that can distort findings, such as random sampling idiosyncrasies (cf. Hunter & Schmidt, 2004). Consequently, there is no doubt that readers can locate individual studies with findings contrary to large-sample findings. However, we believe that an overview of the findings of large-scale studies and meta-analytic syntheses will give the clearest picture of the cumulative body of knowledge on the set of issues outlined above.

Assertion 1: Tests Predict Badly

Prototypically, admissions tests correlate about .35 with first-year grade point average (GPA), and employment tests correlate about .35 with job training performance and about .25 with performance on the job. One reaction to these findings is to square these correlations to obtain the variance accounted for by the test (.25 accounts for 6.25%; .35 accounts for 12.25%) and to question the appropriateness of giving tests substantial weight in selection or admissions decisions given these small values (e.g., Sternberg, Wagner, Williams, & Horvath, 1995; Vasquez & Jones, 2006).

One response to this reaction is to note that even if the values above were accurate (and we make the case below that they are, in fact, substantial underestimates), correlations of such magnitude are of more value than critics recognize. As long ago as 1928, Hull criticized the small percentage of variance accounted for by commonly used tests. In response, a number of scholars developed alternate metrics designed to be more readily interpretable than "percentage of variance accounted for" (Lawshe, Bolda, & Auclair, 1958; Taylor & Russell, 1939). Lawshe et al. (1958) tabled the percentage of test takers in each test score quintile (e.g., top 20%, next 20%, etc.) who met a set standard of success (e.g., being an above-average performer on the job or in school). A test correlating .30 with performance can be expected to result in 67% of those in the top test quintile being above-average performers (i.e., 2 to 1 odds of success) and 33% of those in the bottom quintile being above-average performers (i.e., 1 to 2 odds of success). Converting correlations to differences in odds of success results both in a readily interpretable metric and in a positive picture of the value of a test that "only" accounts for 9% of the variance in performance. Subsequent researchers have developed more elaborate models of test utility (e.g., Boudreau & Rynes, 1985; Brogden, 1946, 1949; Cronbach & Gleser, 1965; Murphy, 1986) that make similar points about the substantial value of tests with validities of the magnitude commonly observed. In short, there is a long history of expressing the value of a test in a metric more readily interpretable than percentage of variance accounted for.

Another response to the practice of squaring and derogating observed validity coefficients is to note that observed validity coefficients are typically substantial underestimates of the operational validity of a test. There are two key truths, well-known to test researchers but often either not recognized or rejected by test critics: first, that studies of highly select samples underestimate validity by restricting the range of observed scores and, second, that unreliable criterion measures result in an underestimation of the true operational validity. We discuss each of these in turn.

Range Restriction Leads to Underestimates of Validity

Consider an employer hiring only individuals with test scores above the 50th percentile. The sample of selected individuals will have a smaller standard deviation of test scores than the standard deviation in the full applicant pool,

Matthew J. Borneman

resulting in a lower correlation in the restricted sample than would be found in the full applicant pool. For example, a correlation of .50 in an unrestricted sample drops to .33 when only the top 50% of the sample have been included (as might be expected in selection or admissions settings, where the top 50% of scorers are selected).

The primary strategy for dealing with this problem of range restriction is the use of psychometric formulas to estimate the correlation in the population of interest (see Sackett & Yang, 2000, for a summary). The most accurate corrections can be made when one knows the precise mechanism by which restriction occurred (e.g., direct restriction due to use of the test for selection vs. indirect restriction due to use of another predictor correlated with the test as the basis for selection). However, if the researcher has information about the test standard deviation in the population of interest (e.g., the applicant pool) as well as in the restricted sample, estimates of the correlation in the population of interest can be made; Hunter, Schmidt, and Le (2006) showed that such corrections may be conservative underestimates. The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) endorses the use of these corrections. Note that range restriction is not limited to the realm of psychometrics; rather, it is an important issue in statistics and economics under the labels "sample selection bias" and "censoring of data" (e.g., Berk, 1983; Heckman, 1979).

Failure to take range restriction into account can dramatically distort research findings. One common scenario where this takes place is in the comparison of the predictive value of two measures, one of which was used for selection, and hence is range restricted, and the other of which was not. For example, Duckworth and Seligman (2005) were interested in the relative impact of self-discipline and IQ on a variety of indices of academic performance. Because standardized IQ scores were used to select the sample, the IQ measure was range restricted. The study's abstract states that self-discipline accounted for more than twice as much variance in each of six outcomes than did IQ. That conclusion, however, was based on observed correlations and did not take range restriction into account. It is interesting that Duckworth and Seligman acknowledged the range restriction issue and documented the degree of restriction on the IQ measure in their discussion. They applied a range restriction correction to one of the six outcome measures (GPA) and reported that whereas the corrected IQ–GPA correlation (.49) was larger than the uncorrected value (.32), it remained lower than the self-discipline–GPA correlation (.67). Although this is true, note that their conclusion (self-discipline accounts for more than twice as much variance as IQ) no longer holds after one takes range restriction into account. In addition, we applied range restriction corrections to other outcomes; in the case of predicting procrastination (as measured by the time homework was begun), for example, IQ had a higher correlation (−.28 corrected, −.18 observed) with procrastination than did self-discipline (−.26) after we corrected for range restriction, which is clearly at odds with the authors' conclusion. Thus, whereas Duckworth and Seligman (2005) made a strong case for the value of studying self-discipline as an additional predictor of academic outcomes, a clear picture of the relative value of self-discipline and IQ requires careful attention to range restriction. In selection settings, the question of interest is generally one of estimating a value for an applicant pool, and thus range restriction corrections are often an essential aspect of evaluating selection systems.

Unreliable Criterion Measures Underestimate Validity

The second key truth is that unreliable criterion measures result in an underestimation of validity. Consider a group of call center customer service representatives, each rated by a group of supervisors who unobtrusively listen in on a sample of phone calls. All else being equal, ratings based on a sampling of 100 calls will be more reliable than ratings based on a sampling of 10 calls. Psychometric theory shows that as the reliability of the criterion measure decreases, the observed correlation of the test with the criterion decreases. Clearly, however, the true validity of the test as a predictor of the construct of job performance is not lowered because a researcher chooses to use the less reliable 10-call sample as the performance measure instead of the more reliable 100-call sample. Thus, it is important to ensure that conclusions about the validity of the test are not obscured by the use of unreliable performance measures. Well-known psychometric theory offers a procedure for correcting observed validity coefficients for unreliability in performance measures. Flaws in criterion measures other than unreliability can also distort validity estimates if they

Brian S. Connelly

add sources of error that are not related to the construct measured by the test.

We offer two concrete examples of ways in which flawed performance measures affect validity. In employment settings, the most prevalent method of evaluating job performance is the use of supervisory ratings. The meta-analytic mean correlation between ratings of overall performance by two raters is .52 (Viswesvaran, Ones, & Schmidt, 1996); thus, .52 is the reliability of a performance rating made by a single supervisor. The Spearman–Brown formula can be used to estimate the reliability that would be obtained if a composite of 2 (.68), 3 (.76), 10 (.92), or any number of raters is used. Conceptually, correcting an observed validity coefficient for criterion unreliability can be thought of as estimating the correlation that would be expected between the test and a composite performance rating from an infinite number of raters. The greater the number of raters, the smaller the impact of any individual rater's idiosyncrasies. Note that this approach views raters as randomly selected from a population of potential raters. Newly selected employees may be assigned to any of a number of supervisors, and the supervisor in subsequent years may be different than the supervisor in the first year. So, the view of a supervisor as a random draw from a population of potential supervisors is a reasonable one.

In educational settings, the rater idiosyncrasy problem is overcome by averaging across multiple courses, thus minimizing grader effects. However, courses differ in difficulty, and two students of equal ability may obtain very different GPAs if one consistently chooses easy courses while another consistently chooses difficult courses. Failure to take this into account will result in an underestimate of validity (Stricker, Rock, Burton, Muraki, & Jirele, 1994).

Illustrating Correcting for Range Restriction and for Criterion Deficiencies

Berry and Sackett (2008) looked at SAT–grade relationships for over 165,000 students at 41 colleges. They found a correlation of .35 between observed SAT combined scores (Verbal + Math) and first-year GPA. These data were range restricted, as the colleges used the SAT as part of the admissions process. Correcting for range restriction resulted in a corrected correlation of .47. Berry and Sackett had access to individual course grades and thus were able to compute separate validity coefficients for each course at each college, which resulted in a meta-analysis of over 147,000 course-specific validity coefficients. Although validity estimates based on first-year GPA suffer from course selection problems, estimates based on individual course grades, by definition, do not (i.e., course difficulty is held constant, as all students are taking the same course). From the individual course data, Berry and Sackett were able to estimate the correlation that would be obtained between the SAT and GPA if all students took a common set of courses (i.e., course difficulty held constant): That correlation was .55. Thus, .55 is our best estimate of the relationship between the SAT and academic performance in this large sample; squaring this correlation gives us .30, instead of .12, which was obtained by squaring the observed correlation. Although there is no doubt that there is criterion variance unexplained by the SAT, we think these findings frame the search for additional predictors as a search for supplements to a strong predictor, rather than as a replacement for a weak predictor.

Meta-Analytic Findings for Ability–Performance Relationships

A number of researchers have presented meta-analytic syntheses of relationships between ability and academic performance, job training performance, and on-the-job performance. Ones and Dilchert (2004) reviewed and integrated these various meta-analyses and offered the following summary of validity findings taking range restriction and criterion unreliability into account (we added the Berry & Sackett, 2008, findings to theirs):

$r = .48$: SAT Verbal + Math with GPA, with no correction for course difficulty ($N = 1.2$ million)

$r = .55$: SAT Verbal + Math with GPA, with correction for course difficulty ($N = 165{,}000$; Berry & Sackett, 2008)

$r = .68$: general cognitive ability with performance in job training ($N = 1.1$ million)

$r = .47$: general cognitive ability with job performance ($N = 300{,}000$).

We offer a number of observations about these summary findings. First, these mean correlations are substantially larger than the observed correlations in the .20s and .30s. Taking range restriction and criterion deficiencies into account presents a very different picture of test–criterion relationships. Second, note the relative similarity between predicting GPA in an educational setting and predicting performance in training in the employment setting ($r = .55$ vs. $r = .68$). Conceptually, the settings are similar: Ability tests are used to predict knowledge acquisition in samples of young adults.

In short, if one's goal is either to make a projection about what will happen in the future if the test is used with a new applicant pool or to contribute to scientific understanding of how the test in question functions, the question of interest is "What is the correlation in the population of interest, using sound performance measures?" Based on these large-scale studies and meta-analyses, these population correlations suggest that these tests do not predict badly; in contrast, these validity coefficients are quite strong.

Assertion 2: Tests Do Not Measure All Important Determinants of All Important Criteria

There are a number of differing versions of this assertion, which we review in sequence:

Tests Predict Only Short-Term Criteria

The assertion here is that tests have predictive power only for first-year grades or in the early stages of employment. However, a large body of research produces findings contrary to this assertion. Regarding the belief that admissions tests predict only first-year grades, admissions tests have been shown to be predictive of grades throughout schooling. Vey et al. (2003) provided meta-analytic evidence that the SAT predicts first-, second-, third-, and fourth-year grades about equally well, with only slight decrements over time. In addition, admissions tests predict other learning outcomes besides course grades. Kuncel and Hezlett (2007) summarized findings relating a number of admissions tests to outcomes beyond first-year grades, including meta-analytic evidence that the Graduate Record Examination (GRE) predicts comprehensive examination performance ($r = .40$, corrected for range restriction), faculty ratings of student performance ($r = .50$), and subsequent citation counts ($r = .23$; Kuncel, Hezlett, & Ones, 2001). Wightman (1998) found that Law School Admission Test (LSAT) scores are predictive of passing the bar examination ($r = .30$, without correction for range restriction). Lubinski, Benbow, Webb, and Bleske-Rechek (2006) reported that the SAT predicts getting a doctoral degree, getting tenure, and getting patents in a gifted sample. Thus, these meta-analyses and large-scale samples provide strong support for the idea that the usefulness of these test scores in academic settings is not limited to predicting first-year grades; indeed, these scores are predictive of a variety of long-term indicators of academic and career success.

In employment settings, there is strong evidence of the predictive power of tests of developed abilities beyond the short term. Armor and Roll (1994) examined scores on the Armed Forces Qualification Test as predictors of soldier performance in samples of soldiers in their first, second, third, and fourth years of a tour of duty. In Year 1, test scores clearly predicted performance levels. Individuals in the lowest category of test scores did indeed show higher performance over time; by Year 4, they achieved the same mean level of performance as was achieved in Year 1 by those in the highest category of test scores. However, by Year 4, individuals in the high test score category also improved, such that the Year 1 gap between high- and low-scoring individuals remained through Year 4. Although low scorers improved over time, so did high scorers; thus, the predictive power of the test remained.

Large-sample evidence for long-term stability of validity in a civilian work environment comes from a study by Farrell and McDaniel (2001), who examined test–performance correlations for intervals ranging from the first six months to 10 years of employment in an occupation for a multi-occupational sample of 6,449 individuals in low-complexity jobs and 2,555 individuals in high-complexity jobs. Although validity was higher in high-complexity jobs, for both high- and low-complexity jobs, test validity was relatively constant through the 10-year period. In short, there is extensive evidence from educational, military, and civilian employment settings against the position that tests predict only short-term outcomes.

Tests Do Not Predict Earnings

One criticism of tests posits lifetime earnings as the criterion of real interest and asserts that test scores bear little relationship to earnings. However, there are indeed relationships between measures of developed abilities and earnings in the nationally representative data set of the National Longitudinal Study of Youth (Johnson & Neal, 1998) and in W. G. Bowen and Bok's (1998) study of students in 28 academically selective colleges. More critically, though, we caution against the premise that earnings are the criterion of real interest. We note that students pursue different careers with differing wage trajectories. When one talented student chooses to be a high school teacher and a second equally able student chooses medical school, the wage trajectories of the two are quite different; however, it is hard to argue that this somehow reflects a failure of a test (i.e., same test score, differing earnings). We suggest that meaningful evaluations of tests should focus on individuals pursuing the same objectives, a strategy almost universally used in employment testing. Correlations between test scores and measures of job performance are computed for individuals applying for the same job, thus addressing the question "Among individuals who present themselves as candidates for this job, how well can the test predict their relative performance?"

Some Relevant Criteria Are Not Predicted Well by Tests of Developed Abilities

We agree that there are criteria of interest to colleges and universities and to work organizations for which tests of developed abilities are not strong predictors. On conceptual grounds, tests of developed abilities have the potential to be useful predictors of criteria for which those abilities are relevant, such as academic performance and performance on work tasks with a cognitive component. There is no reason to expect such tests to be predictive of criteria such as whether a student will or will not be satisfied with the social environment at college or whether an employee will or will not go out of his or her way to be helpful to coworkers.

Empirical findings support these conceptual expectations. For example, the Kuncel et al. (2001) meta-analysis of relationships between the GRE and various outcomes reports a range-restriction-corrected correlation of .41 with first-year grades but a much smaller corrected correlation of .22 with degree completion. The decision to leave a graduate program is often a function of motivation or of a shift in interests: Although a test of developed ability may be a good predictor of whether someone is capable of completing a graduate program, it will not predict whether someone is interested in doing so or willing to do so. In the work domain, Rotundo and Sackett (2002) reviewed and integrated a number of perspectives on the dimensionality of job performance, and they offered task performance (i.e., performance of core required job activities), citizenship performance (i.e., behaviors contributing to the organization's social and psychological environment), and counterproductive work behavior as the three major domains of job performance. Cognitively loaded predictors are generally the strongest correlates of task performance, and noncognitive predictors are generally the best predictors in the citizenship and counterproductive behavior domains (e.g, McHenry, Hough, Toquam, Hanson, & Ashforth, 1990).

Thus, schools and firms may indeed value criteria not predicted well, if at all, by tests of developed abilities. Such a situation clearly calls for research into predictors of these additional criteria and inclusion of these new predictors in the selection system. The interest in predicting multiple criteria (e.g., fit to the college's social and cultural environment as well as academic performance) is one reason why colleges commonly use personal statements and letters of recommendation in addition to predictors more directly linked to the prediction of academic performance. Similarly, employers may use noncognitive tests (e.g., of traits such as conscientiousness) to predict criteria such as citizenship or counterproductivity.

However, it is not clear that valuing additional criteria is a basis for criticism of the use of tests of developed abilities in situations where the interest in these additional criteria leads to the use of a multifaceted selection system aimed at these multiple criteria. Such criticism would be warranted if these new criteria were the only ones valued by the organization. But we doubt that schools really are indifferent to academic performance even if they identify, say, degree completion as another criterion of interest. It is hard to imagine a school saying, "We are indifferent as to whether we have a class of predominantly C students or a class with many A and B students." In short, it is common for there to be multiple criteria of interest and for selection systems to be multifaceted. There is no reason to expect all aspects of the selection system to be related to all criteria.

Important Determinants of Traditional Criteria Are Missed

The concerns above were with failing to consider some additional outcome criteria when evaluating tests. Here the focus is on failing to consider all important determinants of the criteria traditionally used in evaluating tests (e.g., the prediction of grades in an academic setting or the prediction of task performance in work settings). The assertion here is that there are additional determinants of these outcomes beyond developed abilities and thus that the tests of developed abilities are flawed. The first half of this assertion is certainly true; there is extensive evidence for the relationship between a variety of measures outside the domain of developed abilities and grades and task performance. For example, in the employment context, a variety of measures, including self-report measures of personality, situational judgment tests, and scored biographical inventories, have been found predictive of job performance (see Schmidt & Hunter, 1998, for a summary of meta-analytic evidence for a wide array of measures in the employment domain). In the educational domain, Crede and Kuncel (2006) presented meta-analytic evidence that study skills show substantial relationships with academic performance. The added finding that study skills are essentially uncorrelated with measures of developed abilities means that study skills have incremental validity over ability measures. In a similar vein, Robbins et al. (2004) presented meta-analytic evidence that achievement motivation and academic self-efficacy have incremental validity over admissions tests and high school GPA. Thus, there is evidence that constructs in the noncognitive domain have the potential for improved prediction.

The assertion that a test is flawed because it does not measure all determinants of the criterion of interest reflects more of a rhetorical device than a serious argument: Any measure can be criticized because of what it is not (e.g., this ruler is flawed because it only measures height, not weight). As noted in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCMA, 1999), a test is appropriately evaluated in terms of how well it meets the claims made for it. It is not reasonable to fault the test for (a) not measuring characteristics other than those it purports to measure or (b) not being the sole and complete determinant of job or educational success.

Tests are typically a part of broader selection systems. Selection systems typically incorporate multiple measures (e.g., test-based measures in the cognitive and noncognitive domains, or records of past job or educational experiences). Selection systems commonly attempt to include measures of the major known relevant, measurable, operationally valid, administratively feasible, and economically affordable determinants of the identified criteria of interest. Assertions that all relevant determinants of the outcomes of interest must be included in a selection system may be countered with responses such as (a) valid measures of the trait of interest have yet to be developed, (b) measures not vulnerable to compromise by commercial coaching programs have yet to be developed, (c) operational use of a measure is not administratively feasible (e.g., face-to-face interviews with thousands of geographically scattered candidates are not feasible), (d) operational use is too costly (e.g., a complex simulation costing thousands of dollars per candidate is not feasible to administer to all applicants to a large state university), or (e) the organization has made a strategic decision that the characteristic in question will be acquired via training after selection (e.g., a decision to test for prior knowledge of a particular piece of computer software vs. to provide training in the use of the software after selection). In sum, while typically incorporating multiple measures, selection systems do commonly exclude some performance determinants, and investigations into the feasibility of broadening selection systems in various ways are common.

Assertion 3: Tests Do Not Predict Above a Minimum Threshold

The assertion here is that while tests may have value in screening out those with very low levels of ability, increments in ability do not lead to increments in performance for those above a threshold. As such, it is not appropriate to prefer high-scoring individuals over lower-scoring individuals once this threshold is reached. However, there is strong evidence that higher test scores are associated with higher criterion scores throughout the test score range.

In the employment setting, Coward and Sackett (1990) examined 174 studies (N = 36,000) of relationships between scores on a cognitive composite from the General Aptitude Test Battery and supervisor ratings of job performance. Curvilinear models did not out-predict linear models at greater than chance levels. In an educational setting, Arneson (2007) examined relationships between test scores and college GPA in three large data sets (i.e., SAT–grade relationships in a College Board sample of over 165,000 students from 41 colleges and universities, the National Educational Longitudinal Study of 1988, and Project TALENT) and found monotonic test–grade relationships throughout the entire score range in each data set. Cullen, Hardison, and Sackett (2004) found no departure from linearity in SAT–grade relationships in a large sample of over 40,000 students. Overall, these findings show that throughout the entire test score distribution, one would expect higher performance levels for any increase on the predictor test.

Assertion 4: The Appearance of Validity Is Due to SES

A common assertion is that test scores are influenced by socioeconomic factors, such as parents' earnings and education. The stronger versions of this criticism assert that tests measure nothing more than SES. This criticism is mainly leveled at educational admissions tests, as in Guinier's claim that "in the interest of truth in advertising, the SAT should simply be called a wealth test" (as cited in Zwick, 2002b, p. 311), Kohn's (2001) claim that "the SAT merely measures the size of students' houses" (p. B12), and Colvin's (1997) statement that the "only thing the SAT predicts well now is socioeconomic status" (p. B2).

Crosby, Iyer, Clayton, and Downing (2003) stated that "it has now been documented with massive data sets from the University of California that SAT I scores lose any ability to predict freshman year grades if the regression analyses control for socioeconomic status" (p. 100); Biernat (2003) made a similar claim. Both of these *American Psychologist* articles relied on the same source for their claim (Geiser & Studley, 2001). That work did not, in fact, focus on the validity of SAT I when controlling for SES, but rather on the incremental validity of SAT I scores over SAT II scores when controlling for SES (Johnson, 2004; Zwick, Brown, & Sklar, 2004). To address the issue of whether SES was, in fact, responsible for the validity of tests of developed abilities in predicting academic performance, Sackett, Kuncel, and colleagues sought out multiple sources of data, which permitted partialing out the effect of SES from the test–grade correlation (Arneson, Waters, Sackett, Kuncel, & Cooper, 2006; Cooper, Kuncel, Sackett, Arneson, & Waters, 2006; Sackett, Kuncel, Arneson, Cooper, & Waters, 2007).

These researchers summarized results from eight data sets, including a meta-analysis, a College Board SAT data set on over 165,000 students from 41 colleges, data on all individuals entering an accredited law school in 1991, and three large longitudinal studies following samples of high school students through college. They found that SES is indeed related to test scores. In broad unrestricted populations, this correlation is quite substantial (i.e., r = .42 among the population of SAT takers). Consistent with our earlier discussion of range restriction, it is considerably smaller in restricted samples (i.e., rs = .15–.20 among samples of students enrolled in a single institution). Second, test scores are indeed predictive of academic performance, as indexed by grades. Observed correlations in samples of admitted students average about .35; applying range restriction corrections to estimate the validity for school-specific applicant pools results in an estimate of .47 as the operational validity. Third, the test–grade relationship is not an artifact of common influences of SES on both test scores and grades. Partialing SES from the above estimate of the operational validity of tests (r = .47) reduces the estimated validity to .44. The assertion that the predictive power of tests disappears once the effects of SES are removed is at odds with the findings from these multiple sources of data.

Thus while the claim that tests measure "nothing but" SES is not supported, there clearly is a substantial relationship between SES and test scores. However, the relationship between test scores and academic performance is affected only to a small degree by controlling for the effects of SES on both test scores and grades. Thus it is not the case that test validity is an artifact of SES (i.e., that high SES artificially raises both test scores and grades, resulting in the appearance of a correlation between test scores and grades). Rather, SES is linked to the development of abilities that are predictive of academic performance.

These findings highlight important social issues, as egalitarian values are at odds with the findings that SES is linked to higher levels of important developed abilities. Better understanding of this linkage is important. We note the recent issuance of the report of APA's Task Force on Socioeconomic Status as an important compilation of research findings on the meaning and impact of SES (Saegert et al., 2006).

Assertion 5: Tests Are Readily Coached

In the educational admissions domain, multiple firms offer test preparation services, accompanied by claims that large gains can be expected. Major test preparation firms claim average score gains of 120–140 points on the SAT (about 0.70–0.80 *SD*); in fact, one firm guarantees students will increase their SAT scores by 200 points (Princeton Review, 2006). Zwick (2002a) offered a useful critique of research designs that serve as the basis for such claims. These claims of large gains are at odds with the findings of a large body of research. Becker (1990) presented a meta-analysis of the coachability of the SAT. Published studies with a control group showed a mean coaching effect of 9 points on the Verbal section and 16 on the Math section. One large-scale study done by Educational Testing Service researchers since Becker's review produced virtually identical findings (Powers & Rock, 1999). Research on other admissions tests (e.g., LSAT, GRE, ACT) is more sparse, but the findings are consistent with those for the SAT (Zwick, 2002a). We do note some students improve by sizable amounts upon retesting even if not participating in a coaching program. About 15% of such students in the Powers and Rock (1999) study improved by 100 or more points on a given section of the test. Thus, it is not the case that sizable score change cannot occur; the issue is the amount of change that can be expected that is due to coaching. We also note that there has been a general increase in access to coaching, as books with preparation advice and practice tests abound, and test producers are increasingly offering access to online coaching. As Zwick (2002a) noted, ready access to free coaching for all is the best antidote to claims that access to coaching gives unfair advantage to those who can afford it. In sum, coaching does appear to produce small mean improvements in test scores, though gains in controlled studies are far smaller than those claimed by test preparation firms. As access to free or inexpensive preparation materials becomes more readily available, the test preparation landscape will change, and continued studies will be valuable.

Assertion 6: Lower Minority Group Mean Scores Show That Tests Are Biased

This assertion consists of two parts. The first is that mean scores are lower for certain minority groups. This has long been known to be true; we offer details below. The second is that these mean differences can be interpreted as evidence of bias in the tests; this inference is unequivocally rejected within mainstream psychology.

Mean differences on tests of developed abilities of roughly 1.0 *SD* between Whites and African Americans and of roughly 0.67 *SD* between Whites and Hispanics have been consistently reported. The largest scale summary of this literature is a meta-analysis by Roth, Bevier, Bobko, Switzer, and Tyler (2001). Regarding the African American–White mean difference, they reported large-scale meta-analytic mean *d* values (with sample sizes of at least several hundred thousand, except as noted) of 0.99 for the SAT, 1.02 for the ACT, 1.34 for the GRE, 0.99 for employment tests of general ability, and 1.10 for military tests of general ability. Regarding the Hispanic–White mean difference, they reported meta-analytic mean *d* values of 0.77 for the SAT, 0.56 for the ACT, 0.72 for the GRE, 0.58 for employment tests of general ability ($N = 6{,}133$), and 0.85 for military tests of general ability.

Mean differences by gender are much smaller. The general finding is differences around 0.10–0.25 *SD* favoring women on measures of verbal ability and differences of similar magnitude favoring men in quantitative ability. However, there is variability across studies and among subtests and item types within the verbal and quantitative domains (Hedges & Nowell, 1995; Hyde, Fennema, & Lamon, 1990; Hyde & Linn, 1988; Willingham & Cole, 1997).

We offer several observations about these mean differences. Particularly with regard to race and ethnicity, the differences are of a magnitude that can result in substantial differences in selection or admission rates if the test is used as the basis for decisions. Employers and educational institutions wanting to benefit from the predictive validity of these tests but also interested in the diversity of a workforce or an entering class encounter the tension between these validity and diversity objectives. A wide array of approaches have been investigated as potential mechanisms for addressing this validity–diversity trade-off; see Sackett and Wilk (1994) and Sackett, Schmitt, Ellingson, and Kabin (2001) for reviews.

We also note the tendency to equate the presence of a mean difference between two groups as signaling test bias. We believe the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) offers several valuable insights here. One is the observation that "the idea that fairness requires equality in overall passing rates for different groups has been almost entirely repudiated in the professional testing literature" (p. 74). Groups may differ in experience, in opportunity, or in interest in a particular domain; absent additional information, one cannot determine whether mean differences reflect true differences in the developed ability being measured or bias in the measurement of the ability. Thus, the *Standards* states that "most testing professionals would probably agree that while group differences in testing outcomes should in many cases trigger heightened scrutiny for possible sources of test bias, outcome differences across groups do not in themselves indicate that a testing application is biased or unfair" (p. 75). The extensive literature on attempts to understand and reduce group differences reflects this suggested high degree of scrutiny (cf. Sackett et al., 2001). Finally, the *Standards* presents the key idea that has driven the examination of test bias, namely, that "a more widely accepted view (of fairness) would be that examinees of equal standing with regard to the construct the test is intended to measure should on average earn the same test score regardless of group membership" (p. 74).

A recent exception to the above position was offered by Helms (2006), who argued that a correlation between a group-related variable (e.g., racial identity) and a test reflects construct-irrelevant variance that must be removed in order to yield fair test scores. The problem is that such a correlation may reflect construct-irrelevant variance (e.g., racial identity might be found to interfere with a test-taker's ability to show his or her true understanding of math) or it may reflect construct-relevant variance (e.g., racial identity might be found to be associated with undervaluing academic achievement and investing less time in one's studies). It is precisely because of these potential alternative explanations that the dominant view in the testing field rejects the position that a finding of a relationship between race (or, in Helms's model, a race-related variable) and test scores can be directly interpreted as signaling bias or unfairness. See Griffore (2007) and Newman, Hanges and Outtz (2007) for additional comments on Helms's position.

The notion that if a test is unbiased, a given score should have the same meaning regardless of group membership is reflected in the widely used Cleary (1968) model of test bias. This model examines regression lines relating test scores to criterion performance to determine whether a given test score predicts the same level of criterion performance for each group. If, for example, the regression equation relating scores on an admission test to college GPA shows that the mean GPA for White students with a verbal score of 500 is 2.7, one asks whether the mean GPA for Black students with a score of 500 is also 2.7. If it were the case that the test was biased against Black students, then their measured score of 500 would be lower than the score they would obtain on an unbiased test. Thus their score would "underpredict" their GPA: Their "true" score is greater than 500, and thus Black students with a score of 500 might subsequently obtain, say, a mean GPA of 2.9.

There is extensive evidence on the question of whether tests underpredict minority group performance. Examining race and ethnicity in the educational admissions domain, Linn (1973) reviewed 22 studies, and Young (2001) reviewed 28 studies since Linn's review. Both concluded that the consistent finding is overprediction (the predicted GPA is higher than the actual obtained GPA), rather than underprediction, for Black and Hispanic students; Young reported mixed evidence, but potentially a small amount of overprediction, for Asian American students. Findings for Blacks and Hispanics in the employment domain parallelled those in educational admissions (Bartlett, Bobko, Mosier, & Hannan, 1978; Hartigan & Wigdor, 1989).

In contrast, underprediction of women's performance is commonly found in the academic domain at the undergraduate level, with women obtaining GPAs about 0.1 point higher than would be predicted (Leonard & Jiang, 1999). A sizable portion of this gender difference has been found to reflect differences in the course-taking patterns of men and women (Ramist, Lewis, & McCamley, 1990). An additional portion has been linked to differences in study habits such that women devote more effort to their academic work than men with comparable test scores (Stricker, Rock, & Burton, 1993). Our assessment is that underprediction for women is minimal when course-taking and study-habit differences are taken into account. Furthermore, most admissions systems include at least a partial measure of study habits—namely, previous GPA. Thus, although this underprediction may be present for the admissions test, this effect is abated when combined with other predictors in a well-developed admissions system.

Assertion 7: Minority Group Performance Matches Majority Group Performance

The section above documented the differences in mean test performance. The assertion here is that these differences in test performance do not correspond to differences in job or academic performance: Minority group members perform just as well as majority group members once admitted or hired.

In the domain of gender differences in academic performance, Mau and Lynn (2001) found in a nationally representative sample of nearly 10,000 students that women tended to have higher GPAs than men ($d = 0.30$). In the employment domain, C. C. Bowen, Swim, and Jacobs (2000) meta-analyzed gender differences in job performance and found a 0.05-*SD* difference in favor of women when they controlled for key demographic variables; Chang (1993) found a 0.02-*SD* difference in favor of men in a larger meta-analysis. Thus, there are near-zero gender differences in job performance.

In the domain of race, we analyzed a data set gathered by the College Board on three entering cohorts at 41 colleges and universities and found that Whites tended to have higher GPAs than Blacks (mean $d = 0.39$ across schools) and Hispanics (mean $d = 0.31$). In the employment domain, a meta-analysis by McKay and McDaniel (2006) reported a mean White–Black difference in measures of overall job performance of 0.35 *SD* (0.46 when corrected for error of measurement) that was based on 302 samples. Similar differences were found for objective (e.g., production records) versus subjective (i.e., supervisor ratings) measures. There are limited data on other groups; Roth, Honeycutt, and Bobko (2003) were able to locate only 11 samples permitting a White–Hispanic comparison, and their findings are quite varied (a difference of 0.04 *SD* for ratings, but a difference of 0.67 *SD* for objective measures.)

The common finding is that there are racial and ethnic group mean differences in academic and job performance measures. Note that this is consistent with the findings in the prior section on the examination of over- and underprediction of minority group performance: If this assertion that test differences by group did not correspond to performance differences by group were true, the result would be underprediction of minority group performance. The finding that minority group performance is not systematically underpredicted implies that there are indeed performance differences.

Note that the magnitude of group differences on the test and the magnitude of group differences on the measure of performance are not the same (e.g., White–Black test $d = 1.0$; White–Black job performance $d = 0.35$). An intuitive notion may be that the two should be the same and the fact that they are not signals bias in the test. We note that there is generally no reason to expect the two to be comparable unless the test in question is the sole determinant of the outcome of interest. Consider a setting where there are three determinants of job performance: (a) the ability measured by the test, on which group differences exist; (b) aspects of the work settings, such as the degree of mentoring and support from one's supervisor and peers, on which no group differences exist and which are not knowable at time of hire; and (c) conscientiousness, a test of which shows no group differences. Because there are multiple influences on job performance, which differ in the degree of group difference, there is no reason to expect the mean group difference in job performance to correspond to the mean difference on any single one of the determinants (ability, job situation, or conscientiousness). This is at the heart of the basis for rejecting proposed definitions of test bias that rest on comparability of test and criterion differences (such as that of Thorndike, 1971).

Jencks (1998) offered a perspective that moves from examining bias in a single test to what he termed *selection system bias*. Such bias occurs when a selection system relies on measures of some valid predictors but ignores other valid predictors on which smaller group differences are found. For example, note that in the scenario above there are two valid and measurable predictors, namely, ability and conscientiousness. Sole reliance on the ability test as the basis for selection, even if it were a perfectly unbiased measure of ability, would constitute selection system bias given that a valid and available measure with no group difference was excluded from the system.

Jencks (1998) posited, though, that a selection system is biased whenever group mean differences on the composite selection system used are larger than differences on the criterion. However, although his conceptual discussion focuses on missing important predictors that could be included as part of the selection system, group differences on the criterion are affected not only by individual-difference predictors that might be included in a selection system but also by aspects of the workplace or the school setting not knowable at the time when selection decisions are made. Our example above illustrates this, as social support received from peers is a determinant of performance. As the group of peers with whom an individual will choose to associate at work or at school is not knowable at time of selection, we cannot agree with labeling a selection system biased for excluding performance determinants that it could not possibly include. Thus we are in accord with Jencks's call for including other measureable and valid measures in selection systems but not with his proposed method of identifying the presence of selection system bias.

Assertion 8: Motivational Factors Explain Group Differences

One proposed alternative to the position that group mean differences reflect real differences in the developed abilities in question is that they are artifacts of differences in test-taking motivation. Various literatures address this issue. In the employment domain, measures of test-taking motivation have been developed (Arvey, Strickland, Drauden, & Martin, 1990). Arvey et al. (1990) reported lower test-taking motivation for Black test takers and a modest reduction in group differences on tests when motivation was statistically controlled. It is unclear, however, whether these motivational factors are causes of test performance or the by-products of individuals' test-taking histories.

This uncertainty about causality is overcome in a more recent approach to motivational issues. Steele and Aronson (1995) put forward the theory of stereotype threat, which holds that awareness of the common finding of lower mean test scores for members of their group leads to concerns on the part of minority group members as to whether their own test performance might confirm this stereotype; this experienced threat leads to a reduction in test performance. Steele and Aronson made use of an experimental paradigm in which majority and minority students took the same test under one of two conditions. In the first, the test was labeled an intelligence test, with the intent that this label would lead to the experience of stereotype threat; in the second, the test was described as a new experimental task developed by the researcher, with the intent that this innocuous description would not be threatening. Steele and Aronson found that minority students performed better in the nonthreat condition than in the threat condition and suggested that stereotype threat may play a role in group differences on high-stakes tests. See Steele, Spencer, and Aronson (2002) for a review of the quite substantial literature on this topic, and see Nguyen (2006) for a meta-analysis of threat effects in laboratory settings.

Two lines of evidence argue strongly, in our opinion, against the notion that motivational mechanisms, such as stereotype threat, are the prime determinants of group differences in high-stakes testing contexts. The first is the evidence on over- and underprediction, reviewed earlier. Cullen et al. (2004) argued that if motivational mechanisms result in minority group members obtaining observed scores lower than their true scores, then the performance of minority group members should be underpredicted. The failure to find underprediction by race and ethnicity argues against such motivational effects of test scores. The second is an observation put forward by Zwick (2005) that if motivational mechanisms had sizable effects on scores on high-stakes tests, we should see differences in the magnitude of majority–minority mean differences on high-stakes tests versus less threatening low-stakes tests. Yet she noted that the magnitude of differences does not vary systematically across testing contexts. Thus, while motivational issues certainly merit attention, and while they may play a large role in the testing outcomes of individual test takers, the evidence to date does not support the position that motivational mechanisms are a prime explanatory factor for group mean differences in high-stakes testing settings.

Discussion

In this article, we have summarized evidence regarding eight common critical assertions made about the tests of developed abilities that play a significant role in selection systems in employment and educational admissions contexts. While each of these assertions, if true, would constitute a serious challenge to the continued use of such tests, our conclusion is that none of them are supported by the preponderance of the evidence. The notion of preponderance of evidence is central to our position. Again, we have focused on large-scale studies, on national samples, and on meta-analytic syntheses of the literature in drawing our conclusions. Given the extensive research in this area, it is certainly the case that individual studies can be found with contrary findings. We suggest, though, that readers should be wary of attempts to draw broad conclusions from small studies and that the broader literature should be used as the basis for conclusions.

We offer a very positive appraisal of the evidence (a) that tests of developed abilities are generally valid for their intended uses in predicting a wide variety of aspects of short-term and long-term academic and job performance, (b) that validity is not an artifact of SES, (c) that coaching is not a major determinant of test performance, (d) that tests do not generally exhibit bias by underpredicting the performance of minority group members, and (e) that test-taking motivational mechanisms are not major determinants of test performance in these high-stakes settings.

This positive appraisal must be hedged with two caveats. First, there is the persistent finding of subgroup mean differences on developed ability measures. The evidence reviewed here that tests generally do not systematically underpredict minority group performance is important—it directs us to factors other than biased measurement as major determinants of group differences. But many businesses and many educational institutions value a diverse workforce or entering class, and group differences affect the pursuit of diversity. Identifying causes of group differences and developing effective interventions to reduce group differences are among the most pressing contemporary social issues.

Second, businesses and universities may be interested in criteria other than those predicted by tests of developed abilities. For example, when task performance is the outcome of interest, higher test scores will on average result in higher performance; this relation may not hold for other outcomes though. Some employment testing programs view very high scores as undesirable in applicants for jobs of low complexity, arguing that the job will not be sufficiently challenging for such high-ability individuals, who may become restless and leave. Should empirical data in a given setting support this argument, an organization might make a case for this practice. Thus, although research has answered the question "Will high scorers generally perform better than low scorers?" with an affirmative, the question "Should high scorers always be preferred to low scorers?" is one to be answered by test users rather than test creators. It is important to differentiate between technical questions, such as those about how well and under what conditions tests predict various criteria, and values-based questions, such as those about the relative importance of one criterion versus another. Whether an organization chooses to emphasize, say, maximizing mean task performance rather than minimizing turnover as the goal of its selection system is a matter of values.

Although our overall assessment is positive, we emphasize that we are not offering a blanket endorsement of any and all tests. There are certainly bad tests (e.g., tests offered by developers with no evidence of validity or reliability, or any of the other desirable attributes outlined in the *Standards for Educational and Psychological Testing*), and there are good tests used inappropriately (e.g., tests with strong validity evidence for one type of usage put to a different use for which there is no supporting evidence). We do, though, concur with the opening sentences of the *Standards*: "Educational and psychological testing and assessment are among the most important contributions of behavioral science to our society . . . . There is extensive evidence documenting the effectiveness of well-constructed tests for uses supported by validity evidence" (AERA, APA, & NCME, 1999, p. 1).

REFERENCES

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

Armor, D. J., & Roll, C. R., Jr. (1994). Military manpower quality: Past, present, and future. In B. F. Green & A. S. Mavor (Eds.), *Modeling cost and performance for military enlistment: Report of a workshop* (pp. 13–34). Washington, DC: National Academies Press.

Arneson, J. J. (2007). *An examination of the linearity of ability–performance relationships among high-scoring applicants.* Unpublished doctoral dissertation, University of Minnesota, Twin Cities Campus.

Arneson, J., Waters, S. D., Sackett, P. R., Kuncel, N. R., & Cooper, S. R. (2006, May). *A meta-analytic investigation of the role of SES in ability test–grade relationships*. Poster session presented at the annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695–716.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31,* 233–241.

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

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review, 48,* 386–398.

Berry, C. M., & Sackett, P. R. (2008, March). *The validity of the SAT at the individual course level.* Paper presented at the American Educational Research Association Conference, New York, NY.

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist, 58,* 1019–1027.

Boudreau, J. W., & Rynes, S. L. (1985). Role of recruitment in staffing utility analysis. *Journal of Applied Psychology, 70,* 354–366.

Bowen, C. C., Swim, J. K., & Jacobs, R. R. (2000). Evaluating gender biases on actual job performance of real people: A meta-analysis. *Journal of Applied Social Psychology, 30,* 2194–2215.

Bowen, W. G., & Bok, D. (1998). *The shape of the river.* Princeton, NJ: Princeton University Press.

Brogden, H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology, 37,* 65–76.

Brogden, H. E. (1949). A new coefficient: Application to biserial correlation and to estimation of selective efficiency. *Psychometrika, 14,* 169–182.

Chang, C. C. (1993). *Gender and performance appraisals in work settings: A meta-analysis.* Unpublished doctoral dissertation, Pennsylvania State University.

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

Colvin, R. L. (1997, October 1). Q & A: Should UC do away with the SAT? *Los Angeles Times,* p. B2.

Cooper, S. R., Kuncel, N. R., Sackett, P. R., Arneson, J., & Waters, S. D. (2006, May). *The role of SES in the ability–performance relationship: Results from national longitudinal studies.* Poster session presented at the annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Coward, W. M., & Sackett, P. R. (1990). Linearity in ability–performance relationships: A reconfirmation. *Journal of Applied Psychology, 75,* 297–300.

Crede, M., & Kuncel, N. R. (2006). *Study habits, study skills, and study attitudes: A meta-analysis of their relationship to academic performance among college students.* Manuscript submitted for publication.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Crosby, F. J., Iyer, A., Clayton, S., & Downing, R. A. (2003). Affirmative action: Psychological data and the policy debates. *American Psychologist, 58,* 93–115.

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT–grade and ability–job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89,* 220–230.

Duckworth, A. L., & Seligman, M. E. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science, 16,* 939–944.

Farrell, J. N., & McDaniel, M. A. (2001). The stability of validity coefficients over time: Ackerman's model and the General Aptitude Test Battery. *Journal of Applied Psychology, 86,* 60–79.

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science, 15,* 373–378.

Geiser, S., & Studley, R. (2001). *UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California.* Retrieved July 8, 2006, from http://www.ucop.edu/sas/research/researchandplanning/pdf/sat_study.pdf

Griffore, R. J. (2007). Speaking of fairness in testing. *American Psychologist, 62,* 1081–1082.

Hartigan, J., & Wigdor, A. K. (1989). *Fairness in employment testing.* Washington DC: National Academies Press.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47,* 153–161.

Hedges, L. V., & Nowell, A. (1995). Sex difference in mental test scores, variability, and number of high scoring individuals. *Science, 269,* 41–45.

Helms, J. E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist, 61,* 845–859.

Hull, C. L. (1928). *Aptitude testing.* Yonkers-on-Hudson, NY: World Book.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Thousand Oaks, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 93,* 594–612.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107,* 139–155.

Hyde, J. S., & Linn, M. C. (1988). Gender difference in verbal ability. *Psychological Bulletin, 104,* 53–69.

Jencks, C. (1998). Racial bias in testing. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 55–85). Washington, DC: Brookings Institute Press.

Johnson, J. W. (2004). Not all affirmative action rewards merit. *American Psychologist, 59,* 123–124.

Johnson, W. R., & Neal, D. (1998). Basic earnings and the Black–White earnings gap. In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 480–497). Washington DC: Brookings Institute Press.

Kohn, A. (2001, March 9). Two cheers for an end to the SAT. *Chronicle of Higher Education,* p. B12.

Kuncel, N. R., & Hezlett, S. A. (2007, February 23). Standardized tests predict graduate students' success. *Science, 315,* 1080–1081.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127,* 162–181.

Lawshe, C. H., Bolda, R. L., & Auclair, G. (1958). Expectancy charts III: Their theoretical development. *Personnel Psychology, 11,* 545–599.

Leonard, D. K., & Jiang, J. M. (1999). Gender bias and the college predictions of the SATs: A cry of despair. *Research in Higher Education, 40,* 375–407.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43,* 139–161.

Lubinski, D., Benbow, C. P., Webb, R. M., & Bleske-Rechek, A. (2006). Tracking exceptional human capital over two decades. *Psychological Science, 17,* 194–199.

Mau, W. C., & Lynn, R. (2001). Gender differences on the Scholastic Aptitude Test, the American College Test, and college grades. *Educational Psychology, 21,* 133–136.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43,* 335–354.

McKay, P. F., & McDaniel, M. A. (2006). A re-examination of Black–White differences in work performance: More data, more moderators. *Journal of Applied Psychology, 91,* 538–554.

Murphy, K. R. (1986). When your top choice turns you down: Effect of rejected offers on the utility of selection tests. *Psychological Bulletin, 99,* 133–138.

Newman, D. A., Hanges, P. J., & Outtz, J. L. (2007). Racial group and test fairness, considering history and construct validity. *American Psychologist, 62,* 1082–1083.

Nguyen, H. H. D. (2006). *Does stereotype threat differentially affect cognitive ability test performance of minorities and women? A meta-analytic review of experimental evidence.* Unpublished doctoral dissertation: Michigan State University.

Ones, D. S., & Dilchert, S. (2004, October). *Practical vs. general intelligence in predicting success in work and educational settings.* Paper presented at the University of Amsterdam.

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

Princeton Review. (2006). "Our guarantee to you." Retrieved July 20, 2006, from http://www.princetonreview.com/college/testprep/testprep.asp?TPRPAGE=575&TYPE=NEW-SAT-PREPARE

Ramist, L., Lewis, C., & McCamley, L. (1990). Implications of using freshman GPA as the criterion for the predictive validity of the SAT. In W. W. Willingham, C. Lewis, R. Morgan, & L. Ramist (Eds.), *Predicting college grades: An analysis of institutional trends over two decades* (pp. 253–288). Princeton, NJ: Educational Testing Service.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130,* 261–288.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54,* 297–330.

Roth, P. L., Honeycutt, A. I., & Bobko, P. (2003). Ethnic differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88,* 694–706.

Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance for supervisor ratings of overall performance: A policy capturing study. *Journal of Applied Psychology, 87,* 66–80.

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2007). *Socio-economic status and the relationship between the SAT and freshman GPA: An analysis of data from 41 colleges and universities* (Tech. Rep. 2007–5). New York: The College Board.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56,* 302–318.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49,* 929–954.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112–118.

Saegert, S. C., Adler, N. E., Bullock, H. E., Cauce, A. M., Liu, W. M., & Wyche, K. F. (2007). *Report of the APA Task Force on Socioeconomic Status.* Washington, DC: American Psychological Association.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications from over 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811.

Steele, C., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379–440). San Diego, CA: Academic Press.

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist, 50,* 912–927.

Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in prediction of college grades from Scholastic Aptitude Test scores. *Journal of Educational Psychology, 85,* 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,* 178–183.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical validity of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8,* 63–70.

Vasquez, M. J. T., & Jones, J. M. (2006). Increasing the number of psychologists of color: Public policy issues for affirmative diversity. *American Psychologist, 61,* 132–142.

Vey, M. A., Ones, D. S., Hezlett, S. A., Kuncel, N. R., Vanelli, J. R., Briggs, K. H., & Campbell, J. P. (2003, April). *Relationships among college grade indices: A meta-analysis examining temporal influences.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557–574.

Wightman, L. F. (1998). LSAC National Longitudinal Bar Passage Study. Retrieved August 7, 2006, from University of North Carolina, Chapel Hill, The Educational Diversity Project Web site: http://www.unc.edu/edp/pdf/NLBPS.pdf

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Erlbaum.

Young, J. W. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (College Board Research Report No. 2001–6). New York: College Board.

Zwick, R. (2002a). *Fair game? The use of standardized admissions tests in higher education.* New York: Routledge Falmer.

Zwick, R. (2002b). Is the SAT a “wealth test”? *Phi Delta Kappan, 84,* 307–311.

Zwick, R. (2005, April) *Evaluating the association between socioeconomic status and educational achievement measures: The role of methodological and societal factors.* Paper presented at the annual meeting of the National Council for Measurement in Education, Montreal, Quebec, Canada.

Zwick, R., Brown, T., & Sklar, J. C. (2004, July). *California and the SAT: A reanalysis of University of California admissions data* (Research and Occasional Papers Series, Paper CSHE-8–04). Berkeley: Center for Studies in Higher Education, University of California. (Available at http://repositories.cdlib.org/cshe)