Differential Validity, Differential Prediction, and College Admission Testing: A Comprehensive Review and Analysis

John W. Young with the assistance of Jennifer L. Kobrin

Differential Validity, Differential Prediction, and College Admission Testing: A Comprehensive Review and Analysis

John W. Young with the assistance of Jennifer L. Kobrin

John W. Young is an associate professor of Educational Statistics and Measurement and the director of Research and Development at the Graduate School of Education at Rutgers University in New Brunswick, New Jersey. He received his Ph. D. in educational research with a specialization in psychometrics from Stanford University in 1989. He is the recipient of the 1999 Early Career Contribution Award from the American Educational Research Association's Committee on the Role and Status of Minorities in Educational Research and Development for his research on the academic achievement of minority students.

Jennifer L. Kobrin is an assistant research scientist with the College Board. She received her Ed. D. in educational statistics and measurement from Rutgers University in 2000. She was a finalist for the 2001 outstanding dissertation award from the National Council on Measurement in Education and the recipient of the 2001 best dissertation award from the Graduate School of Education at Rutgers University.

The College Board: Expanding College Opportunity

The College Board is a national nonprofit membership association dedicated to preparing, inspiring, and connecting students to college and opportunity. Founded in 1900, the association is composed of more than 3,900 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three million students and their parents, 22,000 high schools, and 3,500 colleges, through major programs and services in college admission, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT™, the Advanced Placement Program® (AP®), and Pacesetter®. The College Board is committed to the principles of equity and excellence, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, contact www.collegeboard.com

Acknowledgments

Dedication

For Carol and all our little friends.

Contents

Figures

Abstract

This research report is a review and analysis of all of the published studies during the past 25+ years (since 1974) in the area of differential validity/prediction and college admission testing. More specifically, this report includes 49 separate studies of differences in validity and/or prediction for different racial/ethnic groups and/or for men and women. All of the studies that were reviewed originated as journal articles, book chapters, conference papers, or research/technical reports. The breadth of studies range from single-institution studies based on a single cohort of several hundred students to large-scale compilations of results across hundreds of institutions that included several thousand students in all. The typical research design in these studies used first-year grade point average (FGPA) as the criterion and test scores (usually SAT® scores) and high school grades as predictor variables in a multiple regression analysis. Correlation coefficients were also usually reported as evidence of predictive validity.

The main contribution of this report is contained in sections 3 and 4 with a focus on racial/ethnic differences and on sex differences, respectively. With regard to racial/ethnic differences, the minority groups that have been studied include Asian Americans, blacks/African Americans, Hispanics, and Native Americans. Some studies used a combined sample of minority students that was usually composed primarily of African American and Hispanic students. Overall, there was no common pattern to the results for validity and prediction for the different minority groups. Correlations between predictors and criterion were different for each minority group with generally lower values (for both blacks/African Americans and Hispanics) or similar values (for Asian Americans) when compared to whites. Too few studies of Native Americans or of combined samples of minority students are available to reliably determine typical validity coefficients for these groups. In terms of grade prediction, the common finding was one of overprediction of college grades for all of the minority groups (except for Asian Americans), although the magnitude differed for each group. With Asian American students, studies that employed grade adjustment methods found that underprediction of grades occurred.

With respect to sex differences, the correlations between predictors and criterion were generally higher for women than for men. In terms of prediction, the typical finding in these studies was that women's college grades were underpredicted. However, in the most selective universities, the correlations for men and women appear to be equal, while the degree of underprediction for women's grades appears to be somewhat less than in other institutions. Compared to earlier research on this topic, sex differences in validity and prediction appear to have persisted, although the magnitude of the differences seems to have lessened.

The concluding section of the report provides a summary of the results, states several conclusions that can be drawn from the research reviewed, and postulates a number of different avenues for further research on differential validity/prediction that could yield useful additional information on this important and timely topic.

I. Introduction

For any educational or psychological test, the validity of the instrument for its intended purposes should be the primary consideration for users of that test. However, questions regarding test validity often yield complex answers. In particular, given populations of examinees that differ on important demographic variables such as race, ethnicity, sex, or socioeconomic status, is the validity of the test invariant across groups? This topic of research, commonly referred to as differential validity, has gained greater prominence, as the composition of examinee pools has become increasingly diverse.

Research on the validity of test scores for selection purposes in higher education has been conducted over several decades. More recently, within the past 30 years, the study of possible differences in test validity for different groups of examinees has gained momentum because of demographic changes that have altered test-taking populations, making them more heterogeneous. Based on this research, some of the findings appear to be more definitive, while other findings are still tentative, often due to small samples and the lack of replication studies.

Test validation is a complicated undertaking that relies on both logical arguments and empirical support. Validity is not an inherent fixed characteristic of any test; instead, validity must be established for each test usage for all populations of interest. The original conception of test validity was one of a trinity of facets: content, criterion-related (which subsumes concurrent and predictive), and construct (American Psychological Association, 1954, 1966). In the field of educational measurement, the present consensus is that all test validation is a form of construct validation (see, e.g., American Psychological Association, 1999). The writings of Messick (1989) and Shepard (1993) are the best examples by way of explanation of this line of reasoning. At present, a unified validity framework can be constructed so as to obtain the four-fold classification

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/Utility
Consequential Basis	Value Implications	Social Consequences

Figure 1. Messick's Facets of Validity Framework.

shown in Figure 1 above (Messick, 1980, 1989). Empirical test validation, as reported in this report, would fall into the top left cell as a form of construct validity because it constitutes one form of evidence for the proper interpretation of test scores.

For historical and scientific reasons, the most common approach used to validate an admission test for educational selection has been through the computation of validity coefficients and regression lines. Validity coefficients are the computed correlation coefficients between predictor variables and criterion variables. By choosing an appropriate criterion (or outcome measure), the predictive validity of a selection test can be determined. A large correlation indicates high predictability from the test to the criterion; however, a large correlation by itself does not satisfy all facets required of test validity.

A cautionary note about the interpretation of validity coefficients is in order. Because these coefficients are usually calculated on only those individuals who are selected for admission, the resulting values are based on a restricted (or censored) distribution of test scores. Since admission decisions are based to some degree on test performance, the validity coefficients obtained are generally substantially lower than what would be expected from an unrestricted population. Using validity coefficients as the main indicator for evaluating the utility of selection tests is a practice that may underestimate the true test validity and is not supported in the literature (see Cronbach and Gleser, 1965). However, validity coefficients can still be useful as a basis for comparative inferences across populations (Wainer, Saka, and Donoghue, 1993).

College Admission Testing

One of the major uses in the United States of educational tests is for selection into higher education. Not all institutions require test scores for admission; however, the large majority of four-year colleges and universities that have admission requirements do. The primary tests for undergraduate admission are ACT's Assessment Program tests of educational development and the College Board's SAT (formerly known as the Scholastic Aptitude Test and the Scholastic Assessment Test). In 1996, the American College Testing Program's corporate name was formally changed to ACT. The ACT tests originated in 1959, while the forerunner to the SAT dates back to 1926. Until 1994, this latter test was called the Scholastic Aptitude Test.

The ACT Assessment reports four subtest scores: in English, Mathematics, Reading, and Science Reasoning, as well as a Composite score. The ACT tests are curriculum-based exams that measure educational development in the four areas represented by the scores. SAT I: Reasoning Test, the admission testing component of the SAT, measures academic aptitude and reports two test scores: a verbal score and a mathematical score. Over the years, both the ACT and the SAT have changed considerably in both content and item format. The SAT has separate achievement tests in specific subject areas, presently called SAT II: Subject Tests, that are also used in admission by some institutions. SAT I is the largest admission testing program in the country, with current annual testing volume of over 1.3 million examinees (College Board, 1999). SAT I is taken by 43 percent of U.S. high school graduates and by students in more than 100 foreign countries. The total across all components of the SAT testing program, including SAT I, SAT II, and the Advanced Placement Program® (AP®) Exams, were 2.2 million students in 1997-98. ACT's volume is almost as large, with over 900,000 students tested annually (ACT, 1997). Most institutions will generally accept scores from either testing program for admission purposes.

Until the early 1960s, the demographic and socioeconomic backgrounds of SAT test-takers were relatively homogeneous. As a result of societal changes, including the civil rights movement of the 1960s and the women's movement of the 1970s, higher education became more accessible to broad segments of the population that had been previously denied this opportunity. More recently, due to shifting immigration patterns and the greater demand for college-educated workers, as well as the implementation of affirmative action and need-based financial aid policies, the degree of racial, ethnic, and linguistic diversity in the backgrounds of college students is greater than ever before.

This increased diversity is also reflected in the demographic characteristics of students who now take the ACT or the SAT. The self-reported sex and racial/ethnic composition of the examinee populations is shown in Figure 2. It is apparent that the diversity of students who currently take one of the college admission tests is greater than at any time previously (ACT, 1997; College Board, 1999).

Since 1964, the College Board has offered its Validity Study Service (VSS), administered by the Educational Testing Service (ETS), to its member institutions. In 1998, VSS was replaced by the Admitted Class Evaluation Service™ (ACES™). This ongoing service enables each college or university to conduct its own internal validity

	ACT Examinees 1995-96	SAT Examinees 1997-98	SAT Examinees 1987-88
Women	56%	54%	52%
Men	44	46	48
African Americans	9	11	9
Asian Americans	3	9	6
Hispanics	5	8	5
Native Americans	1	1	1
Whites	71	67	77
Others	2	4	1

Figure 2. Percentage of examinees by demographic groups.

studies on the admission process and to determine the relationship of SAT scores and high school grades to first-year college grades. Studies conducted through the VSS and ACES comprise the majority of the information on the predictive validity of the SAT in individual institutions (Willingham, 1990). The results from these numerous studies have been documented by Schrader (1971), Ford and Campos (1977), and Ramist (1984). In a similar fashion, validity studies on ACT scores are conducted with the assistance of ACT's Prediction Research Service (American College Testing Program, 1987; ACT, 1997). Many of the findings regarding differential validity and differential prediction are based on these institutional validity studies. In addition, a separate body of work on these topics resulted from investigations carried out by independent researchers.

Some Basic Terms and Concepts

Before proceeding further, a glossary of commonly used terms and concepts is necessary:

- **Correlation Coefficient:** a statistical index of the linear relationship between two variables or measures. Coefficients range from –1.00 to +1.00 with values near zero indicating no relationship and values far away from zero indicating a strong relationship; positive correlations mean that high values on both variables occur jointly while negative correlations mean an inverse relationship exists between the variables. In test validity studies, correlation coefficients between a predictor and a criterion are often called validity coefficients. The value of a particular validity coefficient can be spuriously altered by factors such as restriction of range and/or unreliability in one or both variables.

- **Criterion:** an outcome or dependent variable or test score. In institutional validity studies, the criterion most frequently used is the first-year college grade point average (see FGPA following). Other criteria used include cumulative college grade point average and completion of a degree.

- **Predictor:** an independent variable or test score used to forecast or to predict a criterion. In institutional validity studies, the most commonly used predictors are one or more test scores and high school grade point average (see HSGPA following). Typically, the predictor scores are temporally available before the criterion scores.

- **Prediction Equation:** the resulting equation obtained from a linear regression analysis with a single criterion and one or more predictors computed from a sample of students.

- **Predictive Validity:** one of the aspects of test validity as originally defined by the American Psychological Association. Most commonly used to describe the relationship between a predictor such as a test score and a later criterion such as a grade point average.

- **Race/Ethnicity:** one of the classification variables (the other being sex) used in differential validity studies to identify groups of examinees. The principal populations of interest are African Americans, Asian Americans, Hispanics, Mexican Americans, and whites. There are few studies involving Native Americans due to the lack of samples of adequate size.

- **Asian American/Pacific Islander:** the term currently used for federal race classification. In validity studies, Asian Americans include individuals with origins from any Asian country unless separately identified. Oriental is an older and outdated term.

- **Black/African American:** terms often used interchangeably in the literature. Black is the term currently used for federal race classification, although African American is the preferred usage.

- **Chicano/Mexican American:** Chicano is the term commonly used in California, although Mexican American appears to be the preferred term elsewhere.

- **Hispanic:** the term currently used for federal race classification but actually refers to ethnic origin and can apply to a person of any race. In validity studies, Hispanics include Cuban Americans, Mexican Americans, Puerto Ricans, and other Hispanics unless separately identified.

- **Anglo/White:** Anglo is the term commonly used in validity studies to describe white populations when compared to Chicanos or Mexican Americans. White (or Caucasian) is the term commonly used in comparisons with all other race groups.

- **SAT M:** SAT mathematical, the test section or the score.

- **SAT V:** SAT verbal, the test section or the score.

- **ACT:** American College Testing Assessment Program, the tests or the scores.

- **HSGPA:** high school grade point average.

- **HSR:** high school rank in class.

- **ICG:** individual course grade.

- **QGPA:** first-quarter college grade point average.

- **SGPA:** first-semester college grade point average.

- **FGPA:** first-year college grade point average.

- **CGPA:** cumulative college grade point average.

- **Differential Validity:** refers to a finding where the computed validity coefficients are significantly different for different groups of examinees.

- **Differential Prediction:** refers to a finding where the best prediction equations and/or the standard errors of estimate are significantly different for different groups of examinees.

- **Over/Underprediction:** refers to a comparative finding where the use of a common prediction equation yields significantly different results for different groups of examinees. More specifically, overprediction means that the residuals (computed as actual GPA minus predicted GPA) from a prediction equation based on a pooled sample are generally negative for a specific group, and underprediction means that the residuals are generally positive. The use of these terms is only meaningful when comparing the results of two or more groups. Overprediction and underprediction are sometimes collectively referred to as misprediction. Note that in some studies, residuals were defined differently, but the results reported in this report used the standard definition as given here.

Significance of Differential Validity

It is important to distinguish between differential validity and differential prediction, two terms that are commonly used in the literature. As described by Linn (1978), differential validity refers to differences in the magnitude of the correlation coefficients for different groups of test-takers, and differential prediction refers to differences in the best-fitting regression lines or in the standard errors of estimate between groups of examinees. Differences in regression lines are measured as differences in the slopes and/or intercepts. Comparing standard errors of estimate is preferable to comparing

correlations because any differences are directly related to differences in the degree of predictability. Differential validity and differential prediction are obviously related but are not identical issues. In any validity study encompassing two or more groups, differential validity can and does occur independently of differential prediction. Of the two issues, differential prediction is the more crucial because differences in prediction have a more direct bearing on considerations of fairness in selection than do differences in correlation (Linn, 1982a, 1982b).

In addition to questions of a psychometric nature, differential validity as a topic of research is important because it has relevance for the issues of test bias and fair test use. Bias can be best conceptualized in the manner described by Shepard (1982) as "invalidity, something that distorts the meaning of test results for some groups" (p. 26). Although fairness is a social rather than a technical concept, judgments about whether a test is fair to all examinees necessarily involve reference to the psychometric properties of the test and how the scores are used. Thus, a test that is differentially valid for different groups of examinees may be used in a manner that is consistently unfair to certain groups of examinees.

Research on differential validity has a history spanning over six decades with published reports of sex differences in the prediction of college grades dating back to the 1930s (Abelson, 1952). Originally, the term differential validity encompassed both differential validity and differential prediction. In the 1960s, differential validity became a topic of wide research interest due to racial differences in observed test validity. Theories about validity differences between groups took one of two forms: single-group validity and differential validity (see, for example, Boehm, 1972). Single-group validity means that a test is valid for one group (usually whites) but is invalid (that is, has zero validity) for other groups (typically members of minority groups). Differential validity refers to a situation where a test is predictive for all groups but to different degrees. Single-group validity has been shown to be a special case of differential validity (Hunter and Schmidt, 1978; Linn, 1978).

In the 1970s, as more evidence became available, the existence of differential validity was called into question. Schmidt, Berner, and Hunter (1973) challenged the notion of differential validity, describing it as a "pseudo-problem," and discounted reports of its existence as the result of Type I errors or the incorrect use of statistical procedures. Currently, there is a divergence of opinions about the pervasiveness of differential validity, depending on whether the tests in question are used in educational or employment settings. For example, numerous authors have documented the existence of differential validity for admission tests (e.g., Linn, 1990; Young,

1993). In contrast, no support was found for differential validity in employment tests between whites and blacks in an analysis of 39 studies by Hunter, Schmidt, and Hunter (1979) or between whites and Hispanics in an analysis of 16 studies by Schmidt, Pearlman, and Hunter (1980). Furthermore, the Society for Industrial and Organizational Psychology (SIOP), in its 1987 Principles for Validation and Use of Personnel Selection Procedures, discounted the notion of differential prediction for major ethnic groups (SIOP, 1987).

It should be noted that differences across institutions, majors, courses, and instructors may moderate the findings relative to differential validity and differential prediction in higher education. A comprehensive review of methods developed to adjust for grading differences is given in Young (1993). When these factors are not accounted for, as is true in most differential validity/ prediction studies, the results are spuriously confounded. In those studies where these factors are taken into account, the results are often substantially different. Any interpretation of differential validity/prediction results must bear this point in mind. For example, several studies of sex differences in validity and prediction have found conflicting results depending on whether adjustments have been applied to course grades (see Elliott and Strenta, 1988; Young, 1991a). Any results that were reported based on grade adjustment methods are included for the studies reviewed in this report.

In general, the presumption of differential validity is considered more tenable for educational tests (particularly those used for selection in undergraduate admission) than tests used for personnel identification and selection in the military and the private sector. Given the many unanswered questions about differential validity, its root causes and its impacts, it is not surprising that the topic continues to be actively investigated. Linn has called for continuing efforts to investigate the possibility of differential prediction where feasible (Linn, 1984) and has recommended that differential prediction continue to be a topic on the validation research agenda (Linn, 1994).

Theories of Differential Prediction

Several theories have been advanced that purport to explain why differential prediction occurs for different examinee populations. Misprediction, in the form of either over- or underprediction, is an indication of test bias under the most commonly accepted model of test fairness, the regression model of Cleary and Hilton (1968). This model defines a test as unfair to a group of examinees if it predicts lower average scores on the criterion than the members of the group actually achieve. In other words, test bias exists when the test

underpredicts the performance of that group. One complication in interpreting misprediction findings is that it is also often true that the different examinee groups have significantly different average scores on both the predictor and the criterion. Lower average predictor scores for one group (typically, a minority group) often translates into lower selection rates, a condition known as "adverse impact" for the affected group.

Findings of overprediction or underprediction may occur as a result of large differences between groups on the criterion measure combined with the problem of regression to the mean. Given that the correlations between predictors and criterion must be less than perfect in real admission situations, misprediction may arise if group differences on the criterion are less than differences on the predictors. For example, assuming a correlation of +.50 between predictors and criterion, group differences would have to be twice as large on the predictors as on the criterion in order to obtain unbiased prediction results. Greater or lesser differences would invariably contribute to observed misprediction to some degree.

One theory of differential prediction, reported earlier, is that it is falsely assumed to occur and is due predominantly to statistical and research design artifacts. A second theory states that differential prediction may not be detected because both the predictor (or predictors) and criterion are biased in the same direction against a group or groups of examinees. For example, the same factors that cause bias in admission test scores can also operate to lower the college grades for certain categories of students. In this situation, differential validity goes undetected because bias impacts (positively or negatively) all of the measures for one group.

Assuming that differential prediction is a real phenomenon, one explanation is that the predictor(s) is biased against some examinees and not others while the criterion is valid for everyone. In this scenario, differential prediction is caused by the differential validity of the predictor(s), and therefore the use of this predictor(s) could potentially be unfair to certain examinees. A somewhat different explanation is that both the predictor(s) and criterion are biased, although not necessarily to the same degree, against some examinees. Differential prediction is therefore the result of varying degrees of validity for the variables across examinee groups.

Average Scores by Groups

Although the focus of this report is on differential validity and differential prediction, a few comments about group differences in average performance are necessary. It has been observed for a number of years that substantial differences exist in the average level of performance for

	SAT V	SAT M	SAT Total
Total	505	511	1016
Women	502	495	997
Men	509	531	1030
African Americans	434	422	856
Asian Americans	498	560	1058
Latin Americans	463	464	927
Mexican Americans	453	456	909
Puerto Ricans	455	448	903
Native Americans	484	481	965
Whites	527	528	1055
Others	511	513	1024

Figure 3. Average scores by demographic groups.

various demographic groups. Although the trends have been toward a narrowing of these differences, significant differences continue to occur. A number of theories have been advanced to explain these differences, although no single explanation appears to be sufficient. No attempt will be made here to articulate all of the competing hypotheses. The reader interested in these topics is referred to other sources including Hawkins (1993), Murphy (1992), Wilder and Powell (1989), and Young and Fisler (2000).

In order to indicate the magnitude of the differences in average performance, data on the mean scores for various groups on the SAT in 1998-99 is presented in Figure 3. Note that the scores are reported on the new recentered score scale in use since 1995. Although differential validity/prediction is a separate topic from group differences in average performance, the two issues are necessarily intertwined. Knowledge of these group differences will help the reader better understand the statistical and policy issues inherent in differential validity/prediction research.

Organization of this Report

The most recent research synthesis regarding the validity of college admission measures was published more than 20 years ago by Breland (1979). The purpose of this report is to provide an up-to-date comprehensive review and analysis of the research regarding differential validity and differential prediction, principally for the Scholastic Assessment Test and its predecessor, the Scholastic Aptitude Test. This review focuses primarily on the published scholarly research from the past 25+ years (since 1974) on the criterion-related (principally predictive) validity of the SAT. More specifically, this report examines those studies that investigated possible differences in validity for different racial/ethnic groups and/or for men and women. Differential validity/prediction research on the American College Testing Assessment Program tests is also included.

This report is organized into five sections and is

preceded by an abstract and followed by references and an appendix with summaries of the studies reviewed. The current section provided an introduction to the research on differential validity/prediction. Section 2 provides a review of important earlier summaries on group differences in the validity and predictive ability of college admission measures. In particular, the works by Breland (1979), Duran (1983), Linn (1973, 1982b), and Wilson (1983) are highlighted.

Sections 3 and 4 present the main information of this report, with the focus of Section 3 on racial/ethnic differences in validity and prediction and the focus of Section 4 on sex differences in validity and prediction. Note that analyses of the studies reported in Sections 3 and 4 do not conform to the standards for a true meta-analysis. The analyses in these two chapters are based on quantitative summaries of the information reported by each study's author(s) (usually, correlation and regression results) with qualitative judgments about the nature of each study. Effect sizes were never computed, and there was no attempt to derive estimates of them. Summaries of the results are weighted by the sample sizes for each study so that the units of analysis are individuals rather than institutions or studies. Instances where a study was based on a combination of predictors other than the common approach using SAT scores and high school grades are identified. In addition, studies that reported a different set of results due to the use of one or more grade adjustment methods are highlighted. Section 5 provides a synthesis of the research reviewed, conclusions that can be drawn from what is known to date, and some ideas for further work in this area.

II.   Prior Summaries of Differential Validity and Differential Prediction

To provide necessary background for the information in later sections, this section presents an overview of the differential validity studies conducted prior to 1980. In particular, five important research reviews are presented: Breland (1979), Duran (1983), Linn (1973, 1982b), and Wilson (1983). These earlier summaries are described below in the order of their publication.

Linn (1973)

In his 1973 "Review of Educational Research" article, Linn summarized the results from four studies of differential prediction (Cleary, 1968; Davis and Kerner-Hoeg, 1971; Temp, 1971; Thomas, 1972) which included data from a total of 32 institutions. The first three studies were of race differences between white and black (or African American) students in 22 institutions, and the Thomas study was of sex differences in 10 colleges. Cleary's 1968 study presented the first published regression comparisons involving African American and white students and was based on the only three racially integrated colleges with a large enough number of African American students prior to 1965 to make statistical analysis feasible.

In the Cleary, Davis and Kerner-Hoeg, and Temp studies, the criterion variable was FGPA, the predictors were SAT V and SAT M scores, and the comparisons made were between the prediction equations for a sample of white students versus a sample of black students (no other racial groups were included). The comparisons were conducted sequentially: first, for homogeneity of the errors of estimate for the two groups; second, for equality of the slopes; and third, for equality of the intercepts. This method for determining significant group differences in regression systems is known as the Gulliksen-Wilks procedure (Gulliksen and Wilks, 1950). For each institution, if a significant difference was found for one of the comparisons, then the remaining comparisons were not carried out. For 14 of the 22 institutions, at least one significant difference was found in the regression equation. Linn concluded from these results that the regression systems for white and black students should not routinely be assumed to be similar.

At these 22 institutions, the general finding was one of overprediction for the black students if the prediction equation based on white students was used. That is, the actual FGPAs for blacks were generally lower than those predicted from the equation for whites at that institution. Using test scores one standard deviation below the mean for black students, at the mean for black students, and one standard deviation above the mean for black students, the median overprediction figures were, respectively, .08, .20, and .31 (on a four-point grade scale). At these test score levels, the equations at 16, 18, and 18, respectively, of the 22 institutions would have overpredicted black students' grades. Overprediction occurred at all three levels of test scores in 13 of the 22 institutions, while underprediction at all three score levels occurred at only one institution. Despite the relatively small samples (in five of the institutions, the number of black students included was 43 or fewer), the results consistently pointed to a finding of overpredicted grades for the black students.

Similar methods were employed by Thomas to compare the prediction equations for men and women at 10 colleges using data from the College Board's Validity Study Service. In this study, the results were strikingly consistent across institutions: At all 10 colleges, the equations for men always underpredicted the actual FPGAs of the women. In other words, the women achieved higher grades than would be predicted from the equation based on the men at that college. Using test scores one standard deviation below the mean for women, at the mean for women, and one standard deviation above the mean for women, the median underprediction values were, respectively, .22, .36, and .36 (on a four-point grade scale). The amount of underprediction for women was substantial: The difference in predictions based on the equation for men compared to the equation for women was equal to the difference in predicted FGPA for a woman with average SAT scores compared to a woman with scores a full standard deviation below the mean (at about the 16th percentile) (Linn, 1982b). Note also that the degree of misprediction for women's grades was greater than that for black students in the studies cited above. Underprediction ranged from a low of .08 to a high of .75 which is equivalent to three-quarters of a letter grade or almost one standard deviation (0.98, to be exact) in the distribution of FGPAs.

The significance of Linn's article is that this was the first review documenting the overprediction of black students' grades and the underprediction of women's grades when an equation based on whites or men was used. These results were highly consistent across the institutions that were studied. The findings regarding black students are noteworthy because they do not support the notion that the use of SAT scores in predicting FGPA is biased against blacks, at least as measured by the regression approach used in the Cleary, Davis and Kerner-Hoeg, and Temp studies. For a given test score, the actual grades earned by black students were generally lower than were predicted. In later studies, the overprediction finding for black students (and sometimes for other minority students) and the underprediction finding for women was widely replicated across a number of colleges and universities (with varying institutional characteristics) and in different time periods.

Breland (1979)

In his 1979 College Board research monograph, Breland reviewed a number of studies on differential validity and differential prediction dating back to 1964. With respect to differential prediction, Breland summarized 35 regression studies, most of which focused on race differences. The few studies that examined sex differences appeared inconclusive regarding differential prediction. Of these 35 studies, two are actually review articles (Cleary, Humphreys,

Kendrick, and Wesman, 1975; Linn, 1973) and eight of the studies were of a single racial group, blacks. The three studies that examined race differences cited in Linn's 1973 review article were also included in Breland's summary. The remaining 25 studies compared two or more racial/ethnic groups with respect to their regression results. In most of these studies, the predictors were SAT scores and HSGPA and the criterion was FGPA. Other predictors used included ACT scores and College Board achievement test scores, while some studies used longer-term criteria such as sophomore-year, junior-year, or senior-year GPAs.

Of the 25 studies, 17 are included in a latter summary table of significant differences. Most of these 17 studies are of comparisons either between blacks and whites or between Chicanos and Anglos (many of the studies encompassed several institutions). Comparisons of the regression equations (based on standard errors of estimate, slopes, and/or intercepts) found 19 instances of a significant difference between blacks and whites and six instances of no difference. The corresponding figures for the comparisons between Chicanos and Anglos were 10 instances of a significant difference and 14 instances of no difference.

Breland's report also contained five separate tables that listed differential prediction studies for different combinations of predictors (e.g., HSR only, SAT V score only, etc.). For each table, the results from studies using the specified predictor(s) and the degree of misprediction were given. In these tables, all of the comparisons are listed together so that results for comparisons of blacks versus whites only or of Chicanos versus Anglos were not available. In general, use of the minority group means in a common or nonminority regression equation consistently led to overprediction of the minority students' grades. The amount of overprediction tended to be substantially larger for blacks than for Chicanos; for Chicano students, the amount of overprediction was often small and close to zero. Overprediction was largest when HSR alone was used as a predictor, moderate for SAT V or SAT M (used separately or combined as a total test score), and smallest when HSR and test scores were used as multiple predictors. For all comparisons listed, the median overprediction value for HSR alone was .28; for one or both test scores was .16; and for HSR and test scores together was .05 (all figures are based on a four-point grade scale). Breland's tables of results clearly showed that the regression systems differ systematically between minorities and nonminorities and that the performance of minorities in college is consistently overpredicted by equations based on either nonminority or combined samples. Overprediction occurred for any combination of academic predictors but was substantially reduced when HSR and test scores were used in combination as predictors.

Breland also reviewed a number of differential validity studies by examining correlational values. Correlation coefficients were summarized and compared for two situations: (1) across studies regardless of whether group comparisons were made, or (2) within studies that reported correlations for at least two groups. For the first situation, Breland reported on 335 samples that yielded at least one correlation between an academic predictor and either FGPA or CGPA. Correlations were reported broken down by race and sex for different combinations of predictors.

For whites, the correlations for individual predictors were generally higher for women than for men and with HSR yielding higher correlations than test scores. The multiple correlations of HSR and test scores with a criterion were similar for men and women (with median values of .55 and .56, respectively). For blacks, the correlations for test scores were similar for both men and women (the median values ranged from .40 to .43 for each section of the SAT). However, the correlations for HSR were substantially higher for women than for men (with median values of .57 versus .42) which yielded, for women, somewhat higher multiple correlations based on all predictors (with median values of .64 and .57, respectively).

When all groups were considered, the following conclusions can be drawn: The correlations of test scores with a criterion are of similar magnitude for white women, black men, and black women, and are lower for white men. The correlations for HSR are more variable with black men generally having the lowest median value and black women the highest. The multiple correlations for all predictors are similar for white men, white women, and black men, and somewhat higher for black women. In addition to blacks, only a few other studies based on minority samples (all of Chicanos) were located. When these studies were combined with those based on black students, the results for minority students were essentially identical to those for black students only.

The second set of correlational results was based only on studies with two or more groups. Correlations were compared among Anglo, black, and Chicano samples of students. In general, the median correlations exhibited the following patterns: For Anglos, correlations for HSR and test scores with a criterion were similar in magnitude (the median values ranged from .33 to .37). For blacks, SAT V had the highest correlations (median of .41), followed by SAT M (median of .33), then HSR (median of .27). For Chicanos, HSR had the highest correlations (median of .36), followed by SAT V (median of .25) and SAT M (median of .17). In terms of multiple correlations, the values for Anglos and blacks were similar (.48 and .47, respectively) but appreciably lower for Chicanos (.38). All of the values reported here for correlations were the median figures based on the appropriate samples.

In his report, Breland reached a number of important conclusions including:

- The summaries of regression studies indicated a consistent overprediction of college performance for minority students when the regression equation for predicting grades was based on a white or combined sample.

- The degree of overprediction was much more pronounced for black students than for Chicano students. However, the results for Chicanos are less conclusive due to the limited number of studies conducted to date. No other racial/ethnic groups have been studied sufficiently to warrant drawing any conclusions.

- For women, an opposite type of prediction error tended to occur: Consistent underprediction was the rule if a regression equation for predicting grades was based on males or on a sample combining males and females. It should be noted that the number of studies on sex differences that Breland reviewed is much smaller than the number of studies on race differences.

- Of individual predictors, HSR produced the largest overprediction for minority students when used alone. These overpredictions occurred for both short-term (e.g., FGPA) and longer-term criteria (e.g., senior-year GPA).

- Overpredictions were minimized when HSR is used in combination with test scores in predicting college performance.

- In terms of validity coefficients, the median values of the predictors for women are generally equal to or higher than for men. This was true for both black and white samples.

- With respect to race differences, validity coefficients were highly variable, and no discernible pattern emerged with regard to the best predictors across race groups.

Linn (1982b)

As part of the National Academy of Science's report on ability testing (Wigdor and Garner, 1982), Linn's chapter on individual differences examined the topics of differential validity and differential prediction in educational and employment settings. Linn drew his findings about sex and race differences in predictive validity from several sources: American College Testing (1973), Breland (1978, an earlier version of Breland, 1979), and Schrader (1971). Linn stated that, "Correlations of SAT and ACT scores with freshman GPA are typically somewhat higher for women than men" (p. 368). Based on Schrader's reported distributions of correlations of SAT scores with FGPA and multiple correlations of SAT scores and HSR, the values of the correlations are generally higher for women than for men. Results for the ACT show a similar tendency for FGPA to be slightly more predictable from test scores and HSGPA for women than for men (American College Testing, 1973).

With regard to race differences, FGPA was reported to be more predictable from test scores alone and from a combination of HSR and test scores for whites than for either blacks or Chicanos. The summaries by ACT and Breland yielded comparisons of 28 pairs of multiple correlations of HSR and either ACT or SAT scores with FGPA for blacks and whites and 18 pairs of multiple correlations for Chicanos and Anglos (all comparisons are based on samples within the same college). Linn reported that the median multiple correlation was .430 for blacks and .548 for whites; the corresponding value for Chicanos was .388 and .440 for Anglos. Although no explanation was given for the discrepancy in the figures for whites in the two different sets of samples, sampling variability may be sufficient to account for the difference.

In terms of differential prediction by sex, the use of test scores and HSR to predict FGPA generally resulted in smaller standard errors of estimates for women than men (American College Testing, 1973). This result follows from the typical differential validity finding that correlations are usually higher for women than for men. Based on results reported earlier in Linn (1973), the use of the regression equation for men with SAT scores as predictors of FGPA led to consistent underprediction of women's grades. For women with average SAT scores at the 10 colleges studied, their predicted GPAs ranged from about a quarter (.24) to a full (.98) standard deviation below the actual mean GPA for women. On a four-point grade scale, the equation for men typically underpredicted women's GPAs by .36. Results reported by ACT (American College Testing, 1973) were similar in magnitude. In 19 colleges, the use of ACT scores as predictors in a equation for men and women combined yielded an average underprediction for women of .27. When ACT scores were supplemented by HSR as predictors, the average underprediction was reduced to .20.

Reviewing the studies cited in Linn (1973) and Breland (1978), Linn concluded that an equation based on white students tended to overpredict black students' GPAs irrespective of test scores. The amount of overprediction increased with higher SAT scores, reflecting the tendency of the regression slope between test scores and grades to be somewhat smaller for blacks than for whites. Thus, the largest gap between actual and predicted grades for blacks occurred at the upper extreme of the test score distribution. These results were consistent with those reported using ACT scores (American College Testing, 1973).

In 24 comparisons summarized by Breland (1978), a combined equation based on blacks and whites, with test scores and HSR as predictors and using the mean predictor values for blacks, was found to overpredict black students' GPAs by an average of .15 (on a four-point scale). In contrast, this overprediction finding did not generalize to Chicanos. In the 10 comparisons cited by Breland (1978), a combined equation was as likely to underpredict as to overpredict the FGPA of Chicano students.

Duran (1983)

Duran's 1983 College Board volume presented an overview of findings on the background characteristics and academic achievement of Hispanic students with an emphasis on the transition from high school to college. The main Hispanic subpopulations that were included are Mexican Americans, Puerto Ricans, and Cuban Americans (although validity studies of this last group are virtually nonexistent). Of particular interest in Duran's book is Chapter 5, which is a review of predictive validity studies based on Hispanic populations. A total of 10 differential validity/differential prediction studies, all of which were either reported in journals or appeared as dissertations, were described. All of the studies were published between 1974 and 1981, and nine of the studies (all except for Mestre, 1981) involved Hispanics who are most likely to be predominantly Mexican Americans. This assumption is based on descriptive information reported and on the location of the institutions in the studies (usually California or Texas).

In general, some of the studies indicated the presence of differential validity with Hispanic students having lower correlations of test scores and HSR with FGPA than Anglos. However, this finding was true in only about half of the studies that reported results by racial group; nonsignificant differences were reported in the other studies. One study (Calkins and Whitworth, 1974) reported sex differences in validity coefficients with women having higher correlations than men (in both the Anglo and minority samples); however, two other studies did not find differential validity by sex.

Differential prediction by race was found in only one of the eight studies that investigated the use of an Anglo or a combined Anglo/Chicano equation to predict Hispanic students' GPAs (overprediction of Mexican Americans' GPAs was found by Goldman and Richards, 1974). Differential prediction was not detected in the other studies. However, it should be noted that some of the Hispanic samples were small, which resulted in limited statistical power. Differential prediction by sex (with underprediction of women's GPAs) was found only by Calkins and Whitworth (1974) but did not occur in two other studies.

Wilson (1983)

Wilson's 1983 College Board research report did not focus specifically on differential validity/prediction but rather on the prediction of longer-term academic performance criteria. Few studies have been conducted which investigated the prediction of grades beyond the first year of college. Wilson's review summarized the findings from 32 studies, some dating back to the 1940s, that employed longer-term criteria such as two-year, three-year, and four-year CGPAs, or second-year GPA. Three of the studies reported separate validity coefficients for men and women; a fourth study reported separate coefficients for black males and females and white males and females. Overall, the pattern of validity coefficients for SAT scores and HSR was mixed with respect to higher reported values for men or women.

The one study that examined race by sex differences (Farver, Sedlacek, and Brooks, 1975) found significantly lower multiple correlations for black males than for the other three groups using SAT V, SAT M, and HSR as predictors and FGPA, two-year CGPA, and three-year CGPA as separate outcome variables. For FGPA, the multiple correlation for black males was approximately .10 lower than for the other groups; for two-year CGPA, at least .15 lower; and for three-year CGPA, at least .25 (and as much as .33) lower. For black males, these results clearly showed the declining predictability over time of black male students' grades. The findings were based on two cohorts of black students entering the University of Maryland in the early 1970s and comparative samples of white students from the same cohorts.

Synopsis

These five summaries of earlier research (studies conducted before the mid-1970s) on differential validity and differential prediction were all published during a 10-year period from 1973 to 1983. The information contained within provides an important foundation for understanding and interpreting the research on differential validity/prediction using academic predictors that subsequently followed.

III. Racial/Ethnic Differences in Validity and Prediction

In this section, all of the 29 studies conducted since 1974 that investigated racial/ethnic differences in validity and

prediction are reviewed. The 29 studies can be categorized into one of three types: single institutions (19 studies), multiple institutions, which generally involved several campuses from the same state higher education system (6 studies), and compilations of findings from a large number of institutions, which were usually based on several years of results (4 studies). These compilations were each authored by one or more ACT or ETS researchers with results from each involving at least 80 institutions and samples of over 100,000 students.

All of the studies reviewed appeared as either journal articles or as conference papers. Note that some of the journal articles appeared in an earlier form as an ACT or ETS research report; in those instances, it is the journal article that is referenced. All of the studies were located through computerized searches of relevant journals and sources such as ERIC databases or from the references of targeted journal articles. Table 1 provides a summary of the important characteristics of each of the 29 studies. In addition, a brief description of each study is provided in the Appendix.

TABLE 1

Studies Reviewed in Section 3

Authors	Year	Type	Institution	Classes	Sample N	DV/DP	Groups	Criterion	Predictors
Arbona & Novy	90	S	Houston*	E87	746	DP	B,H	FGPA	SAT V, SAT M
Baggaley	74	S	Pennsylvania	E69	529	DP	B	CGPA	SAT V, SAT M, HSGPA
Bridgeman et al.	2000	M	23 colleges	E94,95	93139	DV/DP	A,B,H	FGPA	SAT V, SAT M, HSGPA
Chou & Huberty	90	S	Georgia	E87	3378	DP	B	QGPA	SAT V, SAT M, HSGPA
Cowen & Fiori	91	S	CSU, Hayward	E88,89	972	DV/DP	A,B,H	FGPA	SAT V, SAT M, HSGPA
Crawford et al.	86	S	W. Virginia State*	AY85-86	1121	DV/DP	B	FGPA	ACT, HSGPA
Elliott & Strenta	88	S	Dartmouth	G86	927	DV/DP	B	ICG,CGPA	SAT V+M, HSGPA, ACH
Farver et al.	75	S	Maryland	E68,69	559	DV/DP	B	CGPA	SAT V, SAT M, HSGPA
Hand & Pranther	85	M	31 GA colleges	E83	45067	DV	B	CGPA	SAT V, SAT M, HSGPA
Hogrebe et al.	83	S	Georgia*	AY77-79	345	DP	B	FGPA	SAT V, SAT M, HSGPA
Maxey & Sawyer	81	C	271 colleges	AY73-77	156844	DP	B,H	FGPA	ACT subtests, HS grades
McCornack	83	S	San Diego State	E79,80	5870	DV/DP	A,B,H,N	SGPA	SAT V+M, HSGPA
Moffatt	93	S	Atlanta Christian	Not Given	570	DV/DP	B	CGPA	SAT V+M
Morgan	90	C	198 colleges	E78,81,85	278074	DV/DP	A,B,H	FGPA	SAT V, SAT M, HSGPA
Nettles et al.	86	M	30 colleges	Not Given	4094	DP	B	CGPA	SAT V+M, HSGPA, other vars.
Noble et al.	96	C	>80 colleges	Not Given	Not Given	DP	B	ICG	ACT subtests, HS grades
Pearson	93	S	Miami	E88	1594	DP	H	CGPA	SAT V, SAT M, HSR
Pennock-Román	90	M	6 universities	E82,86	24637	DV/DP	H	FGPA	SAT V, SAT M, HSGPA
Ramist et al.	94	M	45 colleges	E82,85	46379	DV/DP	A,B,H,N	ICG,FGPA	SAT V, SAT M, HSGPA
Sawyer	86	C	200 colleges	AY74-77	105502	DP	M	FGPA	ACT subtests, HS grades
Sue & Abe	88	M	8 UC campuses	E84	5113	DV/DP	A	FGPA	SAT V, SAT M, HSGPA
Tracey & Sedlacek	84	S	Maryland	E79,80	1973	DV	B	SGPA,CGPA	SAT V+M
Tracey & Sedlacek	85	S	Maryland	E79,80	2742	DV	B	SGPA,CGPA	SAT V+M
Wainer et al.	93	S	Hawaii	E82,89	2791	DV	A	FGPA	SAT V, SAT M, HSGPA
Wilson	80	S	Penn State Univ.*	E71	1275	DV/DP	M	FGPA, CGPA	SAT V, SAT M, HSGPA
Wilson	81	S	Not Given	E70-73	1254	DV	M	FGPA, CGPA	SAT V, SAT M, HSGPA
Young	91b	S	Stanford	E82	1462	DP	M	CGPA	SAT V, SAT M, HSGPA
Young	94	S	Rutgers	E85	3703	DV/DP	A,B,H	CGPA	SAT V, SAT M, HSR
Young & Koplow	97	S	Rutgers	E90	214	DP	M	CGPA	SAT V, SAT M, HSR

*An asterisk after the institution's name means that the study did not identify the institution but is likely based on the description in the study. **Type:** C = compilation, M = multiple campuses, S = single institution. **Classes:** AY = academic year, E = entering year, G = graduation year. **DV/DP:** DV = differential validity, DP = differential prediction. **Groups:** A = Asian Americans, B = Blacks/African Americans, H = Hispanics, M = combined minority group, N = Native Americans. **Criterion:** CGPA = cumulative GPA, FGPA = first-year GPA, ICG = individual course grades, QGPA = quarter GPA, SGPA = semester GPA. **Predictors:** ACH = College Board Achievement Test Scores, ACT = ACT Composite score, SAT V+M = SAT total score, HSR = HS Rank, HS grades = individual course grades.

TABLE 1 *(Continued from page 11)*

Studies Reviewed in Section 3

Authors	*Differential Validity Results*	*Differential Prediction: Grade Prediction Results*
Arbona & Novy	R:B = .08, H = .20, W = .17	
Baggaley	R:B = .25, W = .41	
Bridgeman et al.	R:BM = .45, BF = .44, AM = .44, AF = .43, HM = .38, HF = .44	BM = -.14, BF = +.01, AM = -.07, AF = +.03, HM = -.15, HF = -.02
Chou & Huberty		B = -.15
Cowen & Fiori	R:MM = .42, MF = .57, WM = .47, WF = .43	A = -.06, B = -.06, H = +.07
Crawford et al.	R2:B = .25, W = .22	B: significant overpostdiction
Elliott & Strenta	R:B = .55, W = .50	B = -.03
Farver et al.	R(CGPA):BM = .52, BF = .42, WM = .55, WF = .67	
Hand & Pranther	med adj R2:BM = .36, BF = .44, WM = .45, WF = .47	
Hogrebe et al.	R2:B = .29, W = .19	
Maxey & Sawyer	R:B = .48, H = .55, W = .56	B = -.05, H = .00
McCornack	mean R:A = .56, B = .38, H = .43, N = .41, W = .40	A = -.17, B = -.21, H = -.19, N = +.07 (mean)
Moffatt	r (CGPA):B = .16, W = .54	
Morgan	median R:A = .48, B = .39, H = .42, W = .52	
Nettles et al.		
Noble et al.		
Pearson		H: underpredicted (+.14 using SAT V, +.15 using SAT M)
Pennock-Román	median R:H = .40, W = .44	H = -.02, -.08, -.08, -.15, -.25, -.31 (6 universities)
Ramist et al.	R:A = .48, B = .39, H = .43, N = .55, W = .45	A = +.04, B = -.16, H = -.13, N = -.24
Sawyer		M = -.09
Sue & Abe	R:A = .50, W = .45	A = +.02
Tracey & Sedlacek	R:B = .33, W = .39	
Tracey & Sedlacek	R:B = .26, W = .40	
Wainer et al.	r, 3 predictors: A = .19, .10, .32, W = .43, .35, .51	
Wilson	R:M = .69, W = .57	
Wilson	R:M = .38, W = .55	
Young		M = -.17
Young	R:A = .44, B = .33, H = .47, PR = .34, W = .38	A = -.09, B = -.17, H = -.08, PR = +.01
Young & Koplow		M = -.12

Results: R = multiple correlation, R2 = multiple correlation squared, r = simple correlation.

Most of the 29 studies are of differential prediction only or of differential validity and differential prediction. That is, the studies reported prediction results based on regression analysis along with validity coefficients. Furthermore, most of the studies (21 of the 29) involved a comparison of only one minority group (usually blacks, but sometimes all minority students were combined into a single group) with whites. The most studied minority group was blacks (20 studies), followed by Hispanics (10), and Asian Americans (8). Five additional studies reported on a combined minority group composed mostly or exclusively of blacks and Hispanics. Finally, two studies had large enough samples to report results for Native Americans.

In the remainder of this chapter, the findings on differential validity are reported first followed by the findings on differential prediction. Within each set of findings, results for each racial/ethnic group are described separately. A section that summarizes the results appears at the end of the chapter.

Differential Validity Findings

The differential validity findings, based on reported multiple correlation coefficients (or squared multiple correlations) of predictors with a criterion, are inconsistent with respect to comparisons of minority groups with white students. In general, multiple correlations computed from samples of black or Hispanic students (or samples that combined the two groups) are somewhat lower than for Asian American or white students. However, several studies (generally with small samples) yielded results that

are not consistent with this trend, with black or minority students having higher multiple correlations than whites (see e.g., Crawford, Alferink, and Spencer, 1986; Elliott and Strenta, 1988; Hogrebe, Ervin, Dwinell, and Newman, 1983; Wilson, 1980).

Differential Validity: Asian Americans

Differential validity results for Asian Americans were reported in seven studies (Table 2): Bridgeman, McCamley-Jenkins, and Ervin (2000), McCornack (1983), Morgan (1990), Ramist, Lewis, and McCamley-Jenkins (1994), Sue and Abe (1988), Wainer, Saka, and Donoghue (1993), and Young (1994). All of these studies used the standard combination of SAT scores and HS grades as predictors. Differences in the Asian American samples in these studies due to geographical and socioeconomic variations (i.e., East Coast residents versus California residents) may have been a confounding factor but not enough is known to determine its impact on the results reported.

Wainer, Saka, and Donoghue reported substantially lower correlations of SAT V, SAT M, and HSGPA with FGPA for students who attended Hawaiian secondary schools than for those from the mainland United States and also as compared with national figures. Since approximately three-fourths of Hawaiian high school students are of Asian descent, it can be assumed that the lower correlations are based predominantly on Asian American students. Unfortunately, the authors did not report self-identified race information for students in their study so the actual proportion of Hawaiian students who are Asian Americans cannot be verified. The summary by Morgan (1990), based on 198 institutions, indicated a median multiple correlation of SAT scores plus HSGPA with FGPA that was slightly lower for Asian Americans (.48) than for whites (.52) but higher than for blacks (.39) or Hispanics (.42). In the remaining five studies, the multiple correlations of SAT scores plus HSGPA with FGPA were the same or higher for Asian Americans than for whites (and also usually higher

than for the other minority groups studied). When compared with whites, the multiple correlations ranged from .00 to .16 higher for Asian Americans. In the Bridgeman, McCamley-Jenkins, and Ervin study, the original multiple correlations were essentially identical for Asian Americans and whites but were slightly higher for Asian Americans when FGPA was adjusted for course difficulty.

Based on these seven studies which involved over 200 institutions, it is probably accurate to conclude that the individual and multiple correlations of SAT scores and HSGPA with FGPA are quite similar in magnitude for Asian American and white students and may possibly be slightly lower for Asian Americans. This finding is principally determined by the large sample size used in the Morgan (1990) study.

Differential Validity: Blacks/African Americans

A greater number of differential validity and differential prediction studies have been conducted on blacks/African Americans than on any other minority group. For differential validity, a total of 16 studies reported results for blacks/African Americans (Table 3). Of these, eight studies (Baggaley, 1974; Maxey and Sawyer, 1981; Moffatt, 1993; Morgan, 1990; Ramist, Lewis, and McCamley-Jenkins, 1994; Tracey and Sedlacek, 1984; Tracey and Sedlacek, 1985; Young, 1994) reported significantly lower multiple correlations of SAT scores plus HSGPA with FGPA or CGPA for blacks than for whites. The median multiple correlation was .33 for blacks and .43 for whites, and was larger for whites in all eight studies. The difference in multiple correlations ranged from a low of .05 (Young, 1994) to a high of .38 (Moffatt, 1993). A ninth study, Arbona and Novy (1990), was primarily about differential prediction but also reported a lower multiple correlation of SAT scores with FGPA for blacks than for Hispanics or whites. Note, however, that the Moffatt and Arbona and Novy studies only used SAT scores as predictors,

TABLE 2

Differential Validity Results: Asian Americans

Authors	Criterion	Predictors	Results
Bridgeman et al.	FGPA	SAT V, SAT M, HSGPA	R:AM=.44, AF=.43
McCornack	SGPA	SAT V+M, HSGPA	mean R:A=.56, W=.40
Morgan	FGPA	SAT V, SAT M, HSGPA	median R:A=.48, W=.52
Ramist et al.	ICG, FGPA	SAT V, SAT M, HSGPA	R:A=.48, W=.45
Sue & Abe	FGPA	SAT V, SAT M, HSGPA	R:A=.50, W=.45
Wainer et al.	FGPA	SAT V, SAT M, HSGPA	r:A=.19, .10, .32, W=.43, .35, .51
Young	CGPA	SAT V, SAT M, HSR	R:A=.44, W=.38

Criterion: CGPA = cumulative GPA, FGPA = first-year GPA, ICG = individual course grades, SGPA = semester GPA. Predictors: SAT V+M = SAT total score, HSR = HS Rank. Results: R = multiple correlation, r = simple correlation.

TABLE 3

Differential Validity Results: Blacks/African Americans

Authors	Criterion	Predictors	Results
Arbona & Novy	FGPA	SAT V, SAT M	R:B = .08, W = .17
Baggaley	CGPA	SAT V, SAT M, HSGPA	R:B = .25, W = .41
Bridgeman et al.	FGPA	SAT V, SAT M, HSGPA	R:BM = .45, BF = .44
Crawford et al.	FGPA	ACT, HSGPA	R2:B = .25, W = .22
Elliott & Strenta	ICG, CGPA	SAT V+M, HSGPA, ACH	R:B = .55, W = .50
Farver et al.	CGPA	SAT V, SAT M, HSGPA	R(CGPA):BM = .52, BF = .42, WM=.55, WF=.67
Hand & Pranther	CGPA	SAT V, SAT M, HSGPA	med. adj. R2:BM = .36, BF = .44, WM = .45, WF = .47
Hogrebe et al.	FGPA	SAT V, SAT M, HSGPA	R2:B = .29, W = .19
Maxey & Sawyer	FGPA	ACT subtests, HS grades	R:B = .48, W = .56
McCornack	SGPA	SAT V+M, HSGPA	mean R:B = .38, W = .40
Moffatt	CGPA	SAT V+M	r(CGPA):B = .16, W = .54
Morgan	FGPA	SAT V, SAT M, HSGPA	median R:B = .39, W = .52
Ramist et al.	ICG, FGPA	SAT V, SAT M, HSGPA	R:B = .39, W = .45
Tracey & Sedlacek	SGPA, CGPA	SAT V+M	R:B = .33, W = .39
Tracey & Sedlacek	SGPA, CGPA	SAT V+M	R:B = .26, W = .40
Young	CGPA	SAT V, SAT M, HSR	R:B = .33, W = .38

*Criterion: CGPA = cumulative GPA, FGPA = first-year GPA, ICG = individual course grades, SGPA = semester GPA. **Predictors:** ACH = College Board Achievement Test scores, ACT = ACT Composite score, SAT V+M = SAT total score, HSR = HS Rank, HS grades = individual course grades. **Results:** R = multiple correlation, R2 = multiple correlation squared, r = simple correlation.*

and this may have magnified the differences in correlations. Another study, McCornack (1983), reported essentially similar multiple correlations for four groups (blacks, Hispanics, Native Americans, and whites) but a higher value for Asian Americans. Results similar to McCornack's study were found by Bridgeman, McCamley-Jenkins, and Ervin (2000) in comparing African Americans to whites. However, in this study somewhat lower correlations were found for African Americans after each of several grade adjustment methods were applied to FGPA.

Two other studies, Farver, Sedlacek, and Brooks (1975) and Hand and Pranther (1985), reported results by race and sex and found lower values for black males and females than for their white counterparts. Two additional studies, Crawford, Alferink, and Spencer (1986) and Hogrebe, Ervin, Dwinell, and Newman (1983), found higher squared multiple correlations of .03 and .10, respec-

tively, for blacks than for whites. Elliott and Strenta (1988) reported a higher multiple correlation of SAT scores plus HSGPA with four-year CGPA for blacks (.55) than for whites (.50). Their results differed markedly from those reported in the other studies although no obvious explanations are apparent. For GPAs in years 1 to 3 for these students, the multiple correlation was higher for whites than for blacks but was reversed for year 4. This was sufficient to cause the multiple correlations for four-year CGPA to be higher for blacks. It is possible that the high degree of selectivity at Dartmouth College, coupled with the use of four-year CGPA as the criterion, may have led to this anomaly.

Differential Validity: Hispanics

Differential validity results for Hispanics were reported in eight studies (Table 4): Arbona and Novy (1990), Bridgeman, McCamley-Jenkins, and Ervin (2000), Maxey

TABLE 4

Differential Validity Results: Hispanics

Authors	Criterion	Predictors	Results
Arbona & Novy	FGPA	SAT V, SAT M	R:H = .20, W = .17
Bridgeman et al.	FGPA	SAT V, SAT M, HSGPA	R:HM = .38, HF = .44
Maxey & Sawyer	FGPA	ACT subtests, HS grades	R:H = .55, W = .56
McCornack	SGPA	SAT V+M, HSGPA	mean R:H = .43, W = .40
Morgan	FGPA	SAT V, SAT M, HSGPA	median R:H = .42, W = .52
Pennock-Román	FGPA	SAT V, SAT M, HSGPA	median R:H = .40, W = .44
Ramist et al.	ICG, FGPA	SAT V, SAT M, HSGPA	R:H = .43, W = .45
Young	CGPA	SAT V, SAT M, HSR	R:H = .47, PR = .34, W = .38

*Criterion: CGPA = cumulative GPA, FGPA = first-year GPA, ICG = individual course grades, SGPA = semester GPA. **Predictors:** SAT V+M = SAT total score, HSR = HS Rank, HS grades = individual course grades. **Results:** R = multiple correlation.*

and Sawyer (1981), McCornack (1983), Morgan (1990), Pennock-Román (1990), Ramist, Lewis, and McCamley-Jenkins (1994), and Young (1994). In general, the results for Hispanics are closer to the findings for blacks/African Americans than to those for whites. In four of the five studies with the largest sample sizes (Maxey and Sawyer, 1981; Morgan, 1990; Pennock-Román, 1990; Ramist, Lewis, and McCamley-Jenkins, 1994), the multiple correlation values are slightly (by .01) to notably (by .10) smaller for Hispanics than for whites; in the fifth study (Bridgeman, McCamley-Jenkins, and Ervin, 2000), the values are essentially equal. All of the studies used SAT scores as predictors except for Maxey and Sawyer who based their results on ACT subtest scores; only Arbona and Novy did not additionally include HS grades.

Only the study by Young (1994) reported separate results for Puerto Ricans and for a combined group of non-Puerto Rican Hispanics. In this study, the multiple correlation of the three academic predictors with CGPA for Puerto Ricans was .34; this contrasts with the corresponding figures for non-Puerto Rican Hispanics of .47, for Asian Americans of .44, for blacks of .33, and for whites of .38. Although the sample sizes for the two Hispanic groups were relatively small (N=70 for each group), the difference in the multiple correlation for Puerto Ricans versus non-Puerto Rican Hispanics appears to be substantial.

Differential Validity: Native Americans

Only two studies were located that reported findings on Native Americans: McCornack (1983) and Ramist, Lewis, and McCamley-Jenkins (1994). This is not surprising since few institutions enroll a large enough sample of Native Americans to allow separate analyses of this group. In fact, the McCornack study had 24 and 25 Native Americans in the two cohorts that were analyzed. The Ramist, Lewis, and McCamley-Jenkins study was based on data from 45 colleges, 34 of which had Native American students. From these 34 colleges, the total sample of Native Americans was 184, or an average of fewer than 6 per institution. Thus, it is evident that the empirical base for understanding the performance of Native Americans is extremely limited.

The average multiple correlation of SAT scores plus HSGPA with SGPA for the two cohorts of Native Americans in McCornack (1983) was .41, a figure comparable to that for blacks, Hispanics, and whites and lower than for Asian Americans. In Ramist, Lewis, and McCamley-Jenkins (1994), the multiple correlation with FGPA was .55 for Native Americans, the highest value for any of the five racial/ethnic groups examined and substantially larger than the corresponding value of .48 for the next closest group, Asian Americans.

Differential Validity: Combined Minority Groups

Two studies, both conducted by Wilson (1980, 1981), reported findings for a combined group of minority students (largely blacks, but included Hispanics and Native Americans). The results from the two studies are in conflict with reported multiple correlations of .69 and .38 for the minority students and .57 and .55 for white students (the first figure for each group came from the 1980 study). If the values for each group are averaged, the resulting means are similar (.535 for minority students and .56 for white students). Since the relative compositions of the minority samples were not given, it is difficult to compare these results with earlier ones for separate racial/ethnic groups.

Differential Prediction Findings

Differential prediction findings are derived from analyses of residuals from either one of two designs: (1) a multiple regression equation based on a combined sample of students, or (2) from an equation computed from a sample of white students and then applied to groups of minority students. In general, with few exceptions, the findings consistently point to an overprediction of black/African American and Hispanic students' grades. Overprediction results in a residual value for an individual that is negative when predicted FGPA is subtracted from actual FGPA. In other words, it is generally the case that the actual grades earned by black/African American and Hispanic students are lower than those predicted from test scores and HSGPA. This is true whether the regression equation used came from the first or second design cited above. It should be noted that the magnitude of the overprediction varied considerably across studies and racial/ethnic groups.

The situation for Asian American students is more complex, with results ranging widely from substantial overprediction to no misprediction to slight underprediction. Furthermore, one study that computed adjusted grades found that since Asian Americans are more likely to major in fields with more difficult courses, the results after grade adjustments tended to reflect underprediction rather than oveprediction as is the case with unadjusted grades. This is consistent with the results (not included here) found in Young (1991b).

Differential Prediction: Asian Americans

Six studies (Table 5) reported differential prediction results for Asian Americans (Bridgeman, McCamley-Jenkins, and Ervin, 2000; Cowen and Fiori, 1991;

TABLE 5

Differential Prediction Results: Asian Americans

Authors	Criterion	Predictors	Results
Bridgeman et al.	FGPA	SAT V, SAT M, HSGPA	AM = -.07, AF = +.03
Cowen & Fiori	FGPA	SAT V, SAT M, HSGPA	A = -.06
McCornack	SGPA	SAT V+M, HSGPA	A = -.17 (mean)
Ramist et al.	ICG, FGPA	SAT V, SAT M, HSGPA	A = +.04
Sue & Abe	FGPA	SAT V, SAT M, HSGPA	A = +.02
Young	CGPA	SAT V, SAT M, HSGPA	A = -.09

*Criterion: CGPA = cumulative GPA, FGPA = first-year GPA, ICG = individual course grades, SGPA = semester GPA. **Predictors:** SAT V+M = SAT total score.*

McCornack, 1983; Ramist, Lewis, and McCamley-Jenkins, 1994; Sue and Abe, 1988; Young, 1994). All of these studies used the standard combination of SAT scores and HS grades as predictors; the outcome measures included SGPA, FGPA, and CGPA. Of the six studies, two reported (Ramist, Lewis, and McCamley-Jenkins, 1994; Sue and Abe, 1988) slight underprediction (+.04 and +.02, respectively), while the other four studies reported more substantial overprediction ranging from -.02 to -.17. The figure of -.02 is an estimate for the Bridgeman, McCamley-Jenkins, and Ervin study since results were reported separately by sex.

Two important points should be noted regarding these results: (1) The studies by Ramist, Lewis, and McCamley-Jenkins and Sue and Abe involved a total of over 50,000 students at 53 institutions and are much larger that the samples for the other studies. Thus, the slight underprediction for Asian Americans found in these two studies seems to be the more plausible outcome. (2) The Bridgeman, McCamley-Jenkins, and Ervin study applied several grade adjustment methods to their sample of 23 colleges and found that the original overprediction for Asian Americans was changed to slight underprediction (typically, +.04 to +.05) after grade adjustments were applied. These results are consistent with those found by Ramist, Lewis, and McCamley-Jenkins and Sue and Abe. Given these some-

what variable results from only six studies, it is difficult to draw firm conclusions about differential prediction for Asian Americans, but slight underprediction of grades appears to be the most plausible outcome.

Differential Prediction: Blacks/African Americans

A total of nine studies (Table 6) (using QGPA, SGPA, FGPA, or CGPA as the criterion) reported differential prediction results for black/African American students (Bridgeman, McCamley-Jenkins, and Ervin, 2000; Chou and Huberty, 1990; Cowen and Fiori, 1991; Elliott and Strenta, 1988; Maxey and Sawyer, 1981; McCornack, 1983; Nettles, Theony, and Gosman, 1986; Ramist, Lewis, and McCamley-Jenkins, 1994; Young, 1994). All of these studies except for Maxey and Sawyer (who used ACT subtest scores and HS grades) employed the standard combination of SAT scores and HS grades as predictors (although Elliott and Strenta and Nettles, Theony, and Gosman added other predictors in their studies). In all nine studies, African American students' grades were overpredicted to some degree. Note that the study by Nettles, Theony, and Gosman reported that the grades of African Americans were overpredicted but did not include summary statistics. The amount of overprediction ranged

TABLE 6

Differential Prediction Results: Blacks/African Americans

Authors	Criterion	Predictors	Results
Bridgeman et al.	FGPA	SAT V, SAT M, HSGPA	BM = -.14, BF = +.01
Chou & Huberty	QGPA	SAT V, SAT M, HSGPA	B = -.15
Cowen & Fiori	FGPA	SAT V, SAT M, HSGPA	B = -.06
Crawford et al.	FGPA	ACT, HSGPA	
Elliott & Strenta	ICG, CGPA	SAT V+M, HSGPA, ACH	B = -.03
Maxey & Sawyer	FGPA	ACT subtests, HS grades	B = -.05
McCornack	SGPA	SAT V+M, HSGPA	B = -.21(mean)
Nettles et al.	CGPA	SAT V+M, HSGPA, other vars.	
Noble et al.	ICG	ACT subtests, HS grades	
Ramist et al.	ICG,FGPA	SAT V, SAT M, HSGPA	B = -.16
Young	CGPA	SAT V, SAT M, HSGPA	B = -.17

*Criterion: CGPA = cumulative GPA, FGPA = first-year GPA, ICG = individual course grades, QGPA = quarter GPA, SGPA = semester GPA. **Predictors:** ACH = College Board Achievement Test scores, ACT = ACT Composite score, SAT V+M = SAT total score, HS grades = individual course grades.*

from a low of -.03 in the study by Elliott and Strenta to a high of -.21 in McCornack's study. The mean and median overprediction for these studies was -.11 and is the largest value observed for any group. The results for the three studies with the largest samples (Bridgeman, McCamley-Jenkins, and Ervin, 2000; Maxey and Sawyer, 1981; Ramist, Lewis, and McCamley-Jenkins, 1994) showed slightly less overprediction than for the five smaller studies. Furthermore, there does not appear to be any discernable trend over time as the degree of overprediction appears to be similar for earlier and more recent studies.

Two other studies (Crawford, Alferink, and Spencer, 1986; Noble, Crouse, and Schulz, 1996) reported results on grade prediction in terms of rates on success outcomes. Crawford, Alferink, and Spencer found that the CGPAs of blacks/African Americans were significantly overpostdicted (from a retrospective prediction study) from ACT composite score and HSGPA. Noble, Crouse, and Schulz reported that blacks/African Americans had significantly lower rates of obtaining a grade of B or better in four first-year college courses than was predicted from ACT subtest scores and HS course grades.

Differential Prediction: Hispanics

Eight studies reported differential prediction results for Hispanic students (using SGPA, FGPA, or CGPA as the criterion) (See Table 7). The eight studies include Bridgeman, McCamley-Jenkins, and Ervin (2000), Cowen and Fiori (1991), Maxey and Sawyer (1981), McCornack (1983), Pearson (1993), Pennock-Román (1990), Ramist, Lewis, and McCamley-Jenkins (1994), and Young (1994). All of these studies except for Maxey and Sawyer (who used ACT subtest scores and HS grades) employed the standard combination of SAT scores and HS grades as predictors.

Of these, one (Cowen and Fiori, 1991) reported a modest underprediction of +.07. The remaining six studies (all except Pearson, which is not included here) reported either no misprediction or overprediction of Hispanic students' grades. The amount of overprediction ranged from a minimum of .00 (Maxey and Sawyer, 1981) to a maximum of -.31 (Pennock-Román, 1990). For these seven studies, the misprediction values were calculated to be a median of -.08 and a mean of -.10. Note that since the Pennock-Román study involved six universities, separate values were reported for each institution. Thus, the median and mean figures reported are actually based on the values from 12 separate samples. In addition, Pennock-Román's study was one of the few that used a prediction equation based on white students to forecast grades for minority students. Thus, the overprediction values are slightly larger than what would have resulted from a common equation based on all students. As is the case with black/African American students, there did not appear to be any discernable trend over time for Hispanic students because the degree of overprediction appears to be similar for earlier and more recent studies.

In addition, Young's study was the only one that reported separate results for Puerto Rican students and non-Puerto Rican Hispanics. Because the sample of non-Puerto Rican Hispanics is more similar to the ones used in other studies, the overprediction figure of -.08 was included instead of the +.01 underprediction value found for Puerto Rican students. Since this was the only study that reported results for Puerto Ricans, there was not enough information available for a separate discussion of these students.

Pearson's study was the only one that reported a substantial underprediction of Hispanic students' grades. The amount of underprediction was given as +.14 using SAT V as a predictor and +.15 using SAT M. (No data were presented for any other combinations of predictors.) The main reasons for excluding this study from the analysis of Hispanic students are: (1) her sample differed substantially from those in other studies in several important aspects, and (2) she did not include HS grades as one of the predictors (using only test scores is likely to have distorted the prediction findings). Her study was conducted using data from the University of Miami where the majority of Hispanics are of Cuban descent. In contrast to other Hispanic subgroups such as Mexican Americans, Cuban American students closely resemble the norming samples for national tests in terms of

TABLE 7

Differential Prediction Results: Hispanics

Authors	Criterion	Predictors	Results
Bridgeman et al.	FGPA	SAT V, SAT M, HSGPA	HM = -.15, HF = -.02
Cowen & Fiori	FGPA	SAT V, SAT M, HSGPA	H = +.07
Maxey & Sawyer	FGPA	ACT subtests, HS grades	H = .00
McCornack	SGPA	SAT V+M, HSGPA	H = -.19 (mean)
Pearson	CGPA	SAT V, SAT M, HSR	H:underpredicted (+.14 SAT V, +.15 SAT M)
Pennock-Román	FGPA	SAT V, SAT M, HSGPA	H = -.02, -.08, -.08, -.15, -.25, -.31(6 univ.)
Ramist et al.	ICG,FGPA	SAT V, SAT M, HSGPA	H = -.13
Young	CGPA	SAT V, SAT M, HSR	H = -.08, PR = +.01

Criterion: CGPA = cumulative GPA, FGPA = first-year GPA, ICG = individual course grades, SGPA = semester GPA. Predictors: SAT V+M = SAT total score, HSR = HS Rank, HS grades = individual course grades.

income levels, educational preparation, and other socioeconomic indicators. Unlike Hispanic populations elsewhere, the Miami Latin community (of which over 60 percent are of Cuban origin) is predominately middle and upper middle class. Given the academic and socioeconomic similarities between the Hispanic students and the comparison group of white students, it is not surprising that Pearson's results differed markedly from the other studies of Hispanic students. Pearson attributes the underprediction for the Hispanic students to the fact that although all were bilingual, for some English is the second and weaker language. Being bilingual may have a negative impact on test scores (especially on tests of verbal ability) but may be an advantage (or at least less of a disadvantage) in an educational environment. In this case, the poorer test performance of the Hispanic students did not forecast poor academic performance.

Differential Prediction: Native Americans

The same two studies that reported differential validity results on Native Americans (McCornack, 1983; Ramist, Lewis, and McCamley-Jenkins, 1994) also reported differential prediction findings. The two studies yielded contradictory results with McCornack reporting an underprediction of +.07 while Ramist, Lewis, and McCamley-Jenkins reported an overprediction of -.24. Given the small sample sizes in both studies, any interpretation must be quite tentative. However, given the much larger sample in the Ramist, Lewis, and McCamley-Jenkins study, along with the fact that Native American students are often similar to other minority students in terms of academic preparation and socioeconomic status, the figure from this study may be more representative for Native Americans.

Differential Prediction: Combined Minority Groups

There are three studies that reported results for a combined group of minority students composed of African Americans and Hispanics (Sawyer, 1986; Young, 1991a; Young and Koplow, 1997). A combined group was used in order to increase sample size and power in order to detect significant differences. All three studies reported overprediction of the minority students' grades with values given as -.09 (Sawyer, 1986), -.12 (Young and Koplow, 1997), and -.17 (Young, 1991a), which yielded a mean of -.13. These figures are consistent with the results reported separately for African American and Hispanic students. Note that when college grades were adjusted for course difficulty in Young's study, the mean overprediction for minority students was reduced from -.17 to -.12,

a value more consistent with other studies using samples of African American and Hispanic students.

Summary

Analysis of the differential validity and differential prediction results is challenging, given that none of the groups studied appear to share the same patterns of findings. With respect to differential validity, studies of Asian Americans generally indicated that this group has similar to slightly lower zero-order correlations and multiple correlations of predictors with the criterion than for whites. Studies with blacks/African Americans and Hispanics demonstrated the opposite finding, with these groups having generally lower correlations than for whites. There were too few studies of Native Americans and of combined minority groups to comment about correlations based on these groups.

The differential prediction results for minority groups are also quite complex. For Asian Americans, the prediction results were quite varied, with different studies reporting overprediction, no misprediction, and underprediction. The degree of overprediction typically found was less than that for other minority groups. In addition, adjusting the college grades of Asian American students for course difficulty moderated the overprediction results such that slight underprediction appears to be a more reasonable finding.

For the remaining groups (blacks/African Americans, Hispanics, combined minority groups, and possibly Native Americans), the grades of students from these groups were generally overpredicted. The degree of overprediction ranged from somewhat for Hispanic students (with representative values around -.08) to slightly greater for blacks/African Americans and combined minority groups (with typical values around -.11). Bear in mind that the combined minority groups are composed primarily of African American students so that the values for the two groups should be quite similar. As stated earlier, these overprediction figures are based on the commonly used grade scale of 0 to 4. Given the consistency of the findings for blacks/African Americans and Hispanics, it is evident that the overprediction of grades for these minority students is a well-established phenomenon and not an isolated event. However, it is accurate to say that the causes of this phenomenon are not yet completely known or understood.

IV. Sex Differences in Validity and Prediction

In this section, all of the 37 studies conducted since 1974 that investigated sex differences in validity and

prediction are reviewed. The 37 studies can be categorized into one of three types: single institutions, (21 studies), multiple institutions, which generally involved several campuses from the same state higher education system (11 studies), and compilations of findings from a large number of institutions, which were usually based on several years of results (5 studies). Each compilation included results from 80 or more institutions and samples of over 100,000 students. All of the studies reviewed appeared as either journal articles or as conference papers. Note that some of the journal articles appeared in an earlier form as an ACT or ETS research report; in those instances, it is the journal article that is referenced. Table 8 provides a summary of the important characteristics of each of the 37 studies. In addition, a brief description of each study is provided in the Appendix.

Most of the 37 studies are of differential prediction

TABLE 8

Studies Reviewed in Section 4

Authors	Year	Type	Institution	Classes	Sample N	DV/DP	Criterion	Predictors
Baggaley	74	S	Pennsylvania	E69	529	DP	CGPA	SAT V, SAT M, HSGPA
Baron & Norman	92	S	Pennsylvania	E83,84	3816	DP	CGPA	SAT V+M, HSR, ACH
Boli et al.	85	S	Stanford	AY77-78	1154	DV	ICG	SAT M
Bridgeman & Lewis	96	M	43 colleges	E85	33139	DP	ICG	SAT M, HSGPA
Bridgeman et al.	2000	M	23 colleges	E94,95	93139	DV/DP	FGPA	SAT V, SAT M, HSGPA
Bridgeman & Wendler	91	M	9 universities	E86	12124	DP	ICG	SAT M, HS grades
Chou & Huberty	90	S	Georgia	E87	3378	DP	QGPA	SAT V, SAT M, HSGPA
Clark & Grandy	84	C	41 colleges	E79	Not Given	DV/DP	FGPA	SAT V, SAT M, HSGPA
Cowen & Fiori	91	S	CSU, Hayward	E88,89	972	DV/DP	FGPA	SAT V, SAT M, HSGPA
Crawford et al.	86	S	W. Virginia State*	E85	1121	DV/DP	CGPA	ACT, HSGPA
Dalton	76	S	Indiana	E61-74	17533	DV	SGPA	SAT V+M, HSGPA
Elliott & Strenta	88	S	Dartmouth	G86	927	DV/DP	ICG,CGPA	SAT V+M, HSGPA, ACH
Farver et al.	75	S	Maryland	E68,69	559	DV/DP	CGPA	SAT V, SAT M, HSGPA
Fincher	74	M	29 GA colleges	E58-70	Not Given	DV	FGPA	SAT V, SAT M, HSGPA
Gamache & Novick	85	S	Iowa*	E78	2160	DV/DP	CGPA	ACT, ACT subtests
Hand & Pranther	85	M	31 GA colleges	E83	45067	DV	CGPA	SAT V, SAT M, HSGPA
Hogrebe et al.	83	S	Georgia*	AY77-79	345	DP	FGPA	SAT V, SAT M, HSGPA
Houston & Sawyer	88	M	17 colleges	AY83-87	11821	DP	ICG	ACT, ACT subtests, HSGPA, HS grades
Larson & Scontrino	76	S	U. Washington*	G66-73	1457	DV	CGPA	SAT V SAT M, HSGPA
Leonard & Jiang	95	S	UC, Berkeley	E86,87,88	~10000	DP	CGPA	SAT V, SAT M, HSGPA, ACH
McCornack & McLeod	88	S	San Diego State	AY85-86	57119	DP	ICG	SAT V, SAT M, HSGPA
McDonald & Gawkoski	79	S	Marquette	E63-72	402	DV	Honors Pr	SAT V, SAT M, HSGPA
Morgan	90	C	198 colleges	E78,81,85	278074	DV	FPGA	SAT V, SAT M, HSGPA
Nettles et al.	86	M	30 colleges	Not Given	4094	DP	CGPA	SAT V+M, HSGPA, other vars.
Noble et al.	96	C	>80 colleges	Not Given	Not Given	DP	ICG	ACT subtests, HS grades
Pennock-Román	94	M	4 universities	E88?	14868	DP	FGPA	SAT V, SAT M, HSGPA
Ramist et al.	94	M	45 colleges	E82,85	46379	DV/DP	ICG,FGPA	SAT V, SAT M, HSGPA
Ramist & Weiss	90	C	253 colleges	AY73-88	Not Given	DV	FGPA	SAT V, SAT M, HSGPA
Rowan	78	S	Murray State	Not Given	2289	DV	CGPA	ACT
Saka	91	S	Hawaii	E88	1345	DV	FGPA	SAT V, SAT M, HSGPA
Sawyer	86	C	256 colleges	AY74-77	134600	DP	FGPA	ACT subtests, HS grades
Stricker et al.	93	S	Rutgers	E88	4351	DP	SGPA	SAT V, SAT M, HSGPA
Sue & Abe	88	M	8 UC campuses	E84	5113	DV/DP	FGPA	SAT V, SAT M, HSGPA
Wainer & Steinberg	92	M	51 colleges	AY82-86	46920	DP	ICG	SAT M
Wilson	80	S	Penn State Univ.*	E71	1275	DV	FGPA,CGPA	SAT V, SAT M, HSGPA
Young	91a	S	Stanford	E82	1462	DV/DP	CGPA	SAT V, SAT M, HSGPA
Young	94	S	Rutgers	E85	3703	DV/DP	CGPA	SAT V, SAT M, HSR

*An asterisk after the institution's name means that the study did not identify the institution but is likely based on the description in the study.
Type: C = compilation, M = multiple campuses, S = single institution. **Classes:** AY = academic year, E = entering year, G = graduation year.
DV/DP: DV = differential validity, DP = differential prediction. **Criterion:** CGPA = cumulative GPA, FGPA = first-year GPA, ICG = individual course grades, QGPA = quarter GPA, SGPA = semester GPA. **Predictors:** ACH = College Board Achievement Test scores, ACT = ACT Composite score, SAT V+M = SAT total score, HSR = HS Rank, HS grades = individual course grades.

Studies Reviewed in Section 4

Authors	*Differential Validity Results*	*Differential Prediction Results: Grade Prediction*
Baggaley	R(CGPA):F=.65, M=.52	
Baron & Norman		F: underpredicted CGPA
Boli et al.		
Bridgeman & Lewis		F: underpredicted course grades
Bridgeman et al.	R:F=.45, M=.44	F=+.07, M=-.08
Bridgeman & Wendler		F: underpredicted course grades
Chou & Huberty		F=+.04, M=-.05
Clark & Grandy	mean R:F=.54, M=.50	F=+.05, M=-.04
Cowen & Fiori		F=-.01, M=+.04
Crawford et al.	R2:F=.28, M=.21	
Dalton	median R:F=.56, M=.52	
Elliott & Strenta	R:F=.53, M=.56	F=+.03, M=-.02
Farver et al.	R(CGPA):BM=.52, BF=.42, WM=.55, WF=.67	
Fincher	unweighted mean R:F=.69, M=.58	
Gamache & Novick	median R2:F=.215, M=.184	median for F=+.18 (design 2)
Hand & Pranther	med. adj. R2:BM=.36, BF=.44, WM=.45, WF=.47	
Hogrebe et al.		WM=+.33
Houston & Sawyer		F=+.01, -.02, +.07 (3 first-year courses)
Larson & Scontrino	median R:F=.73, M=.68	
Leonard & Jiang		F=+.10
McCornack & McLeod		F: small amount of underprediction
McDonald & Gawkoski	r:F=.14, .32, .16, M=.00, .17, .18	
Morgan	R:F=.56, .54, .53, M=.53, .49, .48 (3 years)	
Nettles et al.		F: significant underprediction
Noble et al.		
Pennock-Román		median: AF=+.04, BF=+.12, HF=+.05, WF=+.09
Ramist et al.	R:F=.50, M=.46	F=+.06, M=-.06
Ramist & Weiss	med corr r:F=.57, .59, M=.52,.55	
Rowan		
Saka	R2:F=.15, M=.11	
Sawyer		F=+.05, M=-.05
Stricker et al.		F=+.10, M=-.11
Sue & Abe	R:AF=.50, WF=.47, AM=.50, WM=.44	AF=.00, AM=+.03
Wainer & Steinberg		
Wilson	R:MF=.72, WF=.57, MM=.69, WM=.57	
Young	r:SAT V & HSGPA same, SAT M higher for M	F=+.04
Young	R:F=.44, M=.38	F=+.04, M=-.04

Results: *R = multiple correlation, R2 = multiple correlation squared, r = simple correlation.*

or of differential validity and differential prediction. That is, prediction results based on regression analysis were usually reported along with validity coefficients. In the remainder of this section, the findings on differential validity are reported first, followed by the findings on differential prediction. A summary of the results appears at the end of the section.

Differential Validity Findings

The differential validity findings, based on reported multiple correlation coefficients (or squared multiple correlations) of predictors with a criterion are quite consistent with respect to comparisons of male and female students. In general, the magnitude of the correlation coefficients for women is larger than for men. This is true for any single predictor or combinations of predictors including the

Studies Reviewed in Section 4

Authors	Differential Prediction Results: Other
Baggaley	
Baron & Norman	
Boli et al.	Beta in SEM = .00, -.02 for men in 2 courses
Bridgeman & Lewis	F: Std course grade diff: .05 to .22
Bridgeman et al.	
Bridgeman & Wendler	F: d = +.14, +.13, -.01 for 3 math courses
Chou & Huberty	
Clark & Grandy	F: d = +.06
Cowen & Fiori	
Crawford et al.	F: significant underpostdiction
Dalton	
Elliott & Strenta	
Farver et al.	
Fincher	
Gamache & Novick	
Hand & Pranther	
Hogrebe et al.	
Houston & Sawyer	
Larson & Scontrino	
Leonard & Jiang	
McCornack & McLeod	F: 7 courses underpred., 3 courses overpred.
McDonald & Gawkoski	
Morgan	
Nettles et al.	
Noble et al.	F: p(grade of B or better) = +.02 to +.10
Pennock-Román	
Ramist et al.	
Ramist & Weiss	
Rowan	F: higher succ. prob. and survival rate
Saka	
Sawyer	
Stricker et al.	
Sue & Abe	
Wainer & Steinberg	median of -33 SAT M points for women
Wilson	
Young	
Young	

Results: *d = effect size.*

most common set of predictors used in differential validity studies: SAT V and SAT M scores and HSGPA. A total of 12 studies (Table 9) (Baggaley, 1974; Bridgeman, McCamley-Jenkins, and Ervin, 2000; Clark and Grandy, 1984; Dalton, 1976; Elliott and Strenta, 1988; Farver, Sedlacek, and Brooks, 1975; Larson and Scontrino, 1976; Morgan, 1990; Ramist, Lewis, and McCamley-Jenkins, 1994; Sue and Abe, 1988; Wilson, 1980; Young, 1994) reported multiple correlations for men and women using SAT scores plus HS grades (or a slight variation) with either FGPA or CGPA as the criterion measure. A total of 17 coefficients were reported for each sex since several studies reported separate values for different race by sex groups. The median multiple correlation was .51 for men and .54 for women with corresponding means of .52 for men and .55 for women.

Four other studies (Crawford, Alferink, and Spencer, 1986; Gamache and Novick, 1985; Hand and Pranther, 1985; Saka, 1991) reported a total of five squared multiple correlations each for men and for women. The median value of the squared multiple correlations was .21 for men and .28 for women. These squared multiple correlations convert to multiple correlation values of approximately .46 for men and .53 for women and are similar in magnitude to those computed from the studies listed above. Because of rounding, the converted values may be slightly different than that found using more accurate figures. Two additional studies (McDonald and Gawkoski, 1979; Ramist and Weiss, 1990) reported correlations of individual predictors with other criteria (graduating from an honors program in the McDonald and Gawkoski study, individual course grades in the Ramist and Weiss study). In all instances, the magnitude of the correlations for men was smaller than for women.

One additional point worth noting is that in the most selective institutions, the multiple correlations for men are generally higher than those found in less selective institutions such that the values of these correlations are as high as or higher than the comparable values for women at the same institution. This is the opposite of the more common finding in most studies of sex differences where the correlations are generally higher for women. Analysis by degree of institutional selectivity in the studies of Bridgeman, McCamley-Jenkins, and Ervin (2000) and Ramist, Lewis, and McCamley-Jenkins (1994) found that the multiple correlations of the standard set of predictors with FGPA was slightly lower for women than for men when only the most selective colleges were included. This is consistent with the findings reported in studies at two highly selective private institutions: (1) by Elliott and Strenta (1988) on a cohort of Dartmouth College graduates where the multiple correlation with CGPA was slightly higher for men (.56) than for women (.53), and (2) by Young (1991a) on a cohort of Stanford University students where two of the predictors (SAT V and HSGPA) were similarly correlated with CGPA for both men and women, while the third predictor, SAT M, had a substantially higher correlation for men.

Differential Prediction Findings

Differential prediction findings are derived from analyses of residuals from either one of two designs: (1) a multiple

TABLE 9

Differential Validity Results: Men and Women

Authors	Criterion	Predictors	Results
Baggaley	CGPA	SAT V, SAT M, HSGPA	R(CGPA):F=.65, M=.52
Bridgeman et al.	FGPA	SAT V, SAT M, HSGPA	R:F=.45, M=.44
Clark & Grandy	FGPA	SAT V, SAT M, HSGPA	mean R:F=.54, M=.50
Crawford et al.	CGPA	ACT, HSGPA	R2:F=.28, M=.21
Dalton	SGPA	SAT V+M, HSGPA	median R:F=.56, M=.52
Elliott & Strenta	ICG,CGPA	SAT V, SAT M, HSGPA, ACH	R:F=.53, M=.56
Farver et al.	CGPA	SAT V, SAT M, HSGPA	R(CGPA):BM=.52, BF=.42, WM=.55, WF=.67
Gamache & Novick	CGPA	ACT, ACT subtests	median R2:F=.215, M=.184
Hand & Pranther	CGPA	SAT V, SAT M, HSGPA	med. adj. R2:BM=.36, BF=.44, WM=.45, WF=.47
Larson & Scontrino	CGPA	SAT V, SAT M, HSGPA	median R:F=.73, M=.68
McDonald & Gawkoski	Honors Pr	SAT V, SAT M, HSGPA	r:F=.14, .32, .16, M=.00, .17, .18
Morgan	FGPA	SAT V, SAT M, HSGPA	R:F=.56, .54, .53, M=.53, .49, .48 (3 years)
Ramist et al.	ICG,FGPA	SAT V, SAT M, HSGPA	R:F=.50, M=.46
Ramist & Weiss	FGPA	SAT V, SAT M, HSGPA	med. corr. r:F=.57, .59, M=.52, .55
Saka	FGPA	SAT V, SAT M, HSGPA	R2:F=.15, M=.11
Sue & Abe	FGPA	SAT V, SAT M, HSGPA	R:AF=.50, WF=.47, AM=.50, WM=.44
Wilson	FGPA, CGPA	SAT V, SAT M, HSGPA	R:MF=.72, WF=.57, MM=.69, WM=.57
Young	CGPA	SAT V, SAT M, HSGPA	r:SAT V & HSGPA same, SAT M higher for M
Young	CGPA	SAT V, SAT M, HSR	R:F=.44, M=.38

Criterion: CGPA = cumulative GPA, FGPA = first-year GPA, ICG = individual course grades, SGPA = semester GPA. *Predictors:* ACH = College Board Achievement Test scores, ACT = ACT Composite score, SAT V+M = SAT total score, HSR = HS Rank. *Results:* R = multiple correlation, R2 = multiple correlation squared, r = simple correlation.

regression equation based on a combined sample of students, or (2) from an equation computed from a sample of male students and then applied to female students. In general, with rare exceptions, the findings consistently point to a significant underprediction of women's grades. This is true whether the regression equation used came from the first or second design cited above. In other words, it is generally the case that the actual grades earned by women are higher than that predicted from test scores and HSGPA.

A total of 21 studies examined differential prediction of college grades by sex (Table 10). Of these, 14 studies (Bridgeman, McCamley-Jenkins, and Ervin, 2000; Chou and Huberty, 1990; Clark and Grandy, 1984; Cowen and Fiori, 1991; Elliott and Strenta, 1988; Gamache and Novick, 1985; Leonard and Jiang, 1995; Pennock-Román, 1994; Ramist, Lewis, and McCamley-Jenkins, 1994; Sawyer, 1986; Stricker, Rock, and Burton, 1993; Sue and Abe, 1988; Young, 1991a; Young, 1994) reported differential prediction results in sufficient detail that could be further analyzed. All of these studies except for Gamache and Novick and Sawyer used the standard set of predictors (SAT scores and HSGPA) to forecast either FGPA or CGPA. Gamache and Novick used ACT subtest and composite scores, and Sawyer used ACT subtest scores and HS course grades.

Five additional studies (Baron and Norman, 1992; Bridgeman and Lewis, 1996; Bridgeman and Wendler, 1991; McCornack and McLeod, 1988; Nettles, Theony, and

Gosman, 1986) only reported that women's grades (either CGPA or individual course grades) were underpredicted without providing summary statistics. The results from two other studies (Hogrebe, Ervin, Dwinell, and Newman, 1983; Houston and Sawyer, 1988) were not included in the analysis of grade prediction because their methods appeared to depart significantly from the other studies. In the study by Hogrebe, Ervin, Dwinell, and Newman(1983), a significant sex difference in regression intercepts was reported, but the direction of the difference was not given. Furthermore, the sample in this study consisted of students in a developmental studies program (for students who were admitted through a nonstandard admission process) and thus may differ from other samples of students studied. The study by Houston and Sawyer used ACT subtest and composite scores as well as HSGPA and individual HS course grades to predict grades in three college courses. In this study, the mispredictions were small, although women received slightly better grades than was predicted.

Based on the 14 studies with differential prediction results, a total of 17 values were available for analysis (Pennock-Román reported four values, one for each racial/ethnic group in her study). For women, the median amount of underprediction is +.05 (based on a 0-4 grade scale) with a mean of +.06. Of the 17 values, only one was for overprediction for women (a negligible amount at -.01) and another was for zero misprediction. An examination of the three studies with the largest sample sizes (Bridgeman, McCamley-Jenkins, and Ervin, 2000; Ramist,

TABLE 10

Differential Prediction Results: Men and Women

Authors	Criterion	Predictors	Results
Baron & Norman	CGPA	SAT V+M, HSR, ACH	W: underpredicted CGPA
Bridgeman & Lewis	ICG	SAT M, HSGPA	W: underpredicted course grades
Bridgeman et al.	FGPA	SAT V, SAT M, HSGPA	W = +.07, M = -.08
Bridgeman & Wendler	ICG	SAT M, HS grades	W: underpredicted course grades
Chou & Huberty	QGPA	SAT V, SAT M, HSGPA	W = +.04, M = -.05
Clark & Grandy	FGPA	SAT V, SAT M, HSGPA	W = +.05, M = -.04
Cowen & Fiori	FGPA	SAT V, SAT M, HSGPA	W = .01, M = +.04
Elliott & Strenta	ICG, CGPA	SAT V+M, HSGPA, ACH	W = +.03, M = -.02
Gamache & Novick	CGPA	ACT, ACT subtests	median for W = +.18 (design 2)
Hogrebe et al.	FGPA	SAT V, SAT M, HSGPA	WM = +.33
Houston & Sawyer	ICG	ACT, ACT subtests, HSGPA, HS grades	W = +.01, -.02, +.07 (3 first-year courses)
Leonard & Jiang	CGPA	SAT V, SAT M, HSGPA, ACH	W = +.10
McCornack & McLeod	ICG	SAT V, SAT M, HSGPA	W: small amount of underprediction
Nettles et al.	CGPA	SAT V+M, HSGPA, other vars.	W: significant underprediction
Pennock-Román	FGPA	SAT V, SAT -M, HSGPA	median: AW = +.04, BW = +.12, HW = +.05, WW = +.09
Ramist et al.	ICG, FGPA	SAT V, SAT M, HSGPA	W = +.06, M = -.06
Sawyer	FGPA	ACT subtests, HS grades	W = +.05, M = -.05
Stricker et al.	SGPA	SAT V, SAT M, HSGPA	W = +.10, M = -.11
Sue & Abe	FGPA	SAT V, SAT M, HSGPA	AW = .00, AM = +.03
Young	CGPA	SAT V, SAT M, HSGPA	W = +.04
Young	CGPA	SAT V, SAT M, HSR	W = +.04, M = -.04

Criterion: CGPA = cumulative GPA, FGPA = first-year GPA, ICG = individual course grades, QGPA = quarter GPA, SGPA = semester GPA. *Predictors:* ACH = College Board Achievement Test scores, ACT = ACT Composite score, SAT V+M = SAT total score, HSR = HS Rank, HS grades = individual course grades.

Lewis, and McCamley-Jenkins, 1994; Sawyer, 1986) yielded the same results. As is the case with differential validity, the findings from the most selective institutions appears to be somewhat different from those found at less selective institutions. Four studies at highly selective institutions, Elliott and Strenta (at Dartmouth), Leonard and Jiang (at the University of California, Berkeley), Sue and Abe (at the eight University of California undergraduate campuses), and Young (at Stanford), found on average slightly less underprediction of women's grades (mean of +.04).

In addition to the results above on predicting GPAs, seven additional studies (Boli, Allen, and Payne, 1985; Clark and Grandy, 1984; Crawford, Alferink, and Spencer, 1986; McCornack and McLeod, 1988; Noble, Crouse, and Schulz, 1996; Rowan, 1978; Wainer and Steinberg, 1992) reported results on grade prediction in terms of effect sizes or rates on success outcomes (see Table 11). In addition to the grade prediction results reported above, Bridgeman and Wendler and Bridgeman and Lewis also reported small-to-moderate effect sizes in favor of women

TABLE 11

Other Prediction Results: Men and Women

Authors	Criterion	Predictors	Results
Boli et al.	ICG	SAT M	Beta in SEM = .00, -.02 for men in 2 courses
Bridgeman & Lewis	ICG	SAT M, HSGPA	W: Std. course grade diff.: .05 to .22
Bridgeman & Wendler	ICG	SAT M, HS grades	W: d = +.14, +.13, -.01 for 3 math courses
Clark & Grandy	FGPA	SAT V, SAT M, HSGPA	W: d = +.06
Crawford et al.	CGPA	ACT, HSGPA	W: significant underpostdiction
McCornack & McLeod	ICG	SAT V, SAT M, HSGPA	W: 7 courses underpred., 3 courses overpred.
Noble et al.	ICG	ACT subtests, HS grades	W: p (grade of B or better) = +.02 to +.10
Rowan	CGPA	ACT	W: higher succ. prob. and survival rate
Wainer & Steinberg	ICG	SAT M	median of -33 SAT M points for women

Criterion: CGPA = cumulative GPA, FGPA = first-year GPA, ICG = individual course grades. *Predictors:* ACT = ACT Composite score, HSR = HS Rank, HS grades = individual course grades. *Results:* d = effect size.

in predicting individual college course grades.

Boli, Allen, and Payne reported a small negative effect for men in a structural equation model used to predict grades in two science courses at Stanford University. Clark and Grandy reported a small effect size in favor of women in predicting FGPA in a study of 41 colleges. Crawford, Alferink, and Spencer found that women's CGPAs were significantly underpostdicted (from a retrospective prediction study) from ACT composite score and HSGPA. McCornack and McLeod reported that women's grades in seven first-year courses at San Diego State University were underpredicted from SAT scores and HSGPA but overpredicted in three other courses. Noble, Crouse, and Schulz reported that women had higher rates of obtaining a grade of B or better in four first-year college courses than was predicted from ACT subtest scores and HS course grades. Rowan, in a study at Murray State University, found that women had a higher rate of obtaining a CGPA greater than 2.0 and of graduating than was predicted from ACT composite scores. Finally, Wainer and Steinberg reported that in a study of first-year college mathematics courses, women had scored, on average, about 33 points lower on SAT M than men who had taken the same course and received the same grade.

Summary

The differential validity results indicated that the magnitude of correlations between predictors and several different grade criteria are slightly, but consistently, higher for women than for men (although this appears to be less true at the most selective institutions). From the differential prediction studies, we can state that underprediction of women's GPAs is the most common finding, although the degree of misprediction is less than what is generally found for racial/ethnic minority groups such as blacks/African Americans and Hispanics. At the most selective colleges and universities, underprediction was still found, although the magnitude may be somewhat less than that at other institutions.

V.   Summary, Conclusions, and Future Research

Summary

In this report, all studies of differential validity and/or differential prediction in college admission testing published since 1974 were reviewed. A total of 49 studies found in journal articles, research reports, or conference papers are included. Of these, 29 are studies of racial/ethnic differences in differential validity/prediction and 37 are studies of sex differences (17 studies are of both types of differences). The studies that were located are classified according to the number of institutions from which the data originated: single institutions, multiple institutions (typically, several campuses of the same higher education system), and compilations based on a large number of (usually unrelated) institutions. Sample size in the studies ranged from a minimum of 214 to a maximum of 278,074. The samples for single-institution studies typically consisted of several hundred to a few thousand students; for multiple-institution studies, the samples are generally from around 5,000 to 20,000 students; and for compilations of many institutions, the samples include over 100,000 students.

With respect to racial/ethnic differences, the minority groups examined include Asian Americans, blacks/African Americans, Hispanics, Native Americans, and combined samples of minority students. In studies of racial/ethnic differences, whites or Caucasians are used as the reference group. In studies of sex differences, males are usually considered the reference group, while females are the focal group. In the studies reviewed, the most frequently used criterion measure was the first-year grade point average (FGPA) in college. Other outcome measures included two-, three-, or four-year cumulative GPA (CGPA), semester or term GPA, and individual course grade. The set of predictor variables most commonly used was SAT verbal score, SAT mathematical score, and high school GPA (HSGPA). Occasionally, test scores alone were used as predictors as well as total SAT score (SAT V+M). ACT Composite score and ACT subtest scores also functioned as predictors, either together or separately.

The studies of minority students yielded mixed results for differential validity; in contrast, the findings are more consistent in terms of differential prediction. The pattern of correlations between predictors and criterion differs by group with generally lower values (for blacks/African Americans and Hispanics) and similar values (for Asian Americans) when compared to whites. Of course, specific studies may exhibit results at variance from this general pattern; however, the previous statement is an accurate summary of the studies that were reviewed. To date, too few studies with Native American samples have been conducted to allow for meaningful statements concerning differential validity/prediction.

For differential prediction, the common finding is one of overprediction of college grades for all of the

minority groups studied. The degree of overprediction varied by group with, on average, the greatest overprediction observed for blacks/African Americans and combined minority groups and slightly less overprediction for Hispanics and possibly Asian Americans (although underprediction was found using adjusted grades for this group). In comparison to the earlier results reported by Breland (1979) and Duran (1983), the degree of overprediction for minority groups appears to have diminished somewhat compared to studies published two or three decades ago. However, overprediction is still the rule rather than the exception in the majority of the studies reviewed here.

The results from the studies of sex differences are easier to summarize. In terms of differential validity, it is generally the case that the correlations between predictors and criterion are higher for women than for men. In other words, there is a stronger association between the commonly used academic predictors and subsequent college grades for women than for men. The differences between men and women in the magnitude of the correlations are small but persistent. With regard to differential prediction, the general finding from these studies is one of underprediction of women's college grades. That is, women generally earn higher grades than predicted from their prior academic records. The magnitude of the underprediction typically averaged around +.05 to +.06 (on a 4-point grade scale). As a basis for comparison, this is about one-half of the average overprediction for blacks/African Americans and somewhat less than the overprediction for Hispanics. Note that in the most selective colleges and universities, the correlations for men and women appear to be equal, while the degree of underprediction for women's grades appears to be somewhat less than in other institutions. For women, the magnitude of underpredicted grades is smaller than that reported in earlier studies (from the 1960s and early 1970s), but the phenomenon has clearly persisted.

One additional set of analyses deserves mention: The seven studies (Crawford, Alferink, and Spencer, 1986; Gamache and Novick, 1985; Houston and Sawyer, 1988; Maxey and Sawyer, 1981; Noble, Crouse, and Schulz, 1996; Rowan, 1978; Sawyer, 1986) that used ACT test scores (composite scores, subtest scores, or both) were examined separately to determine if these results differed from the studies that used SAT scores.

Comparative analysis between the two admission tests is difficult for two critical reasons: (1) the validation approaches used for the ACT studies differed in important ways from the other studies, and (2) the samples of colleges and universities for which ACT results are based are often quite different since there are geographical differences in the use of the two tests. With respect to the

first point, ACT subtest scores were commonly used as predictors (sometimes with composite scores) along with individual HS course grades or HSGPA. In contrast, there is no comparable set of predictors for studies using the SAT. In fact, only one of the seven studies used a standard set of predictors, ACT composite scores and HSGPA. In addition, some of the studies focused on forecasting success rates in specific college courses rather than on composite grades. With regard to the second point, differences in the samples of institutions using the two tests is a confounding factor. This is already true within any testing program so comparisons across programs are quite tenuous. For example, none of the seven studies reported results on Asian Americans, and only one study gave results for Hispanic students. Given these caveats, a tentative conclusion is that the predictive validity for the two admission tests appears to be of similar magnitude, but much more research is required before one can comment further on this point.

Conclusions

An inspection of Tables 1 and 8 indicates the large degree of variation in the characteristics of the studies reviewed in this report. The studies span an important period in American higher education (from the mid-1970s to the present), one marked by significant changes in student composition as well as evolving educational policies that were subjected to legal challenges at times. The studies differed on several important characteristics such as year published, type and number of institutions involved, sample size, definition and number of cohorts, minority groups studied (in the case of racial/ethnic differences), predictor and criterion variables used, and type of results reported. It would be accurate to state that no two studies were conducted in exactly the same fashion. In some cases, the issue of differential validity/prediction was not central to the author's larger research questions. Thus, these studies did not lend themselves easily to neat summaries of their findings.

The first main conclusion that can be drawn from this review of research is that group differences do occur in validity and prediction. Based on the evidence from studies conducted over this period of 25+ years, small-to-moderate differences in the magnitude of validity coefficients and in the accuracy of prediction equations have been consistently observed. This is true for studies of racial/ethnic and of sex differences.

A second conclusion that can be drawn is that these differences varied considerably depending upon the group of interest. Among the racial/ethnic groups studied, no two groups shared the same pattern of validi-

ty/prediction results. Furthermore, substantial differences in the results within a single racial/ethnic group were sometimes observed. By lumping together all of the studies for a single group, potential differences on other variables such as socioeconomic status, native language, or geographical location are ignored. For example, individuals from a variety of backgrounds (such as Cuban Americans, Mexican Americans, and Puerto Ricans) are collectively labeled as Hispanics. However, there are considerable differences in the educational and social experiences of students from these different groups. Yet, they are treated as homogeneous entities in educational research studies. As another example, studies involving Asian Americans typically focus on institutions on either the East Coast or the West Coast (usually California). However, the immigration patterns and socioeconomic status of Asian American families in these two areas of the country are radically different. These differences may partly explain the inconsistency of validity/prediction results for Asian American students.

A third conclusion is that group (racial/ethnic and sex) differences have not remained fixed and appear to have moderated somewhat during the time period covered in this review (and possibly continuing an earlier trend). This is a tenuous conclusion since the entire universe of studies is so small that trends are difficult to discern. It is unknown whether this trend towards smaller differences will continue so that at some point in the future, group differences will disappear entirely. It is possible that some influence, as yet unknown, may alter the present trend. One could speculate that recent legal challenges to affirmative action policies in higher education admission might radically alter the results of future studies of differential validity/prediction.

A fourth conclusion is that the major causes of group differences in validity/prediction studies are not yet well known or understood. Some tentative hypotheses have been advanced in the professional literature regarding grade underprediction for women and grade overprediction for minority students. However, it is accurate to state that there is currently no single theory that is widely accepted for either of these phenomena. Racial/ethnic differences are usually attributed to one or more of the following reasons: (1) psycho–social differences in the collegiate experiences of minority students (such as in personal adjustment), (2) differences in precollege academic preparation between minority and white students, (3) institutional factors which may differentially impact minority students' grades either positively or negatively, and (4) statistical and research design artifacts inherent in the manner in which most differential validity/prediction studies are conducted.

Of these rationales, the first and third are the most likely explanations from this author's vantage point. That is, differences in the collegiate experiences of white and minority students, coupled with societal and institutional factors that differentially affect students, may have a greater negative impact on the academic performance of some minority students. In other words, minority students will more likely experience adjustment difficulties in a predominantly white campus environment than is true for most white students. These difficulties may lead to a number of potential outcomes, one of them being lower grades than would be expected based on prior academic achievement.

In contrast, sex differences in validity/prediction have been hypothesized to be the result of one or more of the following factors: (1) differences in the choices of college courses and majors by men and women, (2) differences in the construct validity of grades for men and for women (that is, the assignment of grades is based on different combinations of factors for the two sexes), and (3) differences in the construct validity of admission tests for men and for women (that is, a gender bias in the meaning of test scores). Presently, all of these theories are considered plausible, although none appears to be a complete explanation for the results in the studies reviewed. Results from studies that adjusted grades for course difficulty lend support to the first hypothesis. Sex differences in validity/prediction are smaller or nonexistent in these studies, since men and women choose courses and majors at different rates.

At the most selective institutions, grades of both men and women are more predictable from the traditional predictors of test scores and high school grades, and misprediction is not as pronounced. One explanation for this is that behaviors unrelated to those measured by admission tests, such as failing to attend class or completing assignments in a timely fashion, may be more common among men and thus makes predicting men's grades more difficult. In highly competitive colleges and universities, since it is more likely that men and women will attend classes and complete assignments faithfully, the grades of men and women are equally valid. Thus, the utility of admission information should be equal for both sexes (Stricker, Rock, and Burton, 1993). It follows then that in less selective institutions, the hypothesis of sex differences in the construct validity of college grades may be a plausible explanation for observed differences in validity/prediction.

Future Research

A number of possible avenues for additional research on differential validity/prediction is evident based on the review conducted here: (1) The number of published studies for most racial/ethnic groups is small; consequently, it is difficult to draw definitive conclusions about differences in validity and/or prediction. In particular, more studies of Asian Americans, Hispanics, and Native Americans are needed to further advance our understanding of the academic achievement of these groups. Furthermore, it may be necessary to refine our definitions of these groups, as there is evidence that lumping together various subgroups under a single racial/ethnic classification tends to confound validity/prediction results. (2) The main causes of observed sex differences are still to be discovered. Given the importance and pervasiveness of these differences, much more needs to be learned about why sex differences still persist after so many decades of investigation. (3) New methodologies for exploring differential validity/prediction (beyond correlation/regression studies) may aid our understanding of these topics. For example, the approach perfected by Noble, Crouse and Schulz (1996) may help shed new light apart from earlier studies. In addition, other methods, perhaps to be developed at some future date, for studying validity/prediction may eventually lead to a higher level of understanding of group differences and bring us closer to the democratic goal of equal opportunity and access to higher education for students of all backgrounds.

References

Abelson, R.P. (1952). Sex differences in predictability of college grades. *Educational and Psychological Measurement, 12*, 638–644.

ACT. (1997). *ACT Assessment technical manual.* Iowa City, IA: Author.

American College Testing Program (1973). *Assessing students on the way to college: Technical report for the ACT Assessment Program.* Iowa City, IA: Author.

American College Testing Program (1987). *The ACT Assessment Program technical manual.* Iowa City, IA: Author.

American Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51* (2, Part 2).

American Psychological Association (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: Author.

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

Arbona, C., & Novy, D. M. (1990). Noncognitive dimensions as predictors of college success among black, Mexican American, and white students. *Journal of College Student Development, 31*, 415–422.

Baggaley, A. R. (1974). Academic prediction at an Ivy League college, moderated by demographic variables. *Measurement and Evaluation in Guidance, 6*, 232–235.

Baron, J., & Norman, M. F. (1992). SATs, achievement tests, and high-school class rank as predictors of college performance. *Educational and Psychological Measurement, 52*, 1047–1055.

Boli, J., Allen, M. L., & Payne, A. (1985). High-ability women and men in undergraduate mathematics and chemistry courses. *American Educational Research Journal, 22*, 605–626.

Boehm, V. R. (1972). Negro–white differences in validity of employment and training selection procedures. *Journal of Applied Psychology, 56*, 33–39.

Bowers, J. (1970). The comparison of GPA regression equations for regularly admitted and disadvantaged freshmen at the University of Illinois. *Journal of Educational Measurement, 7*, 219–225.

Breland, H. M. (1978). *Population validity and college entrance measures.* College Board Research and Development Report. RDR 78-79, No. 2. Princeton, NJ: Educational Testing Service.

Breland, H. M. (1979). *Population validity and college entrance measures.* Research Monograph No. 8. New York: College Board.

Bridgeman, B., & Lewis, C. (1996). Gender differences in college mathematics grades and SAT M scores: A reanalysis of Wainer and Steinberg. *Journal of Educational Measurement, 33*, 257–270.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning Test* (College Board Report No. 2000-1). New York: College Board.

Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and grades in college mathematics courses. *Journal of Educational Psychology, 83*, 275–284.

Brown, J. L., & Lightsey, R. (1970). Differential predictive validity of SAT scores for freshman college English. *Educational and Psychological Measurement, 30*, 961–965.

Calkins, D. S., & Whitworth, R. (1974). *Differential prediction of freshman grade-point average for sex and two ethnic classifications at a southwestern university.* El Paso, TX: University of Texas (ERIC Document No. 102 199).

Chou, T., & Huberty, C.J. (1990). *A freshman admissions prediction equation: An evaluation and recommendation.* Athens, GA: University of Georgia (ERIC Document Reproduction Service No. ED 333 081).

Clark, M. J., & Grandy, J. (1984). *Sex differences in the academic performance of SAT takers* (College Board Report No. 84-8). New York: College Board.

Cleary, T. A. (1968). Test bias: Prediction of grades for Negro

and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cleary, T. A. & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28,* 61–75.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A. & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30,* 15–41.

Clewell, B. C., & Joy, M. F. (1988). *The national Hispanic scholar awards program: A descriptive analysis of high-achieving Hispanic students* (College Board Report No. 88-10). New York: College Board.

College Board. (1999). *1999 college-bound seniors: A profile of SAT program test takers.* New York: Author.

Cowen, S., & Fiori, S. J. (1991, November). Appropriateness of the SAT in selecting students for admission to California State University, Hayward. Paper presented at the annual meeting of the California Educational Research Association, San Diego, CA (ERIC Document Reproduction Service No. ED 343 934).

Crawford, P. L., Alferink, D. M., & Spencer, J. L. (1986). *Postdictions of college GPAs from ACT composite scores and high school GPAs: Comparisons by race and gender.* West Virginia State College (ERIC Document Reproduction Service No. ED 326 541).

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

Dalton, S. (1976). A decline in the predictive validity of the SAT and high school achievement. *Educational and Psychological Measurement, 36,* 445–448.

Davis, J. A., & Kerner-Hoeg, S. (1971). *Validity of pre-admission indices for blacks and whites in six traditionally white public universities in North Carolina.* Project Report PR-71-15. Princeton, NJ: Educational Testing Service.

Dittmar, N.(1977). *A comparative investigation of the predictive validity of admissions criteria for Anglos, Blacks, and Mexican Americans.* Unpublished doctoral dissertation, The University of Texas at Austin.

Drasgow, F., & Kang, T. (1984). Statistical power of differential validity and differential prediction analyses for detecting measurement nonequivalence. *Journal of Applied Psychology, 69,* 498–508.

Duran, R. P. (1983). *Hispanics' education and background: Predictors of college achievement.* New York: College Board.

Ekstrom, R. B. (1994). *Gender differences in high school grades: An exploratory study* (College Board Report No. 94-3). New York: College Board.

Elliott, R., & Strenta, A. C. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement, 25,* 333–347.

Farver, A. S., Sedlacek, W. E., & Brooks, G. C. (1975). Longitudinal prediction of university grades for blacks and whites. *Measurement and Evaluation in Guidance, 7,* 243–250.

Fincher, C. (1974). Is the SAT worth its salt? An evaluation of the use of the Scholastic Aptitude Test in the university system of Georgia over a thirteen-year period. *Review of Educational Research, 44,* 293–305.

Ford, S. F. & Campos, S. (1977). *Summary of validity data from the Admission Testing Program Validity Study Service.* New York: College Entrance Examination Board.

Gamache, L. M., & Novick, M. R. (1985). Choice of variables and gender differentiated prediction within selected academic programs. *Journal of Educational Measurement, 22,* 53–70.

Goldman, R. D., & Hewitt, B. N. (1975). An investigation of test bias for Mexican American college students. *Journal of Educational Measurement, 12,* 187–196.

Goldman, R. D., & Hewitt, B. N. (1976). Predicting the success of black, Chicano, oriental, and white college students. *Journal of Educational Measurement, 13* (2), 107–117.

Goldman, R. D., & Richards, R. (1974). The SAT prediction of grades for Mexican American versus Anglo American students at the University of California, Riverside. *Journal of Educational Measurement, 11,* 129–135.

Goldman, R. D., & Widawski, M. H. (1976a). An analysis of types of errors in the selection of minority college students. *Journal of Educational Measurement, 13,* 185–200.

Goldman, R. D., & Widawski, M. H. (1976b). A within-subjects technique for comparing college grading standards: Implications in the validity of the evaluation of college achievement. *Educational and Psychological Measurement, 36,* 381–390.

Grant, C. A., & Sleeter, C. E. (1986). Race, class, and gender in education research: An argument for integrative analysis. *Review of Educational Research, 56,* 195–211.

Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika, 15,* 91–114.

Hand, C. A., & Prather, J. E. (1985, April) *The predictive validity of Scholastic Aptitude Test scores for minority college students.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL (ERIC Document Reproduction Service No. ED 261 093).

Hawkins, B. D. (1993). Socio-economic family background: Still a significant influence on SAT scores. *Black Issues in Higher Education, 10,* 14–16.

Hogrebe, M. C., Ervin, L., Dwinell, P. L., & Newman, (1983). The moderating effects of gender and race in predicting the academic performance of college developmental students. *Educational and Psychological Measurement, 43,* 523–530.

Houston, W., & Sawyer, R. (1988). *Central prediction systems for predicting specific course grades* (Research Report No. 88-4). Iowa City, IA: American College Testing.

Hunter, J. E. & Schmidt, F. L. (1978). Differential and single group validity of employment tests by race: A critical analysis of three recent studies. *Journal of Applied Psychology, 63,* 1–11.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721–735.

Jones, L. V., & Appelbaum, M. I. (1989). Psychometric methods. *Annual Review of Psychology, 40,* 23–43.

Khan, S. B. (1973). Sex differences in predictability of academic achievement. *Measurement and Evaluation in Guidance, 6,* 88–91.

Larson, J. R., & Scontrino, M. P. (1976). The consistency of high school grade point average and of the verbal and mathematical portions of the Scholastic Aptitude Test of the College Entrance Examination Board, as predictors of college performance: An eight year study. *Educational and Psychological Measurement, 36,* 439–443.

Leonard, D. K., & Jiang, J. (1995, April). *Gender bias in the college predictions of the SAT.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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

Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology, 63,* 507–512.

Linn, R. L. (1982a). Admissions testing on trial. *American Psychologist, 37,* 279–291.

Linn, R. L. (1982b). Ability testing: Individual differences, prediction and differential prediction. In R. L. Linn (Ed.), *Ability testing: Uses, consequences, and controversies.* Washington, DC: National Academy Press.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21,* 3–47.

Linn, R. L. (1990). Admissions testing: Recommended uses, validity, differential prediction, and coaching. *Applied Measurement in Education, 3,* 313–329.

Linn, R. L. (1994). Fair test use: Research and policy. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel Selection and Classification,* (pp. 363–375). Hillsdale, NJ: Lawrence Erlbaum.

Lowman, R., & Spuck, D. (1975) Predictors of college success for the disadvantaged Mexican American. *Journal of College Student Personnel, 16,* 40–48.

Maxey, J., & Sawyer, R. (1981, July). *Predictive validity of the ACT Assessment for Afro-American/Black, Mexican-American/Chicano, and Caucasian-American/White students* (ACT Research Bulletin 81-1). Iowa City, IA: American College Testing.

McCornack, R. L. (1983). Bias in the validity of predicted college grades in four ethnic minority groups. *Educational and Psychological Measurement, 43,* 517–522.

McCornack, R. L., & McLeod, M. M. (1988). Gender bias in the prediction of college course performance. *Journal of Educational Measurement, 25,* 321–331.

McDonald, R. T., & Gawkoski, R. S. (1979). Predictive value of SAT scores and high school achievement for success in a college honors program. *Educational and Psychological Measurement, 39,* 411–414.

Mestre, J. P. (1981). Predicting academic achievement among bilingual Hispanic college technical students. *Educational and Psychological Measurement, 41,* 1255–1264.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.), pp. 13–103. New York: Macmillan.

Merritt, R. (1972). The predictive validity of the American College Test for students from low socioeconomic levels. *Educational and Psychological Measurement, 32,* 443–445.

Moffatt, G. K. (1993, February). *The validity of the SAT as a predictor of grade point average for nontraditional college students.* Paper presented at the annual meeting of the Eastern Educational Research Association, Clearwater Beach, FL (ERIC Document Reproduction Service No. ED 356 252).

Morgan, R. (1990). Analyses of predictive validity within student categorizations. In Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L., *Predicting college grades: An analysis of institutional trends over two decades* (pp. 225–238). Princeton, NJ: Educational Testing Service.

Murphy, S. H. (1992). Closing the gender gap: What's behind the differences in test scores, What can be done about it? *The College Board Review, (163),* 18–25, 36.

Nettles, M. T., Thoeny, R., & Gosman, E. J. (1986). Comparative and predictive analyses of black and white students' college achievement and experiences. *Journal of Higher Education, 57,* 289–318.

Noble, J. P., (1991). *Predicting college grades from ACT Assessment scores and high school course work and grade information* (Research Report No. 91-3). Iowa City, IA: American College Testing.

Noble, J., Crouse, J., & Schulz, M. (1996). *Differential prediction/impact on course placement for ethnic and gender groups* (Research Report No. 96-8). Iowa City, IA: American College Testing.

Noble, J. P., & Sawyer, R. L. (1989). Predicting grades in college freshman English and mathematics courses. *Journal of College Student Development, 30,* 345–353.

Novick, M. R. (1982). Educational testing: Inferences in relevant subpopulations. *Educational Researcher, 11,* 6–10.

Pearson, B. Z. (1993). Predictive validity of the Scholastic Aptitude Test for Hispanic bilingual students. *Hispanic Journal of Behavioral Sciences, 15,* 342–356.

Pennock-Román, M. (1988). *The status of research on the Scholastic Aptitude Test and Hispanic students in postsecondary education* (Research Report No. 88-36). Princeton, NJ: Educational Testing Service.

Pennock-Román, M. (1990). *Test validity and language background: A study of Hispanic-American students at six universities.* New York: College Board.

Pennock-Román, M. (1994). *College major and gender differences in the prediction of college grades* (College Board Report No. 94-2). New York: College Board.

Pfeifer, C. M., & Sedlacek, W. E. (1971). The validity of academic predictors for black and white students at a predominantly white university. *Journal of Educational Measurement, 8,* 253–260.

Ramist, L. (1984). Predictive validity of the ATP tests. In T. F. Donlon (Ed.), *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests* (pp. 141–170). New York: College Board.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Report No. 93-1). New York: College Board.

Ramist, L., & Weiss, G. (1990). The predictive validity of the SAT, 1964 to 1988. In Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L., *Predicting college grades: An analysis of institutional trends over two decades* (pp. 117–140). Princeton, NJ: Educational Testing Service.

Reynolds, C. R. (1982). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), *Handbook of Methods for Detecting Test Bias* (pp. 199–227). Baltimore, MD: Johns Hopkins University Press.

Rowan, R. W. (1978). The predictive value of the ACT at Murray State University over a four-year college program. *Measurement and Evaluation in Guidance, 11,* 143–149.

Saka, T. T. (1991). High school GPA, SAT scores and college academic achievement for University of Hawaii freshmen. *Pacific Educational Research Journal, 7,* 19–32.

Sanber, S. R., & Millman, J. (1987). *Gender and race effects on standardized tests predictive validity: A meta-analytical study.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C. (ERIC Document Reproduction Service No. ED 286 914).

Sawyer, R. (1986). Using demographic subgroup and dummy variable equations to predict college freshman grade average. *Journal of Educational Measurement, 23,* 131–145.

Schmidt, F. L., Berner, J. G. & Hunter, J. E. (1973). Racial differences in validity of employment tests: Reality or illusion? *Journal of Applied Psychology, 53,* 5–9.

Schmidt, F. L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior, 33,* 272–292.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology, 33,* 705–724.

Schrader, W. B. (1971). The predictive validity of College Board admissions tests. In W. H. Angoff (Ed.), *The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests.* New York: College Board.

Scott, C. (1976). *Longer-term predictive validity of college admission tests for Anglo, Black, and Mexican American students.* New Mexico Department of Educational Administration, University of New Mexico.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9–30). Baltimore, MD: Johns Hopkins University Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

Siegelman, M. (1971). SAT and high school average predictions of four year college achievement. *Educational and Psychological Measurement, 31,* 947–950.

Society for Industrial and Organizational Psychology (SIOP). (1987). *Principles for the validation and use of personnel selection procedures (3rd ed.).* College Park, MD: American Psychological Association.

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,* 710–718.

Sue, S., & Abe, J. (1988). *Predictors of academic achievement among Asian American and white students* (Research Report No. 88–11). New York: College Board.

Sue, S., & Zane, N. W. S. (1985). Academic achievement and socioemotional adjustment among Chinese university students. *Journal of Counseling Psychology, 32,* 570–579.

Temp, G. (1971). Test bias: Validity of the SAT for blacks and whites in 13 integrated institutions. *Journal of Educational Measurement, 6,* 203–215.

Thomas, C. L. (1972, April). *The relative effectiveness of high school grades and standardized test scores for predicting college grades of black students.* Paper presented at the annual convention of the National Council on Measurement in Education, Chicago.

Tracey, T. J., & Sedlacek, W. E. (1984). Noncognitive variables in predicting academic success by race. *Measurement and Evaluation in Guidance, 16,* 171–178.

Tracey, T. J., & Sedlacek, W. E. (1985). The relationship of noncognitive variables to academic success: A longitudinal comparison by race. *Journal of College Student Personnel, 26,* 405–410.

Wainer, H., Saka, T., & Donoghue, J. R. (1993). The validity of the SAT at the University of Hawaii: A riddle wrapped in an enigma. *Educational Evaluation and Policy Analysis, 15,* 91–98.

Wainer, H., & Steinberg, L. S. (1992). Sex differences in performance on the Mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review, 62,* 323–335.

Warren, J. (1976). *Prediction of college achievement among Mexican American students in California.* College Board Research and Development Report. Princeton, NJ: Educational Testing Service.

Wigdor, A. K., & Garner, W. R. (Eds.) (1982). *Ability testing: Uses, consequences, and controversies.* Washington, D.C.: National Academy Press.

Wilder, G. Z., & Powell, K. (1989). *Sex differences in test performance: A survey of the literature* (College Board Report No. 89-3). New York, NY: College Board.

Willingham. W. W. (1990). Introduction: Interpreting predictive validity. In *Predicting college grades: An analysis of institutional trends over two decades.* Princeton, NJ: Educational Testing Service.

Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L. (1990). *Predicting college grades: An analysis of institutional trends over two decades.* Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1980). The performance of minority students beyond the freshman year: Testing a "late-bloomer" hypothesis in one state university setting. *Research in Higher Education, 13,* 23–47.

Wilson, K. M. (1981). Analyzing the long-term performance of minority and nonminority students: A tale of two studies. *Research in Higher Education, 15,* 351–375.

Wilson, K. M. (1983). *A review of research on the prediction of academic performance after the freshman year* (College Board Report No. 83–2 and Educational Testing Service Research Report No. 83–11). New York: College Board.

Wright, R. J., & Bean, A. G. (1974). The influence of socioeconomic status on the predictability of college performance. *Journal of Educational Measurement, 11,* 277–283.

Young, J. W. (1991a). Gender bias in predicting college academic performance: A new approach using item response theory. *Journal of Educational Measurement, 28,* 37–47.

Young, J. W. (1991b). Improving the prediction of college performance of ethnic minorities using the IRT-based GPA. *Applied Measurement in Education, 4,* 229–239.

Young, J. W. (1993). Grade adjustment methods. *Review of Educational Research, 63,* 151–165.

Young, J. W. (1994). Differential prediction of college grades by gender and by ethnicity: A replication study. *Educational and Psychological Measurement, 54,* 1022–1029.

Young, J. W., & Fisler, J. L. (2000). Sex differences on the SAT: An analysis of demographic and educational variables. *Research in Higher Education, 41,* 401–416.

Young, J. W., & Koplow, S. L. (1997). The validity of two questionnaires for predicting minority students' college grades. *Journal of General Education, 46,* 45–55.

Differential Validity/Prediction Studies Cited in Sections 3 and 4

Arbona, C., & Novy, D. M. (1990). Noncognitive dimensions as predictors of college success among black, Mexican American, and white students. *Journal of College Student Development, 31,* 415–422.

Baggaley, A. R. (1974). Academic prediction at an Ivy League college, moderated by demographic variables. *Measurement and Evaluation in Guidance, 6,* 232–235.

Baron, J., & Norman, M. F. (1992). SATs, achievement tests, and high-school class rank as predictors of college performance. *Educational and Psychological Measurement, 52,* 1047–1055.

Boli, J., Allen, M. L., & Payne, A. (1985). High-ability women and men in undergraduate mathematics and chemistry courses. *American Educational Research Journal, 22,* 605–626.

Bridgeman, B., & Lewis, C. (1996). Gender differences in college mathematics grades and SAT M scores: A reanalysis of Wainer and Steinberg. *Journal of Educational Measurement, 33,* 257–270.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning Test* (College Board Report No. 2000-1). New York: College Board.

Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and grades in college mathematics courses. *Journal of Educational Psychology, 83,* 275–284.

Chou, T., & Huberty, C.J. (1990). *A freshman admissions prediction equation: An evaluation and recommendation.* Athens, GA: University of Georgia (ERIC Document Reproduction Service No. ED 333 081).

Clark, M. J., & Grandy, J. (1984). *Sex differences in the academic performance of SAT takers* (College Board Report No. 84-8). New York: College Board.

Cowen, S., & Fiori, S. J. (1991, November). *Appropriateness of the SAT in selecting students for admission to California State University, Hayward.* Paper presented at the annual meeting of the California Educational Research Association, San Diego, CA (ERIC Document Reproduction Service No. ED 343 934).

Crawford, P. L., Alferink, D. M., & Spencer, J. L. (1986). Postdictions of college GPAs from ACT composite scores and high school GPAs: Comparisons by race and gender. West Virginia State College (ERIC Document Reproduction Service No. ED 326 541).

Dalton, S. (1976). A decline in the predictive validity of the SAT and high school achievement. *Educational and Psychological Measurement, 36,* 445–448.

Elliott, R., & Strenta, A. C. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement, 25,* 333–347.

Farver, A. S., Sedlacek, W. E., & Brooks, G. C. (1975). Longitudinal prediction of university grades for blacks and whites. *Measurement and Evaluation in Guidance, 7,* 243–250.

Fincher, C. (1974). Is the SAT worth its salt? An evaluation of the use of the Scholastic Aptitude Test in the university system of Georgia over a thirteen-year period. *Review of Educational Research, 44,* 293–305.

Gamache, L. M., & Novick, M. R. (1985). Choice of variables and gender differentiated prediction within selected academic programs. *Journal of Educational Measurement, 22,* 53–70.

Hand, C. A., & Prather, J. E. (1985, April) The predictive validity of Scholastic Aptitude Test scores for minority college students. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL (ERIC Document Reproduction Service No. ED 261 093).

Hogrebe, M. C., Ervin, L., Dwinell, P. L., & Newman, (1983). The moderating effects of gender and race in predicting the academic performance of college developmental students. *Educational and Psychological Measurement, 43,* 523–530.

Houston, W., & Sawyer, R. (1988). *Central prediction sys-*

tems for predicting specific course grades (Research Report No. 88-4). Iowa City, IA: American College Testing.

Larson, J. R., & Scontrino, M. P. (1976). The consistency of high school grade point average and of the verbal and mathematical portions of the Scholastic Aptitude Test of the College Entrance Examination Board, as predictors of college performance: An eight year study. *Educational and Psychological Measurement, 36,* 439–443.

Leonard, D. K., & Jiang, J. (1995, April). *Gender bias in the college predictions of the SAT.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Maxey, J., & Sawyer, R. (1981, July). *Predictive validity of the ACT Assessment for Afro-American/Black, Mexican-American/Chicano, and Caucasian-American/White students* (ACT Research Bulletin 81-1). Iowa City, IA: American College Testing.

McCornack, R. L. (1983). Bias in the validity of predicted college grades in four ethnic minority groups. *Educational and Psychological Measurement, 43,* 517–522.

McCornack, R. L., & McLeod, M. M. (1988). Gender bias in the prediction of college course performance. *Journal of Educational Measurement, 25,* 321–331.

McDonald, R. T., & Gawkoski, R. S. (1979). Predictive value of SAT scores and high school achievement for success in a college honors program. *Educational and Psychological Measurement, 39,* 411–414.

Moffatt, G. K. (1993, February). *The validity of the SAT as a predictor of grade point average for nontraditional college students.* Paper presented at the annual meeting of the Eastern Educational Research Association, Clearwater Beach, FL (ERIC Document Reproduction Service No. ED 356 252).

Morgan, R. (1990). Analyses of predictive validity within student categorizations. In Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L., *Predicting college grades: An analysis of institutional trends over two decades* (pp. 225–238). Princeton, NJ: Educational Testing Service.

Nettles, M. T., Thoeny, R., & Gosman, E. J. (1986). Comparative and predictive analyses of black and white students' college achievement and experiences. *Journal of Higher Education, 57,* 289–318.

Noble, J., Crouse, J., & Schulz, M. (1996). *Differential prediction/impact on course placement for ethnic and gender groups* (Research Report No. 96-8). Iowa City, IA: American College Testing.

Pearson, B. Z. (1993). Predictive validity of the Scholastic Aptitude Test for Hispanic bilingual students. *Hispanic Journal of Behavioral Sciences, 15,* 342–356.

Pennock-Román, M. (1990). *Test validity and language background: A study of Hispanic-American students at six universities.* New York: College Board.

Pennock-Román, M. (1994). *College major and gender differences in the prediction of college grades* (College Board Report No. 94-2). New York: College Board.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994).

Student group differences in predicting college grades: Sex, language, and ethnic groups (College Board Report No. 93-1). New York: College Board.

Ramist, L., & Weiss, G. (1990). The predictive validity of the SAT, 1964 to 1988. In Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L., *Predicting college grades: An analysis of institutional trends over two decades* (pp. 117–140). Princeton, NJ: Educational Testing Service.

Rowan, R. W. (1978). The predictive value of the ACT at Murray State University over a four-year college program. *Measurement and Evaluation in Guidance, 11,* 143–149.

Saka, T. T. (1991). High school GPA, SAT scores and college academic achievement for University of Hawaii freshmen. *Pacific Educational Research Journal, 7,* 19–32.

Sawyer, R. (1986). Using demographic subgroup and dummy variable equations to predict college freshman grade average. *Journal of Educational Measurement, 23,* 131–145.

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,* 710–718.

Sue, S., & Abe, J. (1988). P*redictors of academic achievement among Asian American and white students* (Research Report No. 88–11). New York: College Board.

Tracey, T. J., & Sedlacek, W. E. (1984). Noncognitive variables in predicting academic success by race. *Measurement and Evaluation in Guidance, 16,* 171–178.

Tracey, T. J., & Sedlacek, W. E. (1985). The relationship of noncognitive variables to academic success: A longitudinal comparison by race. *Journal of College Student Personnel, 26,* 405–410.

Wainer, H., Saka, T., & Donoghue, J. R. (1993). The validity of the SAT at the University of Hawaii: A riddle wrapped in an enigma. *Educational Evaluation and Policy Analysis, 15,* 91–98.

Wainer, H., & Steinberg, L. S. (1992). Sex differences in performance on the Mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review, 62,* 323–335.

Wilson, K. M. (1980). The performance of minority students beyond the freshman year: Testing a "late-bloomer" hypothesis in one state university setting. *Research in Higher Education, 13,* 23–47.

Wilson, K. M. (1981). Analyzing the long-term performance of minority and nonminority students: A tale of two studies. *Research in Higher Education, 15,* 351–375.

Young, J. W. (1991a). Gender bias in predicting college academic performance: A new approach using item response theory. *Journal of Educational Measurement, 28,* 37–47.

Young, J. W. (1991b). Improving the prediction of college performance of ethnic minorities using the IRT-based GPA. *Applied Measurement in Education, 4,* 229–239.

Young, J. W. (1994). Differential prediction of college grades by gender and by ethnicity: A replication study. *Educational and Psychological Measurement, 54,* 1022–1029.

Young, J. W., & Koplow, S. L. (1997). The validity of two questionnaires for predicting minority students' college grades. *Journal of General Education, 46,* 45–55.

Appendix: Descriptions of Studies Cited in Sections 3 and 4

Arbona and Novy (1990)(3)

Examined the validity of SAT scores and the Non-Cognitive Questionnaire (NCQ) in predicting grades and persistence for black, Mexican American, and white freshman students at a predominantly white southern university (presumably the University of Houston) entering in 1987. Hierarchical multiple regression analyses were performed to examine whether, and to what extent, SAT scores predicted FGPA. A discriminant analysis was performed to examine the predictive power of these variables on enrollment status after the first year in college. Neither SAT scores nor the NCQ was predictive of black students' cumulative GPAs. For Mexican American students, SAT M scores were predictive of FGPA; for white students, both SAT M and SAT V scores were predictive of FGPA. SAT scores (neither math nor verbal) did not predict persistence in college for any group of students.

Baggaley (1974) (3,4)

Studied differential characteristics of regressions of cumulative GPA for three semesters on SAT V and SAT M scores and high school rank (HSR) for various demographic groups at the University of Pennsylvania entering in 1969. Females' GPAs were somewhat more predictable than males; SAT scores showed greater predictive validity for females than males. No gender differences were found when using HSR as predictor, but HSR showed more predictive validity for whites than blacks (but not significantly). HSR tended to be more valid than test scores for predicting CGPA for white students, particularly males; test scores seemed to have no predictive validity for black males.

Baron and Norman (1992) (4)

Looked at the validity of high school rank (HSR), SAT scores, and an average score on three College Board Achievement Tests in predicting the college GPA of students entering the University of Pennsylvania in 1983 and 1984. Once HSR and the average Achievement Test score were entered into the multiple regression equation, SAT scores did not add significant prediction. The authors conclude that the SAT makes a relatively small contribution to prediction that is even smaller when Achievement Tests and HSR are known.

Boli, Allen, and Payne (1985) (4)

Investigated the performance (course completion and grades) and perceptions of performance of high-ability males and females in introductory chemistry and mathematics courses at Stanford University in the fall of 1977. A questionnaire was used to obtain information on perceptions of performance. Men outperformed women in both courses, even when high school calculus preparation was held constant. However, when SAT M scores were controlled for, the performance difference was substantially reduced. In a multiple regression path analysis, gender had no direct effect on course performance, but it did have a sizable indirect effect by way of mathematics background (i.e., SAT scores).

Bridgeman and Lewis (1996) (4)

A re-analysis of the data set used by Wainer and Steinberg (1992) which was comprised of the freshman class of 1985 at 43 colleges. Analyzed gender differences in SAT M within individual courses within colleges; evaluated gender differences when SAT M is used with high school record. Even within individual courses, on average men had higher SAT M scores than women with same course grades, yet the HSGPA of women was greater than that of men with the same calculus grades. Slight underprediction of women's grades in precalculus and calculus courses occurred using a standardized composite of SAT M and HSGPA.

Bridgeman, McCamley-Jenkins, and Ervin (2000) (3,4)

This study examined the impact of revisions in the content of the SAT and adoption of a new, recentered score scale on the predictive validity of the SAT. Data from the 1994 and 1995 entering classes at 23 colleges (13 public and 10 private) were used to determine the validity of SAT scores and HSGPA in predicting FGPA. Changes in the test content and use of the new score scale had virtually no impact on predictive validity. Correlations of SAT scores and HSGPA with FGPA were generally higher for women than for men, although this was not the case at colleges with very high SAT scores. Consistent with many earlier studies, using a single prediction equation led to underprediction of the grades of women. The grades of minority students were found to be generally overpredicted; however, adjusting for course difficulty changed the slight overprediction to underprediction in the case of Asian American students. Validity coefficients adjusted for course difficulty and range restriction were substantially higher than the corresponding unadjusted values.

Bridgeman & Wendler (1991) (4)

Investigated sex differences in grades and SAT M scores within a sample of algebra, precalculus, and calculus courses based on the entering class of 1986 at nine universities. Within each course, it was found that women typically had equal or higher grades, whereas men had higher SAT M scores. If a single regression equation was used to predict course grades of men and women from SAT M scores, underprediction of women's grades would result with a weighted average effect size of +.14 for algebra, +.13 for precalculus, and -.01 for calculus in favor of women.

Chou and Huberty (1990) (3,4)

Investigated the effectiveness of different freshman admission prediction equations at the University of Georgia for the entering class of 1986. Used SAT V and SAT M scores, HSGPA, sex, race, and high school grouping to predict FGPA. Evaluated 11 different regression equations comprised of different combinations of predictors. The evaluation of the models was based on the mean residual, mean absolute residual, standard deviation of residuals, and misclassification rates. It was found that the inclusion of gender, race, and high school grouping did not improve the predictive accuracy in terms of mean absolute residual, residual standard deviation, and misclassification rates; some improvement in reducing the mean residual was observed, however. The authors suggest using the misclassification error rate as a criterion for evaluating the effectiveness of a prediction model.

Clark and Grandy (1984) (4)

Summarized research on the academic performance of women and men by examining sex differences among all SAT takers, test-takers grouped by anticipated major field of study, and college freshman year courses and grades. Investigated whether there are consistent differences in the intellectual abilities of men and women, whether precollege admission variables predict college performance with equal accuracy for women and men, and whether the contents or structure of the SAT have contributed to observed sex differences in performance on the test. Reviewed a large body of literature on sex differences, and reported three empirical investigations. The empirical studies indicated that the test scores of women have declined more than the scores of men over the past 15 years, and the characteristics of the test-taking groups have changed, but it is not clear that the demographic changes account for the score declines.

Concluded that the evidence in the research is not sufficient to account for all of the observed sex differences in performance on the SAT. Also reported validity and prediction results for 41 institutions that participated in the 1980 College Board Validity Study Service.

Cowen and Fiori (1991) (3,4)

Examined the claims that the SAT adds little incremental validity to the prediction of first-year college performance and the claim that the SAT is biased. Looked at regular progressing versus slower progressing students after one year and two years of those matriculating in 1988 at California State University, Hayward. The criterion variables were FGPA and a quantitative GPA, comprised of math, science, and other quantitative courses. In the regression of FGPA on HSGPA and SAT, for most groups, the SAT contributed an additional .04 to .06 to the multiple correlation after HSGPA, which was the most important predictor. For slower progressing students, neither SAT scores nor HSGPA were significant. The SAT was a better predictor for the quantitative GPA. The addition of SAT did not significantly reduce the difference between predicted and actual GPAs for all groups studied, nor was there significant over- or under-prediction for any group.

Crawford, Alferink, and Spencer (1986) (3,4)

Compared students' FGPA with their "postdicted" GPA, based on ACT scores and HSGPA. Examined race (blacks, whites) and sex subgroups for students entering a West Virginia college (assumed to be West Virginia State College) in 1985. Found that postdiction accuracy was increased by including HSGPA with ACT in the prediction model. Female performance was underpostdicted and males were over-postdicted; however, this decreased somewhat when HSGPA was added to the model. Statistics on residuals from regression equations were not reported. Instead, frequency counts of over- and under-postdicted GPAs were analyzed by race and sex using a chi-square test of independence.

Dalton (1976) (4)

Examined the predictive validity of SAT Total and HSR for predicting first-semester college grades for five entering cohorts over a 13-year period (from 1961 to 1974) at Indiana University. Females were more predictable than men with regard to GPA. There was a decline in predictive validity over the years, which could not be attributed to restriction of range in the predictor variables.

Elliott and Strenta (1988) (3,4)

Investigated the impact of an adjusted CGPA based on within–as well as between–department grading standards on the predictive validity of the SAT, College Board Achievement Test scores, and HSR to predict CGPA. Data came from the Dartmouth College graduating class of 1986. Also looked at the difference in the prediction of independently and annually computed GPAs, and the effect of criterion adjustment by sex and race. The addition of the within-department and between-department adjustments had only a small empirical effect. The prediction of grades by SAT scores for black students was improved when the GPA criterion was made more reliable either by adjustment or by confining prediction to one or two courses having fairly reliable standards. However, the adjustment increased black–white differences in grades, because it served to enhance the grades of those who took more science courses. The adjustment reduced, but did not eliminate, the underprediction of grades for women.

Farver, Sedlacek, and Brooks (1975) (3,4)

Compared the prediction of freshman, sophomore, junior, and senior and cumulative GPAs for blacks and whites, and female and male students for two separate entering years (1968 and 1969) at the University of Maryland. The predictors SAT V, SAT M, and HSGPA showed significant zero-order correlations with freshman through upper-class university grades. HSGPA was more important in the prediction of freshman grades than in the prediction of later university grades, and was a consistently poor predictor for black males. Black males were less predictable beyond their freshman year compared to the other race/sex subgroups. White females were the most predictable subgroup for the two years. The 1968 and 1969 entrants showed differential prediction patterns. A common regression equation for all students was not employed.

Fincher (1974) (4)

Studied the incremental effectiveness of the SAT in predicting college grades in the University System of Georgia (29 institutions) over a period of 13 years (from 1958 to 1970). A frequency count of the times that SAT scores contributed to the prediction equations developed for separate institutions showed that the SAT V contributed to the prediction of college grades in almost three out of four equations, and the SAT M made a significant contribution slightly less than half of the time. There was consistently better prediction for female students' GPAs when compared to male students. Over the 13 years, there was a fairly consistent gain in predictive efficiency between regression equations using HSGPA alone and the equations including both HSGPA and SAT scores. Efficiency indices were reported which could be converted to multiple correlation coefficients. Discussed efforts to determine the cost-effectiveness in using the SAT.

Gamache and Novick (1985) (4)

Examined gender bias in prediction of two-year CGPA at a large state university (assumed to be the University of Iowa) from ACT subtest and composite scores within four major programs (to control for differential coursework) for students entering in 1978. Used the Johnson-Neyman technique to detect sex differences in the regression equations. Differential prediction existed (with women underpredicted), but was reduced with the use of a subset of the original four predictors. In almost all instances, the use of gender differentiated equations increased the predicted criterion value for women.

Hand and Pranther (1985) (3,4)

Examined the predictive validity of the SAT for predicting GPAs for white males, white females, black males, and black females enrolled in 1983 across 31 institutions of a state college system (in Georgia). Used the unstandardized regression coefficients which the authors say can be compared across populations. Regression equations were derived for each of the institutions, by sex and race, and the coefficients for each predictor variable and constant in the regression equations were plotted and compared. The authors conclude that GPAs are least predictable for black males due to the lower weights of SAT V and HSGPA for predicting CGPA.

Hogrebe, Ervin, Dwinell, and Newman (1983) (3,4)

Looked at the predictive validity of SAT scores and HSGPA for predicting the performance of Developmental Studies students at a large southern university (possibly the University of Georgia) during the 1977-78 and 1978-79 academic years. A significant slope difference was found for blacks versus whites (with a larger slope for blacks). In addition, there was an intercept difference for sex for white students but not for black students. The SAT M was a significant predictor of FGPA only for black students.

Houston and Sawyer (1988) (4)

Investigated two central prediction models based on small sample sizes, which used collateral information across institutions to obtain refined within-group parameter estimates. Two different prediction equations were studied: an eight-variable equation based on the four ACT subjects and four HS grades, and a two-variable equation based on ACT composite and HSGPA. For each prediction equation, regression coefficients and residual variances were estimated using three different models: within-college least squares (WCLS), pooled least squares with adjusted intercepts (ANCOVA), and empirical Bayesian m-group regression. It was found that both models employing collateral information with a sample size of 20 resulted in crossvalidated prediction accuracy comparable to that obtained using the within-college least squares procedure with sample sizes of 50 or more.

Larson and Scontrino (1976) (4)

Evaluated the consistency of HSGPA and SAT scores as predictors of four-year cumulative college GPA over an eight-year period (from 1966 to 1973) at a small West Coast university (possibly the University of Washington). The multiple correlations were consistently high with yearly values ranging from .53–.80 for females, .65–.79 for males, and .60–.73 for all students combined. Inclusion of SAT scores in the prediction equation slightly improved predictability for males in all years, but did not increase predictability for females when the equations were crossvalidated.

Leonard and Jiang (1995) (4)

Presented data that demonstrated the underprediction of women's college performance (using CGPA as the criterion) at the University of California, Berkeley for freshman admits between 1986 and 1988. The University of California's Academic Index Score (AIS), which is made up of HSGPA and five test scores (SAT V, SAT M, and three College Board Achievement Tests) was found to underpredict the undergraduate grades of women and to overpredict those of men. When field of study as well as selection bias were controlled for, this underprediction of women's grades persisted.

Maxey and Sawyer (1981) (3)

Reported the results for 271 institutions that participated in ACT's Prediction Research Service in 1977-78 and in an earlier year. The variables used to predict college grades were four ACT test scores and four high school grades. The prediction equation for each college was cross-validated against actual 1977-78 data for the total group, and for separate ethnic/racial groups. On average, black students' college grades were overpredicted slightly. The grades of Chicano students were neither over- nor under-predicted. The mean absolute errors in grade prediction for Chicanos and blacks were somewhat larger than that for whites, implying lower validity coefficients for these groups.

McCornack (1983) (3)

Looked at the accuracy of a regression equation for predicting the GPAs of white, Asian, Hispanic, black, and Indian students based on white students entering San Diego State University in 1979. Found that the GPAs of black, Hispanic, and Asian students were overpredicted but that of Native Americans were underpredicted. Although the samples were small (N=24 in 1979 and N=25 in 1980), this was one of the few studies that examined the performance of Native American students.

McCornack and McLeod (1988) (4)

Examined whether gender bias existed in the prediction of individual college course grades from SAT scores and HSGPA, and compared the prediction accuracy using individual course grades and CGPA as the criterion variable. Three prediction models were studied for each of 88 introductory courses at San Diego State University in the 1985-86 academic year. These models included the common equation with no gender effects, including high school GPA, SAT V, and SAT M as predictors; the different intercepts model with a dummy-coded gender predictor added to permit separate intercepts but identical slopes for HSGPA, SAT V, and SAT M; and the gender-specific model, which permitted both separate intercepts and different slopes. For the individual courses, models with gender effects tended to be less accurate than the common equation. For the majority of courses, the prediction was the same for women and men. In the few courses in which gender bias was found, it most often involved the overprediction of women in a course in which men earned a higher average grade. When a single equation was used to predict CGPA, a small but significant amount of underprediction occurred for women.

McDonald and Gawkoski (1979) (4)

Examined the validity of SAT scores and HSGPA in predicting success in the Honors Program at Marquette University between 1963 and 1972. Success was defined as receiving an honors degree (minimum GPA of 3.0 and the completion of at least 46 credits in specially designed, challenging honors courses). HSGPA was the variable with the strongest predictive validity, but significant relationships were also found between success or lack of success for the entire group and both SAT V and SAT M scores. For men, the relationship between SAT V and the success criterion was not significant, but for women SAT M was the only relatively strong predictor of success.

Moffatt (1993) (3)

Examined the predictive validity of SAT total for older, nontraditional college students at Atlanta Christian College (year of the study's sample was not given). SAT total was found to be a significant predictor of CGPA for white students under 30, but not for black students of any age. SAT total was not a significant predictor of CGPA for students who had not taken the SAT prior to age 30, regardless of race.

Morgan (1990) (3,4)

Analyzed the predictive validity of the SAT, TSWE, and College Board Achievement Tests within subgroups based on sex, race, and intended college major for enrolling classes at 198 colleges in 1978, 1981, and 1985. Raw correlations and correlations corrected for restriction of range were estimated along with regression weights. All correlation estimates were higher for females than males. For both sexes, SAT M was the best single predictor of FGPA, followed by SAT V and then TSWE. The SAT correlation declines for all students were similar to those for each sex. All racial groups studied (Asian Americans, blacks, Hispanics, and whites) showed a decline in the raw multiple correlation of SAT scores with FGPA over the years studied. However, the corrected multiple SAT correlation did not drop significantly for Asian Americans and rose for Hispanics. SAT scores were better predictors of FGPA for blacks. Analyses of predictive validity by intended major did not show any patterns. The author concluded that with a few possible exceptions, declines of SAT correlations with FGPA are characteristic of freshmen in general, and not attributable to any specific subgroup.

Nettles, Theony, and Gosman (1986) (3,4)

Compared black and white students' college performance (using CGPA) and their academic, personal, attitudinal, and behavioral characteristics. Determined the predictive validity of a variety of students' academic, personal, and attitudinal characteristics, as well as of faculty attitudes and behaviors. Data are based on the survey responses of students and faculty from 30 colleges and universities in the southern and eastern United States. Found many variables that were significant predictors of CGPA, which for the most part were equally effective predictors for black and white students. Four variables—SAT scores, student satisfaction, peer relationships, and interfering problems—had differential predictive validity. Significant racial differences on several of the predictor variables helped explain racial difference in college performance.

Noble, Crouse, and Schulz (1996) (3,4)

Predicted success in four standard college courses from ACT scores or high school subject area grade averages (SGA) using data from over 80 institutions and 11 different courses. Linear regression analyses were performed to determine whether there was differential prediction of course grades for females and males, or for African Americans or Caucasian Americans. Using an approach developed by Sawyer, logistic regression was used to predict specific course outcomes (grade of B or higher, or C or higher). The results showed that ACT scores and SGAs slightly underpredicted the course grades of females, with a smaller difference using SGA. ACT scores and SGA both overpredicted English composition grades of African Americans. Adding ACT scores to SGA in a two-predictor model slightly reduced this overprediction.

Pearson (1993) (3)

Compared SAT scores and four-semester cumulative college GPA for Hispanic and non-Hispanic white students who entered the University of Miami in the fall of 1988. Hispanic students had significantly lower SAT scores (both verbal and math), despite equivalent college grades. Both ethnic groups showed similar sex differences. In stepwise regression analyses, ethnicity was found to be a significant predictor when only SAT scores were in the model, but was not significant when

high school performance (reported as decile rank) was entered in the model. Separate regressions for Hispanics and non-Hispanics showed that the percentage of variance in college GPA accounted for by SAT scores and the raw regression weights were similar for the two groups. However, the intercepts differed. Hispanic students' GPAs were overpredicted, with a regression equation based on both ethnic groups.

Pennock-Román (1990) (3)

Examined whether differences in the prediction of FGPA occurred for Hispanic students as compared with white students at six universities. Two of the universities were located in California, one in Florida, one in Massachusetts, one in New York, and one in Texas. For the California schools, the data were from entering first-year students in 1982; for the other institutions, the data were from students entering in 1985. Students' language background was also examined to determine if measures of English proficiency improved grade prediction for the Hispanic students. Across all six universities, there was slight-to-moderate overprediction of Hispanic students' FGPAs, and lower multiple correlations of preadmissions predictors with FGPA for Hispanics than for whites.

Pennock-Román (1994) (4)

Four institutions from the Pennock-Román (1990) data set were used to examine sex differences in the prediction of FGPA after controlling for differential course grading based on college major. Used SAT V, SAT M, HSGPA, and a variable called "MAJSCAL" to reflect the degree of grading toughness/leniency by major. Overall, females were underpredicted using the males' equation, both with and without MAJSCAL. However, MAJSCAL improved the predictive accuracy, reducing the intercept difference and the amount of female underprediction. The largest underprediction occurred for females, with the SAT M as the only predictor, even after using MAJSCAL. Author supports the use of the standard model (SAT scores plus HSGPA) rather than HSGPA only.

Ramist, Lewis, and McCamley-Jenkins (1994) (3,4)

Using a database of entering freshmen in 1982 and 1985 at 38 institutions, the authors looked at possible causes for the increasing decline in the correlation of SAT scores and FGPA. Differences by sex and for four minority groups (Asian Americans, blacks, Hispanics, and Native Americans) in validity and prediction were investigated.

Found better predictions of course grades for females; the SAT added more incremental information over HSGPA for females than for males. Also found better predictions for Asian Americans than for any other group, but the SAT added more incremental information over HSGPA for blacks than for any other racial/ethnic group. Females were underpredicted overall, but were overpredicted in technical courses other than math. Nonnative English speakers were underpredicted, except in English courses. American Indians were overpredicted overall, while Asian Americans were underpredicted, especially in math and science. Black and Hispanic students' grades were overpredicted using any combinations of predictors.

Ramist and Weiss (1990) (4)

Analyzed SAT predictive validity studies of schools participating in the College Board Validity Study Service from 1964 to 1988. Matched earlier and later studies for the same institutions to make comparisons by years and by groups of years (periods). Looked at the correlations of SAT scores and freshman grade point average (FGPA), corrected for restriction of range to make them comparable from year to year. Found that the correlations increased from pre-1973 (1964–1972) to 1973–1976, and decreased from 1973–1976 to 1985–1988. Both the increase and the decrease were greater for males than for females. The college characteristic that was the best predictor of change in the SAT correlation was the SAT mean level.

Rowan (1978) (4)

Investigated the validity of the ACT in predicting FGPA and CGPA (for successive intervals) and in predicting college completion in four years for females and males entering Murray State University (KY) starting about 1969. It was found that the ACT was a significant predictor of GPA at yearly intervals over the four-year span for the two classes studied, although the magnitude of the validity coefficient decreased over time. The ACT was also found to be a significant predictor of college completion. The findings were inconclusive with regard to gender differences in predictability. Expectancy tables revealed that success probability and survival rate were higher for females than for males, but it was not clear whether this prediction difference could be attributed to the ACT or to other factors.

Saka (1991) (4)

Studied the relationship among FGPA, SAT scores, and HSGPA for freshmen attending the University of Hawaii at Manoa in 1988-89. Found that HSGPA and SAT scores

were better predictors of FGPA for students attending mainland or foreign high schools than for students attending Hawaiian public or private schools. HSGPA accounted for the greatest amount of unique variation in FGPA, and SAT M was not a significant predictor of FGPA for Hawaii public school students. The caveat is included that the results should be viewed as purely descriptive due to some limitations that were not considered.

Sawyer (1986) (3,4)

Analyzed three data sets constructed from freshman grade information submitted by colleges to the ACT predictive research services. The first data set consisted of 105,500 student records from 200 colleges; the second consisted of 134,600 student records from 256 colleges; and the third consisted of 96,500 student records from 216 colleges. At each college, multiple linear regression prediction equations were calculated on a set of "base year" data, and the equations were applied to a set of "cross-validation year" data. Five different sets of predictor variables were used to predict freshman grade average at each college. The standard prediction equation consisted of four ACT subtest scores in English, mathematics, social studies, and natural sciences, and four self-reported HS grades. Four alternative prediction equations included a reduced set of predictors (ACT Composite score and HSGPA), and demographic information, either in the form of dummy variables or separate subgroup equations. From the cross-validation year data, two measures of predication accuracy were calculated for each college, prediction method, and subgroup: the observed mean squared error and bias (the average observed difference between predicted and earned grade average). The results showed that, across all colleges, the standard total group prediction equations underpredicted the grade averages of females and older students, and overpredicted the grade averages of males, minority students, and students age 17–19. The alternate prediction equations reduced the underprediction for older students and females, and reduced the overprediction for males. However, the alternate equations produced large negative biases for minority students.

Stricker, Rock, and Burton (1993) (4)

Appraised two explanations for sex differences in over- and underprediction of college grades by the SAT: sex-related differences in the nature of the grade criterion, and sex-related differences in variables associated with academic performance. Data consisted of 4,351 full-time students in the fall 1988 entering class at Rutgers University. Predictor variables identified through a literature search

on sex differences were taken from a longitudinal database and two academic questionnaires, one administered to students during freshman orientation, and the other administered in November of 1988. Two criterion variables were examined: the raw first-semester GPA, and an adjusted GPA that controlled for grading standards in individual courses. Analyses were conducted for a residualized GPA criterion predicted by SAT scores. The results indicated that sex had very similar correlations with the raw and adjusted GPA residualized criteria. A small but statistically significant sex difference occurred in over- and underprediction, with women being underpredicted. Regression analyses for 15 sets of predictor variables, sex, and the interaction between the explanatory variables and sex with respect to the GPA residualized criterion were conducted. The results indicated that sex differences in over- and underprediction were reduced when other differences between women and men (such as academic preparation, studiousness, and attitudes about mathematics) were eliminated. Course differences in grading standards had no noticeable impact on sex differences in over- and underprediction.

Sue and Abe (1988) (3,4)

Examined various predictors of academic performance for Asian American and white first-year students enrolled at the eight University of California campuses in fall 1984. The purpose of the study was to determine whether HSGPA, SAT scores, and College Board Achievement Test scores predicted FGPA, and to determine whether the predictors varied according to membership within different Asian American groups, major, language spoken, and gender. Regression analyses were conducted with two sets of predictor variables. The first set consisted of SAT scores and HSGPA, and the second consisted of Achievement Test scores and HSGPA. Marked differences for the various Asian subgroups were found. The regression equation based on white students underpredicted the FGPA of Chinese, Other Asians, and Asian Americans for whom English was not the best language, and overpredicted for Filipinos, Japanese, and Asian Americans for whom English was the best language.

Tracey and Sedlacek (1984) (3)

Examined the reliability, construct validity, and predictive validity of the Non-Cognitive Questionnaire (NCQ). Two separate random samples of first-year students entering the University of Maryland in 1979 and 1980 were given the NCQ. The construct validity of the instrument was examined using principal components factor analysis, with separate analyses done for each

race. The predictive validity of the NCQ and SAT scores on SGPA and CGPA was examined using stepwise multiple regression, and the predictive validity of the NCQ and SAT scores on persistence was examined using stepwise discriminant analyses. The results of the separate factor analyses conducted showed fairly similar structures for each racial group. In all analyses, the NCQ items were either very similar or more highly predictive of the criteria examined than SAT scores alone. The NCQ was found to be more predictive of first-semester grades for whites than for blacks in both years. In contrast, a strong relationship was found between the NCQ and college success for blacks but not for whites.

Tracey and Sedlacek (1985) (3)

Compared the relationship of SAT scores and Non-Cognitive Questionnaire (NCQ) subscale scores to academic success (GPA and persistence) over four years for black and white students. The data were based on all first-year students entering the University of Maryland in 1979, and a random sample of 25 percent of entering students in 1980. Stepwise multiple regressions were run separately for each year and race group using the NCQ subscales and SAT scores as predictors of CGPA at varying points over four years. The relationship of the NCQ and SAT scores to persistence was examined for each year and race group separately using stepwise discriminant analysis. The NCQ provided relatively accurate predictions of grades for both whites and blacks, typically equal to or better than predictions using SAT scores alone. The specific noncognitive subscales that were predictive of grades at all points in a student's academic career were those that reflected positive self-concept and realistic self-appraisal. SAT scores showed little relationship to persistence for either blacks or whites; none of the NCQ subscales were significantly related to persistence for whites but a number of NCQ subscales was significant for blacks.

Wainer, Saka, and Donoghue (1993) (3)

Examined a phenomenon regarding the predictive validity of the SAT for students entering in 1982 and 1989 at the University of Hawaii–Manoa. The relationship between SAT scores and FGPA is somewhat lower than the national average, although the performance of high school students on the SAT entering the university is higher than the national mean, and HSGPA is almost as high as the nationwide data would predict. By 1989, the SAT–FGPA correlations diminished considerably, while HSGPA still performed reasonably well as a predictor. The authors tested the hypothesis that this phenomenon

occurred due to heterogeneity of the population on the traits being measured. According to this hypothesis, if the population were divided properly based on important traits, each subgroup would show a strong relationship between SAT and FGPA. Employed differential item functioning analysis and bivariate Gaussian decomposition to attempt to uncover the subgroups. There was clear evidence of two different groups of students in the population. However, the SAT–FGPA correlations for these groups was still much lower than would be expected.

Wainer and Steinberg (1992) (4)

Examined sex differences on SAT M by comparing the scores of men and women who performed similarly in first-year college math courses. Analyzed data from about 47,000 first-year students attending 51 colleges and universities between 1982 and 1986. In a retrospective analysis, the authors found that women scored lower on the SAT M than men matched by grade and course type. Using a forward regression analysis in which sex and SAT M scores were used to predict course grades, men's SAT M scores were predicted to be, on average, 33 points higher than the scores of women in the same class receiving the same grades. The authors concluded with a discussion of how educators might respond to possible inequities in test performance.

Wilson (1980) (3,4)

Examined the validity of standard admission variables (SAT scores and HSR) for predicting the long-term performance of minority and nonminority students at the main campus of a complex state university system, possibly Penn State. Analyzed data from 272 minority students and a random sample of 1,003 nonminority students entering the university in the fall of 1971, and continuing through the fall of 1976. Tested the "late bloomer" hypothesis, in which the GPAs of minority students show greater improvement than those of nonminority students. Found that, especially for minority students, the validity of the admission variables was greater with respect to CGPA than with respect to short-term GPA criteria. The validity coefficients of the admission variables with respect to GPA criteria were consistently higher for minority than for nonminority students.

Wilson (1981) (3)

Conducted a comparative longitudinal analysis of the performance of minority (n=121) and nonminority (n=1,133) students in four successive entering classes (1970 through 1973) at a highly selective college for

men. Assessed the predictive validity of SAT scores, College Board Achievement Tests, and HSR with respect to long-term and short-term GPA. For nonminority students, the predictor variables individually and in best-weighted combination had a higher correlation with four-year CGPA than with FGPA. For minority students, the validity was somewhat lower, regardless of the GPA criterion, and the observed coefficients were slightly lower for four-year CGPA than for the FGPA. When the data for minority and nonminority students were pooled, the validity coefficients were higher than in either sample alone, and were generally higher for four-year CGPA than for FGPA.

Young (1991a) (4)

Investigated the use of Item Response Theory to develop an adjusted CGPA, the IRT-based GPA, to equate grades across courses with different grading standards. Data came from first-year students entering Stanford University in 1982. Conducted analysis of covariance to predict the IRT-based GPA and CGPA, using SAT V, SAT M, and HSGPA as predictors and sex as an indicator variable. Significant underprediction of women occurred using CGPA as the criterion measure. In contrast, the use of the IRT-based GPA indicated no significant underprediction for men or women, and the IRT-based GPA was more predictable from preadmission measures than CGPA. A single regression equation worked best in predicting both men's and women's IRT-based GPA.

Young (1991b) (3)

Investigated whether the use of the IRT-based GPA as the criterion measure would increase the validities of preadmission predictors for minority students, and would decrease the degree of overprediction of minority students' grades. Data were based on first-year students entering a selective, private university in the western United States in 1982. Prediction equations for a combined sample of all students using multiple regression analyses were computed for three traditional preadmissions measures (SAT V, SAT M, and HSGPA) as predictors, with the IRT-based GPA and CGPA as separate outcome measures. In addition, separate prediction equations were also computed for minority students (African Americans and Hispanics) and a combined group of Asian American and white students. The use of the IRT-based GPA improved the predictability of minority students' performance according to some statistical criteria but was found to be similar to CGPA on others. When the IRT-based GPA replaced CGPA as the criterion, there was a significant decrease in the standard error of estimate, and there was a significant decrease in the degree of overprediction of the minority students' grades.

Young (1994) (3,4)

Investigated whether differential predictive validity, as detected in previous studies, existed for a diverse sample of first-year students entering Rutgers University in 1985. Computed a prediction equation for the total sample of students using SAT V, SAT M, and HSR as predictor variables and CGPA as the outcome variable. Also computed separate prediction equations for men and women, and for each ethnic group. On average, the CGPAs of women were slightly underpredicted. Sex differences in course selection in this cohort may explain, to some degree, the observed underprediction of women. For minority students, significant overprediction occurred for African Americans and Asian Americans, but not for Puerto Ricans or Hispanics (non-Puerto Ricans). However, this overprediction did not appear to be related to course selection.

Young and Koplow (1997) (3)

Investigated whether adding measures of nonacademic constructs would lead to more accurate predictions of minority students' grades. Data were based on 214 respondents (98 minority students, 116 white students) in their fourth year at Rutgers University who entered in the fall of 1990. Nonacademic constructs were measured by the Student Adaptation to College Questionnaire (SACQ), and the Non-Cognitive Questionnaire, Revised (NCQR). A regression analysis indicated that significant overprediction occurred using only preadmission measures (SAT scores and HSR) to predict four-year CGPA. However, one SACQ subscale, Academic Adjustment, contributed significantly to the prediction model, and reduced the overprediction of minority students' CGPAs.