CollegeBoard
connect to college success®

Research Report
No. 2008-4

Differential Validity and Prediction of the SAT®

Krista D. Mattern, Brian F. Patterson, Emily J. Shaw, Jennifer L. Kobrin, and Sandra M. Barbuti

www.collegeboard.com

College Board Research Report No. 2008-4

Differential Validity and Prediction of the SAT®

Krista D. Mattern, Brian F. Patterson, Emily J. Shaw,
Jennifer L. Kobrin, and Sandra M. Barbuti

The College Board, New York, 2008

Krista D. Mattern is an assistant research scientist at the College Board.

Brian F. Patterson is an assistant research scientist at the College Board.

Emily J. Shaw is an assistant research scientist at the College Board.

Jennifer L. Kobrin is a research scientist at the College Board.

Sandra M. Barbuti is a data analyst at the College Board.

The College Board: Connecting Students to College Success

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 5,400 schools, colleges, universities, and other educational organizations. Each year, the College Board serves seven million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #080482567) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Acknowledgments

The authors wish to acknowledge many contributors to this research. Wayne Camara, Mary-Margaret Kerns, Andrew Wiley, Robert Majoros, and Helen Ng were crucial to planning and securing the resources necessary to undertake such a large-scale study. Stephen Frustino, Pooja Kosunam, and Mylene Remigio expertly prepared the database for analysis. Maureen Ewing, Barbara Dodd, and Nathan Kuncel provided valuable reviews and feedback. The College Board's regional staff greatly assisted by recruiting institutions for participation. Finally, the College Board's Research Advisory Committee and Psychometric Panel provided important guidance along the way.

Contents

Abstract

In March 2005, substantial changes were made to the SAT®, most notably the addition of a writing section. Due to these changes, it is imperative that data from the revised test are examined to guarantee that the psychometric quality of the SAT has been preserved. The differential validity and prediction, or whether the test functions differently for various subgroups of students (e.g., males versus females), are characteristics of the test that should be assessed. Previous research has found, in general, smaller correlations between SAT scores and first-year grade point average (FYGPA) for African American and Hispanic students compared to white students (see Young, 2001, for a review). Alternatively, the correlation between SAT scores and FYGPA is generally slightly higher for females than for males. As for differential prediction, academic success, measured by FYGPA, was overpredicted for minority students but underpredicted for female students (Young, 2001). The purpose of the current study is to examine the differential validity and prediction of the SAT using a nationally representative sample of first-year college students admitted with the revised version of the test. The findings demonstrate that there are similar patterns of differential validity and prediction by gender, race/ethnicity, and best language subgroups on the revised SAT compared with previous research on older versions of the test. Future research is discussed.

Introduction

In March 2005, substantial revisions were made to the SAT, most notably the addition of a writing section, to better align test specifications with K–12 curriculum (Lawrence, Rigol, Van Essen, and Jackson, 2003). The writing section is comprised of two parts: a student-written essay and multiple-choice items that require students to identify sentence errors and improve sentences and paragraphs. There were also other, smaller changes made to the test. For the critical reading (formerly verbal) section, the changes included the elimination of analogies and the addition of shorter reading passages. Changes to the mathematics section included the removal of quantitative comparisons and the addition of third-year math content, such as exponential growth, absolute value, functional notation, and negative and fractional exponents. Due to these changes, total testing time increased from 3 hours to 3 hours and 45 minutes.

Due to these test revisions, it is imperative that the psychometric properties of the SAT be reassessed to determine what, if any, impact these changes had on the validity of test scores. Specifically, the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) state, "When substantial changes are made to a test, the test's documentation should be amended, supplemented, or revised to keep information for users current and to provide useful additional information or cautions" (Standard 6.13, p. 70). The validity of test scores must be evaluated in the context of the intended use of those test scores. The main purpose of the SAT is to serve as a college entrance test by providing an index of one's potential to succeed in college. Therefore, the relationship between SAT scores and college performance should be well documented (see Kobrin, Patterson, Shaw, Mattern, and Barbuti, 2008, for the most recent validity evidence of the SAT for the prediction of FYGPA).

Test Fairness

In addition to gathering evidence on the overall relationship between test scores and college success, the *Standards* (AERA/APA/NCME, 1999) stress the importance of assessing the fairness of a test. Two analyses are used to evaluate how the test functions across subpopulations (Drasgow and Kang, 1984). First, during the test development process, all items should be evaluated for differential item functioning[1] (DIF) (Raju and Ellis, 2003). All SAT items are pretested for DIF. Any items exhibiting moderate DIF are excluded from operational forms to ensure measurement equivalence, and those few items that are discovered to exhibit DIF after a full administration are excluded from scoring and/or equating, where appropriate.

Second, tests should be evaluated with regard to equivalent relations with criterion variables (e.g., first-year GPA), also referred to as differential validity and differential prediction (Drasgow and Kang, 1984). Differential validity exists if the magnitude of the test-criterion relationship varies by subgroup. In other words, if the correlation between SAT scores and first-year GPA (FYGPA) is different when examining only males than when examining only females, then the SAT exhibits differential validity by gender.

Furthermore, it is also common practice to examine whether a single regression equation is sufficient for the various subgroups. Differential prediction occurs when a test systematically over- or underpredicts the criterion (e.g., FYGPA) by subgroups. This is calculated

[1] An item is said to exhibit differential item functioning when individuals from different subpopulations (e.g., females versus males) have the same standing on the latent trait but have a different probability of answering the item correctly (Raju and Ellis, 2003). If different subgroups have a different probability of getting an item correct, this does not constitute DIF. To test for DIF, true ability is controlled for and then expected probabilites are estimated.

by subtracting predicted FYGPA from earned FYGPA (i.e., residual = $FYGPA_{earned}$ – $FYPGA_{predicted}$). Negative (residual) values indicate overprediction, and positive values indicate underprediction. For example, if a specific subgroup (e.g., females) tends to earn higher FYGPAs than what is predicted by a regression equation using SAT scores, then the SAT exhibits differential prediction by gender, namely underprediction for females. Furthermore, it should be noted that a test is not necessarily biased or unfair if it exhibits differential validity or prediction but rather it indicates that the relationship between the test score and FYGPA varies by subgroup.

Differential Validity and Prediction in College Admission

There has been a substantial amount of research examining the differential validity and prediction of the SAT. For example, a study by Ramist, Lewis, and McCamley-Jenkins (1994) examined the differential validity and prediction of the SAT using data on 46,379 students from the freshman classes of 1982 and 1985 at 45 colleges and universities. Students were grouped on four dimensions: gender, English as best language, racial/ethnic group, and academic composite. They found that for the total group, the SAT Mathematics (SAT-M) and SAT Verbal (SAT-V), in combination, correlated 0.53[2] with FYGPA. However, they found that this value varied by subgroup. For academic composite, which categorized students into three ability levels (high, medium, low) based on SAT scores and high school grade point average (HSGPA), they found that the SAT was more predictive for academically able students, with a correlation of 0.59 for the high-ability group, 0.53 for the medium-ability group, and 0.43 for the low-ability group. As for gender, the SAT was more predictive of FYGPA for females (r = 0.58) than males (r = 0.52). For students with English as their best language, the SAT correlated 0.53 with FYGPA compared to 0.49 for students whose best language was not English. Finally, for race/ethnicity, the correlation between SAT and FYGPA was highest for Asian American students (r = 0.54) and white students (r = 0.52) and lowest for Hispanic students (r = 0.40). For both American Indian and African American students, the correlation was 0.46. A similar pattern of results was found when SAT scores and HSGPA were used, in combination, as predictors.

Results for differential prediction paralleled those for differential validity. For the academic composite, the high-ability group was underpredicted (mean residual = 0.19), the medium-ability group was accurately predicted (mean residual = -0.01), and the low-ability group was overpredicted (mean residual = -0.19). A mean residual of 0.19 translates to roughly 0.3 of a standard deviation since the mean standard deviation of FYGPA was 0.71 in the Ramist et al. (1994) study. An examination of the gender subgroups showed that females were underpredicted (mean residual = 0.09), and males were overpredicted (mean residual = -0.10). Students whose best language was English were accurately predicted with a mean residual of 0.00; however, for students whose best language was not English, the average underprediction was 0.18. For race/ethnicity, Ramist et al. found Asian American (mean residual = 0.08) and white (mean residual = 0.01) students were underpredicted while American Indian (mean residual = -0.29), black (mean residual = -0.23), and Hispanic (mean residual = -0.13) students were overpredicted.

In 2000, Bridgeman, McCamley-Jenkins, and Ervin compared the differential validity and prediction of the revised and recentered 1995 SAT to the SAT prior to these changes. Based on individual-level data on nearly 100,000 students for the entering classes of 1994 and 1995 at 23 colleges and universities, differential validity was examined by gender as well as by gender and ethnicity combined. The differential validity results showed significantly higher correlations between SAT scores and FYGPA for females (rs ranging from 0.50 to 0.56) compared to males (rs ranging from 0.46 to 0.51). As for gender by ethnicity analyses, results revealed a similar pattern, with larger correlations for females compared to males for almost all racial/ethnic group comparisons. Results were not presented for ethnic/racial groups alone.

As for differential prediction, Bridgeman et al. (2000) found that SAT-V, SAT-M, and the combination of SAT sections all result in underprediction for females, with mean residuals ranging from 0.07 to 0.12, and overprediction for males, with mean residuals ranging from -0.13 to -0.08 for the revised and recentered 1995 SAT. This was similar to the results for the previous version of the SAT. As for gender by ethnicity analyses, the SAT overpredicted African American students' FYGPAs; however, overprediction was greater for African American males, with mean residuals ranging from -0.24 to -0.20 compared to African American females, with mean residuals ranging from -0.13 to -0.04 for the 1994 version of the SAT. Interestingly, the 1995 version of the SAT revealed a reversal of results, with underprediction for African American females for SAT-M (mean residual = 0.02) and SAT composite (mean residual = 0.03) and slightly more overprediction for African American males. For Asian American students, SAT-M resulted

[2] Correlations were corrected for restriction of range.

in overprediction for males (mean residuals of -0.13 and -0.14 for the 1994 and 1995 versions of the SAT, respectively) and underprediction for females (mean residuals of 0.05 and 0.03 for the 1994 and 1995 versions of the SAT, respectively). As for SAT-V, Asian American males were slightly overpredicted, with a mean residual of -0.01 for both the 1994 and 1995 SAT versions, whereas Asian American females were underpredicted, with a mean residual of 0.09 for the 1994 SAT and 0.07 for the 1995 SAT. As for Hispanic students, both the 1994 and 1995 versions of the SAT resulted in overprediction for both genders but to a larger extent for males (mean residuals ranging from -0.22 to -0.17) compared to females (-0.09 to -0.02). Finally, for white students, results showed similar patterns for the 1994 and 1995 versions of the SAT, with overprediction for males (mean residuals ranging from -0.11 to -0.07) and underprediction for females (mean residuals ranging from 0.11 to 0.18).

In 2001, Young reviewed the literature on differential validity and prediction by gender and race/ethnicity in college admission, which comprised 49 validity studies and/or prediction studies including the Ramist et al. (1994) and Bridgeman et al. (2000) studies described above, and arrived at similar conclusions. Using multiple predictors, which entailed HSGPA and SAT scores for the majority of studies, the predictive validity for college success (i.e., FYGPA, individual course grades, and cumulative GPA) was almost always higher for females (median = 0.54) than males (median = 0.51). The one exception was a study by Farver, Sedlacek, and Brooks (1975) examining the predictive validity for cumulative GPA by gender within race/ethnicity groups and found that the correlation for black males (r = 0.52) was higher than for black females (r = 0.42). For race/ethnicity, the predictive validity was similar for white and Asian American students but was 0.05 to 0.03 lower for African American students and 0.01 to 0.10 lower for Hispanic students. Only two of the studies included in Young's review examined American Indian students, and there were conflicting results. It is apparent that much more research should be conducted on this under-studied group. One goal of this study is to address this research gap.

In Young's review of differential prediction, 21 studies examined differential prediction by gender, and the overwhelming majority of them found underprediction for females. Of those 21 studies, 14 provided sufficient data, allowing Young (2001) to replicate the results. He found that the median underprediction for women was 0.05.[3] As for race/ethnicity, most studies found that college performance was overpredicted for African American and Hispanic students while there were mixed results and/or insufficient data for Asian American and American Indian students. In sum, the literature indicates that admissions criteria, such as SAT scores and HSGPA, tend to underpredict college success for females and overpredict college success for underrepresented students, with the exception of Asian Americans, where the results don't provide a clear picture. Furthermore, prior research has yielded little information on American Indian students, prohibiting the formation of any definitive conclusions for this group.

Additionally, there has been some research examining *why* the SAT results in differential prediction for various subgroups. Such research has found that these differences may not indicate a problem with the test but rather may be a function of differential student behavior by subgroup once enrolled in college (e.g., Clark and Grandy, 1984; Hewitt and Goldman, 1975; Stricker, Rock, and Burton, 1991; Wainer and Steinberg, 1992). For example, researchers have speculated that the SAT underpredicts females' college performance because they may enroll in less stringently graded courses in college than males (e.g., Clark and Grandy, 1984; Hewitt and Goldman, 1975). Correcting for course difficulty results in a reduction in the amount of underprediction for females. Furthermore, other researchers have offered additional explanations as to why females tend to earn higher grades in college than what would be predicted by standardized test scores. They propose that, on average, females have more effective study skills, higher class attendance, and greater academic motivation, which are positively related to college performance (e.g., Stricker, Rock, and Burton, 1991; Wainer and Steinberg, 1992). Much less is known as to why the SAT tends to overpredict minority student performance. Unlike gender, research has found that HSGPA tends to overpredict minority performance to a greater extent than SAT scores (e.g., Ramist et al., 1994). Therefore, the explanation for these findings may be different than that for gender. Given the significance of the issue, additional research should be conducted in order to more fully understand this complex phenomenon.

Purpose of the Current Study

The purpose of the current study was to examine the extent to which the revised SAT displays differential validity and differential prediction for various subgroups. Students were categorized and examined by gender, race/ethnicity, and best language. Furthermore, the extent to which the revised SAT exhibits differential validity and prediction in comparison to the SAT before its

[3] Thirteen of these studies used HSGPA and SAT scores in their predictor set. One study used ACT scores.

most recent revisions were evaluated to determine the impact of the changes for various subgroups. Finally, the predictor and/or combination of predictors that resulted in the least amount of differential validity and prediction were examined.

Method

Recruitment and Sample

Colleges and universities across the United States were contacted in order to provide first-year performance data from the fall 2006 entering cohort of first-time students. Preliminary recruitment efforts targeted the 726 four-year institutions that received at least 200 SAT score reports in 2005 to serve as the target population. Based on the College Board's Annual Survey of Colleges (The College Board, 2007), information on these schools' characteristics including control (public/private), region of the country, admissions selectivity, and enrollment size were used to form stratified target proportions for the institutions to be recruited.

Participating institutions were offered a stipend for the work involved in creating the files with students' first-year performance and retention to the second year data. These files were uploaded to the free and secure Admitted Class Evaluation Service™ (ACES™) after the 2006-07 school year concluded. ACES allows institutions to design and receive unique admission validity studies to—among other things—evaluate existing admissions practices. The ACES system served as the data portal for the study, securely transferring data from the institution to the College Board for aggregate analysis. Data collected from each institution included information such as students' course work and grades, FYGPA, and whether or not they returned for the second year. These data were matched to College Board databases that included SAT scores, self-reported HSGPA, and other demographic information.

The original sample consisted of individual level data on 196,364 students from 110 colleges and universities from across the United States. Upon transmission from the ACES system to the College Board, all data were examined for inconsistencies and miscoding to ensure the integrity of the analyses described.[4] The final sample included 151,316 students. For a detailed description of the institutional characteristics of the participating institutions, refer to Kobrin, et al. (2008).

Measures

SAT® Scores

Official SAT scores obtained from the 2006 College-Bound Senior Cohort database were used in the analyses. This database is comprised of the students who participated in the SAT Program and reported that they would graduate from high school in 2006. The student's most recent score was used in the analyses. The SAT is comprised of three sections—critical reading, mathematics, and writing—and the score scale range for each section is 200 to 800.

SAT-Questionnaire Responses

Self-reported gender, race/ethnicity, best language, as well as HSGPA were obtained from the SAT-Questionnaire that each student completes during registration for the SAT.

First-Year GPA (FYGPA)

Each participating institution supplied FYGPA values for their 2006 first-time, first-year students. The range of FYGPA across institutions was 0.00 to 4.27.

Analyses

Differential Validity

Differential validity was assessed by computing the correlation between SAT scores and FYGPA by subgroup. If the correlation coefficient varies by subgroup, then the test is said to exhibit differential validity. The relationship between SAT scores and FYGPA as well as HSGPA and FYGPA was computed by gender, race/ethnicity, and best language. All analyses were conducted at the level of the institution and then pooled for the entire sample. Furthermore, correlations were corrected for restriction of range using the Pearson-Lawley multivariate correction, with the 2006 College-Bound Seniors cohort as the population (Gulliksen, 1950). This correction was applied because correlations based on enrolled students would result in underestimation of the true correlation. This is due to the fact that students are selected based on test scores, which is referred to as restriction of range. The range is considered restricted because admitted students tend to have a narrower range of scores than the larger applicant pool, which artificially reduces the test score–FYGPA relationship.

[4] One check was for institutions with particularly high proportions of students with zero FYGPAs. This was incorporated into the data-cleaning procedures to ensure that these FYGPAs were not miscoded as zero when they should have been coded as missing. Students in the sample that did not have a valid FYGPA from their institution were removed from the sample (n = 6,207, 3 percent). Similarly, students without scores on the revised SAT were not included (n = 31,151, 16 percent). Additional students were removed from the sample because they did not indicate their HSGPA on the SAT-Questionnaire (n = 7,690, 4 percent).

In order to have sufficient information to calculate the multiple correlation for the three SAT sections and HSGPA with FYGPA, the sample size per subgroup had to be at least 5. However, correlations computed with this minimum requirement of subgroup size per institution were unstable, particularly for American Indian students due to their small sample size within institutions. Appendix A provides the results at minimum cut points of 5, 10, 15, 20, 25, 30, and 35 students for the American Indian subgroup to illustrate the instability of the results at various minimum cut points. For all other racial/ethnic groups, the value of a minimum cut point did not appreciably affect the results. That is, the results did not change if a minimum value of 5 students was used compared to a minimum value of 35 students; and therefore, results by other racial/ethnic groups are omitted from Appendix A. Due to the variability of the American Indian results, the minimum sample size was increased to 15.[5] In the results section, correlations by gender, race/ethnicity, and best language are presented.

Differential Prediction

To assess the extent to which the SAT, as well as HSGPA, exhibits differential prediction, regression equations within each institution were calculated. Before running the regression analyses, FYGPAs were standardized within each institution to have a mean of zero and a standard deviation of one to eliminate the impact of differences in grading scales across institutions on the results. This results in residuals that are on the same scale—standard units—across institutions, which eases interpretation and generalizability. Next, regression equations were estimated separately for each school, and the average residual by subgroup was computed across the entire sample. Note that the issue of minimum subgroup size described above does not enter into this analysis since regression equations are estimated for the entire institution, and the residuals are simply aggregated by subgroup. In other words, if an institution has at least one student from a given subgroup, that student and institution are included in the analysis.

In terms of the method of calculation, least squares estimations, the average residual value for the total group always equals zero. However, if the average residual value by subgroup does not equal zero, then the measure is said to exhibit differential prediction. Specifically, if the average residual value is positive for a specific subgroup, then the test tends to underpredict academic success for that group. In other words, students from this group tend to perform better than what is predicted by the regression equation. Likewise, if the average residual value is negative, then the test tends to overpredict academic success for that group, or the students tend to perform worse than what is predicted by the regression equation.

An example is provided for illustrative purposes. Suppose subgroup X has an average standardized residual of 0.10.[6] This indicates that subgroup X tends to be underpredicted by one-tenth of a standard deviation. The standard deviation of FYGPA for this sample (calculated across all institutions) is 0.71. Therefore, one-tenth of a standard deviation equals approximately 0.07, which means that students from subgroup X tend to earn a FYGPA that is 0.07 points higher than their predicted FYGPA (e.g., $FYGPA_{earned}$ = 3.57 versus $FYGPA_{predicted}$ = 3.50). In the results section, standardized mean residuals by gender, race/ethnicity, and best language are presented. Furthermore, Appendix B provides the differential validity results based on unstandardized residuals for comparison with previous findings.

Results and Discussion

Descriptive Statistics

The sample size, mean, and standard deviation for each predictor and predictor set by subgroup, as well as for the total group, are presented in Table 1. Of the 151,316 students, 54 percent are female, which is the same as the 2006 College-Bound Seniors cohort. Similar to the 2006 cohort, the results for this sample reveal that males, on average, score higher on the SAT mathematics section (SAT-M) (M = 602, F = 559) and the SAT critical reading section (SAT-CR) (M = 564, F = 557), whereas females, on average, score higher on the SAT writing section (SAT-W) (F = 557, M = 550) and HSGPA (F = 3.65, M = 3.55).

In terms of race/ethnicity, the sample consists of 69 percent white students, 9 percent Asian American students, 7 percent African American students, 7 percent Hispanic students, less than 1 percent American Indian students, 4.5 percent no response, and 3 percent other.

[5] A minimum value of 15 per subgroup was chosen for two reasons: (1) to strike a balance between estimating correlations with minimal error while including as many institutions and students as possible in the analyses and (2) to ensure that the number of degrees of freedom under this procedure is non-negative because the number of covariance parameters estimated in the separate 5-by-5 covariance matrices for each subgroup at each institution is 15.

[6] The use of the term "standardized residuals" is different than what is provided in the SPSS output for standardized residuals. For SPSS, regression analysis is conducted with the original scale of the variable. Then, the predicted value is subtracted from the observed value. The difference is standardized to have a mean of zero and standard deviation of one. In this report, FYGPA was standardized to have a mean of zero and standard deviation of one. This transformed variable was used in the regression analysis to estimate predicted FYGPA. Next, the predicted value was subtracted from the standardized FYGPA to compute the residual.

Table 1

Descriptive Statistics of Study Variables

Variable		*n*	*SAT-CR*		*SAT-M*		*SAT-W*		*HSGPA*	
			Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Gender	Male	69,765	564.28	95.19	601.59	95.46	550.13	94.98	3.55	0.52
	Female	81,551	556.53	96.08	558.89	93.35	556.81	93.60	3.65	0.48
Race/Ethnicity	American Indian or Alaska Native	798	544.20	88.04	555.29	88.31	529.36	87.90	3.52	0.54
	Asian, Asian American, or Pacific Islander	14,296	561.96	104.75	623.66	97.56	561.65	101.80	3.66	0.47
	Black or African American	10,304	506.13	88.59	503.43	87.87	497.83	87.18	3.39	0.55
	Hispanic, Latino, or Latin American	10,659	524.15	92.93	536.93	94.06	519.64	90.54	3.59	0.51
	No Response	6,738	587.24	101.48	590.43	98.42	575.60	100.72	3.63	0.50
	Other	4,497	557.86	98.67	571.94	98.62	553.31	97.31	3.57	0.50
	White	104,024	567.34	92.17	583.79	91.96	560.46	90.97	3.62	0.49
Best Language	English	140,559	563.00	94.39	578.95	95.50	555.88	93.16	3.60	0.50
	English and Another	7,458	531.43	100.63	569.71	108.83	534.47	100.85	3.61	0.49
	Another Language	1,718	461.74	100.23	604.76	115.13	477.97	102.74	3.61	0.52
	Not Stated	1,581	544.35	106.22	558.84	111.05	535.88	106.85	3.53	0.54
Total		151,316	560.10	95.75	578.58	96.70	553.73	94.30	3.60	0.50

This is similar to the 2006 College-Bound Seniors cohort, with white students being slightly overrepresented in the sample, and African Americans and Hispanic students being slightly underrepresented. Likewise, the pattern of subgroup differences by race/ethnicity resembles the results of the 2006 cohort. Namely, white and Asian American students as well as students who do not report their race/ethnicity generally score higher on all three sections of the SAT and have higher HSGPAs than American Indian, African American, and Hispanic students.

For the best language subgroups, 93 percent of the sample indicated that their best language is English, 5 percent indicated English and another language, 1 percent indicated another language, and 1 percent did not respond. Students whose best language is English are over-represented in this sample compared to the 2006 College-Bound Seniors cohort. These students score higher than students who selected another language or English and another language on SAT-CR and SAT-W. On the other hand, for SAT-M, students whose best language is not English score the highest, with a mean of 605. These patterns of findings are similar to those of the 2006 cohort data. The best language subgroups perform similarly with regard to HSGPA except for those who did not respond. These students perform slightly lower, with a mean HSGPA of 3.53.

Overall, SAT scores are slightly higher for this sample compared to those of the 2006 College-Bound Seniors cohort. For example, the mean SAT-M score for College-Bound Seniors in 2006 was 518 but was 579 for this sample. This was expected, however, because these are enrolled students and not applicants. It should also be pointed out that the sample size per subgroup, which limited much of the previous research on differential prediction and validity, is relatively large for this study. For example, data are available on 798 American Indian students (compared to 184 in Ramist et al., 1994), which may be the most under-studied subgroup due to the limited amount of data available. The sample sizes for the other racial/ethnic groups are also quite large, as well as the sample sizes for each best language subgroup. This large dataset will provide a precise estimation of the extent to which the SAT as well as HSGPA results in differential validity and prediction for various subgroups.

Differential Validity

Gender

For gender, the results show that the SAT is more predictive of FYGPA for females than males. For females, the correlations for the three sections of the SAT as well as the multiple correlation of the SAT with FYGPA range from 0.52 to 0.58. For males, the correlations are smaller, ranging from 0.44 to 0.50. As for HSGPA, there is a similar pattern with a larger correlation for females (r = 0.54) compared to males (r = 0.52), although the differences are smaller. Combining the three sections of the SAT and HSGPA results in a multiple correlation of 0.65 for females and 0.59 for males. Differential validity is the smallest for HSGPA, with a difference of 0.02, followed by the use of all measures, with a difference of 0.06. Differential validity is the largest for the SAT, by section and for the combination of all three sections, with differences ranging from 0.07 to 0.08. The findings are similar to those reported in previous research (Bridgeman et al., 2000; Ramist et al., 1994; Young, 2001). More detailed results are provided in Table 2.

Race/Ethnicity

The results for race/ethnicity show that for the individual SAT sections, the SAT is most predictive for white students,

Table 2

Correlation of SAT Scores and HSGPA with FYGPA by Subgroups (Minimum Sample Size ≥ 15)

Variable		*k*	*n*	*SAT-CR*	*SAT-M*	*SAT-W*	*SAT*	*HSGPA*	*SAT, HSGPA*
Gender	Male	107	69,765	0.44	0.45	0.47	0.50	0.52	0.59
	Female	110	81,551	0.52	0.53	0.54	0.58	0.54	0.65
Race/Ethnicity	American Indian or Alaska Native	16	384	0.41	0.41	0.42	0.54	0.49	0.63
	Asian, Asian American, or Pacific Islander	82	14,109	0.41	0.43	0.44	0.48	0.47	0.56
	Black or African American	83	10,096	0.40	0.40	0.43	0.47	0.44	0.54
	Hispanic, Latino, or Latin American	86	10,486	0.43	0.41	0.46	0.50	0.46	0.57
	No Response	90	6,544	0.46	0.43	0.49	0.53	0.52	0.63
	Other	73	4,175	0.46	0.45	0.48	0.55	0.45	0.61
	White	109	104,017	0.48	0.46	0.51	0.53	0.56	0.63
Best Language	English Only	110	140,559	0.49	0.47	0.52	0.54	0.55	0.63
	English and Another Language	79	7,237	0.41	0.43	0.45	0.50	0.42	0.55
	Another Language	28	1,292	0.28	0.34	0.32	0.42	0.35	0.48
	Not Stated	44	1,171	0.39	0.37	0.45	0.59	0.47	0.69

Note: Pooled within-institution correlations are presented. Correlations are corrected for restriction of range using the 2006 national cohort. Computations were made within institutions for subgroups with at least 15 members. k = number of qualifying institutions and n = subgroup sample size. SAT is the multiple correlation for all three sections.

with correlations ranging from 0.46 to 0.51. The SAT appears less predictive for underrepresented groups, in general, with correlations ranging from 0.40 to 0.46. Likewise, the multiple correlation for three sections of the SAT is higher for white students than for the underrepresented groups (0.53), with the exception of American Indian students (0.54). It should be pointed out that sample size per institution for American Indian students is sometimes quite small (n = 15); therefore, the results should be interpreted with caution. As for HSGPA, the same pattern emerges with higher correlations for white students (0.56) compared to minority groups (correlations ranging from 0.44 to 0.49). The results are consistent with past findings, except the results reveal lower correlations for Asian American students compared to white students. Previous research found more similar correlations.

Best Language

For best language, the correlation between SAT scores and FYGPA is highest for students whose best language is English, with values ranging from 0.47 to 0.54. Correlations for students who select English and another language as their best are in the middle, with correlations ranging from 0.41 to 0.50. The relationship is weakest for students whose best language is not English, with correlations ranging from 0.28 to 0.42. These findings are similar to the results of Ramist et al. (1994); however, in that study, English and another language students and students whose best language is not English were collapsed into one group. These results suggest that these students should not be collapsed into one group but rather they should be analyzed separately due to divergent results between these groups. For HSGPA, a similar pattern to that of the SAT emerges, with highest correlations for English-only students (r = 0.55) and lowest for students whose best language is not English (r = 0.35). More details are provided in Table 2.

Differential Prediction

Gender

The results for gender reveal that the SAT tends to underpredict FYGPA for females, with mean standardized residuals ranging from 0.10 to 0.17 for the three sections and the combined SAT. Conversely, the SAT overpredicts male performance, with mean standardized residuals ranging from -0.11 to -0.20. A similar pattern emerges for HSGPA as well as for the combination of HSGPA and SAT, with females being underpredicted (0.07 and 0.09, respectively) and males being overpredicted (-0.08 and -0.10, respectively). Results of these analyses are provided in Table 3.

The pattern of results is in alignment with previous research on the differential validity of the SAT by gender (Bridgeman et al., 2000; Ramist et al., 1994; Young, 2001), although these values are slightly larger than what is typically found. This is due to the method of computation. Specifically, FYGPAs were standardized to have a mean of zero and a standard deviation of one. Therefore, interpretation of the residuals can be explained in terms of standard units. In order to compare these results with past findings, multiplying the standardized residuals by the standard deviation of FYGPA for this sample (0.71) will provide an estimate of the unstandardized residuals. For example, the mean unstandardized residual for females for the SAT range from 0.07 to 0.11 compared to the mean standardized residuals ranging from 0.10 to 0.17 reported above. For more comparisons, refer to Table 3 for the standardized residuals and Appendix B for the unstandardized residuals.

Recall that Ramist et al. (1994) found that females' FYGPAs were underpredicted (mean unstandardized residual = 0.09) and males' FYGPAs were overpredicted (mean unstandardized residual = -0.10) by the SAT (M + V). The mean standardized residuals for this study

Table 3

Average Overprediction (-) and Underprediction (+) of FYGPA for SAT Scores and HSGPA by Subgroups (Standardized Residuals)

Variable		*k*	*n*	*SAT-CR*	*SAT-M*	*SAT-W*	*SAT*	*HSGPA*	*SAT, HSGPA*
Gender	Male	107	69,765	-0.14	-0.20	-0.11	-0.15	-0.08	-0.10
	Female	110	81,551	0.12	0.17	0.10	0.13	0.07	0.09
Race/Ethnicity	American Indian or Alaska Native	103	798	-0.26	-0.25	-0.22	-0.22	-0.25	-0.20
	Asian, Asian American, or Pacific Islander	109	14,296	0.05	-0.07	0.04	0.01	0.02	0.02
	Black or African American	108	10,304	-0.30	-0.26	-0.26	-0.20	-0.32	-0.17
	Hispanic, Latino, or Latin American	110	10,659	-0.17	-0.16	-0.16	-0.11	-0.27	-0.12
	No Response	110	6,738	-0.01	0.04	0.01	0.00	0.05	0.01
	Other	110	4,497	-0.04	-0.03	-0.04	-0.03	-0.03	-0.01
	White	110	104,024	0.04	0.05	0.04	0.03	0.06	0.03
Best Language	English Only	110	140,559	0.00	0.01	0.00	0.00	0.01	0.00
	English and Another Language	110	7,458	-0.03	-0.09	-0.04	-0.02	-0.13	-0.03
	Another Language	102	1,718	0.40	0.00	0.37	0.33	0.06	0.30
	Not Stated	107	1,581	-0.10	-0.10	-0.08	-0.07	-0.11	-0.07

Note: Mean residuals based on standardized within-institutions FYGPAs are provided. Negative values indicate overprediction. Positive values indicate underprediction. Values are computed by subtracting predicted FYGPA from actual FYGPA. FYGPA prediction equations are calculated for each institution separately. SAT is the multiple correlation for all three sections.

for the SAT (which includes the writing section) are 0.13 and -0.15 for the females and males, respectively. However, the unstandardized residuals perfectly replicate the Ramist et al. study, with 0.09 for females and -0.10 for males (see Appendix B). This is also similar to the results of the Bridgeman et al. (2000) study, which found that the 1995 version of the SAT underpredicted females (mean unstandardized residual = 0.10) and overpredicted males (mean unstandardized residual = -0.11). In sum, the revised version of the SAT appears to result in the same amount of differential validity for gender.

Race/Ethnicity

For race/ethnicity, the pattern of results supports previous findings. Specifically, American Indian, African American, and Hispanic students are overpredicted for all measures and combinations of measures. African American students' FYGPAs tend to be the most overpredicted, with mean standardized residuals ranging from -0.32 to -0.17. White students, along with students who did not state their race/ethnicity, tend to be accurately predicted to slightly underpredicted, with mean standardized residuals ranging from -0.01 to 0.06. Students who selected "other" tend to be slightly overpredicted, with mean standardized residuals ranging from -0.04 to -0.01. For Asian American students, their FYGPAs tend to be accurately to slightly underpredicted for all measures (mean standardized residuals ranging from 0.01 to 0.05), except for SAT-M (mean standardized residual = -0.07), where their performance is overpredicted. Unlike the results for gender, HSGPA tends to result in the most differential validity for most racial/ethnic groups, while the combination of SAT and HSGPA results in the least differential validity. These results support the argument of using multiple measures in the admissions process because of the fact that the combination of the two measures results in the least amount of differential prediction. See Table 3 and Appendix B for more detailed results.

As with gender, the mean standardized residuals are slightly larger than what has been reported in the past. To make comparisons with the past findings using a similar metric, refer to Appendix B, which provides the unstandardized residuals. Comparing the unstandardized residuals for SAT in the current study with those found in 1994 by Ramist et al., the current values tend to be smaller (American Indian = -0.13 versus -0.29, Asian American American = 0.01 versus 0.08, African American = -0.14 versus -0.23, Hispanic = -0.08 versus -0.13, and white = 0.02 versus 0.01). In sum, the revised SAT results in less differential prediction by racial/ethnic group.

Best Language

For the best language subgroups, the SAT results show that students whose best language is English are accurately predicted (mean standardized residual ranging from 0.00 to 0.01), whereas students whose best language is not English tend to be underpredicted by the critical reading and writing sections of the SAT (mean standardized residuals of 0.40 and 0.37, respectively) and accurately predicted by the mathematics section (mean standardized residual = 0.00). These results are similar to prior findings for these groups (Ramist et al., 1994). Interestingly, the SAT tends to overpredict the FYGPAs (mean standardized residual ranging from -0.09 to -0.02) of students whose best language is English and another language. In the Ramist et al. (1994) study, students who did not select English as their best language were combined into a single group; however, these results suggest that students whose best language is not English and students whose best languages are English and another language should be

analyzed separately. Finally, students who did not respond to the best language question were overpredicted, with mean standardized residuals ranging from -0.07 to -0.10 for the SAT. For HSGPA, the results depict a similar pattern except for students whose best language is not English. The amount of underprediction in this group tends to be smaller with a mean standardized residual of 0.06 compared to the results for SAT-CR (mean standardized residual = 0.40) and SAT-W (mean standardized residual = 0.37). Table 3 provides mean standardized residuals for all four groups for each SAT section, combined SAT, HSGPA, and the combination of SAT and HSGPA. As with the other subgroups, refer to Appendix B for the unstandardized residual results.

Future Research

Future research should replicate these findings as well as expand the current study by examining alternative outcomes and different grouping variables. For example, does a similar pattern of results emerge in terms of differential validity and prediction when examining retention to second year, second-year GPA, cumulative GPA, and graduation? This would shed light on whether the same results occur regardless of the indicator of college success. As for alternative subgroups, similar analyses should be conducted by college major. Do SAT scores over- or underpredict college performance for specific majors? It is reasonable to hypothesize that SAT scores would underpredict college performance for students in less rigorous majors and overpredict college performance for students in more rigorous majors based on the score distribution of grades for their required courses; however, empirical data should be collected to test such claims. Finally, similar to the studies by Farver, Sedlacek, and Brooks (1975) and Bridgeman, McCamley-Jenkins, and Ervin (2000), future research should examine more fine-grained subgroups such as African American females versus African American males to see whether different patterns emerge.

In addition, the College Board has a full research agenda planned for the upcoming year. Additional studies will examine the placement validity of the SAT into first-year English and mathematics courses. The predictive validity of the SAT in terms of alternative indicators of college success, such as retention to second year, will also be examined. Numerous studies are planned to examine the relationship between AP® participation and scores with college success. More research will also be conducted on students with discrepant SAT section scores and discrepant HSGPA and SAT scores. In the future, additional data will be collected from participating schools in order to conduct more longitudinal studies and examine more distal outcomes such as cumulative GPA and graduation.

References

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

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 Research Report No. 2000-1). New York: The College Board.

Clark, M. J., & Grandy, J. (1984). *Sex differences in the academic performance of scholastic aptitude test takers* (College Board Research Report 84-8). New York: The College Board.

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.

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*, 243–50.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley and Sons.

Hewitt, B.N., & Goldman, R.D. (1975). Occam's razor slices through the myth that college women overachieve. *Journal of Educational Psychology*, 67, 325-30.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average* (College Board Research Report No. 2008-5). New York: The College Board.

Lawrence, I. M., Rigol, G. W., Van Essen, T., & Jackson, C. A. (2003). *A historical perspective on the content of the SAT* (College Board Research Report No. 2003-3). New York: The College Board.

Raju, N. S., & Ellis, B. B. (2003). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 156–88). San Francisco, CA: Jossey-Bass.

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

Stricker, L., Rock, D., & Burton, N. (1991). *Sex differences in SAT prediction of college grades* (College Board Research Report No. 91-2). New York: The College Board.

The College Board (2007). *The College Board college handbook 2007* (44th ed.). New York: The College Board.

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*, 323–36.

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

Appendix A: Correlation of SAT Scores and HSGPA with FYGPA for American Indian Students at Different Minimum Cut Points

Minimum Students per Subgroup		*5*	*10*	*15*	*20*	*25*	*30*	*35*
	k	55	26	16	9	6	5	2
	n	694	502	384	265	199	174	81
Correlations	SAT-CR	0.33	0.40	0.41	0.45	0.41	0.43	0.37
	SAT-M	0.33	0.40	0.41	0.43	0.35	0.36	0.33
	SAT-W	0.35	0.42	0.42	0.46	0.42	0.40	0.39
	SAT	0.62	0.57	0.54	0.54	0.49	0.46	0.41
	HSGPA	0.45	0.49	0.49	0.53	0.49	0.48	0.38
	SAT, HSGPA	0.74	0.68	0.63	0.64	0.59	0.57	0.47

Note: Pooled within-institution correlations are presented. Correlations are corrected for restriction of range using the 2006 national cohort. Computations were made within institutions. *k* = number of qualifying institutions and *n* = subgroup sample size. SAT is the multiple correlation for all three sections.

Appendix B: Average Overprediction (-) and Underprediction (+) of FYGPA for SAT Scores and HSGPA by Subgroups (Unstandardized Residuals)

Variable		*k*	*n*	*SAT-CR*	*SAT-M*	*SAT-W*	*SAT*	*HSGPA*	*SAT, HSGPA*
Gender	Male	107	69,765	-0.10	-0.13	-0.08	-0.10	-0.06	-0.07
	Female	110	81,551	0.08	0.11	0.07	0.09	0.05	0.06
Race/Ethnicity	American Indian or Alaska Native	103	798	-0.16	-0.15	-0.14	-0.13	-0.15	-0.12
	Asian, Asian American, or Pacific Islander	109	14,296	0.04	-0.04	0.03	0.01	0.02	0.02
	Black or African American	108	10,304	-0.20	-0.17	-0.18	-0.14	-0.21	-0.11
	Hispanic, Latino, or Latin American	110	10,659	-0.12	-0.10	-0.10	-0.08	-0.17	-0.08
	No Response	110	6,738	0.00	0.03	0.00	0.00	0.03	0.01
	Other	110	4,497	-0.02	-0.02	-0.03	-0.02	-0.02	-0.01
	White	110	104,024	0.03	0.03	0.03	0.02	0.04	0.02
Best Language	English Only	110	140,559	0.00	0.00	0.00	0.00	0.00	0.00
	English and Another Language	110	7,458	-0.01	-0.05	-0.02	-0.01	-0.08	-0.02
	Another Language	102	1,718	0.26	0.01	0.25	0.21	0.05	0.19
	Not Stated	107	1,581	-0.07	-0.06	-0.05	-0.04	-0.07	-0.04

Note: Mean residuals are provided. Negative values indicate overprediction. Positive values indicate underprediction. Values are computed by subtracting predicted FYGPA from actual FYGPA. FYGPA prediction equations are calculated for each institution separately. SAT is the multiple correlation for all three sections.