The College Board
Office of Research and Development

Research Summary
RS-04, February 1998

SAT® and Gender Differences[1]

The SAT® I: Reasoning Test actually predicts how well females will do in college better than it predicts for males. Results from validity studies that have been conducted with hundreds of colleges and universities and examined by ETS® and external researchers consistently confirm that the SAT correlation with both freshman GPA (FGPA) and individual course grades are actually higher for females than males. Research shows that the correlation between SAT scores and FGPA is .62 for females and .56 for males (after appropriate statistical adjustments are made for unreliability of FGPA and restriction in the range of SAT scores at colleges). When SAT and high school (HS) GPA are used in combination to predict FGPA, the correlations increase to .71 for females and .65 for males. When individual college course grades are predicted, these correlations are .79 for females and .73 for males. The SAT alone is a better predictor of students' grades in most individual freshman courses than high school grades, with the exception of courses in English and foreign languages. However, again, the combination of SAT and high school grades provides the best prediction of individual college grades. These findings hold up for all subgroups, and the SAT actually has a higher correlation than HS GPA with FGPA for African-American and Asian-American students.

The SAT contributes important supplemental information in predicting college success for women beyond that provided by high school grades. Many critics of testing may argue that the SAT is unnecessary because high school grades correlate nearly as highly with FGPA as SAT and high school grades combined. This argument ignores two important points. First, while high school grades are a good predictor of college success, when SAT scores are added, the prediction increases substantially. For females, adding SAT scores to HS GPA increases the prediction of FGPA by .10 and the prediction of college course grades by .15 (the incremental increases for males are .07 and .12, respectively). Including the SAT in admission decisions increases the accuracy and validity of those decisions for females and males—this results in fairer decisions for individual students. Second, the SAT is the only objective and standard measure available to compare students who attend different schools and complete different curricula. Grades reflect student achievement and motivation, but also incorporate factors such as student attendance, participation, punctuality, and the difficulty and grading standards of a school. The SAT is a measure of developed verbal and mathematical reasoning. The combined use of SAT scores and HS GPA results in the most valid, and consequentially fairest, prediction of college performance for all groups, including women.

The proportion of college-bound students with A averages has increased by nearly one-third since 1987. As grade inflation becomes an increasing problem nationally, the SAT provides an independent and objective source of information for all students, including females. Since 1987, the population of students with A averages in high school has grown from 28 percent to 37 percent, while their scores on the SAT have fallen slightly. At the University of California at Berkeley last year, 12,000 of their 27,000 applicants submitted HS GPAs of 4.00 or above.

The SAT and HS GPA each slightly underpredict FGPA for females. However, when these measures are used together, this underprediction is reduced to a lower level. Use of the same single statistical equation (or what we call regression equation) to predict any future behavior (e.g., FGPA, job performance) will result in overpredic-

[1]A common index used to describe how well a measure like grades or the SAT predicts college grades is the correlation between these measures. A perfect correlation would be "1," a situation where performance on one measure will always result in perfect prediction on a second measure. A correlation of "0" represents a situation where two events are completely unrelated. Arguments concerning the validity, reliability, and fairness of the SAT, grades, and other measures are usually presented in the form of "correlational data."

tion for some groups and underprediction for other groups. Underprediction of FGPA actually means that students will receive slightly higher FGPA than predicted. HS GPA underpredicts how well Asian Americans and females will do in college, while it overpredicts performance for whites, males, African Americans, and Hispanics. If used alone, the SAT also underpredicts FGPA for females and Asian Americans, while overpredicting performance for these same other groups. When the SAT and HS GPA are combined, these effects are substantially reduced—specifically, the underprediction of female FGPA is reduced to .06. To put this into perspective, if the average FGPA at the college were 3.00, then women would be expected to have an average GPA of 3.06.

Differences still persist among females and males in high school course preparation that affect performance on standardized tests and other outcome measures. Females are quickly closing the gap with males that has persisted for decades in the number of advanced math and science course completed. However, important differences still persist in the proportion of males and females completing advanced courses in math, science, and computer programming. Several studies show that gender differences are substantially reduced when we control for these differences in preparation.

Differences in college course selection also account for much of the gender differences found when FGPA is used. Research also shows that a smaller proportion of females will complete rigorous science and math courses and major in these and related fields in college. This difference in course taking between males and females, while small, does account for over half of the score difference for SAT mathematics because science and math courses have been consistently shown to have more stringent grading standards than courses in the humanities, arts, social sciences, and English across a range of universities and colleges. For example, the average high school grades in math and science are 3.00 and 3.12, respectively. Average grades in high school arts and music courses, social science and history, English and foreign language courses range from 3.14 to 3.68. Related research demonstrates that even when males and females complete similar numbers of math and science courses, males are more likely to take more rigorous courses geared for math and science majors (e.g., engineering, chemistry), which are graded more stringently than general science and math courses. Males also achieve higher grades in college math and science courses designated for majors. These findings are supported by research illustrating that underprediction of females is cut in half when grades in college courses are examined, as opposed to FGPA.

Course-taking patterns also reflect substantial differences in aspirations and expectations of males and females. Males and females still aspire to different fields and majors, which in turn accounts for differences in course-taking patterns and affects scores on tests such as the SAT. In 1996, the top four intended majors for males were: (1) Engineering, (2) Business & Commerce, (3) Health & Allied Services, and (4) Social Science and History. The top four intended majors for females were: (1) Health & Allied Services, (2) Social Science and History, (3) Business & Commerce, and (4) Education.

More than 75,500 additional females take the SAT than males, and these "additional" females are less likely to have taken rigorous academic courses than other students. If equal numbers of males and females took the SAT, females would actually have a somewhat higher score than males on the verbal scale, rather than the four-point gap currently found. A much higher proportion of females than males taking the SAT come from families with lower levels of income and parental education. What these and other findings suggest is that a greater proportion of women from lower socioeconomic status (SES) families having less preparation are inclined to attend college than males. Although this increased interest in college for females should be applauded, research has shown that these differences in the self-selected population of males and females taking a test tends to decrease the average score or remain unchanged. These results hold for other undergraduate and graduate admission tests.

The assumption that females should get higher math scores on the SAT because they receive higher math grades in high school and college courses is false. As explained above, differences between males and females in both the number and rigor of math and science courses completed prior to college and upon entering col-

lege still differ somewhat. When these differences in preparation are controlled for, the gender difference on the SAT is reduced by about two-thirds or more. Some studies have shown that when additional differences in terms of aspirations (e.g., intended major, career), interests, and expectations are considered, male and female scores do not differ on the SAT math section. As noted above, men achieve higher grades in math and science courses intended for majors; these tend to be more rigorous courses geared toward engineering and science majors. Other research suggests that females are more likely to receive higher grades because they consistently receive higher teacher ratings for punctuality, attendance, following directions, and participation—components that are not part of the SAT or other standardized tests of achievement or ability.

Twelfth-grade females tend to perform better than males on verbal and writing tests, and males score higher on tests of natural science, mechanical skills, and math. These findings are found in an analysis of 74 different tests. In math specifically, females tend to have a slight advantage in fourth grade, with no difference found by eighth grade and a slight male advantage by twelfth grade on 10 of 12 national tests. However, twelfth-grade females show a clear advantage on most national tests measuring writing, language use, reading, study skills, and perceptual skills.

Fairness does not mean equal outcomes. Groups differ on nearly every major educational outcome and input whether we look at standardized tests, performance assessments, grades, course work, or extracurricular achievements. It is unrealistic to expect equal outcomes on the SAT or any other measure when such major differences persist in terms of preparation, SES, and aspirations.

The SAT is a fair and objective assessment. ETS has instituted multiple layers and procedures to ensure the assessments they develop meet the highest level of content and psychometric requirements for validity and fairness for all groups. ETS certifies individuals to conduct sensitivity reviews to ensure test content does not contain language, symbols, words, phrases, and examples that may favor one group over another or might be regarded as sexist, racist, or negative toward any group. Each SAT form is reviewed by high school and college faculty for content and bias. Extensive formal guidelines exist, and an external panel reviews all aspects of the test development process. In addition, all items on the SAT are pretested prior to inclusion on an operational form.

Differential Item Functioning (DIF) analyses are conducted to compare how five groups (including females) perform in comparison to white students matched to their ability. Calculations determine the likelihood that differences in performance on any question result from overall ability differences or something inherent in the question. Questions that clearly perform differently for any group are carefully reviewed and nearly always eliminated from the pool of potential test questions. A number of additional analyses and quality control procedures are implemented at the question and test level to ensure tests are fair to all groups.

With respect to the addition of a writing component to the PSAT/NMSQT™:

The change was made for educational reasons. We had been exploring whether and how to add writing since development of the New SAT was undertaken in the early 1990s. Educators have warmly welcomed the change.

Differences in mean scores reflect the developed skill levels of the different groups of students who take the test. It is not biased. Stringent quality control during test development ensures that no question lands in the test that might be somehow skewed against a particular group. The test measures and reflects what is out there. The test is an objective measure that is extremely fair to all students.

Data from many different tests show that there are group differences in skill areas. As a general rule, females tend to do somewhat better on verbal and writing areas, while males tend to do better in mathematics.

There are probably many different factors that cause the group differences we see on the test. In the case of math, research has shown that at least some of the difference is caused by course-taking patterns, with females still not taking advanced math and science courses as frequently as males. In the case of writing, we may need to begin to emphasize to young men the importance of writing skills in college and in many other endeavors.

It is important to remember that we are focusing on averages, and that there are many

males who do very well in writing and other verbal areas and many females who excel in mathematical areas. And for self-selected tests like the SAT and PSAT/NMSQT™, there are no equal "samples" of students from all groups taking the test. Therefore, average scores don't necessarily reflect what the differences would be if *all* students had taken the test.

Office of Research and Development
The College Board
45 Columbus Avenue
New York, NY 10023-6992
212 713-8000
research@collegeboard.org