PART 3

Sex and Gender Bias in Testing
Many of the research studies reviewed in this report use test results as key measures of outcomes. Whenever tests are used, issues of sex and gender bias must be addressed. As this section will show, questions of sex bias in the design, construction, and administration of tests are complex. Perfectly good achievement tests can be designed in math, language arts, and other subjects on which girls will tend to score higher than boys. Other equally good tests can be developed on which boys will tend to score higher, and still other tests can be developed on which there will be no sex differences. “No sex differences” does not mean that all girls and boys will get the same scores. Rather, it means that scores will vary according to individual skills, knowledge, and motivation. However, the scores of girls as a group and of boys as a group will be about the same.

**Bias in Tests**

Test results are the basis of most educational research and evaluation and play a large role in research-and-

*The contribution of researchers at the Educational Testing Service (ETS) to what we know about testing and gender—and their willingness to share that data—must be acknowledged. Indeed, because of the accessibility of SAT/GRE data, it has been used so frequently in studies of gender and test bias that care should be taken in generalizing the results to other tests. In research on testing, the terms sex and gender are used interchangeably even more frequently than in other areas. This section reflects that usage.

“[Testing is] overrelied upon, lacks adequate public accountability, sometimes leads to unfairness in the allocation of opportunities, and too often undermines vital social policies.”

evaluation-based policy decisions. Tests are instrumental in deciding who passes, who goes to selective colleges, and who receives scholarships.\(^1\) While girls are more apt than boys to go to college and to get higher grades in both high school and college, scholarships based on test scores are twice as likely to go to boys.\(^2\)

The tests discussed here are standardized tests. Teacher-made tests are not included because there is little research on them and little that can be generalized about them. The discussion on standardized tests considers, as much as possible, the interaction of sex and race/ethnicity, a linkage usually overlooked.

Much of the "test unfairness" cited by the National Commission is seen as being caused by sex and race/ethnic bias. Biased tests, those that favor one sex or one group, are specifically prohibited by Title IX of the Education Amendments of 1972. Discrimination in the use of tests is also forbidden.\(^3\)

While tests contribute to unfairness and discrimination, in most cases tests reflect rather than cause inequities in American education. Differences between groups are often taken as proof of test bias. But some differences "may simply mean that the groups on average know different amounts about what is being tested."\(^4\) Particularly when issues of race/ethnicity and socioeconomic status are addressed, differences in test results often reflect differences in educational opportunities and resources.

A good test must be both reliable and valid. The reliability of a test is the extent to which it measures something consistently. The validity of a test is the extent to which it measures that which it is said to measure.
Validity is specific to a particular use and a particular situational context. In aptitude tests, the most important form of validity is predictive validity, the degree to which the test predicts future performance.

As already noted, the fact that groups score differently on a test does not necessarily mean that the test is biased. If, however, the score differences are related to the validity of the test—for example, say that girls and boys know about the same amount of math but boys’ test scores are consistently and significantly higher—then the test is biased. Bias is also present when the number of references to or characters of one sex exceeds the numbers represented by the other sex or if roles are presented stereotypically.5

Research in this area has centered around the search for bias in individual test items. To look for bias, researchers match females and males on overall test performance and then look at test items where the gender differences are greater than on the test as a whole.

The most obvious source of bias is the number of references to women and men included in items and whether the sexes are portrayed in stereotypical ways. In the 1970s most test companies developed procedures to balance the numbers of references to women and men in their tests and to screen out items that might be offensive. However, a 1984 analysis of tests found twice as many references to men as to women, and more pictures of and references to boys than girls. Even male animals were listed almost twice as often as female animals.6 A later study of the SAT found references to forty-two men and only three women in the reading comprehension passages used in the four 1984–85
exams. Of the forty-two men, thirty-four were famous and their work was cited; one of the three women was famous (Margaret Mead) and her work was criticized. These findings indicate there has not been much of a change from Tittle, McCarthy, and Steckler's 1974 study of the roles of women and men in achievement tests.

It is possible to create or eliminate sex differences in test scores in the content areas or in "intelligence" by the selection of test items. For example, the 1942 revision of the Stanford Binet test had as its aim to "produce a scale which will yield comparable I.Q.s for the sexes." To do this, the authors accepted the hypothesis that large sex differences can exist in experience and training, and "sought to avoid using test items showing large sex differences in percents passing." These test developers were aware that "intellect can be defined and measured in such a manner as to make either sex appear superior."

The Stanford-Binet is not the only test that developers have constructed to eliminate sex differences. CTB/McGraw-Hill staff were quoted as saying that "very little bias was found in the California Achievement Test, and those questions were balanced so that an equal number of items favored each sex." Similarly, researchers at the Educational Testing Service report that the SAT-Verbal reading passages were selected from different disciplines in specific percentages because of empirical evidence that this would help balance scores between the sexes.
Other research concludes that "while specific changes in the content of the SAT could not be associated with score changes, it seems likely that numerous small adjustments have played some part in recent shifts among the average test scores of women and men [to favor men]." Researchers found that, as a result of efforts to make the SAT-Verbal more sex-neutral, the "relative slight advantage [three to ten points] has been shifted from women to men." It is interesting to note that no efforts have been made to balance the SAT-Math, on which males outscore females by about one-half of a standard deviation, or about fifty points.

Tests are balanced or unbalanced by the selection of test items with different characteristics. Test items can differ in terms of

- that which is being tested (skill areas),
- the format of the item (such as essay or multiple choice),
- the item content and context (including the use of gender references and the selection of a reading comprehension passage on, say, child care or football).

The relative emphasis placed on different skill areas within a content area determines if a test will help minimize or maximize sex differences. While much of the work in this area has been related to mathematics, the conclusions can be generalized to other areas. In mathematics, girls outperform boys in computation, while boys outperform girls in some problem solving.
Your high school record is probably the best single predictor of how well you will do in college, but a combination of your high school grades and test [SAT] scores is an even better indicator."


when girls and boys were matched on overall skill areas, girls performed significantly better on mathematics test items that required arithmetic algebra than they did on items requiring arithmetic geometry. Girls' performance was also higher in areas of logic. For the 1987 SAT, where boys outperformed girls on almost all items, the differences were smaller in arithmetic and algebra questions than in geometry questions.

In general, if a mathematics test emphasizes computation, logic, and combined arithmetic-and-algebra skills, girls will do better. If the test emphasizes word problems and combined arithmetic-and-geometry skills, boys will do better. All of these areas are integral parts of mathematical knowledge, but the emphasis test developers place on one area or another area can increase or decrease gender differences.

As noted earlier, some gender differences persist in math and science course taking: boys are more apt to take physics and calculus. As a result, science and math tests that include larger numbers of physics or calculus items tend to favor boys.

On SAT-Verbal items, girls tend to perform better on test material that is general and abstract rather than specific and concrete. Girls also appear to do better when asked to deal with concepts and ideas rather than facts and "things."

The same skills can be tested in very different test contexts. For example, a student's ability to grasp analogies can be tested by having a question that refers to "guns:war" or "pans:dinner." The context in which skills are tested can skew test outcomes.

Generally boys outperform girls in applying mathe-
matics to measurement, sports, and science areas. Girls outperform boys in applying knowledge to areas of aesthetics, interpersonal relationships, and traditionally female tasks. However, girls did not do as well as boys on SAT word problems even when the problems related to food and cooking.

In verbal areas there is a consistent gender pattern in scores across the Graduate Record Examination (GRE) and the SAT. The context of most verbal items strongly favoring men relates to science or sports, while the context for verbal items strongly favoring women relates to aesthetic or philosophical areas or to relationships.

The amount of science content in reading-comprehension passages on the SAT-Verbal jumped from 20 percent before 1978 to 33 percent after. By 1980, two years after the increase in science content, reading sub-scores favoring boys had climbed from three to twelve points.

References to females and males appear less problematic, at least in one study of mathematics testing. Here, researchers found that "reference to male or female names, pronouns, possessions, or occupations in the place of neutral language had no demonstrable effect at all on examinee performance on mathematics word problems." However, students of both sexes were more apt to omit problems with unfamiliar content, and less likely to solve such problems correctly. Familiarity is correlated with stereotypical content for girls but not boys. Girls performed better on items that mirrored curriculum content, while boys tended to perform better on less routine, "real life" word problems.

College students' scores on word problems, however,
did not consistently reflect an item's sex typing as defined by researchers. For example, a question about the amount of stuffing in a chair might have been expected to favor girls, while another about the amount of oxygen in a tank might have been expected to favor boys. However, when students, not researchers, defined an item as "masculine" or "feminine," the sex typing of those items made a major difference in female and male performance. Items identified as favoring women involved "batter in a loaf pan" and "a secretary typing business letters," while items identified as favoring boys involved "timing a male runner," "buying uniforms for a baseball team," and "computing sports averages." 29

The proportion of different item types within a test can determine whether sex differences are found. Girls tend to score higher on essay or open-ended items, while boys tend to score higher on multiple-choice items. 30 There are also sex differences in the number of items completed, with girls completing fewer items. Seventh-grade girls taking the SAT and girls taking the National Assessment of Educational Progress (NAEP) science assessment are more apt than boys to respond "I don't know" and to fail to reach the end of the test. 31

Seventh-grade girls taking the SAT-Math fared better than boys on questions about the amount of information needed to solve a problem (data sufficiency items). Predictably, they found the multiple-choice items harder. No difference was found on those items that contained figures and/or diagrams. 32

Sex differences have been found within the different types of objective items. For example, the use of analogies slightly favors boys. 33 The use of antonyms slightly
favors girls. Girls and boys perform equally well on science inquiry items, where an understanding of process is key, but girls do less well when tested on specific science content.

The impact of affective factors is mixed. While girls report being more anxious about tests than do boys, girls' increased anxiety does not correlate with poorer test performance. Attitudes toward math do not have an impact on the SAT-Math gender gap. Sex differences in SAT-Math scores were found even among students who chose math as their favorite subject or who chose science first and math second. It also appears that attitudes toward science are not central in explaining gender differences in science achievement.

Perhaps most important, SAT scores, which are designed to predict college success as defined by first-year grades, underpredict women's grades and overpredict men's. Young women tend to receive higher college grades than young men with the same SAT scores.

The underprediction of women's college grades does not result from women taking easier courses. In math courses at all levels, grades of females and males are very similar, but male SAT-Math scores are higher than female scores. Attempts to control statistically for differences in courses taken usually diminish—but do not
eliminate—gender differences in over- and underprediction of grades based on SAT scores. When researchers compared the SAT-Math scores of college women and men who had earned the same grades in the same college courses, they found women's scores were thirty-five points lower than those of their male classmates. The authors concluded that there may be "evidence of unfairness to women on the SAT-Math; women score lower than men of comparable academic performance."  

In the American College Testing Program (ACT), girls score slightly higher than boys in English and lower in math. While girls earn slightly higher grades than boys in high school English courses, they earn about the same grades in math. Boys score higher than girls in math even when both have taken four years of high school math. One analysis found that female high school students had a much higher correlation between their SAT and ACT scores than do males. The SAT-Math score of a girl with an ACT math score of 20 is predicted to be 470, while the SAT-Math score of a boy with an ACT score of 20 is predicted to be 540.  

This has serious implications for girls and women. Test scores affect qualifications for scholarships and a variety of educational opportunities. If, in the above example, a score of 500 were the cut-off point for a scholarship, the male would be eligible while the female would not.  

Very little research has been conducted in the United States on bias in performance, or hands on, testing. Because research in other areas has revealed gender bias in observations, there is concern that in the new field of
performance testing the scorer's observation of the examinee's age, sex, race, accent, general appearance, and mode of dress creates the potential for biased scoring.46

In Great Britain, studies of science performance testing at the pre-college level found that some girls chose not to do test tasks related to electricity, expressing the belief (even when that was not the case) that they did not have the necessary knowledge.47 Some boys rejected test tasks with an obvious domestic content such as choosing the most suitable floor surface for a kitchen.48 Furthermore, the researchers found that in some areas girls posed, and then solved, their own problems rather than doing the test task exactly as phrased by the examiner. The girls' problems were frequently more complex and required sophisticated procedural strategies to solve. However, all too often such behavior was labeled as "off task" and "incorrect" and the girls were scored down.49

**Implications**

Research shows that one result of bias in testing is that test scores may provide an inaccurate picture of girls' and boys' abilities. Other factors, such as grades, portfolios of student work, extracurricular achievements, and out-of-school accomplishments, must be considered along with test scores when making judgments about girls' and boys' skills and abilities.