DOCUMENT RESUME

ED 424 337 UD 032 600

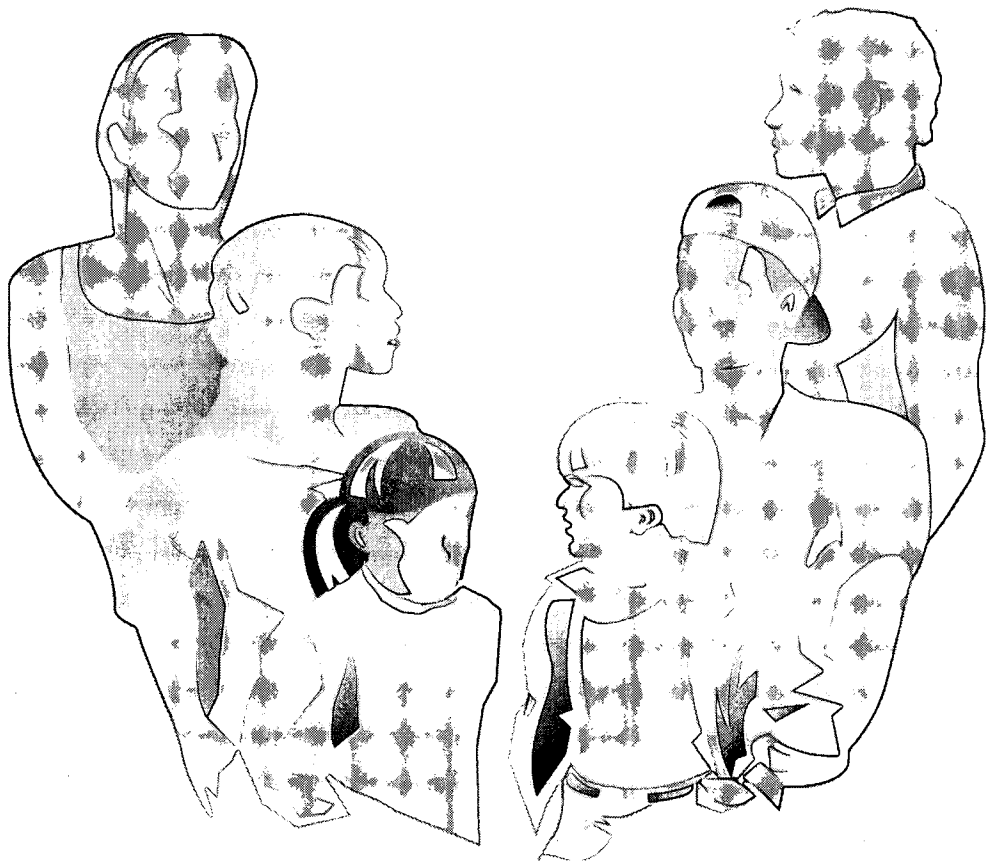
AUTHOR Cole, Nancy S.
TITLE The ETS Gender Study: How Females and Males Perform in Educational Settings.
INSTITUTION Educational Testing Service, Princeton, NJ.
PUB DATE 1997-05-00
NOTE 36p.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; *Age Differences; Elementary School Students; Elementary Secondary Education; *High School Students; Language Arts; Mathematics Education; Science Education; *Sex Differences; *Test Results
IDENTIFIERS Educational Testing Service

ABSTRACT

The Educational Testing Service (ETS) Gender Study is the result of 4 years of work by several researchers using data from more than 400 tests and other measures from more than 1,500 data sets involving millions of students. The study focuses on nationally representative samples that cut across grades (ages), academic subjects, and years in order to control factors that may have introduced confusion and contradictory results into previous studies of gender differences in educational settings. Results indicate that gender differences are not quite as expected. For nationally representative samples of 12th graders, the gender differences are quite small for most subjects, small to medium for a few subjects, and quite symmetrical for females and males. There is not a dominant picture of one gender excelling academically, and in fact, the average performance difference across all subjects is essentially zero. The familiar mathematics and science advantage for males was found to be quite small, significantly smaller than 30 years ago. However, a language advantage for females has remained largely unchanged over that time period. Also, gender differences for component skills of academic disciplines were often different than for the discipline as a whole. Gender differences were shown to change as students grew older and moved to higher grades. Patterns of gender differences in performance are similar to patterns of differences in interests and out-of-school activities, suggesting that a broad constellation of events relates to observed differences. Results show larger gender differences for self-selected groups taking high-stakes tests than for nationally representative samples, reflecting primarily the wider spread of male scores. Results indicate that neither guessing, speededness, nor the multiple-choice format per se accounts for the gender differences. However, results on presently used open-ended questions sometimes reflected no gender effect and sometimes reflected effects in which females' performances exceeded those of males and vice versa. Implications of these findings are discussed. A list of 67 resources for further reading is included. (Contains 8 figures and 19 endnotes.) (SLD)

The ETS Gender Study:

How Females and Males Perform in Educational Settings



ETS Educational Testing Service

THE ETS GENDER STUDY:

How Females and Males Perform in Educational Settings

Nancy S. Cole

Educational Testing Service

May 1997

Nancy S. Cole is president of Educational Testing Service and a principal author with Warren W. Willingham of *Gender and Fair Assessment,* the technical book on which this monograph is based.

Educational Testing Service
Princeton, NJ 08541-0001

EXECUTIVE SUMMARY

The ETS Gender Study is the result of four years of work by several researchers using data from more than 400 different tests and other measures from more than 1,500 data sets involving millions of students. It focuses on nationally representative samples that cut across grades (ages), academic subjects, and years in order to control factors that may have introduced confusion and contradictory results in previous studies.

Findings

The results of the study indicate that gender differences are not quite as people expect. For nationally representative samples of 12th graders, the gender differences are quite small for most subjects, are small to medium for a few subjects, and are quite symmetrical for females and males. There is not a dominant picture of one gender excelling over the other and, in fact, the average performance difference across all subjects is essentially zero. The familiar math and science advantage for males was found to be quite small, significantly smaller than 30 years ago. However, a language advantage for females has remained largely unchanged over that time period. Also, the gender differences for component skills of academic disciplines were often different than for the discipline as a whole.

Gender differences were shown to change as students grew older and moved to higher grades. The gender differences were very small at grade 4. Females increased their small lead on males in some language subjects from grade 4 to grade 8, and males registered small gains over females in math concepts and science from grade 8 to grade 12. The spread in scores was found to change over the grades as well. At grade 4 spread differences were very small but the spread of scores was larger for males than for females at grade 12, a result that especially affects differences in highly selected groups.

Patterns of gender differences in performance are similar to patterns of differences in interests and out-of-school activities, suggesting that a broad constellation of events relates to observed differences.

The results showed larger gender differences for self-selected groups taking high-stakes tests than for nationally representative samples, reflecting primarily the wider spread of male scores. For example, there are more males than females among high-scoring 12th graders in math and science and somewhat larger gender differences on math and science tests among college-going students than among high school seniors generally.

Results indicate that neither guessing, speededness, nor the multiple-choice format *per se* accounts for the gender differences. However, results on presently used open-ended questions sometimes produced no gender effect and sometimes produced effects on which females' performances exceeded that of males and sometimes vice versa.

The study also addressed the common use of the word "bias" associated with any finding of difference. The author notes that "bias" implies a systematic error in measuring knowledge and skill. This study indicates that observed differences are not an error but a correct reflection of differences that occur on many different types of measures, in many different samples of students. The author concludes that content on tests must be guided by how educationally important the content is, not what differences it produces.

Implications

These results indicate the wide breadth of relevant and valuable skills students have and need to have. We believe both females and males need a broader set of skills today to have access to the full range of educational and career options. Even with progress closing some gender gaps, both genders are failing to develop some of the desirable skills necessary for some career options in tomorrow's changing world. A major implication of this study is to call renewed attention to the need for students of both genders to learn a breadth of skills.

Research shows that females have closed the gap significantly on math and science scores, but males continue to lag behind in writing and some language skills. We should not ignore the differences that exist in either direction, and we need serious attention by parents and educators to teaching and measuring the breadth of skills for both genders.

The most significant thing we have learned from studying performance of groups is the importance of considering each student as an individual without stereotypes. The massive overlap in performance between the genders reinforces the most fundamental result of all — that group membership is far less important in performance in educational settings than individual characteristics.

THE ETS GENDER STUDY:

How Females and Males Perform in Educational Settings*

Why are girls more likely to keep a journal and boys more likely to take a radio apart? Why do girls earn, on average, higher grades in school than boys? Why do young women and young men generally choose to major in different academic disciplines in college? The similarities and differences between girls and boys, men and women, continually intrigue and perplex us.

Test performance reflects the kinds of differences noted above, and test makers need to understand differences and how to respond to them in their tests. Yet many key results on differences are confusing or contradictory. We cannot determine if there is a problem or how to address it without clearer knowledge of what the results actually are.

Educational Testing Service (ETS) has completed an extensive four-year study of the similarities and differences in test performance and other forms of academic achievement of females and males. We had two objectives in undertaking this gender study:

- to improve our understanding of the patterns of gender difference and similarity in academic performance
- to examine the implications of such understanding for current and future educational assessments

ETS was in a unique position to bring a key new source of information — information that has been available but not thoroughly examined — to the understanding of gender differences. That information comes from large-scale, nationally representative sets of data and other large data sets on well-known, self-selected samples. With such data, we hoped to bring a new clarity to the picture of gender differences.

Background on Gender and Fair Assessment

In the past quarter century, we have witnessed many important changes in the participation of women and men in American society. According to the National Center for Education Statistics, women and men are now equally likely to complete high school, whereas prior to 1970 women were more likely to graduate. In 1990, women earned 53% of all bachelor's degrees conferred, 52% of master's degrees, but only 37% of doctoral degrees. Although women

* This short monograph provides highlight results about how females and males perform in educational settings from a large gender study conducted by ETS researchers over the past four years. It draws on a broader and more technical study by Warren W. Willingham and Nancy S. Cole, with contributions by several other researchers, to be published in book form by Lawrence Erlbaum.

were still less likely to earn professional degrees than men, there was a dramatic increase in the number of women earning professional degrees in the 20 years between 1970 and 1990. Women earned 30.6% of all dental degrees, 34% of medical degrees, and 42% of law degrees awarded in 1990, as compared to less than 1% of the dental degrees, 8.4% of the medical degrees, and 5.4% of the law degrees only 20 years earlier.[1] It is natural to wonder about the underlying changes that are occurring in the educational skills of females and males to support or limit changes such as these.

In the past few decades, research on gender differences has proliferated. A notable event was Maccoby and Jacklin's 1974 work, *The Psychology of Sex Differences*.[2] Their analyses, based on some 1,600 studies in eight areas of achievement, personality, and social relationships, led Maccoby and Jacklin to several conclusions. They noted "unfounded beliefs" — that girls are more social and suggestible but have less self-esteem and motivation for achievement. They noted some "open questions" such as which gender is more competitive or compliant. Their four main conclusions regarding "sex differences that are fairly well established" were that:

- Girls have greater verbal ability
- Boys excel in visual-spatial ability
- Boys excel in mathematics
- Males are more aggressive

These conclusions have since been qualified in various ways by succeeding research.[3] The essential role of tests to both fairly assess and accurately reflect performance has brought testing closer to the center of work on gender similarities and differences, and there is much new research available on the test performance of males and females. However, there have been inconsistencies in the findings, requiring a closer look. For example, some researchers have contended that there are no longer any gender differences in verbal ability. Yet others have continued to find that females tend to perform better on writing assessments than males.[4]

With the turn of the century approaching, there is national concern about the effect of inadequate and inequitable learning opportunities on our nation's ability to compete effectively in an international economy. Concern that we set high and rigorous standards for what students should learn leads to issues about how to measure whether students have met those standards. Testing is more prominent than ever in policy initiatives to improve education. This prominence was illustrated again by President Clinton's call, in the 1997 State of the Union address, for rigorous national standards and national tests in reading and mathematics to monitor the progress of *all* children.

Making high-quality education available so that all students have the opportunity to meet high and rigorous standards is a vital national goal.

However, accomplishing that goal requires attention to the diversity of individual youngsters with their own special experiences, talents, and skills. We have much to learn about how to use that rich individual diversity to pursue common high standards. To make progress, we need to understand the kinds of experiences — for individuals and groups — that foster high-level achievement as well as those that impede it.

The swift pace of technological innovations and change in the international marketplace is giving rise to a new American economy. More and more jobs require a broad range of high-level skills, and many jobs require rather different skills than jobs of the past. For example, technological skills are increasingly playing a key role in the work force; math and science are more important for more jobs than ever before. Language skills play an increasingly important role in a service economy, and employers regularly complain that the youngsters they employ cannot read, write, or speak adequately.[5]

In this period of change in job requirements, there is a national concern about the effectiveness of education generally and particular concern that all our students have the knowledge and skills they need to meet these new demands. It is vital to our well-being as a society that we shape the learning experiences of all youngsters to prepare them for a wide range of future job opportunities and career options. The traditions of past gender differences raise the possibilities that we might fail to recognize the limits we could be putting on boys or girls if we fail to attend to and counter differences through actions as parents or educators.

Yet, we recognize that data on gender differences can be seen as a double-edged sword. Objective evidence of knowledge and skill can cut through myths as to the careers and social roles to which women and men are well-suited. In another guise the same evidence may risk reinforcing stereotypes. Research methods also tend to emphasize difference rather than similarity. For parents, educators, and policymakers, the challenge is to gain a clearer understanding of the similarities and differences to better ensure that we are preparing all our children for the wide range of opportunities they will encounter in the future.[6] Hence, we see studying gender differences as unavoidable.

Design of the ETS Gender Study

There are four major features of the study's design. First, we attended to key factors that need to be better controlled — the particular skills measured, the comparability of samples, and the differences for different populations. Second, we studied a wide breadth of data and multiple measures to understand general findings and to look at gender differences for particular skills. Third, we used representative samples of different populations (e.g., different ages, different decades) to control for possible sample differences and to address changes in differences over age or time. Finally, we introduce the

measure used to compare differences across different tests, the standard mean difference D.

Factors Needing Special Control. The design was driven by the need to understand and control three main sources of potential misunderstanding and confusion. They are:

- The nature of particular skills (the construct). Various tests, even with the same name, differ in the content of the test questions and hence the particular skills on which they most focus. Our goal was to attend more closely to the particular skills measured (the constructs) in order to better understand the results.

- The comparability of the female and male samples (the samples). Many studies in the literature were unable to match carefully the males and females studied. For example, if only volunteer samples were available to study, it was quite possible for the males studied to differ in significant ways from the females, introducing "noise" into the results. We focused on ensuring that the samples of females and males are comparable.

- Differences in different populations (the cohort). Results are available on youngsters of various ages and from different decades. If gender differences are not the same for some of these different populations then considerable confusion could be introduced by not taking this cohort factor into account.

Breadth of the data. It was essential to cast a wide net if we were to address the construct issues by considering a broad range of types of measures. We drew on information from over 400 different tests and other measures and more than 1,500 data sets. This broad array of data allowed us to analyze gender similarity and difference in multiple subject areas as measured by different types of tests for a much closer look at the particular skills (constructs). For example, we could look for math tests that emphasized reasoning and contrast them with tests that emphasized computation. Similarly, we could explore a variety of verbal skills — writing, language use, reading, and verbal reasoning.

Use of Large and Representative Samples. Of critical importance to the study was the decision to use nationally representative samples of students or samples that were large and widely known. Such samples come from large-scale testing programs (commercial testing programs or state-linked programs), from large federal studies, and from tests used for admission to college (e.g., ACT, SAT) or graduate study. They cover ages from grades 4 through graduate school. Such data is especially critical to the control of samples and the consideration of cohorts.

FIGURE 1 provides a framework of the data used. The first three columns are for large-scale surveys and test batteries used with nationally

representive samples of the general population at grades 4, 8, and 12. The fourth and fifth columns delineate the high-stakes testing programs used in the undergraduate and graduate admissions process and show a link of the two sets through the PSAT/NMSQT given to a national group in the norming study as well as to self-selected groups and also through the ITED and ACT.

Measuring Differences — The Statistic D. The study uses data from hundreds of different tests with a variety of score scales and a variety of samples. We needed to compare gender differences on the five-point scale of the Advanced Placement examinations with differences on the 200-800 scale of the SAT, and with differences on the 1-32 scale of the ACT. To do so we had to have some type of standard index that would give us meaningful compari-

Figure 1

Sources of Data for Nationally Representative and Self-selected Samples

Nationally Representative Samples			Self-selected Samples	
			College Applicants	Grad./Prof. Applicants
Grade 4	**Grade 8**	**Grade 12**		
ITBS	ITBS	TAP		
Stanf	Stanf	TASK		
NAEPr	NAEPr	NAEPr		
NAEPt	NAEPt	NAEPt		
IAEP	IAEP			
	DAT	DAT		
	NELS	NELS		
		HS&B		
		NLS		
		ASVAB		
		NALS		
		PSAT Norms	PSAT/NMSQT	
		ITED	ACT	
			AP	
			SAT	GRE-G
			ATP	GRE-S
				MCAT
				GMAT
				LSAT

sons. The statistic D, the standard mean difference, is the standard index used in the research literature and the primary index we used to compare the size of female-male differences across various test scales.[7]

If there is no gender difference, D is zero. If females have a higher average score, D is positive, and if men have a higher average score, D is negative. Generally, a D value smaller than .20 is considered very small; we typically treat Ds of this size as insignificant. A D value between .2 and .5 is still considered small but worth noting nonetheless. Ds from .5 to .8 are considered medium in size and above .8 is considered large.

To assist in understanding the size and importance of values of D, FIGURE 2 depicts hypothetical data for which the Ds are quite small (D = .20) and for a larger D of .50, though one still only considered of small-to-medium size. Another way to describe the difference is by the proportion of the variation in test scores that is accounted for by the mean differences. For a D of .20, only 1 percent of the variation is accounted for by the mean difference, as indicated by the almost complete overlap of the two distributions. For a D of .50, this translates to 5.9 percent of the variation accounted for by the mean difference, still indicating substantial overlap of the distributions as shown.

Results on Gender Similarities and Differences

Our most common result was that gender differences in performance in educational settings are different from what many people expect. This finding is a theme of the several categories of results noted below.

Real Similarities and Real Differences

There is a cluster of results about similarities and differences:

> ***Result 1.*** For many subjects, the differences are quite small — smaller than people realize.
>
> ***Result 2.*** However, there are some real differences on some subjects.
>
> ***Result 3.*** The results contradict the view that the problem of gender is that the girls need to catch up with the boys. We found that the differences cut both ways and that 12th-grade girls have substantially closed the familiar math and science gap over the past 30 years, but there continues to be a fairly large gap in writing skills that boys have not closed.

Figure 2

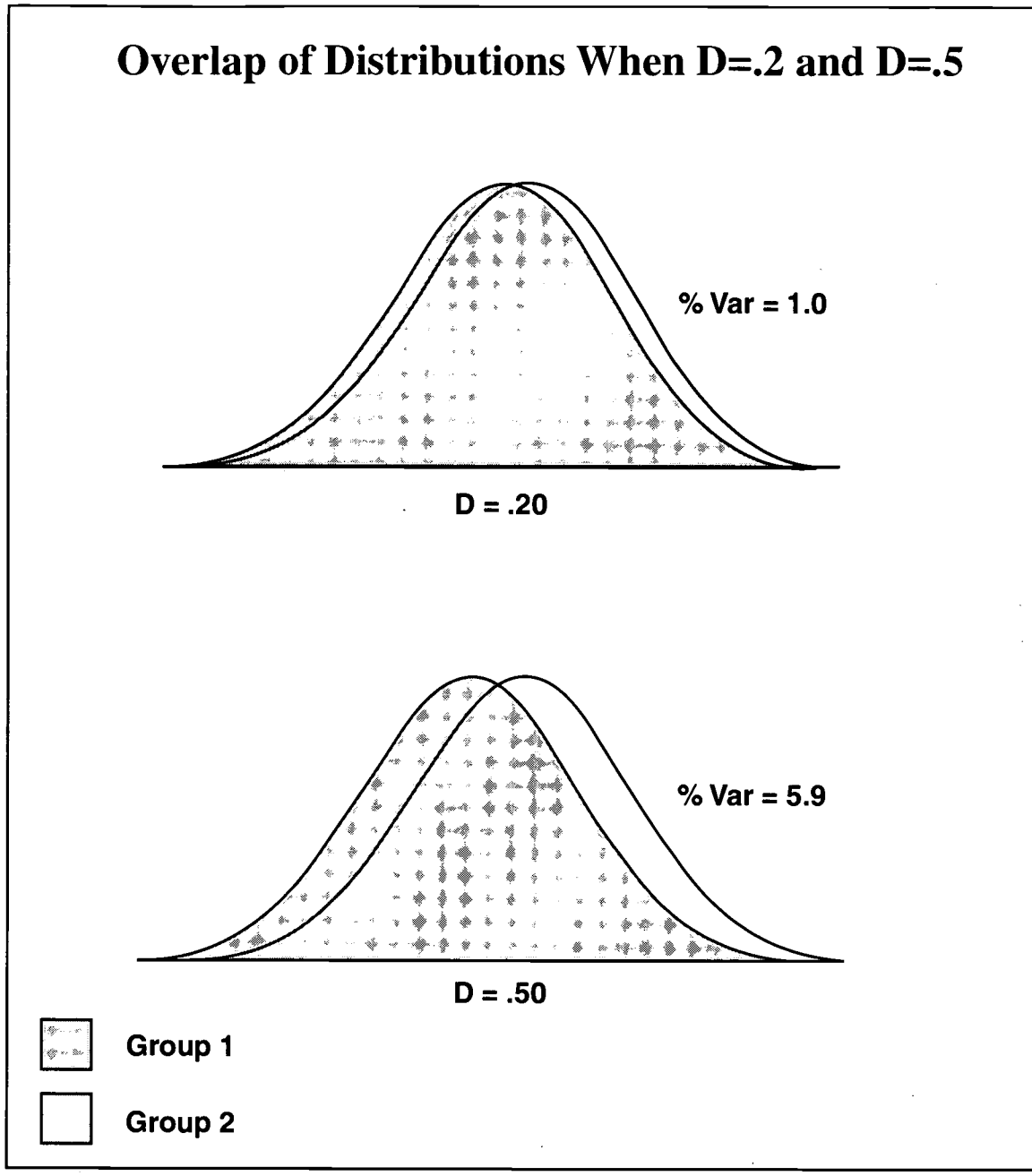

Overlap of Distributions When D=.2 and D=.5
% Var = 1.0
D = .20
% Var = 5.9
D = .50
Group 1
Group 2

Figure 3

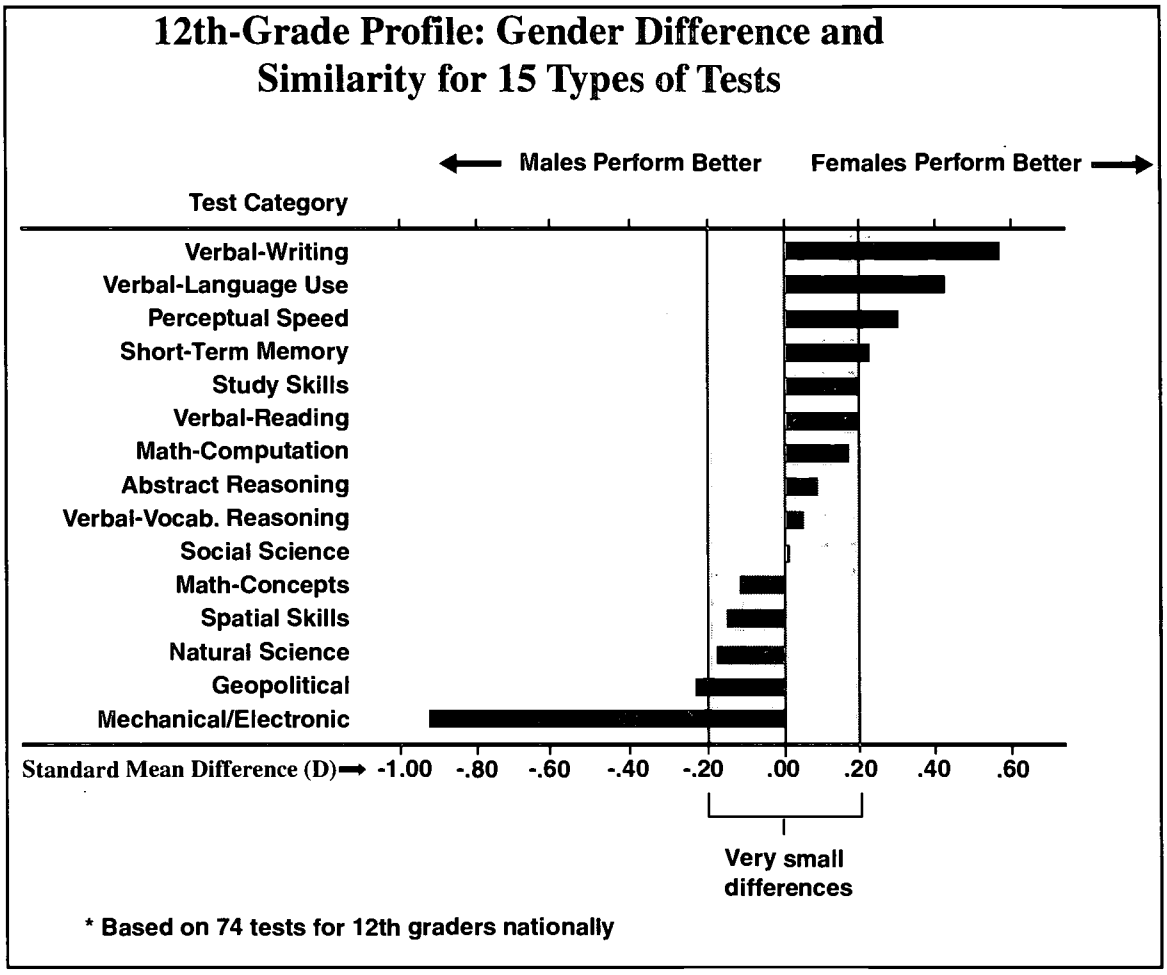


FIGURE 3 provides a profile of 12th-grade students that summarizes the results found from 74 different tests in 15 different subject categories from nationally representative samples. This summary profile of a very large amount of data reveals several key results that support the findings noted above. The results are for 15 categories of tests ranging from verbal-writing at the top to mechanical/electronic at the bottom. The subjects are ordered from those for which females score higher to those for which males score higher.

The first prominent result (Result 1 above) comes from results in this "very small" zone of D from -.2 to +.2. For nine of the 15 test categories — study skills, verbal-reading, math-computation, abstract reasoning, verbal-vocabulary reasoning, social science, math concepts, spatial skills, and natural science — the results are in this zone of very small differences. This zone, for 12th graders nationally, includes the two math categories as well as natural science. So for many important subject categories the male-female differences are quite small, likely smaller than many people realize. (Refer to

Figure 2 above to recall the huge overlap of the distributions that differences this small indicate.)

Result 2 is demonstrated by the bars that reach further to the right and left. These bars indicate there are some real gender differences. For verbal-writing and mechanical/electronic, the bars reach the level of "medium" or "large" differences. Verbal-language use, perceptual speed, and short-term memory are categories on which females perform better than males although the differences would be characterized as "small." Geopolitical subjects (economics, history, geography) show "small" differences, with males scoring higher.

This profile indicates that the expectation is wrong that girls are the ones falling behind, as indicated in Result 3. In fact, the profile in Figure 3 is quite symmetrical, and the average D over all 74 tests in all 15 test categories is very close to zero. Further, the differences that do exist cut both ways — some show higher female performance and some show higher male performance.

FIGURE 4 provides supplementary information on the issue of "catching up." This figure reports gender differences in three subjects (science, mathematics, and writing) from 1960 to 1990.[8] These data show gender difference D in science being reduced from about -.60 to under -.20 from 1960 to 1990, with mathematics showing a similar reduction from -.45 to almost -.10 over the same time. However, females sustained the writing advantage they had from 1960 to 1990, the Ds staying close to .40 for both years.

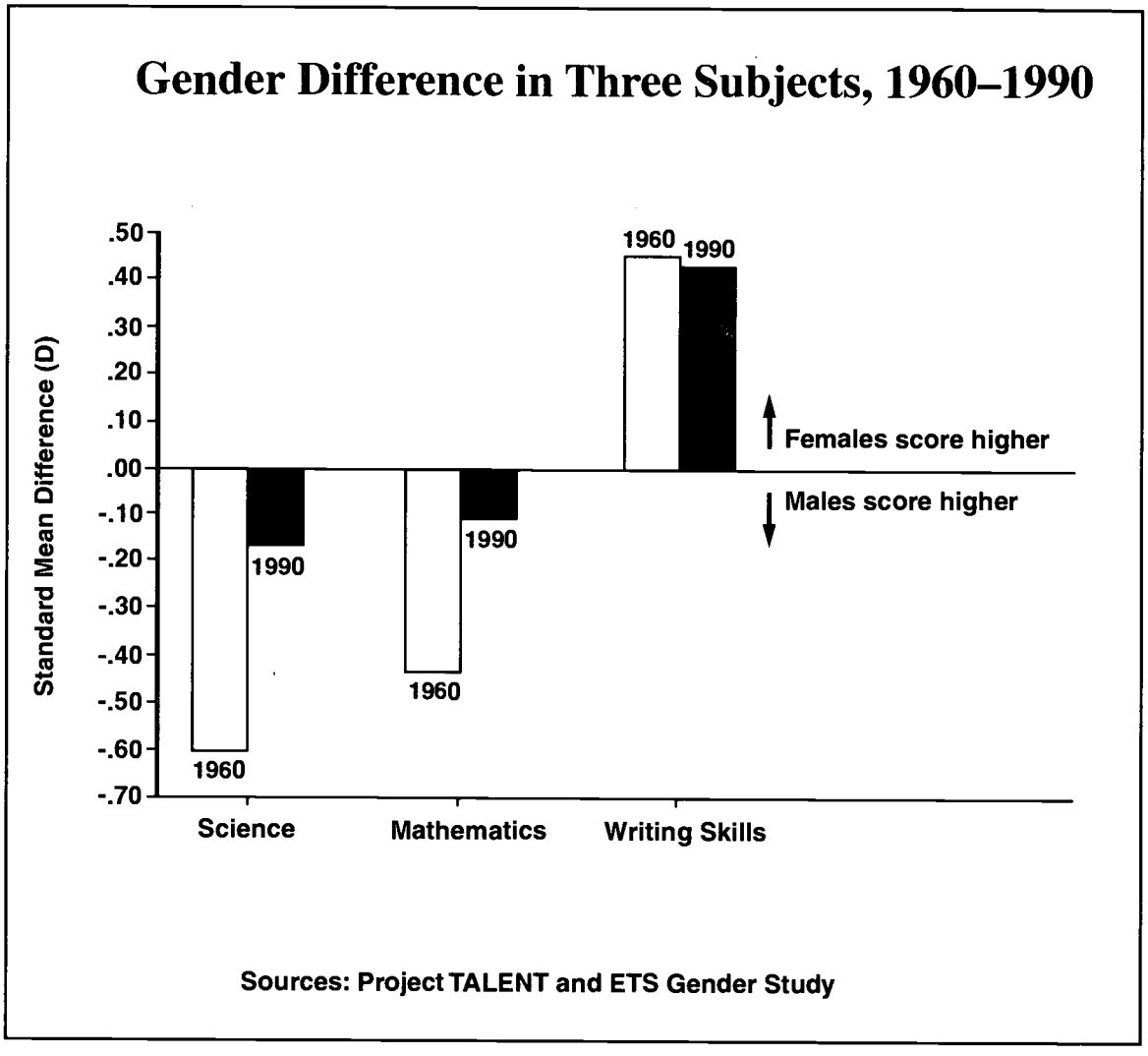
Differences within Subject Matter

Many discussions of gender treat academic subjects as uniform — if one gender is better in the subject, it is presumed that the gender is better in all aspects of the subject. This is not what we found.

> ***Result 4.*** When you break the academic disciplines into component skills, a different picture of gender differences emerges. For example, some subskills within math are stronger for females and others for males. Similarly, females are not better in all aspects of language skills.

The profile of 12th graders (Figure 3) demonstrates that the results for broad subject areas are not uniform. When we examine skills within a broad subject, the gender differences vary quite a lot. Consider, for example, the four categories beginning with the word "Verbal" shown in Figure 3. The results vary from noticeable differences favoring females for writing and language use to very small differences for reading and vocabulary reasoning. A similar difference exists for the two math categories, math computation and math concepts. Although the results for both are in the very small zone, females outperform males on computation, and males outperform females on concepts.

Figure 4



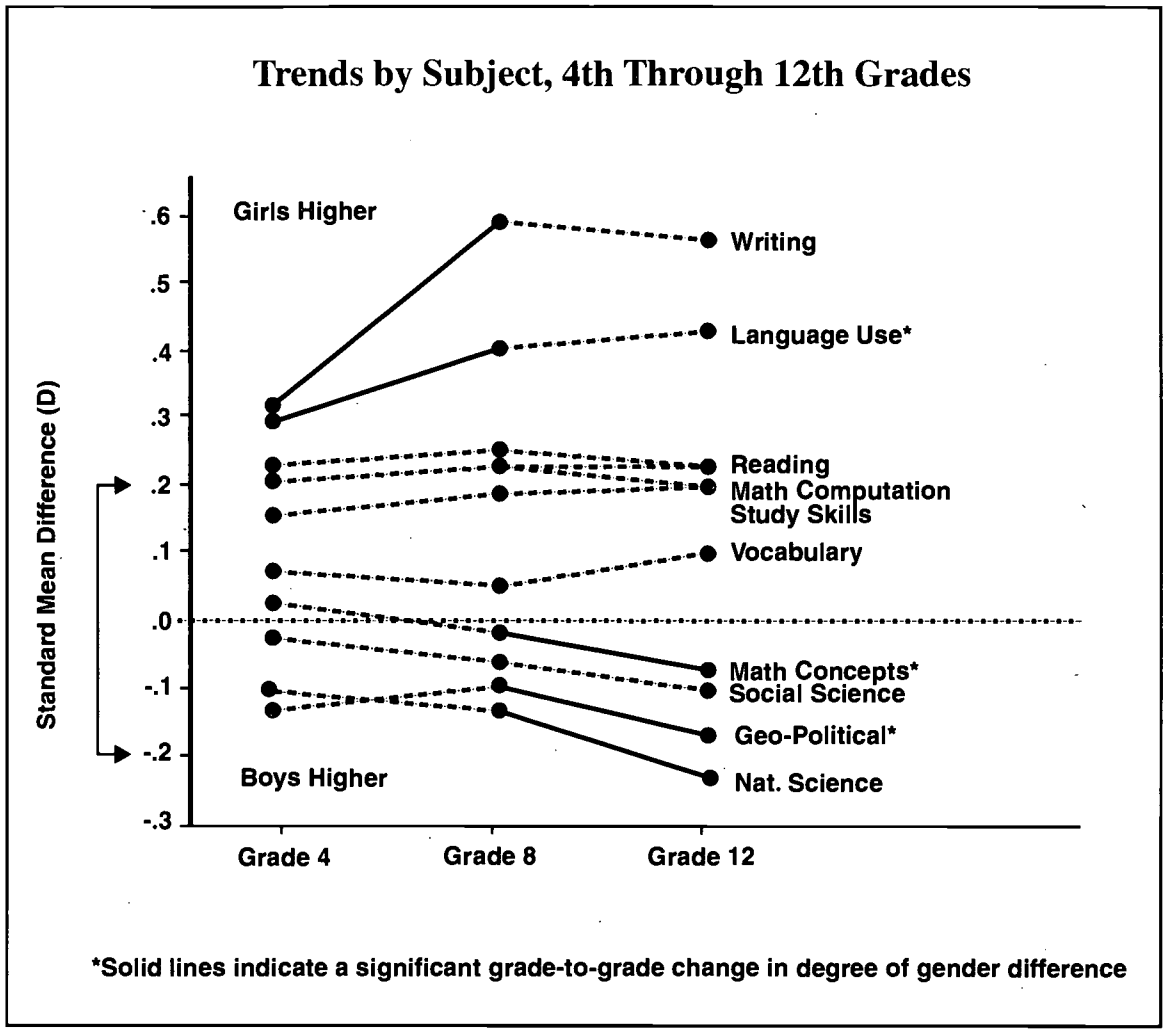
Patterns from Grade 4 to Grade 12

We analyzed important data to examine what happens to gender differences as students become older, recognizing that many people assume that differences are fixed at birth and stay unchanged over time. We found a different picture.

> ***Result 5***. Gender differences grow over the years in school. At 4th grade, there are only minor differences in test performance on a range of school subjects. Larger differences do not occur until later and then at different times for different subjects.

FIGURE 5 provides trends by subject from the 4th through the 12th grade on nationally representative samples to address the issue of changes in gender differences as students grow through the school years. In Figure 5, the D is plotted on the left vertical axis, and the lines show the trends over the three grades. These data are from the subset of data used in the 12th-grade

Figure 5



profile, for which there were representative samples at the other two grades on comparable subjects.

Three aspects of the results in Figure 5 are particularly notable. First, the gender differences are quite small at grade 4. Note that most Ds fall between -.2 and +.2 (only writing, language use, and reading have D values at grade 4 slightly above the .20 level). Second, differences increase after grade 4 as indicated by the spreading upward and downward of the trend lines. Third, the spread occurs at different times for different subjects. The subjects for which the trends are significant (for which one grade is significantly higher or lower than the preceding grade) are shown as solid lines. Thus, females significantly increase their performance advantage over males in writing and language use from the 4th to the 8th grade, whereas males increase their performance advantage over females from grade 8 to grade 12 in math concepts, geopolitical subjects, and natural science.

Relation of Performance Differences to Other Variables

Although people look for simple explanations of gender differences that imply simple "fixes" for those differences, the patterns we found in a variety of measures suggest that the differences and changing them involve complexities.

> ***Result 6.*** Gender differences are not easily explained by single variables such as course-taking patterns or types of tests. They not only occur before course-taking patterns begin to differ and across a wide variety of tests and other measures, but they are also reflected in different interests and out-of-school activities, suggesting a complex story of how gender differences emerge.

Figure 3 indicated the ordering of test performance differences in the 12th-grade profile ranging from those on which females scored higher, such as writing and language, to areas on which males scored higher, such as geopolitical subjects and mechanical/electronic areas. Aspects of this test performance ordering have parallels in patterns of interests. For example, in interest areas most related to school coursework and activities, females score higher on scales that involve the arts, writing, and social service, while males score higher on mechanical areas, athletics, and science.[9]

FIGURE 6 gives data on contrasting activities, awards won, and educational choices. For example, females report leisure activities in art, music, and drama, whereas males report leisure activities in sports and computers. When asked "Have you ever tried to ...?", grade 11 girls responded "yes" more frequently to figuring out what was wrong with an unhealthy plant or animal, whereas grade 11 boys responded "yes" more frequently to fixing something mechanical or electrical. Different experiences of girls and boys are also reflected in the areas in which they excel — girls in writing, leadership, and arts; boys in science and sports.

Further indications of the differences that arise out of the complex of performance and interests come from differences in selection of a college major field of study. Figure 6 indicates large differences in the ratio of females to males across academic fields, in patterns similar to others noted here.

Spread of Male and Female Scores

An important result, although one difficult to understand, concerns the greater spread of male score distributions. This is not a new finding; others have reported it before, but we replicated this finding in data set after data set.[10]

> ***Result 7.*** The spread of scores of males tends to be larger than the spread for females. This means that there are more males among the very highest scorers and also more males among the very lowest scorers.

Figure 6

Differences in Activities, Awards, and Educational Choices (Female/Male Ratio)

	Higher for Females	Higher for Males
Reported Leisure Activities[a]	(1.60) Taking classes (music, art, language, dance)	(0.37) Participation in non-school sports
	(1.22) Religious activities	(0.51) Taking sports lessons
	(1.19) Talking/doing things with parents	(0.70) Using personal computers
Answered "Yes": Have you ever tried to...?[b]	(1.63) Figure out what was wrong with an unhealthy plant?	(0.20) Fix something mechanical
	(1.19) Figure out what was wrong with an unhealthy animal?	(0.17) Fix something electrical
Won High School Award in[a]	(1.45) Writing (1.39) Leadership (1.34) Arts	(0.51) Science (0.42) Sports
Intended College Major[c]	(4.26) Psych./Sociology (3.00) Education (2.33) Health Services (2.23) Languages	(0.23) Engineering (0.39) Math/Computer Sci. (0.56) Architecture (0.59) Physical Science

[a]National Education Longitudinal Study, 1992
[b]NAEP Science Report Card, 1986
[c]*College Bound Seniors, 1996.* College Board

FIGURE 7 shows the common difference in spread of scores found in the high and low ends for nationally representative samples of 12th graders. Below the 10th percentile and above the 90th percentile, there are about 4 females for every 5 males. We see this low-end result perhaps in the presence of more males in some special education classes. We see the high-end result in the greater number of males in certain highest performing categories. The high-end result is especially important for self-selected groups of students, such as those taking high-stakes tests. These groups come from the high end of the distribution and, all other things being equal, we can expect more males than females among such groups and higher average scores for males than for females among such groups.

For example, in national 12th-grade samples, males outnumber females in the top 10 percent on math tests by 1.5 to 1 and in science by 2 to 1. Similarly, as one moves from national samples to self-selected samples, D tends to become more negative by about .20 in both math and science. So our results indicate that females still have some distance to go to achieve equal representation in the top ranks, but that does not alter the quite favorable picture of female achievement overall.

Although these differences in spread are consequential for high-end groups at grade 12, it is important to note two other findings. First, the spread of the distributions for females and males was closest at the 4th grade, with the spread of male scores only very slightly greater; the spread increased to grade 8 and grade 12. Second, the differences in the gender distributions produced by the differences in spread are dwarfed by the large amount of overlap in male and female distributions, as can be seen in Figure 7.

Grades and Test Performance

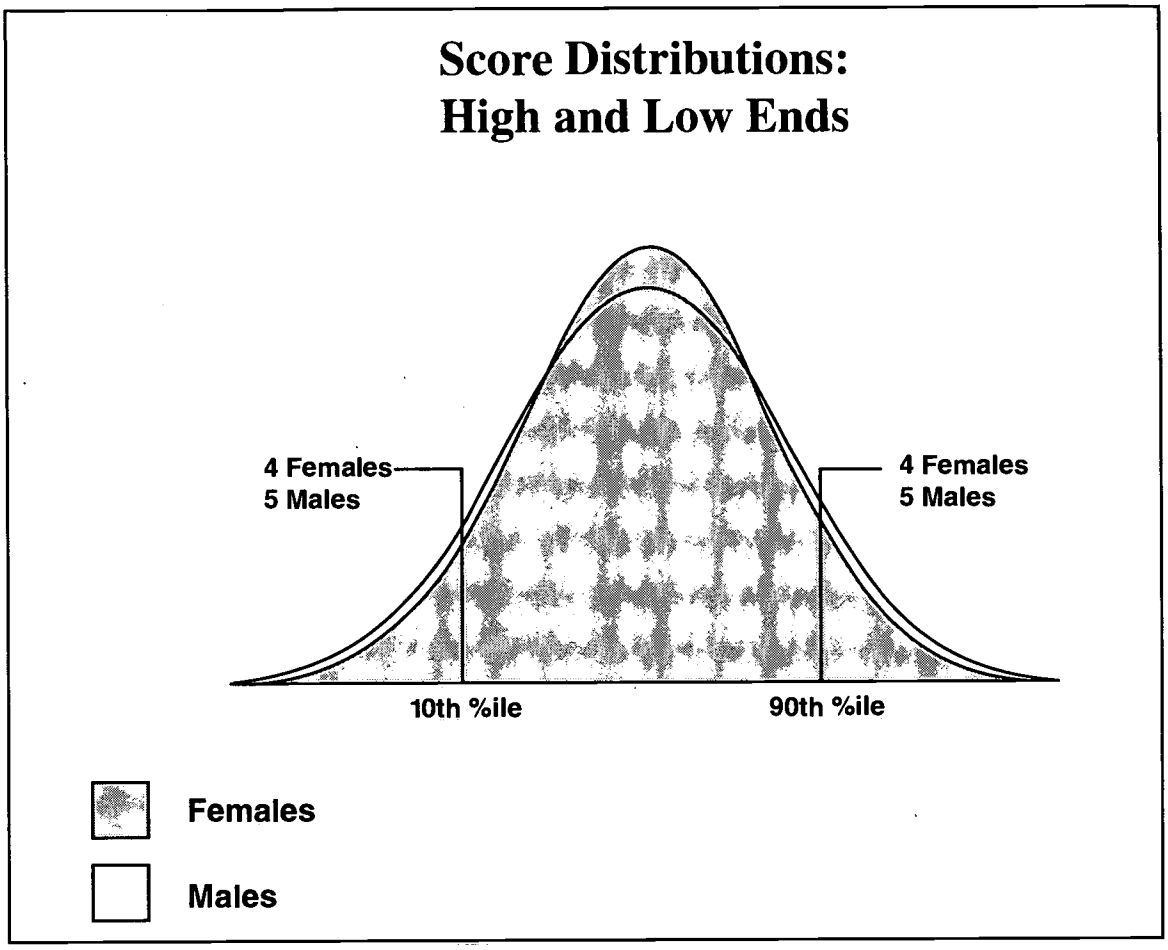
The difference in results between grades and tests fascinates many people and is not well understood. We found some results that relate to this interesting subject.

> ***Result 8.*** Females make, on average, higher grades than males on all major subjects, which contrasts with the symmetry reported in test performance. Tests measure particular, isolated skills; grades measure broad and less well-defined, but important, skills. Tests and grades often complement each other. Neither is biased; both are valuable measures.

We found (as have others before us) that females consistently make better grades on average in all major subjects. Female grades exceed male grades most in English, followed by smaller differences in social studies, science, and math. Our analyses suggest real differences (as well as overlap) in what grades and tests measure.

Tests measure particular skills at particular points in time (on a single day). Grades measure a much wider array of skills, some of which may not

Figure 7



even be well enumerated, and performance over a time frame of perhaps some months. Some people disparage grades as subjective and unreliable and favoring students who are "nice" and "compliant." Given that grades have consistently been found for decades to be one of the best predictors of academic performance after high school, we seriously doubt the appropriateness of the disparagement. In fact, we view grades as likely measuring a constellation of desirable characteristics that we call "studenting" skills — skills that are especially valuable in school or in work. These skills may include characteristics such as persistence, follow-through, doing required work, participating, and performing in different contexts (homework, class participation, teacher tests, etc.).

Tests and grades have proven both to be valuable and often complementary measures. Years of results in predicting college grades have, for example, shown that grades are most often the single best predictor and tests follow a close second. Also, tests have consistently been shown to add to the prediction of college performance beyond that accomplished by grades alone.

Analyses of gender effects of the two predictors reveal that tests and grades work somewhat differently although the effects are typically quite

small. For example, were the SAT used alone, it would slightly underpredict the overall grade-point average of first-year female college students, but when the SAT and high school average are used together, more accurate predictions are produced overall as well as very little gender difference. Specifically, when both measures are used to predict a first-year college GPA that is comparable for females and males, the actual GPA of the women is three-hundredths of a grade point higher than predicted — about as close as one might expect to get.[11]

One subject, calculus, has yielded larger differences than were found for GPA or most other subjects examined. Earlier results had indicated underprediction of college calculus grades when the SAT was used alone.[12] To add to understanding of this result, we found that, like for the smaller GPA differences, adding high school grades corrected the underprediction. In fact, using grades alone would have resulted in underprediction of calculus grades for males in those cases. FIGURE 8 provides the results for calculus considering both grades and tests in the form of the original study.[13]

Figure 8

Gender Difference Among Students Who Earned the Same College Calculus Grade

College Calculus Grade Earned	Mean Score for Females Minus Mean Score for Males		
	SAT-M	HSGPA	Composite (SAT-M + HSGPA)
A	-21	23	2
B	-28	24	-2
C	-29	21	-5
D	-33	31	-1
F	-35	29	-4

All entries are expressed as points on the SAT scale.
Source: Bridgeman & Lewis, 1995

People are quick to reference findings of difference on grades or tests as bias. It is important to recognize that the word "bias" refers to consistent or systematic errors in measuring student skills or accomplishments. Since grades and tests measure skill constellations for which much evidence indicates there are some real differences, they should not be labeled as biased. Grades and tests correctly measure partly different and partly overlapping skills. Both give important information of slightly different types and should be used to complement each other when it is practically feasible to do so.

Results on Gender and Testing

The study was directed to several key questions about gender and testing.

Self-Selected Samples on High-Stakes Tests

We wondered why the gender differences are greater for self-selected groups on high-stakes tests.

> ***Result 9***. We found that differences in self-selected samples on high-stakes tests tended to fall in the direction of higher male performance when compared to results from nationally representative samples. Further, we found the fact of greater spread in male distributions was a dominant factor in this shift.

As can be seen in Figure 7, the greater spread found for males in nationally representative samples results in there being more males with higher scores. Considering highly selected groups, such as those self-selecting to take high-stakes tests, is akin to looking at a right-hand portion of the distribution in Figure 7. That portion may be about half of the distribution for some high-stakes tests or a much more extreme portion (maybe only 10 percent) for other tests. From Figure 7, it is apparent that if there were no gender difference in test performance in the nationally representative group, there would nonetheless be gender difference (favoring males) in the selected group.

This result is further complicated when some gender difference exists in the representative group. If that existing gender difference is one for which males score higher than females on average, then the joint effect of the spread and that difference is to greatly magnify the male performance advantage in the self-selected group. If the original gender difference favors females, the spread effect may greatly mute the higher female performance and may even show male performance advantages for sufficiently extreme groups.[14]

A second, though less dominant, factor in the difference between the results for high-stakes tests and national samples on regular school-based tests is that the skills within subjects in high-stakes tests may, in some instances (such as in math tests for college admissions), focus on skills on which males show higher performance (such as reasoning and concepts).[15]

Each of those content decisions must be judged in its own right, but we note our belief that content is appropriately set on the basis of the educational importance of the content. If reasoning, for example, is critical for college work, then that justifies the decision to include it, even if it leads to gender differences.

A third, though also less dominant, factor is that some skills on which females excel, though important, have either been overlooked or have been difficult and expensive to measure (such as writing). The increasing inclusion of writing in high-stakes tests in recent years means that this factor will not be operating in the future as in the past.

Our analyses show that the impact of these three factors seems to account quite well for the observed differences between gender differences in representative and self-selected samples. Although people are quick to point to results described here as a sign of "bias" in high-stakes tests, it is clear that they are predicted from, and the result of, characteristics of the nationally representative samples. In this sense they are not surprising or an indication of bias but are expected and follow from the results in representative samples.

Guessing and Speededness

> ***Result 10***. We did not find evidence to support the supposition that different guessing habits and different responses to the fact of time limits on tests affect female and male scores differently.

We reviewed previous studies on this topic by ETS researchers and by other researchers. The evidence indicated that whatever gender differences were observed, manipulation of speededness (e.g., adding more time) did not alter the original gender difference, nor did testing students under conditions where guessing played less of a role.[16]

Gender Effects of Multiple-Choice Questions

> ***Result 11***. We found that asking students to produce the correct short answer rather than choose the correct short answer on otherwise similar questions does not affect gender differences.

Many people suppose that the multiple-choice questions favor males. Studies that addressed this issue controlled for the nature of the questions being asked by keeping the questions the same across conditions in which the student was asked to produce or select an answer. In these circumstances, in which the only variable was whether the answer was produced (open ended) or selected (multiple choice), the gender differences were unaffected.[17]

Gender Effects of Open-Ended Questions

The results above apply to open-ended as well as multiple-choice questions when the nature of the question and the general nature of the answer (short answers) is controlled. However, in practice in the real world, open-ended questions are typically used not to duplicate the multiple-choice questions but to gain additional information about student skills. So in use, open-ended questions do not keep the question the same and provide considerable latitude for the nature of the answer. Thus, we looked at performance on open-ended questions in wide use today from Advanced Placement tests of the College Board compared to performance on the multiple-choice section of the same AP subject to get a sense of gender differences.

> ***Result 12.*** When comparing the gender results for the types of open-ended tests in use today, we found mixed results.

For such tests, it seems that about half the time an open-ended test produces the same pattern of gender differences as does the counterpart multiple-choice test of the same subject.[18] When gender differences did appear, they cut both ways. The only consistency noted was that the differences tended to favor females if the response was written and tended to favor males if the response was to produce a figure or part of a figure to explain or interpret information.[19]

Isn't Gender Difference a Sign of Bias?

In this study, we addressed the commonly asked question noted in the heading. Answering the question is not a matter of referring to a specific set of data or a particular analysis. It requires the consolidation of information and logical as well as data analyses. Our answer to this common question is a clear "No." The word "bias" refers to consistent or systematic errors in measuring student skills or accomplishments. If a test produces score differences on skills for which the groups do not really differ, then the word would apply. However, if differences are real and the test correctly reflects them, then the test should not be considered biased. A primary result from this large amount of data we examined was that some of the differences between the genders are real differences — found in many types of measures, by many different approaches, and in many samples. Tests that reflect such widely corroborated differences are not making an error. They are correct, not baised.

Can't We Fix Differences by Fixing the Content?

The notion here is that if we could just remove from the test the questions testing the knowledge and skills on which males do better than females and replace them with questions testing the knowledge and skills on which females do better (or sometimes vice versa), we could "fix" the problem of gender differences. The answer is "Yes, to some degree." By manipulating the test content we could mute differences somewhat. First, recall that in representative samples, the differences are symmetric for males and females, and

for most subjects no differences exist. So the only "fix" that someone might seek would be to change the content for subjects on which the groups differ, such as writing, language use, and geopolitical subjects. The larger differences occur in self-selected groups taking high-stakes tests, and such content manipulation would not eliminate the differences produced by the dominant effect of the greater male spread.

The problem with this manipulation arises if content of less importance replaces content of more importance as would presumably often be the case when the "fix" is driven by a goal of no difference rather than a goal of important content. If the importance of the content is reduced, it would harm the meaningfulness and usefulness of the test. The skills or content that are most important have always and should always drive the make-up of a test. The preemminence of the knowledge and skills is an essential technical characteristic of tests on which public confidence is largely based.

Note, however, that it is not inappropriate to reexamine content periodically and add important content that has been ignored or has been difficult to include in the past. This type of action is for the purpose of including important content and strengthening the test, not to adjust the test to meet a predetermined difference goal. The key is always the importance of the content. Without that, tests will have little meaning or value.

What ETS Is Doing About These Results

There are many implications, partly indicated earlier, of the results on this study for educators, parents, policymakers, and testers. We will be exploring those implications in various ways with other affected parties. However, rather than point here to what everyone else might do in response to these results, we conclude this monograph with a brief summary of the things ETS is doing about them.

Research

ETS continues to sponsor research on issues of group difference as well as on ways to make assessments more useful and fair. ETS research has focused on new forms of assessment, with special attention to performance assessment, writing, and new forms possible through technology. The introduction of computer-based testing opens many possibilities for testing a wide array of skills in a variety of forms that fit well with the learning or work experience of the test takers.

The breadth of tested skills cannot be expanded in practice unless we learn to measure a wider band of skills in practically feasible ways. Writing has been very difficult to include on tests because of the complexity and expense of scoring written answers. ETS researchers have led the way in developing reliable and valid scoring approaches and, most recently, in developing scoring networks so that scoring can occur with greater speed and

efficiency. Similarly, the development of computer delivery of tests allows more practical presentation of complex problems as well as forms of computer scoring to help make such approaches practical.

Although specific results vary, of course, for different groups, issues and principles underlying this treatment of gender differences are much the same as for other important groups. ETS will continue to pursue the many unanswered questions raised not only by this study for gender but also for other groups such as racial and ethnic minorities.

Changes in Assessment

ETS's responses to these results as they became known over recent years has been, with the support of its clients, to make changes in tests. Writing has been added to several major tests in response to the increased recognition of its importance and the increasing practical feasibility of testing it. In 1994, when the new SAT was introduced, a new SAT II Writing test was introduced with it. Also about that time, ETS's teacher licensing test (Praxis) introduced a writing portion on computer. The Graduate Management Admission Test added a writing component also in 1994 and will continue that portion on computer when the GMAT moves to computer-based delivery in the fall of 1997. A writing portion is being added to the PSAT/NMSQT this fall as well, and a writing component is scheduled for addition to the Graduate Record Examinations as part of a redesign, likely in 1999.

The introduction of large-scale computer delivery of tests is a major ETS-sponsored change that will eventually make it practically feasible to measure a new breadth of skills and knowledge. ETS has now given over one million tests on computer including the GRE, Praxis, the NCLEX of the National Council of State Boards of Nursing, and the highly innovative exam of the National Council of Architectural Registration Boards.

Communicating Results

One of ETS's self-imposed responsibilities is to communicate what it learns about issues such as gender and testing. To that end, we are publishing these results in book form to reach the technical testing field and readers with special interest and resolve. We have highlighted the more general results aimed at a broader readership for many public groups. We are sharing both the more general and the more technical results with our various clients. We have scheduled a day-long briefing for test publishers to review our findings in some depth, and we expect to provide briefings to a variety of public or governmental groups as well. Our goal in all of these communications is to help people understand what we have learned and its implications for education and for testing.

ETS is an organization defined by its commitment to lead the production of knowledge on key issues that relate to assessment, to communicate to the public those findings, to respond to what we learn by making changes and

improvements in the assessments we develop for and with many clients, and to lead the development of new assessment possibilities that will open doors for new and better tests in the future to assist in the ever-more-effective education of youngsters. We hope that with this study we have lived up to those responsibilities.

What We Hope People Will Remember

1. There are many similarities and some genuine differences between how females and males perform in educational settings.
2. The differences are the result of many factors, and they widen particularly between the 4th and 12th grades.
3. While research shows that females have closed the gap significantly on math and science scores, males show a continuing gap in writing and language skills. Our attention to gaps needs to cut both ways.
4. There is a breadth of relevant and valuable skills that women and men need to know. Educators and parents need to concentrate on teaching and measuring that breadth of skills for both genders.
5. And finally, while we can learn significant things from studying group behavior, these data remind us to look at each student as a unique individual and not stereotype anyone because of gender or other characteristics.

Notes

1. National Center for Education Statistics, U.S. Department of Education. (1992). *The condition of education 1992* (NCES 93-290). Washington, DC: U.S. Government Printing Office.

   National Center for Education Statistics, U.S. Department of Education. (1994). *Digest of education statistics 1992*. Washington, DC: U.S. Government Printing Office.

2. Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University.

3. Cleary, T. A. (1992). Gender differences in aptitude and achievement test scores. In *Sex equity in educational opportunity, achievement, and testing: Proceedings of the 1991 ETS Invitational Conference* (pp. 51-90). Princeton, NJ: Educational Testing Service.

   Jacklin, C. N. (1989). Female and male: Issues of gender. *American Psychologist*, *44*(2), 127-133.

   Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher*, *18*, 17-27.

   Linn, M. C., & Petersen, A. C. (1985). Facts and assumptions about the nature of sex differences. In S. Klein (Ed.), *Handbook for achieving sex equity through education* (pp. 53-77). Baltimore: Johns Hopkins University.

   Wilder, G. Z., & Powell, K. (1989). *Sex differences in test performance: A survey of the literature* (CB Rep. No. 89-3; ETS RR-89-4). New York: College Entrance Examination Board.

4. Applebee, A. N., Langer, J. A., Jenkins, L. B., Mullis, I. V., & Foertsch, M. A. (1990). *Learning to write in our nation's schools: Instruction and achievement in 1988 at grades 4, 8, and 12* (NAEP 19-W-02). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

   Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53-69.

   Mullis, I. V., Dossey, J. A., Foertsch, M. A., Jones, L. R., & Gentile, C. A. (1991). *Trends in academic progress* (Rep. No. 21-T-01). Washington, DC: U.S. Government Printing Office.

5. Committee for Economic Development. (1996). *American workers and economic change: A statement by the Research and Policy Committee for Economic Development*. New York: Author.

6. Campbell, P. B., & Greenberg, S. (1993). Equity issues in educational research methods. In S. K. Biklen & D. Pollard (Eds.), *Gender and education: Ninety-second yearbook of the National Society for the Study of Education, Part I* (pp. 64-89). Chicago: University of Chicago.

7. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

8. The Project TALENT data in Figure 4 were gathered in 1960 and reported in Flanagan et al., 1964. The 1990 data came from Figure 3 in this report (i.e., Science, Math Concepts, and Verbal-Language Use).

   Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycoft, M. F., Orr, D. V., Goldberg, I., & Neyman, C. A., Jr. (1964). *Project TALENT: The American high-school student* (Final Report for Cooperative Research Project No. 635, U.S. Office of Education). Pittsburgh, PA: University of Pittsburgh.

9. Hansen, J. C., & Campbell, D. P. (1985). *Manual for the SVIB-SCII. Strong-Campbell interest inventory—Form T325 of the Strong Vocational Interest Blank* (4th ed.). Stanford, CA: Stanford University.

10. Most recently, Hedges and Nowell reported findings on differential variability quite similar to our own; i.e., male scores about 10% more variable on average. This degree of differential variability is represented in Figure 7.

    Cleary, T. A. (1992). Gender differences in aptitude and achievement test scores. In *Sex equity in educational opportunity, achievement, and testing: Proceedings of the 1991 ETS Invitational Conference* (pp. 51-90). Princeton, NJ: Educational Testing Service.

    Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62*(1), 61-84.

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

11. Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic group* (CB Rep. No. 93-1; ETS RR-94-27). New York: College Entrance Examination Board.

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

    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*, 323-336.

13. 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.

14. Lewis, C., & Willingham, W. W. (1995). *The effects of sample restriction on gender differences* (ETS RR-95-13). Princeton, NJ: Educational Testing Service.

15. Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24*, 157-166.

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

    Snow, R. E., & Ennis, M. (1996). Correlates of high mathematical ability in a national sample of eighth graders. In C. P. Benbow, & D. Lubinski (Eds.), *Intellectual talent: Psychometric and social issues* (pp. 301-327). Baltimore: Johns Hopkins University Press.

16. Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253-271.

Goodison, M. B. (1982). *A summary of data collected from Graduate Record Examinations test-takers during 1980-1981* (Data Summary Rep. No. 6). Princeton, NJ: Educational Testing Service.

Goodison, M. B. (1983). *A summary of data collected from Graduate Record Examinations test-takers during 1981-1982* (Data Summary Rep. No. 7). Princeton, NJ: Educational Testing Service.

Klein, S. P. (1981). *The effect of time limits, item sequence, and question format on applicant performance on the California Bar Examination.* San Francisco: Committee of Bar Examiners of the State Bar of California and the National Conference of Bar Examiners.

Schmitt, A. P. (1995, April). *Performance of gender, ethnic and language groups on the verbal and math content of the new PSAT/NMSQT.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Wild, C. L., & Durso, R. (1979). *Effect of increased test-taking time on test scores by ethnic and gender subgroups* (GRE No. 76-06R). Princeton, NJ: Educational Testing Service.

17. Beller, M., & Gafni, N. (1995). *International perspectives on the schooling and learning achievement of girls and boys as revealed in the 1991 International Assessment of Educational Progress (IAEP).* Jerusalem: National Institute for Testing and Evaluation.

Beller, M., & Gafni, N. (1996). *Can item format (multiple-choice vs. open-ended) account for gender differences in mathematics achievement?* Jerusalem: National Institute for Testing and Evaluation.

Bridgeman, B. (1993). *A comparison of open-ended and multiple-choice question formats for the quantitative section of the Graduate Record Examinations General Test* (GRE Rep. No. 88-13P; ETS RR-91-35). Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Rock, D. A. (1993). *Development and evaluation of computer-administered analytical questions for the Graduate Record Examinations General Test* (GRE Rep. No. 88-06P; ETS RR-92-49). Princeton, NJ: Educational Testing Service.

Dossey, J. A., Mullis, I. V., & Jones, C. O. (1993). *Can students do mathematical problem solving? Results from constructed-response questions in NAEP's 1992 mathematics assessment.* Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Lawrence, I. M., Lyu, C. F., & Feigenbaum, M. D. (1995). *DIF data on free-response SAT I mathematical items* (ETS RR-95-22). Princeton, NJ: Educational Testing Service.

18. Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1991, April). *Do women perform better, relative to men, on constructed-response tests or multiple-choice tests? Evidence from the Advanced Placement examinations.* Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago.

19. Jones, L. R., Mullis, I. V., Raizen, S. A., Weiss, I. R., & Weston, E. A. (1992). *The 1990 science report card: NAEP's assessment of fourth, eighth, and twelfth graders.* Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Pollack, J. M., & Rock, D. A. (in press). *Constructed response tests in the NELS:88 school effects study.* Washington, DC: National Center for Education Statistics.

Suggestions for Further Reading on Selected Topics

1. Books & Major Reviews of Gender Difference & Similarity

Deaux, K. (1985). Sex and gender. *Annual Review of Psychology, 36*, 49-81.

Halpern, D. F. (1992). *Sex differences in cognitive abilities* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hyde, J. S. (1991). *Half the human experience: The psychology of women* (4th ed.). Lexington, MA: D. C. Heath.

Hyde, J. S., & Linn, M. C. (Eds.). (1986). *The psychology of gender: Advances through meta-analysis*. Baltimore: Johns Hopkins University.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University.

Wilder, G. Z. (1997). Antecedents of gender differences. In *Supplement to Gender and fair assessment*. Princeton, NJ: Educational Testing Service.

2. Summary Analyses of Test Data

Cleary, T. A. (1992). Gender differences in aptitude and achievement test scores. In *Sex equity in educational opportunity, achievement, and testing: Proceedings of the 1991 ETS Invitational Conference* (pp. 51-90). Princeton, NJ: Educational Testing Service.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43*(2), 95-103.

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

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

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104*(1), 53-69.

Linn, M. C., & Petersen, A. C. (1985a). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*, 1479-1498.

Stanley, J. C., Benbow, C. P., Brody, L. E., Dauber, S., & Lupkowski, A. E. (1992). Gender differences on eighty-six nationally standardized aptitude and achievement tests. In N. Colangelo, S. G. Assouline, & D. L. Ambroson (Eds.), *Talent development: Proceedings from the 1991 Henry B. and Jocelyn Wallace National Research Symposium on Talent Development* (pp. 41-48). Unionville, NY: Trillium.

3. National Assessments

Applebee, A. N., Langer, J. A., Mullis, I. V., Latham, A. S., & Gentile, C. A. (1994). *NAEP 1992 writing report card*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycoft, M. F., Orr, D. V., Goldberg, I., & Neyman, C. A., Jr. (1964). *Project TALENT: The American high-school student* (Final Report for Cooperative Research Project No. 635, U.S. Office of Education). Pittsburgh, PA: University of Pittsburgh.

Hammack, D. C., Hartoonian, M., Howe, J., Jenkins, L. B., Levstik, L. S., MacDonald, W. B., Mullis, I. V., & Owen, E. (1990). *The U.S. history report card*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Jones, L. R., Mullis, I. V., Raizen, S. A., Weiss, I. R., & Weston, E. A. (1992). *The 1990 science report card: NAEP'S assessment of fourth, eighth, and twelfth graders*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the National Adult Literacy Survey*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Mullis, I. V., Campbell, J., & Farstrup, A. (1993). *NAEP 1992 reading report card for the nation and the states: Data from the national and trial state assessments* (Report No. 23-ST06). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Mullis, I. V., Dossey, J. A., Campbell, J. R., Gentile, C. A., O'Sullivan, C., & Latham, A. S. (1994). *NAEP 1992 trends in academic progress* (Report No. 23-TR01). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Mullis, I. V., Dossey, J., Owen, E., & Phillips, G. (1993). *NAEP 1992 mathematics report card for the nation and the states—Data from the national and trial state assessments* (Rep. No. 23-ST-02). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Rock, D. A., Pollack, J. M., Owings, J., & Hafner, A. (1990). *Psychometric report for the NELS:88 base year test battery* (NCES 90-468). Washington, DC: U.S. Department of Education, National Center for Education Statistics, Office of Educational Research and Improvement.

4. International Studies

Beaton, A. E., Martin, M. O., Mullis, I. V., Gonzalez, E. J., Smith, T. A. & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Beaton, A. E., Mullis, I. V., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Beller, M., & Gafni, N. (1996b). The 1991 International Assessment of Educational Progress in mathematics and sciences: The gender differences perspective. *Journal of Educational Psychology, 88*(2), 365-377.

Lapointe, A. E., Askew, J. M., & Mead, N. A. (1992). *Learning science*. Princeton, NJ: Educational Testing Service, International Assessment of Educational Progress.

Lapointe, A. E., Mead, N. A., & Askew, J. M. (1992). *Learning mathematics*. Princeton, NJ: Educational Testing Service, International Assessment of Educational Progress.

Murphy, R. J. (1982). Sex differences in objective test performance. *British Journal of Educational Psychology, 52*, 213-219.

Organisation for Economic Co-Operation and Development. (1986). *Girls and women in education: A cross-national study of sex inequalities in upbringing and in schools and colleges*. Paris: Author.

United Nations. (1995). *The world's women 1995: Trends and statistics* (Social Statistics Indicators, Series K, No. 12). New York: Author.

5. Mathematics

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.

Benbow, C. P. (1988b). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. *Behavioral and Brain Sciences, 11*, 169-232.

Chipman, S. F., Brush, L. R., & Wilson, D. M. (1985). *Women and mathematics: Balancing the equation*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24*, 157-166.

Dossey, J. A., Mullis, I. V., & Jones, C. O. (1993). *Can students do mathematical problem solving? Results from constructed-response questions in NAEP's 1992 mathematics assessment*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Eccles, J. S., & Jacobs, J. E. (1986). Social forces shape math attitudes and performance. *Signs: Journal of Women in Culture and Society, 11*, 367-380.

Gallagher, A. M., & De Lisi, R. (1994). Gender differences in the Scholastic Aptitude Test: Mathematics problem solving among high ability students. *Journal of Educational Psychology, 86*, 204-211.

Kupermintz, H., Ennis, M. M., Hamilton, L. S., Talbert, J. E., & Snow, R. E. (1995). Enhancing the validity and usefulness of large-scale educational assessments: I. NELS:88 mathematics achievement. *American Educational Research Journal, 32*, 525-554.

Lubinski, D., & Humphreys, L. G. (1990). A broadly based analysis of mathematical giftedness. *Intelligence, 14*, 327-355.

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*, 323-336.

6. Writing

Breland, H. M. (1996). *Writing skill assessment: Problems and prospects*. Princeton, NJ: Educational Testing Service, Policy Information Center.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*, 37-50.

Camp, R. (1993). Changing the model for the direct assessment of writing. In M. M. Williamson, & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 45-78). Cresskill, NJ: Hampton Press.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability* (Research Monograph No. 6). New York: College Entrance Examination Board.

White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance* (2nd ed.). San Francisco: Jossey-Bass.

7. Grades & Grading

Bejar, I. I., & Blew, E. O. (1981). *Grade inflation and the validity of the Scholastic Aptitude Test* (CB Rep. No. 81-3). New York: College Entrance Examination Board.

Goldman, R. D., & Hewitt, B. N. (1975). Adaptation-level as an explanation for differential standards in college grading. *Journal of Educational Measurement, 12*(3), 149-161.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic group* (CB Rep. No. 93-1; ETS RR-94-27). New York: College Entrance Examination Board.

Strenta, A. C., & Elliott, R. (1987). Differential grading standards revisited. *Journal of Educational Measurement, 24*(4), 281-291.

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.

8. Other Talents & Accomplishments

College Board. (1986). *Measures in the college admissions process: A College Board colloquium*. New York: College Entrance Examination Board.

Gardner, J. W. (1961). *Excellence: Can we be equal and excellent too?* New York: Harper & Row.

Richards, J. M., Jr., Holland, J. L., & Lutz, S. W. (1967). Prediction of student accomplishment in college. *Journal of Educational Psychology, 58*(6), 343-355.

Wallach, M. A. (1976). Psychology of talent and graduate education. In S. Messick & Associates, *Individuality in Learning* (pp. 178-210). San Francisco: Jossey-Bass.

Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability*. New York: College Entrance Examination Board.

9. Antecedents of Gender Difference — Educational, Biological, Social & Experiential

American College Testing Program. (1988). *ACT assessment program technical manual.* Iowa City, IA: Author.

Benbow, C. P. (1988a). Neuropsychological perspectives on mathematical talent. In L. K. Obler & D. Fein (Eds.), *The exceptional brain: Neuropsychology of talent and special abilities* (pp. 48-69). New York: Guilford.

Burton, N. W., Lewis, C., & Robertson, N. (1988). *Sex differences in SAT scores* (CB Rep. No. 88-9; ETS RR-88-58). New York: College Entrance Examination Board.

College Board, Admissions Testing Program. (1993b). *College-bound seniors: 1993 profile of SAT and Achievement Test takers*. New York: College Entrance Examination Board.

Dey, E. L., Astin, A. W., & Korn, W. S. (1991). *The American freshman: Twenty-five year trends*. Los Angeles: University of California, Higher Education Research Institute.

Eccles, J. S. (1987). Gender roles and women's achievement-related decisions. *Psychology of Women Quarterly, 11*, 135-172.

Ekstrom, R. B., Goertz, M. E., & Rock, D. A. (1988). *Education and American youth: The impact of the high school experience*. London: Falmer.

Fausto-Sterling, A. (1985). *Myths of gender: Biological theories about women and men.* New York: Basic Books.

Fennema, E., & Peterson, P. (1985). Autonomous learning behavior: A possible explanation of gender-related differences in mathematics. In L. C. Wilkinson & C. B. Marrett (Eds.), *Gender influences in classroom interaction* (pp. 17-35). Orlando, FL: Academic Press.

Klein, S. S. (Ed.). (1985). *Handbook for achieving sex equity through education.* Baltimore: Johns Hopkins University.

Mann, J. (1994). *The difference: Growing up female in America.* New York: Warner.

Wellesley College Center for Research on Women. (1992). *The AAUW report: How schools shortchange girls: A study of major findings on girls and education.* Washington, DC: American Association of University Women Educational Foundation and National Education Association.

00791 • Y57M8 • 250317 • Printed in U.S.A.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").