Sex Differences in Mental Test Scores, Variability, and Numbers of High-Scoring Individuals

Larry V. Hedges* and Amy Nowell

Sex differences in central tendency, variability, and numbers of high scores on mental tests have been extensively studied. Research has not always seemed to yield consistent results, partly because most studies have not used representative samples of national populations. An analysis of mental test scores from six studies that used national probability samples provided evidence that although average sex differences have been generally small and stable over time, the test scores of males consistently have larger variance. Except in tests of reading comprehension, perceptual speed, and associative memory, males typically outnumber females substantially among high-scoring individuals.

Understanding whether there are sex differences in intellectual abilities—and, if so, to what degree—has long been a concern of scientists in many disciplines. Such differences are relevant to people who are interested in achieving fair representation of women in scientific and technical fields where excellence requires a high degree of ability. Recent work in labor economics has also found that sex differences in ability (particularly mathematical ability) are associated with sex differences in earnings and occupational status (*1*). Studies of sex differences in ability conducted during the late 19th and early 20th centuries involved rather weak empirical evidence (*2*) that was usually collected from samples of people that were not demonstrably representative of any important population. The quality of the evidence has improved in recent decades as a result of the development of modern survey methods, which have made it possible to identify and collect data from nationally representative samples at a reasonable cost (*3*). In the past 35 years, six large-scale surveys have used these methods to collect mental test data from representative samples of adolescents and young adults in the United States. We used these survey data to investigate sex differences in intellectual abilities. Specifically, we examined the magnitude of sex differences in mean scores, in the variance of these scores, and in the numbers of individuals with particularly high or low scores.

The relative frequency of eminent men and women was one of the first lines of evidence used to support the notion of sex differences in cognitive abilities (*4*). Pearson (*5*) criticized this indirect evidence of mental ability and urged the collection of more direct evidence. For several decades, studies have used mental tests to collect data on cognitive sex differences in small samples of people. Such samples were conveniently available but were not chosen on a probability sampling basis to be representative of a specific population. Much of this work is summarized by Maccoby (*6*) and Maccoby and Jacklin (*7*). Quantitative syntheses (meta-analyses) of this work have also been provided (*8*, *9*). A more recent approach to the assessment of cognitive sex differences has been the combination of evidence from test norming samples, which are believed to be more representative than the samples chosen for small-scale research studies (*10*, *11*).

Variability. The assumption that the variance of intellectual abilities among men is greater than among women seems to have arisen in connection with evolutionary theory before 1900 (*2*). Maccoby and Jacklin (*7*) concluded that, compared to the score distributions of females in various mental tests, the distributions of male scores had larger variance for some abilities (mathematical and spatial abilities) but had equal variance for others. At about the same time, Jensen (*12*) reviewed the literature on sex differences in intelligence quotient (IQ) and concluded that the standard deviation of male IQ scores was about 20% larger than that of females. Recently, Feingold (*11*) reviewed test norming summary statistics to study sex differences in variability and concluded that the variance of male test scores was larger than that for females in tests of quantitative and spatial ability but not in tests of verbal ability.

Talent. The connection between variance and the occurrence of unusually talented individuals (individuals with unusually high mental test scores) was recognized by Thorndike (*13*), who felt that the most important consequences of sex differences in variance would occur at the highest percentiles of the ability distribution. Researchers such as Terman (*14*) and Benbow (*15*) who sought out talented individuals found many more males than females among the talented individuals they identified. However, because they examined only individuals selected on the basis of high test scores, their research design could not determine the source of the imbalance favoring males. Such differences could be a consequence of greater variance among males [even in the absence of a mean difference favoring males (*16*)], greater average male scores, the joint effects of differences in mean and variance, or selection bias favoring males (*16*–*18*). Feingold (*11*) and Hedges and Friedman (*19*) evaluated data from test norming studies to determine the likely joint effects of sex differences in mean and variance on the numbers of males and females that would be expected in the tails of the ability distributions. Their findings suggested that the larger numbers of high-scoring males found in studies of talented people might be primarily a consequence of sex differences in test score variance.

Weaknesses of previous research. Most work on sex differences and talent has relied on data collected from samples that were not representative of the nation as a whole. Reviews and meta-analyses of data from nonrepresentative samples are not necessarily any more representative than the studies on which they are based. For example, although the samples in Hyde's (*8*) meta-analysis of cognitive sex differences included one nationally representative sample and other reasonably unselected samples, it also included samples drawn from Harvard undergraduates, other college students (from colleges with less selective entrance requirements), and the Terman study of geniuses. Other studies have made use of tests that are offered nationally but taken selectively, such as the Scholastic Aptitude Test (SAT) and other College Board tests (*20*). Studies of talented individuals have almost exclusively used samples derived from "talent searches" that solicit volunteers and consequently have a potential for bias; this problem was recognized by researchers at least 40 years ago (*18*). The use of test norming samples would seem to mitigate some of these difficulties because such samples are typically large and broad-based. But they are usually quota samples and are almost never nation-

The authors are in the Department of Education, University of Chicago, 5835 South Kimbark Avenue, Chicago, IL 60637, USA.

*To whom correspondence should be addressed.

ally representative in the sense of being true national probability samples. When data from test norming samples have been used to study talent (*11, 19*), the computation of the characteristics of extreme scores from the means and standard deviations requires the assumption that scores are normally distributed. This assumption is sometimes questionable (*21*).

These biases due to selection, sampling, and the use of distributional assumptions may be relatively small, yet they are important because overall sex differences in either mean or variance are themselves small. Hence it is plausible that these sources of bias may have effects that are not negligible compared to real sex differences. Also, the small differences in mean or variance that have been found can lead to very large differences (several to one) in the numbers of males as compared to females in the upper percentiles of the national distribution (*19*). For example, a mean difference of 0.3 standard deviations, which would be judged as "small" by the convention of effect size introduced by Cohen (*22*), coupled with a variance difference of 15%, could lead to 2.5 times as many males as females in the top 5% of the test score distribution and more than 6 times as many males in the top 0.1%.

Method

We performed secondary analyses of six large data sets collected between 1960 and 1992. Each of these surveys used a stratified national probability sample of adolescents and provided sampling weights to permit inferences about specifically defined national populations. The surveys, which used slightly different population definitions and measured mental abilities with slightly different conventional mental tests, are described below and summarized in Table 1.

The Project Talent complete age group (age 15) data set. In 1960, Project Talent collected a national probability sample of high school students (*23*), along with a supplementary sample of 15-year-olds who were not in high school (because they had dropped out, were not in school as a result of serious illness or physical disability, or were mentally retarded or institutionalized). The combined sample of 73,425 examinees was representative of the entire population of 15-year-olds in the United States in 1960 (*24*). The examinees were administered a battery of 23 cognitive tests over a full day of testing. The tests are described in (*23*).

The NLS-72 data set. The National Longitudinal Study of the High School Class of 1972 (NLS-72) collected a national probability sample of high school seniors from public and private high schools. The sample was designed to be representative of the students in the senior class in American high schools in the spring of 1972. A total of 16,860 students were administered a 69-minute battery of tests that measured both verbal and nonverbal abilities. The six specific tests used were vocabulary, reading, mathematics, letter groups (a test of inductive reasoning), picture number (a test of associative memory), and mosaic comparisons (a test of perceptual speed and accuracy) (*25*).

The NLSY data set. The National Longitudinal Study of Youth (NLSY) was conducted to study labor force behavior. The sample we used actually consists of three independent probability samples that, when appropriately combined, yield a cross-sectional sample representing the noninstitutionalized civilian segment of American youth (ages 15 to 22) as of 1 January 1980. In total, 11,914 participants were administered the Armed Services Vocational Aptitude Battery (ASVAB) Form 8A in the spring and summer of 1980 (*26*).

The ASVAB was developed by the U.S. Armed Services as a tool for selecting and sorting new recruits into appropriate training programs and, subsequently, jobs. The ASVAB comprises 10 scales, all of which are timed. Eight of these are "power tests"; the remaining two are "speed tests," that is, quickness in performance is an aspect of the ability being measured. The 10 scales are arithmetic reasoning, mathematics knowledge, word knowledge (vocabulary), paragraph comprehension (reading comprehension), general science, numerical operations (a test of speed in arithmetic computation), coding speed, automotive and shop information (a measure of general knowledge and principles of auto repair, metal and wood shop procedures, and tool use), mechanical comprehension (a test of mechanical principles, including ability to decipher and visualize motion in schematic drawings), and electronics information.

The HS&B data set. High School and Beyond, 1980: A Longitudinal Survey of Students in the United States (HS&B) collected national probability samples in the spring of 1980 for two separate cohorts, senior and sophomore students in public and private high schools. We used the sample of 25,069 high school seniors only. The tests administered were vocabulary, reading, mathematics, spatial ability, picture number (a test of associative memory), and mosaic comparisons (a test of perceptual speed and accuracy). All of these tests are very similar or identical to the corresponding tests used in NLS-72 (*25*).

The NELS:88 data set. The National Educational Longitudinal Study of the Eighth Grade Class of 1988 (NELS:88) used a two-stage national probability sample of 24,599 eighth-grade students who were enrolled in public and private schools in 1988. The students were followed for 4 years and were resurveyed in 1992, when most were in the 12th grade. Some students surveyed were not in school 4 years after the eighth grade because they dropped out or graduated early. In 1992, these students were administered an 85-minute battery of four cognitive tests that were designed to measure achievement in

Table 1. Summary of the characteristics of the six data sets.

Characteristic	Project Talent	NLS-72	NLSY	HS&B	NELS:88	NAEP
Year of assessment	1960	1972	1980	1980	1992	1971–1992
Sample size	73,425	16,860	11,914	25,069	24,599	Varies
Population	All 15-year-olds	12th-grade students	Noninstitutionalized 15- to 22-year-olds	12th-grade students	8th-grade students as of 1988	17-year-olds in school
Abilities measured						
Reading comprehension	◆	◆	◆	◆	◆	◆
Vocabulary	◆	◆	◆	◆		
Mathematics	◆	◆	◆	◆	◆	◆
Perceptual speed	◆	◆	◆	◆		
Science	◆		◆		◆	◆
Social studies	◆				◆	
Nonverbal reasoning	◆	◆				
Associative memory	◆	◆		◆		
Spatial ability	◆			◆		
Mechanical reasoning	◆		◆			
Electronics information	◆		◆			
Auto and shop information			◆			
Writing						◆

Table 2. Sex differences in means, variance, and numbers of extreme scores. Differences in means are expressed as *d* values (in standard deviation units). Differences in variance are expressed as VR values (ratios of male score variance to female score variance). Differences in numbers of extreme scores are expressed as ratios of the number of males to the number of females who scored in the bottom 10%, top 10%, or top 5% of the national distribution. Standard errors are in parentheses (*31*). For each subject area, the surveys are listed in chronological order. Abbreviations for individual Project Talent and NLSY tests within a subject area are as follows: AR, arithmetic reasoning; BS, biological science information; CC, clerical checking; CS, coding speed; MK, mathematics knowledge, NO, numerical operations; OI, object inspection; PS, physical science information; and TR, table reading. Values not given could not be computed. An infinite ratio in the tails reflects the fact that no females scored in the tail examined.

Subject area	*d*	VR	Tail region		
			≤10%	≥90%	≥95%
Reading comprehension					
Project Talent	−0.15 (0.013)	1.16 (0.015)	1.71 (0.088)	0.90 (0.051)	1.00 (0.080)
NLS-72	−0.05 (0.027)	1.03 (0.028)	1.15 (0.13)	0.94 (0.11)	0.81* (0.14)
NLSY	−0.18 (0.031)	1.16 (0.036)	1.50 (0.19)	0.83 (0.12)	—
HS&B	0.002 (0.017)	1.10 (0.024)	1.07 (0.099)	1.03 (0.096)	1.06 (0.14)
NELS:88	−0.09 (0.020)	1.16 (0.023)	1.75 (0.14)	0.80 (0.072)	0.83 (0.11)
Vocabulary					
Project Talent	0.25 (0.013)	1.05 (0.014)	0.89 (0.050)	1.57 (0.081)	1.50 (0.011)
NLS-72	−0.06 (0.027)	1.00 (0.027)	1.02 (0.12)	0.89 (0.11)	0.87 (0.15)
NLSY	−0.03 (0.031)	1.08 (0.034)	1.20 (0.15)	0.87 (0.12)	—
HS&B	0.07 (0.017)	1.05 (0.023)	0.84 (0.082)	1.06 (0.098)	1.06 (0.14)
Mathematics					
Project Talent	0.12 (0.013)	1.20 (0.015)	1.00 (0.055)	1.33 (0.069)	1.50 (0.011)
NLS-72	0.24 (0.027)	1.05 (0.028)	0.72 (0.090)	1.76 (0.019)	2.34* (0.36)
NLSY: AR	0.26 (0.031)	1.25 (0.039)	1.84 (0.23)	1.90 (0.24)	2.20 (0.39)
MK	0.08 (0.031)	1.19 (0.037)	0.99 (0.13)	1.70 (0.21)	1.90 (0.34)
HS&B	0.22 (0.022)	1.16 (0.026)	0.77 (0.078)	1.67 (0.14)	2.06 (0.26)
NELS:88	0.03 (0.020)	1.06 (0.021)	0.97 (0.082)	1.34 (0.11)	1.64 (0.18)
Perceptual speed					
Project Talent: TR	—	—	2.17 (0.11)	0.82 (0.047)	1.00 (0.080)
CC	—	—	1.79 (0.092)	0.73 (0.044)	0.81 (0.068)
OI	—	—	1.50 (0.077)	1.00 (0.055)	1.00 (0.080)
NLS-72	−0.23 (0.027)	1.04 (0.028)	1.54 (0.17)	0.70 (0.089)	0.69 (0.12)
NLSY: CS	−0.43 (0.031)	0.98 (0.031)	1.60 (0.20)	0.41 (0.077)	0.38 (0.11)
NO	−0.23 (0.031)	1.08 (0.034)	1.50 (0.19)	0.69 (0.10)	0.67 (0.14)
HS&B	−0.21 (0.022)	1.15 (0.025)	1.49 (0.13)	0.73 (0.075)	0.79 (0.12)
Science					
Project Talent: PS	0.50 (0.013)	1.28 (0.017)	0.57 (0.038)	2.83 (0.15)	7.00 (0.65)
BS	0.29 (0.013)	1.15 (0.015)	0.78 (0.046)	2.00 (0.10)	—
NLSY	0.38 (0.031)	1.42 (0.044)	0.92 (0.13)	3.40 (0.45)	7.20 (1.6)
NELS:88	0.11 (0.020)	1.14 (0.023)	0.87 (0.076)	2.04 (0.16)	2.50 (0.28)
Social studies					
Project Talent	0.31 (0.013)	1.26 (0.016)	0.89 (0.050)	2.29 (0.12)	3.50 (0.27)
NELS:88	0.04 (0.020)	1.14 (0.023)	1.23 (0.10)	1.59 (0.13)	1.74 (0.19)
Nonverbal reasoning					
Project Talent	0.04 (0.013)	1.04 (0.013)	1.00 (0.055)	1.09 (0.059)	1.00 (0.080)
NLS-72	−0.22 (0.027)	1.15 (0.031)	1.49 (0.16)	0.74 (0.092)	0.67 (0.12)
Associative memory					
Project Talent	−0.32 (0.013)	0.82 (0.011)	1.56 (0.080)	0.50 (0.035)	0.43 (0.048)
NLS-72	−0.26 (0.027)	1.01 (0.027)	1.44 (0.16)	0.70 (0.089)	0.69 (0.12)
HS&B	−0.18 (0.022)	1.14 (0.025)	1.23 (0.11)	—	—
Spatial ability					
Project Talent	0.13 (0.013)	1.27 (0.028)	0.82 (0.047)	1.86 (0.095)	2.33 (0.35)
HS&B	0.25 (0.022)	1.27 (0.028)	0.79 (0.079)	1.90 (0.17)	2.39 (0.30)
Mechanical reasoning					
Project Talent	0.83 (0.012)	1.45 (0.019)	0.36 (0.029)	8.50 (0.059)	11.00 (1.2)
NLSY	0.72 (0.030)	1.74 (0.055)	0.60 (0.094)	8.00 (1.3)	10.90 (2.8)
Electronics information					
Project Talent	1.22 (0.011)	2.72 (0.035)	0.44 (0.033)	15.20 (1.3)	∞
NLSY	0.72 (0.030)	1.56 (0.049)	0.62 (0.096)	8.00 (1.3)	9.90 (2.5)
Auto and shop information					
NLSY	1.02 (0.029)	2.34 (0.073)	0.44 (0.079)	66.3 (27)	464 (702)

*These figures are for the 97th percentile.

reading, mathematics, science, and social studies (history and government).

The NAEP trend data sets. In 1969, Congress established the National Assessment of Educational Progress (NAEP) program to monitor the academic achievement of 9-, 13-, and 17-year-olds. The NAEP program has periodically tested large samples (70,000 to 100,000 students) in the areas of reading, mathematics, science, and writing. NAEP samples are national probability samples of students at the ages of interest who are in school. One part of the NAEP program is the periodic collection of data on equivalent measures, using exactly the same procedures in each assessment wave; these so-called trend data permit the accurate estimation of trends over time (*27*). We used only the 17-year-old samples.

Analysis. For each test in each survey, we used the sampling weights provided by the surveys to construct estimates of the national means and variances of the test score distribution for each sex. We then calculated variance ratios (ratios of male score variance to female score variance) and represented mean differences in standard deviation units by subtracting the estimated national mean score for females from that of males and dividing by the estimated national standard deviation for the entire distribution for both sexes combined. To compute national percentiles for the entire population, we first computed an estimate of the proportions of the test scores of each sex in the national population that were in the top 5%, top 10%, and bottom 10% of each test score distribution. These represent the proportions of "talented" or "untalented" individuals (as defined according to a series of different definitions of degree of talent). We then computed ratios of the estimated numbers of males and females in the national population who fell into each talent category (*28*).

Results

Sex differences in means. We used the standardized mean difference *d* to evaluate sex differences in means. Because *d* was calculated as the mean score for males minus the mean score for females, divided by the standard deviation in the total population, a positive value of *d* implies that males scored higher on average. Data from five of the six surveys (Project Talent, NLS-72, NLSY, HS&B, and the 1992 follow-up of NELS: 88) concerning sex differences in means are presented in Table 2. Virtually all of the mean differences are several times their standard errors and hence are reliably different from 0 at $P = 0.05$. However, because the sample sizes of these surveys were large, even differences too small to be of practical importance could be statistically

significant. We therefore interpreted the sizes of effects according to Cohen's convention (*22*), which construes a standardized mean difference of 0.2 as small, 0.5 as medium, and 0.8 as large.

On average, females exhibited a slight tendency to perform better on tests of reading comprehension, perceptual speed, and associative memory, and males exhibited a slight tendency to perform better on tests of mathematics and social studies. All of the effect sizes were relatively small except for those associated with vocational aptitude scales (mechanical reasoning, electronics information, and auto and shop information) in which average males performed much better than average females. The effect sizes for science were slightly to moderately positive, and those for perceptual speed were slightly to moderately negative. Thus, with respect to the effect size convention, these data suggest that average sex differences are generally rather small.

It is not obvious from these data that sex differences have changed since 1960. However, the population definitions of these five surveys are not identical. Project Talent (in 1960) and NLSY (in 1980) surveyed the total population of adolescents and young adults, both in school and out of school, whereas NLS-72, HS&B, and NELS:88 surveyed only students who were in school (in either the 8th or 12th grade). The NAEP trend studies measured a more limited range of abilities, but because the population definition and mental tests did not vary between assessment waves, the trends were measured with less ambiguity. Table 3 gives the sex differences in means (as *d* values) for the NAEP trend sample. Females performed better in reading and writing, and males performed better in science and mathematics. Average sex differences were small except for writing, in which females performed substantially better than males in every year. Although average sex differences in mathematics and science scores appear to have narrowed somewhat over time, sex differences in reading and writing scores have not.

Sex differences in variance. Examination of the ratios of male score variance to female score variance (VR values) in Table 2 reveals that the variance of male scores is larger than that of female scores (that is, VR $>$ 1) in all but two cases: the Project Talent associative memory (word memory) test and the NLSY coding speed test. In both cases, measures of the same constructs in other surveys showed greater male variability. The difference in variance is small, typically on the order of 3 to 15%. However, male scores had considerably larger variance than female scores on some tests, such as measures of science achievement and the vocational aptitude scales. There is little evidence from the data in Table 2 that sex differences in variance have changed systematically over time. Here again, differences among the population definitions of the surveys might obscure small changes in variance. The data on sex differences in variance computed from the NAEP trend samples (Table 3) suggest that the variance of male scores is typically greater than that of female scores (all of the VRs are $>$1) and that the difference is typically 5 to 20%. Trends over time in the VRs computed from the NAEP data are not striking, but it appears that for mathematics and science scores these ratios have increased over time.

Sex and talent. Sex differences in the proportions of males and females scoring in the extreme ranges of the ability test score distributions are summarized in Table 2. This table gives the ratio of the number of males to the number of females who scored in the bottom 10%, top 10%, and top 5% of the national distribution for both sexes combined; values of this ratio greater than 1 reflect more males than females. For reading comprehension, perceptual speed, and associative memory, more males than females scored in the bottom 10% of the national distribution (ratios of 1.4 to 2.2) and fewer males scored in the top 5 to 10%. In mathematics, science, and social studies, more males than females were in the upper tails of the distribution (ratios of 1.3 to 3.4 in the top 10%) and more females than males were in the lower tails. The differences favoring males were more profound in the vocational aptitude scales, with 8 to 10 times as many males as females scoring in the top 10%.

It has been shown that if scores are normally distributed in two populations and if one population has both a higher mean score and a larger variance than the other, then the ratio of the number of individuals in the population with the higher mean to that of the other population (the tail ratio) increases at higher percentiles in the upper tails of the distribution (*17, 19*). This pattern held for the tests that had sufficiently high ceilings to accommodate the estimation of percentiles above 95. For example, in the Project Talent mathematics (total) scale, the sex ratio was 1.3 for scores in the top 10%, 1.5 in the top 5%, 2.1 in the top 3%, and 7.0 in the top 1% of the overall distribution. For several of the science and vocational aptitude tests, the sex ratio became infinite in the top 3% or 1% of the overall distribution because no females scored in this range.

Implications

These data demonstrate that in U.S. populations, the test scores of males are indeed more variable than those of females, at least for the abilities measured during the 32-year period covered by the six national surveys. Moreover, there is little indication that variance ratios are changing over time. The evidence presented here also helps to resolve an apparent contradiction between the high ratios of males to females in highly talented samples and the generally small mean differences found between the sexes in relatively unselected samples. These data show that high sex ratios (5:1 among the top 3% and 7:1 among the top 1%) are

Table 3. Sex differences in mean and variance computed from NAEP trend sample data for 17-year-old students. Standard errors are in parentheses (*31*). The NAEP writing data were collected from representative samples of children by grade rather than age; in this case, grade 11 corresponds roughly to age 17.

Survey year	Reading		Mathematics		Science		Writing	
	d	VR	*d*	VR	*d*	VR	*d*	VR
1971	−0.27 (0.039)	1.08 (0.020)						
1975	−0.26 (0.032)	1.14 (0.026)						
1977					0.33 (0.036)	1.08 (0.019)		
1978			0.19 (0.041)	1.08 (0.018)				
1980	−0.18 (0.042)	1.11 (0.025)						
1982			0.18 (0.044)	1.07 (0.021)	0.36 (0.041)	1.10 (0.029)		
1984	−0.25 (0.030)	1.10 (0.016)					−0.55 (0.091)	1.14 (0.051)
1986			0.17 (0.051)	1.14 (0.036)	0.28 (0.055)	1.27 (0.047)		
1988	−0.21 (0.057)	1.07 (0.043)					−0.61 (0.087)	1.03 (0.064)
1990	−0.30 (0.049)	1.20 (0.033)	0.11 (0.050)	1.17 (0.043)	0.22 (0.045)	1.27 (0.038)	−0.60 (0.063)	1.05 (0.054)
1992	−0.27 (0.045)	1.12 (0.032)	0.15 (0.052)	1.11 (0.034)	0.23 (0.051)	1.20 (0.043)	−0.53 (0.074)	1.05 (0.042)

found in the upper tails of the ability distributions of unselected (nationally representative) samples. Thus, the high sex ratios found in some highly talented samples need not be attributed to differential selection by sex.

Our analyses suggest that average sex differences in most measured abilities are small, with the possible exception of science, writing, and stereotypically male vocational aptitudes. Contrary to the findings of small-scale studies, these average differences do not appear to be decreasing but are relatively stable across the 32-year period investigated. This finding demonstrates the weakness of relying on data that were not collected from explicitly representative samples to estimate values of small between-group differences or to discern weak trends over time for the nation as a whole.

The large sex differences in writing ability suggested by the NAEP trend data are alarming, particularly because these differences were found on assessments that used actual writing samples. The data imply that males are, on average, at a rather profound disadvantage in the performance of this basic skill. With respect to sex differences in vocational aptitude scores, military research has found that these scales do have substantial predictive validity for the obvious occupations. These occupations have been male-dominated, and attempts to promote fairness in representation may be thwarted by a shortage of females with a basic amount of aptitude relevant to these occupations.

The sex differences in mathematics and science scores, although smaller, are of concern because ability and achievement in science and mathematics may be necessary to excel in scientific and technical occupations. Small mean differences combined with modest differences in variance can have a surprisingly large effect on the number of individuals who excel. There is evidence [for example, from follow-up surveys of occupational behavior in Project Talent (*29*)] that people who have careers in science and engineering are overwhelmingly more likely to have scored in the 90th percentile on mathematics tests in high school. Sex differences in variance and mean lead to substantially fewer females than males who score in the upper tails of the mathematics and science ability distributions and hence are poised to succeed in the sciences. The achievement of fair representation of women in science will be much more difficult if there are only one-half to one-seventh as many women as men who excel in the relevant abilities.

Differences in the representation of the sexes in the tails of ability distributions are likely to figure increasingly in policy discussions about salary equity. Economists have recently begun to use individual differences in ability test scores to explain sex differences in wages and occupational advancement (*30*). Different kinds of abilities are not equally related to economic outcomes; one recent empirical study suggests that quantitative ability test scores (but not verbal ability test scores) "[account] for the observed male-female differences in earnings and occupational choices of recent college graduates" (*1*). The generally larger numbers of males who perform near the bottom of the distribution in reading comprehension and writing also have policy implications. It seems likely that individuals with such poor literacy skills will have difficulty finding employment in an increasingly information-driven economy. Thus, some intervention may be required to enable them to participate constructively in the work force.

Our results shed little light on the origins of sex differences in either mean or variability. However, the largest sex differences occur in areas not generally taught in school (such as mechanical comprehension and other vocational aptitudes). Moderately large differences are associated with performance in subject areas in which there appears to be considerable variability in the amount, content, and difficulty of the curriculum (such as science, social studies, and mathematics). If males are more likely to undertake more, or more challenging, course work in these areas, we would expect the observed pattern of sex differences to emerge. However, our data are not entirely consistent with the hypothesis that substantial sex differences arise only in connection with differences in opportunity to learn, because we found substantial differences in writing performance, which is presumably a skill taught to all students. If, as seems likely, differences in ability arise because of differences in experience and socialization, more work is needed to document that these differences exist and are linked to ability.

REFERENCES AND NOTES

1. M. Paglin and A. M. Rufolo, *J. Labor Econ.* **8**, 123 (1990).
2. S. A. Shields, *Am. Psychol.* **30**, 739 (1975).
3. W. G. Cochran, *Sampling Techniques* (Wiley, New York, 1963); M. H. Hansen, W. N. Hurwitz, W. G. Madow, *Sample Survey Methods and Theory* (Wiley, New York, 1953); L. Kish, *Survey Sampling* (Wiley, New York, 1965).
4. H. Ellis, *Man and Woman: A Study of Human Secondary Sexual Characteristics* (Walter Scott, London, 1894; Scribner, New York, 1984).
5. K. Pearson, *The Chances of Death* (Edward Arnold, London, 1897), vol. 1, chap 8.
6. E. E. Maccoby, Ed., *The Development of Sex Differences* (Stanford Univ. Press, Stanford, CA, 1966).
7. _____ and C. N. Jacklin, *The Psychology of Sex Differences* (Stanford Univ. Press, Stanford, CA, 1974).
8. J. S. Hyde, *Am. Psychol.* **36**, 892 (1981).
9. B. J. Becker and L. V. Hedges, *J. Ed. Psychol.* **76**, 583 (1984); J. S. Hyde and M. C. Linn, *Psychol. Bull.* **104**, 53 (1988); L. Friedman, *Rev. Ed. Res.* **59**, 185 (1989); J. S. Hyde, E. Fennema, S. J. Lamon, *Psychol. Bull.* **107**, 139 (1990).
10. A. Feingold, *Am. Psychol.* **43**, 95 (1988).
11. _____, *Rev. Ed. Res.* **62**, 61 (1992).
12. A. R. Jensen, in *Intelligence: Genetic and Environmental Influences*, R. Cancro, Ed. (Grune and Stratton, New York, 1971), pp. 107–161.
13. E. L. Thorndike, *Educational Psychology* (Teacher's College Press, New York, ed. 2, 1910).
14. L. M. Terman, *Genetic Studies of Genius* (Stanford Univ. Press, Stanford, CA, 1925), vol. 1.
15. C. P. Benbow, *Behav. Brain Sci.* **11**, 169 (1988).
16. B. J. Becker and L. V. Hedges, *ibid.*, p. 183.
17. C. Lewis and W. W. Willingham, "The effects of sample restriction on gender differences" (Research Report ETS RR-95-13, Educational Testing Service, Princeton, NJ, 1995).
18. L. M. Terman and L. E. Tyler, in *Manual of Child Psychology*, L. Carmichael, Ed. (Wiley, New York, ed. 2, 1954), pp. 1064–1114.
19. L. V. Hedges and L. Friedman, *Rev. Ed. Res.* **63**, 94 (1993).
20. J. C. Stanley, C. P. Benbow, L. E. Brody, S. Dauber, A. E. Lupkowski, in *Talent Development: Proceedings from the 1991 Henry B. and Joycelyn Wallace National Research Symposium on Talent Development*, N. Colangelo, S. G. Assouline, D. L. Ambroson, Eds. (Trillium, Unionville, NY, 1992), pp. 42–65.
21. L. S. Hollingworth, *Am. J. Soc.* **22**, 19 (1914); T. Micceri, *Psychol. Bull.* **105**, 156 (1989).
22. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, New York, rev. ed., 1977).
23. J. C. Flanagan *et al.*, *Design for a Study of American Youth* (Houghton Mifflin, New York, 1962).
24. M. F. Shaycroft, J. T. Dailey, D. B. Orr, C. A. Neyman, S. E. Sherman, *Studies of a Complete Age Group—Age 15* (Project Talent Office, University of Pittsburgh, Pittsburgh, PA, 1963).
25. D. A. Rock, T. L. Hilton, J. Pollack, R. B. Ekstrom, M. E. Goertz, *Psychometric Analysis of the NLS and High School and Beyond Test Batteries* (Government Printing Office, Washington, DC, 1985).
26. R. D. Bock and E. Moore, *Advantage and Disadvantage: A Profile of American Youth* (Erlbaum, Hillsdale, NJ, 1986).
27. I. Mullis, J. A. Dossey, M. A. Foertsch, L. R. Jones, M. A. Gentile, *Trends in Academic Progress* (Government Printing Office, Washington, DC, 1991).
28. A complication here is that it will not always be possible to get the exact percentile desired because the test scores are discrete. For example, two items incorrect may correspond to the 96th percentile while three items incorrect corresponds to the 93rd percentile. In cases where the desired percentile did not correspond to any possible test score, the nearest percentile corresponding to an obtainable score was used, and these are noted in the results.
29. L. L. Wise, L. Steel, C. MacDonald, *Origins and Career Consequences of Sex Differences in High School Mathematics Achievement* (American Institutes for Research, Palo Alto, CA, 1979).
30. S. Rosen, *Am. Econ. Rev.* **71**, 845 (1981); E. P. Lazear and S. Rosen, *J. Labor Econ.* **8**, S106 (1990).
31. The design effects used to compute standard errors were 1.75 for Project Talent, HS&B, and NLS-72, 1.56 for NELS:88, and 1.71 for NLSY. Standard errors for NAEP statistics were calculated with the use of a jackknife procedure that incorporated design effects.

32. Supported in part by a grant from the University of Chicago School Mathematics Project Fund for Research in Mathematics Education. We thank two anonymous reviewers for their comments.