Geometry Proof Writing: A New View of Sex Differences in Mathematics Ability

SHARON SENK
Newton [Massachusetts] Public Schools

ZALMAN USISKIN
University of Chicago

A study of 1,364 students in 74 senior high school classes in which geometry proof was taught found equal ability among males and females to write geometry proofs. These results held as well for select high-achieving subsamples. These findings and data from other recent studies suggest that girls and boys perform equally well even on complex mathematical tasks if both in-class and out-of-class exposure to the tasks is equal.

Sex differences in mathematics performance favoring males have been reported for many years (Maccoby and Jacklin 1974). Until recently, studies rather consistently indicated that, although no systematic sex differences in performance are observed in young children, by early adolescence boys begin to surpass girls on many mathematical tasks, and by the end of high school the gap between males and females is both statistically and educationally significant. Yet some recent studies report declines in differences or no differences at all (Armstrong 1981). The largest and most consistent sex differences reported have been on so-called high-level cognitive tasks such as applications of mathematics in real-world situations or problem solving. These differences seem particularly marked among higher-ability students (Benbow and Stanley 1980/81). Often differences in performance are attributed to sex differ-

0036-6773/83/9102-0001$01.00

ences in tests of spatial ability (Macoby and Jacklin 1974; Benbow and Stanley 1980/81).

Given these reported differences, one might expect significant sex differences in performance on doing geometry proofs, which requires some spatial ability, qualifies as a high-level cognitive task, and is considered among the most difficult processes to learn in the secondary school mathematics curriculum. The first purpose of this article is to report that, in the first large-scale study of geometry proof-writing performance ever conducted in the United States, we have found no consistent sex differences. The second purpose of this article is to propose an explanation for the inconsistent patterns of sex differences that characterize recent studies.

Design

The data we present are from the Cognitive Development and Achievement in Secondary School Geometry (CDASSG) project and represent only one of many aspects of geometry learning investigated by the project.[1] The CDASSG sample includes 2,699 students in 99 geometry classes from 13 public high schools in five states (table 1). The schools were chosen to represent a national cross-section of educational and socioeconomic conditions. Black, Hispanic, and Oriental minorities were sizable in a few schools. Within the schools, the subsample for this study includes all students in the geometry classes that had studied proof writing and whose teachers gave permission for testing, a total of 1,520 students in 74 classes from 11 high schools in five states. At the time of the spring testing, more than 95 percent of the students were age 14–17, and the mean age was 16 years, 2 months.

The study was conducted during the 1980–81 school year. During the first week of school, students were given a 25-minute test for entering knowledge of geometry terminology and facts.[2] In the last month of the school year, students took the 40-minute Comprehensive Assessment Program (CAP) (1980) standardized geometry achievement

---

SHARON SENK teaches mathematics at Newton North High School, Newton, Massachusetts, and recently completed her Ph.D. work in education at the University of Chicago. Her current interests are the learning of geometry and the history of mathematics curriculum reform. ZALMAN USISKIN is professor of education at the University of Chicago. His major interests are policy matters relating to the scope, sequence, and underlying structure of the K–12 mathematics curriculum.

TABLE 1

Sex by Track for the Entire CDASSG Project Sample and for All Those Who Took the EG and Proof Tests

	CDASSG Project Sample				Proof Sample			
		Students in Sample				Students in Sample		
Track Description	Number of Classes	Female	Male	Total	Number of Classes	Female	Male	Total
Highest of three tracks	15	165	183	348	15	136	164	300
Higher of two tracks	4	28	48	76	4	20	35	55
Middle of three tracks	27	352	328	680	25	233	212	445
Lower of two tracks	13	154	177	331	11	99	87	186
Lowest of three tracks	14	164	200	364	0	0	0	0
Untracked	26	452	448	920	19	186	192	378
Total	99	1,315	1,384	2,699	74	674	690	1,364

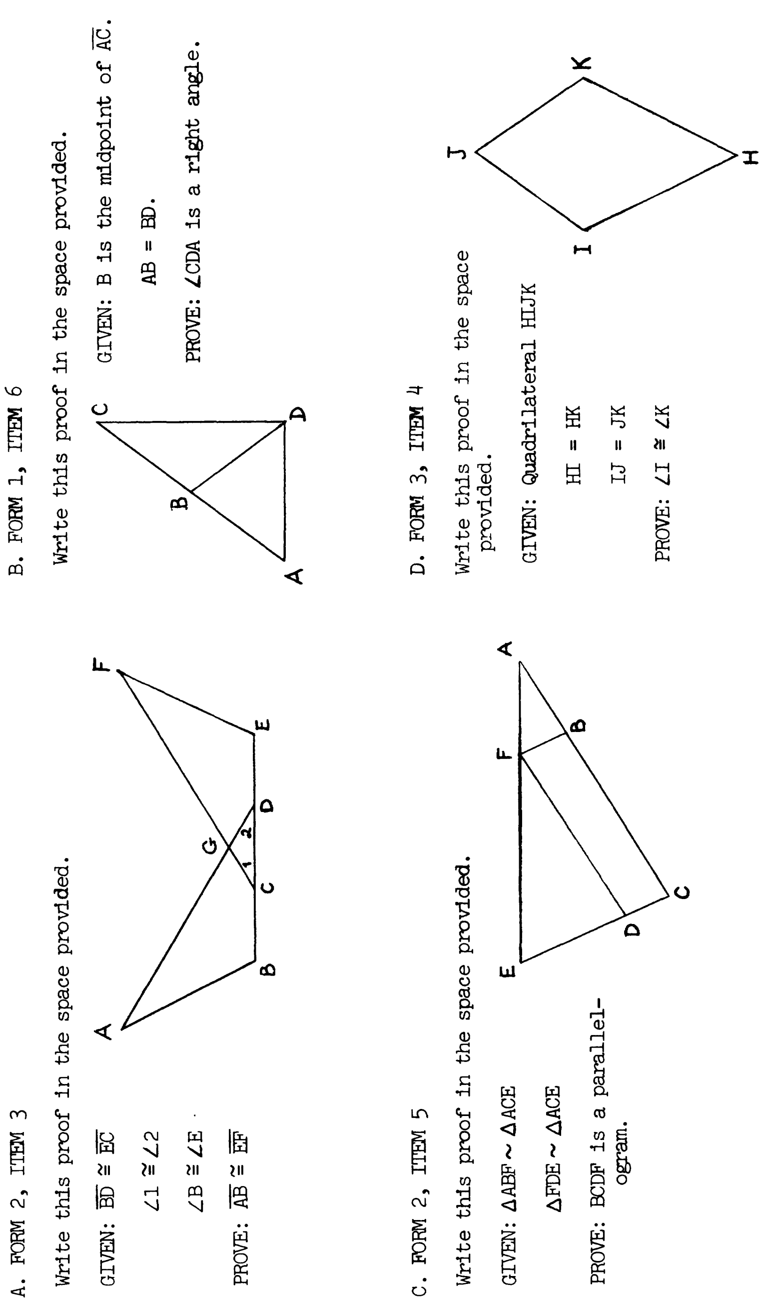

FIG. 1.—Sample items from the CDASSG proof tests

TABLE 2

Scores for Items Shown in Figure 1

Test Item	Mean	Standard Error
A. Form 2, item 3:		
Female (N = 214)	2.98	.11
Male (N = 241)	3.09	.10
B. Form 1, item 6:		
Female (N = 219)	.83	.07
Male (N = 235)	.79	.07
C. Form 2, item 5:		
Female (N = 214)	.84	.09
Male (N = 241)	1.20	.10
D. Form 3, item 4:		
Female (N = 241)	2.27	.12
Male (N = 216)	2.12	.13

NOTE.—Item A was the easiest of the 12 full proofs; item B was the most difficult. Item C most favored the boys; item D most favored the girls.

test and one of three forms of a 35-minute proof test devised by CDASSG project personnel.[3] All tests were administered by classroom teachers during the normal school day and monitored by project representatives. The proof test forms were alternated among the students so that approximately one-third of the students in each class received each form.

Three forms of a proof test were devised so that performance on a greater number of proofs could be analyzed. Each form contained six items: the first required the student to fill in four missing statements or reasons in a proof; the second required translation of a verbal statement into an appropriate figure, "given," and "to prove"; and the last four required the student to write complete proofs. All items were representative of standard geometry proofs, ranging from easy to difficult, covering congruent and similar triangles, parallel lines, and quadrilaterals. Sample proof items are shown in figure 1, and scores on these proofs are presented in table 2. Two pilot studies of the proof tests had been conducted to insure appropriate test length, clarity of instructions, and approximate balance of item difficulty and subject matter across forms, but no effort was exerted to make the forms statistically equivalent.

No large-scale assessment of proof-writing performance had been undertaken prior to this study, perhaps because of perceived difficulties in grading proofs and in finding items that would be fair for students who studied from texts with different terminology and theorem order. Neither of these potential difficulties seems to have arisen, perhaps because of the pilot studies and grading procedures we used.

Eight experienced high school mathematics teachers (six male, two female) were hired to grade the proof tests. Proof items were graded on a scale from 0 to 4 based on general criteria developed by Malone, Douglas, Kissane, and Mortlock (1980).

0—Student writes nothing, writes only the "given," or writes invalid or useless deductions.
1—Student writes at least one valid deduction and gives reason.
2—Student shows evidence of using a chain of reasoning, either by deducing about half the proof and stopping or by writing a sequence of statements that is invalid only because it is based on faulty reasoning early in the steps.
3—Student writes a proof in which all steps follow logically, but in which there are errors in notation, vocabulary, or names of theorems.
4—Student writes a valid proof with at most one error in notation.

Before grading each item, graders discussed the application of the general criteria to that item. Every item on each student's test was scored independently by a different pair of graders who had no access to the student's name, sex, or school. Interrater agreement ranged from 81 percent to 95 percent across the 18 items, averaging 86 percent. Less than 2 percent of the scores of the pair of graders differed by more than one point. When the two graders' scores disagreed, a third independent blind reading was undertaken, and the median of the three scores was chosen as the item score. The grading of the 1,520 test papers was completed in 40 person-days.

Two measures of proof-writing performance were calculated. The first, called "total score," is the customary sum of the item scores, with a maximum possible of 24. The second, called "number of proofs correct," is the number of full proof items upon which the student scored 3 or 4. The maximum possible "number of proofs correct" is 4.

Findings

We report here only on those 1,364 students who took a proof test and the entering geometry (EG) test. Of these, 690 are male and 674 are female, yielding a ratio within one-half percent of sex ratios in both national and school populations at ages 14–17 (U.S. Department of Commerce 1980). The breakdown of sex by track (table 1) shows that more males than females are in higher-track classes. The students range from seventh to twelfth graders, with 63 percent in tenth grade (table 3).

For this sample, total score means and standard deviations for the three forms of the proof test are, respectively, 12.59 ± 5.43, 14.27 ± 5.22, and 12.98 ± 6.37. Differences between these means are significant, as are differences in the shapes of the distributions, so the three proof test forms are not equivalent.[4] As a consequence, data from this study are reported separately by form.

Mean scores on the proof tests are reported by sex in table 4. Raw mean total scores are higher for males on two forms and for females on the third form, but none of the differences is statistically significant. The mean number of proofs correct is higher for males on all three forms, but never significantly.

Mean scores for girls are significantly lower than mean scores for boys on the EG test.[5] When the proof total scores are adjusted using ANCOVA for this entering geometry knowledge, adjusted mean proof total scores for females are higher than for males on all forms and significantly higher on Form 3. When the mean number of proofs correct are similarly adjusted, the results favor the females on all three forms, though never significantly.

For the 18 items viewed individually, mean scores for the sexes are significantly different at the .05 level on two items, on a full proof favoring males, the other a translation favoring females. At this significance level, 2 differences in 18 can be expected by chance, and no pattern in the content of items favoring either sex (even when statistical significance was ignored) was observed.

Thus, although girls enter the high school geometry course with generally less geometry knowledge, at the end of the year there is no consistent difference between the sexes on proof-writing performance. This finding is particularly striking not just because of the widely held

TABLE 3

Sex by Grade in School for Those Taking the EG and Proof Tests

	Grade							
Sex	7	8	9	10	11	12	NA	Total
Male	0	12	94	437	100	31	0	674
Female	1	6	103	426	125	28	1	690

TABLE 4

Mean Proof Scores for All Students Taking the Entering Geometry (EG) and Proof Tests

Form and Sex	Mean Raw Total Score	Mean Total Score Adjusted for EG	Mean Raw Number of Proofs Correct	Mean Number of Proofs Correct Adjusted for EG
1:				
Female (N = 219)	12.34 (.43)	12.91 (.36)	1.50 (.08)	1.61 (.07)
Male (N = 234)	12.87 (.42)	12.33 (.36)	1.55 (.08)	1.45 (.07)
2:				
Female (N = 214)	13.93 (.44)	14.65 (.36)	1.72 (.10)	1.88 (.08)
Male (N = 240)	14.60 (.41)	13.95 (.34)	1.97 (.09)	1.83 (.08)
3:				
Female (N = 241)	13.05 (.49)	13.63 (.41)	1.64 (.10)	1.75 (.09)
Male (N = 216)	12.82 (.52)	12.18 (.43)*	1.75 (.11)	1.62 (.09)

NOTE.—Numbers in parentheses are standard errors.
* Difference is significant at the .05 level.

belief that boys are better than girls at high-level mathematical reasoning but because, on our other measure of geometry performance at the end of the school year, the CAP test, boys' unadjusted means are significantly higher than girls' unadjusted means. Yet when CAP scores are adjusted by ANCOVA for scores on the EG test, adjusted means for girls and boys are nearly identical.[6] Consequently, the differences between boys' and girls' performance on the standardized geometry test at the end of the year result largely from differences in entering knowledge of geometry. That is, when differences in entering geometry knowledge are taken into account, girls and boys learn both geometry problems and proof writing equally well.

Benbow and Stanley's (1980/81) study of mathematically precocious youth (SMPY) led them to conclude that the "greatest disparity between the girls and boys is in the upper ranges of mathematical ability." Because of the publicity surrounding their results, we examined three subsets of high-achieving students, each in some way comparable to Benbow and Stanley's sample. The first subset consists of the top-scoring students on each form of the proof tests. These were the 20 students whose total scores were 22–24 on Form 1 (only two students received perfect scores on this form), the 20 students with perfect total scores on Form 2, and the 31 students with perfect total scores on Form 3. This subset has 37 females and 34 males. A second subset consists of students in grades 7 or 8 during the study and thus accelerated at least two years. Among this subset of 12 girls and 7 boys no significant differences by sex were found between the means on either the total proof score or the number of proofs correct, adjusted or unadjusted. The third subset consists of those in the sample who scored in the top 3 percent nationwide as determined by the CAP norms, comparable to the SMPY study prerequisite that students score in the top 3 percent nationwide on a standardized mathematics achievement test. This subset consists of 89 students—31 females and 58 males in grades 7–10—and indicates that, as in the Benbow and Stanley sample, significantly more males than females score at the higher levels on a multiple-choice test of standard content. But, as shown in table 5, proof-writing performance for this third subset indicates no sex-related differences. Thus our study indicates equal proof-writing performance by high-achieving girls and boys.

In summary, we have found no consistent pattern of statistically significant differences favoring either sex on any form of our proof tests. This finding holds in both our complete sample of 1,364 mixed-ability students and the three highly select subsets we examined. Thus we conclude that there are no sex differences in geometry proof-writing performance.

TABLE 5

Mean Proof Scores for Students Scoring in the Top 3 Percent Nationwide on the CAP Test According to CAP Norms

Form and Sex	Mean Proof Total Score	Mean Number of Proofs Correct
1:		
Female (N = 9)	20.33 (.71)	3.11 (.26)
Male (N = 19)	20.11 (.41)	2.95 (.16)
2:		
Female (N = 12)	22.58 (.57)	3.75 (.13)
Male (N = 25)	22.00 (.50)	3.52 (.15)
3:		
Female (N = 10)	22.60 (.37)	3.80 (.13)
Male (N = 14)	21.93 (.46)	3.57 (.20)

NOTE.—Numbers in parentheses are standard errors. Since about 56 percent of high school seniors have taken geometry (12), this subsample represents about the top 1.5 percent of the age cohort population. It includes 4.2 percent of those in the larger CDASSG study who took the CAP test.

Our findings refute the necessary existence of sex-related differences on geometry tasks requiring high-level reasoning. They cast suspicion on hypotheses of sexual differences in ability to perform other high-level cognitive tasks in mathematics. And they raise the question of what accounts for the inconsistencies in achievement by sex found between older and more recent studies and among recent works.

Related Studies

Our study is not the first to find equal mathematics performance by male and female high school students. For example, Swafford (1980) found no sex differences in algebra achievement among first-year algebra students. The 1977–78 National Assessment of Educational Progress (NAEP) and the 1978 Women in Mathematics survey (Armstrong 1981) concluded that at age 13 girls are better than boys at computation and about equal in algebra and problem-solving skills. Although by the end of high school boys have surpassed girls in problem-solving performance, these two studies found no significant differences between boys' and girls' scores on tests of computation and algebra.

Since the mid-1970s, several studies have reported increased participation by females in mathematics courses and few or no differences on spatial tasks (Becker 1978; Jacklin 1979; Armstrong 1981; Fennema

1981). We agree with Jacklin that older studies and reviews (e.g., Maccoby and Jacklin 1974) may not describe very accurately the world today, and we urge researchers to proceed with caution when basing hypotheses or conclusions on them.

However, sex differences have shown up even in recent studies. Males have outperformed females on tests of problem solving (Armstrong 1981), consumer applications (Swafford 1980), and the mathematics portion of the Scholastic Aptitude Test (SAT-M) (Benbow and Stanley 1980/81).

Confounding the problem, researchers have come to different conclusions regarding sex differences, even when working from the same data. For example, the Project TALENT study originally reported significant sex differences in mathematics scores in grade 12 favoring males (Flanagan, Davis, Dailey, Shaycroft, Orv, Goldberg, and Neyman 1964), yet when Wise, Steel, and MacDonald (1979) reanalyzed these data controlling for the number of years students had studied mathematics, testing a hypothesis of Fennema (1974), no significant sex differences were found. Moreover, the conclusion of Benbow and Stanley (1980/81) that "sex differences result from superior male mathematical ability" disagrees with Fox and Cohn's conclusions (1980) from the same data.

Although Fennema's hypothesis of differential course-taking explains many sex differences, both Benbow and Stanley's study of intellectually gifted students and NAEP data (Armstrong 1981) from a national probability sample show that differential participation in formal courses is not the sole factor.

Resolving Inconsistencies among Recent Studies

Why do some studies show great sex differences in mathematics performance whereas others do not? Our explanation relies on comparing the test items with students' formal and informal educational experiences. When test items cover material that is taught and learned almost exclusively in the classroom, no pattern of sex differences tends to be found. This holds for routine tasks such as computation and algebra exercises (Armstrong 1981; Swafford 1980), and our study shows that it holds even for such a high-level cognitive task as geometry proof writing. In contrast, when test items attempt to be purposely unlike the exercises in commonly used texts, as in tests of problem solving (Armstrong 1981), consumer applications (Swafford 1980), and the SAT-M (Benbow and Stanley 1980/81), males outperform females. Our entering geometry test appears to be somewhere in between, with questions covering content found in junior high school texts but often not taught, and only moderate sex differences arise. Thus the studies of mathematics performance we have cited fit the following general pattern: the more an instrument directly measures students' formal educational experiences in mathematics, the less the likelihood of sex differences.

Benbow and Stanley's conclusion regarding the mathematical ability of talented boys and girls rests on the assumption that the SAT-M is a test whose items are relatively and equally unfamiliar to the sexes. However, unfamiliarity is a quality relating item and student that varies greatly among students, and scores on this kind of test could easily be affected by experiences outside the mathematics classroom. These informal experiences appear to be different for the sexes throughout schooling (Burton 1979). For example, more boys than girls participate in mathematics contests, and more boys than girls work with computers (Tinker 1981). Furthermore, SMPY talented boys have tended to be more interested in mathematics than SMPY talented girls (Tobin and Fox 1980). Since better students are more likely than average students to be involved with school subjects outside the classroom, the differences in interests between boys and girls could easily result in greater differences in knowledge between the sexes among better students than among average or poorer students.

In this regard, geometry proof is a unique topic. Work with mathematics contests, computers, or advanced reading in mathematics seldom involves geometry proofs. So geometry proof writing is unlikely to be encountered even by the most interested student outside of geometry classes. Since the time of Euclid, geometry proof has been considered a model for deductive reasoning. Abstract symbols and laws of inference are often consciously applied in doing these proofs. Geometry proof writing is quite difficult for students; over one-fourth of our sample had zero proofs correct, despite the existence of easy proofs on each form and despite students having spent a significant portion of the year on the topic.[7] No algorithm exists that will handle all geometry proofs. These attributes of geometry proof writing confirm its classification as a high-level cognitive task. Thus geometry proof items provide a hard test of reasoning, yet they are likely to have been experienced by the sexes equally both inside and outside of class.

Given the documented disparity in the social and informal educational experiences of boys and girls relating to mathematics, to define mathematical ability by a score on a test of supposedly unfamiliar content forces a sex bias upon the research design. We propose that mathematical ability not be defined by tests of problem solving, spatial ability, or SATs, for which out-of-class experiences can play such an important role. Instead, we suggest that mathematical ability be defined as the extent to which students learn routine or complex tasks involving topics that are not encountered even by interested students outside the classroom. Proof writing is one of the few topics in the standard curriculum that has sufficient complexity and difficulty to be used as a measure of mathematical ability and with which formal and informal encounters are likely to be equal for the sexes. Our results with proof writing, together with our analysis of other studies, lead us to believe that boys and girls are of equal mathematical ability.

In summary, we have found that, when male and female students are tested on writing geometry proofs, a high-level cognitive task encountered almost exclusively in the classroom, no consistent pattern of sex differences in performance exists. Our results hold for both our total national sample of mixed-ability students and for select high-scoring subsamples. Our findings and data from other recent studies suggest that, when experience can be controlled, regardless of the difficulty or complexity of the items, girls and boys perform equally well.

Notes

This work was partially supported by grant NIE-G-79-0090 from the National Institute of Education. The views expressed here are those of the authors and do not necessarily reflect views of NIE. We thank Betsy Becker, Max Bell, Camilla Benbow, Susan Chipman, Edward Esty, Jacob Getzels, Larry Hedges, and Julian Stanley for helpful comments and criticism on drafts of this paper. We especially appreciate the work of Roberta Dees, who first suggested that we look at our proof data by sex and who ran many side analyses for us.

1. Copies of the final report of the CDASSG project, including all unpublished instruments mentioned here, are available from Zalman Usiskin, University of Chicago, Department of Education, 5835 S. Kimbark Avenue, Chicago, Illinois 60637, for $10, which includes handling and mailing.

2. The entering geometry (EG) test is a 19-item multiple-choice test created by the CDASSG project staff, utilizing the easier items from a 50-item test given in a study of entering geometry knowledge by Jane Macdonald, Ohio State University, 1971. The K-R 20 reliability for the EG test is .77.

3. The K-R 20 reliability for the CAP test is reported as .89; Cronbach's α reliabilities for the three forms of the proof tests are .86, .85, and .88, respectively.

4. One-way ANOVA of total score by form yields $F(2,1361) = 9.09$, $p < .0001$. For the shape of the distributions of total score by form, $\chi^2(48) = 125.17$, $p < .0001$.

5. Mean EG scores for females and males, t values: for Form 1 subsample—10.19, 11.36, 3.35, $p < .001$; Form 2 subsample—10.03, 11.47, 3.93, $p < .001$; Form 3 subsample—10.03, 11.15, 3.16, $p < .01$.

6. Mean ($\pm$ s.e.) CAP scores for females and males in this sample were: unadjusted—20.11 $\pm$ .29, 21.63 $\pm$ .28, $p < .0002$; adjusted with EG as co-

variate—20.02 ± .23, 20.87 ± .23, $p < .8641$. A similar pattern holds for the project's larger sample including students who did not study proof writing.

7. The numbers of students with no proofs correct were as follows: Form 1—55 females, 52 males; Form 2—52 females, 53 males; Form 3—91 females, 77 males.

References

Armstrong, Jane M. "Achievement and Paticipation of Women in Mathematics: Results of Two National Surveys." *Journal for Research in Mathematics Education* 12 (1981): 356–72.

Becker, Betsy Jane. The relationship of spatial ability to sex differences in the performance of mathematically precocious youths on the mathematics sections of the Scholastic Aptitude Test. Unpublished M.A. thesis, Johns Hopkins University, 1978.

Benbow, Camilla Persson, and Julian C. Stanley. "Sex Differences in Mathematical Ability: Fact or Artifact?" *Science* 210 (1980): 1262–64; "Letter to the Editor," *Science* 212 (1981): 118–19.

Burton, Grace M. "Regardless of Sex." *The Mathematics Teacher* 72 (1979): 261–70.

Comprehensive Assessment Program. *High School Subject Tests, Geometry*. Glenview, Ill.: Scott, Foresman and Co., 1980.

Fennema, Elizabeth. "Mathematics Learning and the Sexes: A Review." *Journal for Research in Mathematics Education* 5 (1974): 126–39.

Fennema, Elizabeth. "Women and Mathematics: Does Research Matter?" *Journal for Research in Mathematics Education* 12 (1981): 380–85.

Flanagan, J. C., F. B. Davis, J. T. Dailey, M. F. Shaycroft, D. B. Orv, I. Goldberg, and C. A. Neyman. *The American High School Student*. Pittsburgh: University of Pittsburgh, Project TALENT Office, 1964.

Fox, Lynn H., and Sanford J. Cohn. "Sex Differences in the Development of Precocious Mathematical Talent." In *Women and the Mathematical Mystique*, edited by Lynn H. Fox, Linda Brody, and Dianne Tobin. Baltimore, Md.: Johns Hopkins University Press, 1980.

Jacklin, Carol Nagy. "Epilogue." In *Sex-related Differences in Cognitive Functioning*, edited by M. Andrisin Wittig and Anne C. Peterson. New York: Academic Press, 1979.

Maccoby, Eleanor Emmons, and Carol Nagy Jacklin. *Psychology of Sex Differences*. Stanford, Calif.: Stanford University Press, 1974.

Malone, John A., Graham A. Douglas, Barry V. Kissane, and Roland S. Mortlock. "Measuring Problem-Solving Ability." In *Problem Solving in School Mathematics*, edited by Stephen Krulik and Robert E. Reys. Reston, Va.: National Council of Teachers of Mathematics, 1980.

Peng, Samuel S., William B. Fetters, and Andrew J. Kolstad. *High School and Beyond: A Capsule Description of High School Students*. Washington, D.C.: National Center for Educational Statistics, 1981.

Swafford, Jane. "Sex Differences in First-Year Algebra." *Journal for Research in Mathematics Education* 11 (1980): 335–46.

Tinker, Robert F. "Editorial." In *Hands On!* Cambridge, Mass.: Technical Education Research Centers, Fall 1981.

Tobin, Dianne, and Lynn H. Fox. "Career Interests and Career Education: A Key to Change." In *Women and the Mathematical Mystique*, edited by Lynn H. Fox, Linda Brody, and Dianne Tobin. Baltimore, Md.: Johns Hopkins University Press, 1980.

U.S. Department of Commerce. *Statistical Abstract of the U.S.* Washington, D.C.: Government Printing Office, 1980.

Wise, L. K., L. Steel, and C. MacDonald. Origins and Career Consequences of Sex Differences in High School Mathematics. Report to the National Institute of Education, grant no. NIE-G-78-001. Palo Alto, Calif.: American Institute for Research, 1979.